UNIVERSITÄT SALZBURG

Visual Evaluation, Scaling and Transport of Secure Videos

by
Heinz Hofbauer

Cumulative dissertation submitted to the
Faculty of Natural Sciences, University of Salzburg
in partial fulfillment of the requirements
for the Doctoral Degree.

Thesis Supervisor
Uhl Andreas Univ.-Prof. Mag.rer.nat. Dr.rer.nat.

Department of Computer Sciences
University of Salzburg
Jakob Haringer Str. 2
5020 Salzburg, AUSTRIA

Salzburg, April 2013

Abstract

Due to the pervasive access to broad band internet at home and on mobile devices *Video-on-Demand* (VoD) and video streaming become more and more popular. Consumers of VoD want to use such systems without locational restrictions, as is embodied in the idea of "ubiquitous computing", and on any device ranging from cell phones via 3G connection to home cinema system. This requires that VoD providers store videos in various resolutions and qualities, adapted to the screen resolution and connection speed of consumer systems. This in turn leads to an increase in storage and consequently an increased cost for providers.

However, there is a solution for this kind of problem, the *Universal Multimedia Access* (UMA). UMA means a video is encoded in a way which allows the adaptation to consumer requirements from a single source video. This leads not only to a reduction in required storage but also allows to adapt the VoD system to target platforms which were not originally taken into account.

The drawback of such a system is the computational cost required to adapt the video. Fortunately, this computational cost can be shifted away from the original server since modern network technologies allow adaptation in the network *Just-in-Time* (JIT), i.e., at the last possible moment in the network. This can be done by using *multimedia aware network elements* (MANE), and can theoretically reduce the actual transfer rate at the original server by using multicast and performing the adaptation JIT at the appropriate point in the network.

Wavelet based video codecs are inherently scalable and thus fit this application scenario perfectly. Additionally, the performance in terms of quality is the same as for current state of the art, non wavelet based, video codecs. The *Motion Compensated Embedded Zero Bit Codec* (MC-EZBC) is a state of the art wavelet based video codec, and thus meets all the requirements for its use for UMA.

To use a wavelet based codec for UMA seems deceptively easy. However, the direct application is prevented by MANEs, network protocols and standards which are designed in regard to the current standardised codec, i.e., H.264. Additionally, providers of VoD want a secure end-to-end communication with their customers to prevent piracy. This is in conflict with the idea to use JIT scaling in the network since full encryption would also prevent MANEs from accessing the video for adaptation.

In this cumulative thesis we will thus deal with the following problems: We have to ensure that MC-EZBC based video codecs can be transported via existing technology; We have to develop encryption methods for secure end-to-end connection which allow JIT adaptation in the network; and we have to develop methods to ensure and evaluate these encryption methods.

Abstract (German)

Durch die weite Verbreitung von Breitbandinternet im Heim und Mobilen Bereich werden *Video-on-Demand* (VoD) und Streamingsysteme immer beliebter. Konsumenten von VoD wollen diese Systeme überall nutzen, Stichwort "ubiquitous computing", von Smartphones mittels 3G Verbindung bis zu Homecinema Systemen über Breitbandinternet. Dies verlangt von VoD Anbietern das Videos in diverser Form gespeichert werden, also in verschiedener Auflösung und Qualität, angepasst an das jeweilige Endgerät und die mögliche Verbindungsgeschwindigkeit. In weiterer Folge bedeutet das eine gesteigerten Speicherverbrauch und damit höhere Kosten für Anbieter.

Für dieses Problem gibt es eine Lösung, den *Universal Multimedia Access* (UMA). Gemeint ist damit das ein Video auf eine Art gespeichert wird die eine Adaption an die eventuellen Anforderungen eines Benutzers aus dem gleichen Quellmaterial erlaubt. Damit wird der benötigte Speicherbedarf gesenkt und die Möglichkeit gewährleistet auf neue Anforderungen einzugehen die im Ursprungssystem nicht vorgesehen waren.

Der Nachteil dieser System ist die benötigte Rechenleistung zur Adaption der Videos. Allerdings führt dies nicht zu einem Flaschenhals beim Anbieter da moderne Netzwerktechnologien erlauben diese Adaption *Just-in-Time* (JIT), also erst an der letzt möglichen Stelle im Netzwerk, zu erledigen. Durch diese *multimedia aware network elements* (MANE) ist theoretisch sogar eine Einsparung der Bandbreite beim Anbieter möglich da verschiedene Endnutzer mit einem einzigen Stream seitens des Anbieters bedient werden können indem die Adaption JIT im Netzwerk ausgeführt wird.

Wavelet basierte Video Codecs sind durch ihre inhärente Skalierbarkeit gut für diese Anwendung geeignet und bieten die gleiche Qualität wie herkömmliche Video Codecs. Der *Motion Compensated Embedded Zero Bit Codec* (MC-EZBC) ist ein State of the Art Video Codec der auf Wavelets basiert und damit den Anforderungen von UMA grundsätzlich genügt.

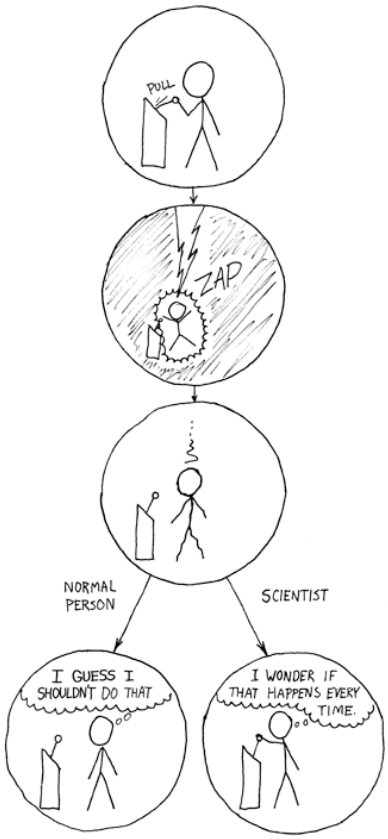
Die Lösung für das UMA Problem einfach einen Wavelet basierten Codec zu verwenden klingt offensichtlich wird aber vorerst dadurch verhindert das MANEs und die zugehörigen Netzwerkprotokolle nur Standardisierte Codecs, nämlich H.264, verstehen. Zusätzlich wollen Anbieter von VoD Systemen auf eine sichere Art mit ihren Endnutzern kommunizieren um Piraterie zu vermeiden, was eine JIT Skalierung im Netzwerk ausschließt da die MANEs vollen Zugriff auf das gestreamte Video zur Adaption benötigen.

In dieser kumulativen Dissertation geht es also darum folgende Probleme zu lösen: Es muss gewährleistet werden das MC-EZBC Video Ströme über die Existierenden Transporttechnologien übertragen werden können; Es müssen Methoden zur sicheren End-zu-End Verbindung entwickelt werden die eine JIT Skalierung erlauben; und es müssen Methoden zur Evaluierung der Sicherheit des Visuellen Inhalts von Videos entwickelt und evaluiert werden.

Acknowledgments

Writing this thesis was not always an easy undertaking. All the more grateful I am for the people who encouraged, inspired and motivated me during my studies and work at the University of Salzburg.
I am very grateful for the tutelage, guidance and assistance from my advisor Andreas Uhl throughout my PhD.
I greatly appreciated discussions with my colleagues at the University which were exhilarating, interesting and inspiring. For the interesting time and fruitful work together I would especially thank my co-authors (in alphabetical order) Robert Kuschnig, Christian Rathgeb, Thomas Stütz and Peter Wild. Furthermore, I would like to thank Michael Gschwandtner, Stefan Huber and Peter Meerwald for thought provoking discussions outside work which nonetheless served to make life more interesting.
Finally, I would like to thank my family and friends for their encouragement and understanding and their continued support in what I do.
This thesis has been funded in part by the Austrian Science Fund (FWF) project P19159-N13.

Salzburg, April 2013 *Heinz Hofbauer*



Comic by Randall Munroe (xkcd.com)

Contents

1. Introduction

This cumulative dissertation covers my research with respect to selective encryption and transportation of wavelet based video codecs and image metrics and their use for visual security. These two topics, while on the surface somewhat different, share a strong connection when it comes to digital rights management (DRM). Encryption for DRM does not deal with the traditional notion of cryptographic security as coined by Shannon [47]. In the context of visual media, images and videos both, it is not necessary to prevent an adversary from gaining any knowledge about the media but rather to prevent unauthorized consumers from accessing the high quality content. Image metrics are utilized to estimate quality resulting from encryption since they, by design, model the human visual system (HVS). The reason to use image metrics instead of actual human observers, which would be optimal, is that they are very time and cost efficient.

In the following two sections the reasons, nomenclature and techniques used for video transportation and encryption as well as the visual evaluation with image metrics will be explained in more detail.

1.1. Selective Encryption and Transport of Videos

Video coding in its usual form targets a certain resolution and output size. Within these parameters it exploits redundancies in the visual domain to optimize the output quality versus the bit rate. When the application target is fixed this approach is well suited, e.g., encoding a video to be delivered on a blue-ray means the spatial resolution is usually HD1080 and the size of the output is limited by the amount of data which can be stored on the physical disk. However, when the target platform is not fixed this approach has some drawbacks. An example of this would be a streaming service where the recipient platforms can range from cell phones to home theatre systems. While this affects spatial resolution there is also the bit rate dependency in the form of network throughput, which again can range from low to high, e.g., wireless connection or broadband access. In this case the streaming server either has to store a coded video for each combination of output parameters or do a reencoding on the fly from a higher quality source. This leads to either an overhead in storage or computational resources.

A possible solution to this is to store a video in such a manner that it can be adapted to current end user requirements in a timely manner. This notion is called universal multimedia access (UMA) [55] and the prime enabling technology is the use of scalable video codecs. The state of the art codec H.264/AVC [24, 27] has a scalable video extension H.264/SVC [28, 46], but there are some drawbacks to using it. The most important is the fact that it requires scaling targets to be specified when first encoding the video, i.e., possible application scenarios have to be known when the video is first created. Wavelet based codecs on the other hand are inherently scalable, while the options are limited by the decomposition structure of the wavelet all the options are available all the time. Furthermore, wavelet based codecs are similar in performance to H.264 [35, 14]. For this reason we chose the motion compensated embedded zeroblock coder (MC-EZBC) [22, 5, 6, 60]. The MC-EZBC uses a T+2D wavelet decomposition, i.e., temporal wavelet coding (with 5/3 CDF wavelets) is followed by spatial wavelet coding (with 9/3 CDF wavelets).

The reason for using a wavelet based codec is rather straightforward, however it has its drawbacks. Since H.264 is the current ITU-T standard, a lot of hardware and software de-

sign is focused on transporting H.264/AVC. Likewise, there exists a lot of work regarding the transmission [56, 1, 56] and adaptation [53, 30, 31] of H.264/SVC. As such, a major part of this thesis is dealing with transporting a wavelet based video by utilizing hardware designed for H.264/SVC, see section 2.1.1.

Closely tied in with transportation is encryption of the video stream. In order to securely transport a stream from the server to the client encryption has to be used, especially when considering multicast applications. The easiest option for this scenario is to use transport security by using a secure streaming protocol like the secure real-time transport protocol (SRTP) as defined in RFC3711 [2].

A drawback of this approach especially in context of UMA is key distribution. If the UMA principle is used adaptation to each end user requirement is done and the stream is then transported to the client. This would essentially result in a unilateral connection for each client and all the computational work, slight as it may be, being done on the server. A better approach to this scenario is to use multicast and at the last possible point in the network adapt to user requirements. This can be done by utilizing multimedia aware network elements (MANE) [31]. However, in order to be able to scale on a MANE, the MANE needs to be able to decrypt the transport stream in order to adapt it, which necessitates a key distribution system between server, MANEs and clients. This introduces an overhead in workload for the server and network and introduces a number of potential points of attack on the MANEs.

This problem of key distribution and transport encryption can be circumvented by utilizing encryption on the bitstream. However, in order to be able to perform scaling on MANEs we need to leave certain information of the bitstream in plain text ,i.e., markers in the bitstream which facilitate scaling. An extension of this idea to leave codec relevant data in plain text is the notion of format compliance. In order to be format compliant an encrypted bitstream has be decodable by a standard conform decoder without a fatal error.

Since we moved away from traditional cryptographic security in the sense of Shannon [47] it is possible to choose the encryption strength based on the amount of bitstream data which is encrypted. There are two basic options for this: One, reduction in the encrypted amount to increase encryption speed; Two, careful choice of the encrypted data in order to facilitate certain goals.
The first option is rather intuitive and straightforward, in order to reduce encryption time less data is encrypted. More interesting is the second option were carefully selection of the encrypted data can facilitate other goals, especially in the context of digital rights management.

The following possible application scenarios are typical for selective encryption:

Confidentiality Encryption Means MP security (message privacy). The formal notion is that if a system is MP-secure an attacker cannot efficiently compute any property of the plain text from the cipher text [3]. This can only be achieved by the conventional encryption approach.

Content Confidentiality Is a relaxation of confidential encryption. Side channel information may be reconstructed or left in plaintext, e.g. header information, packet length, but the actual visual content must be secure in the sense that the image content must not be intelligible / discernible [51].

Sufficient Encryption Means we do not require full security, just enough security to prevent abuse of the data. The content must not be consumable due to high distortion (e.g. for DRM systems) by destroying visual quality to a degree which prevents a pleasant viewing experience or destroys the commercial value. This implicitly refers to message quality security (MQ), which requires that an adversary cannot reconstruct a higher quality version of the encrypted material than specified for the application scenario [50].

Perceptual / Transparent Encryption Means we want consumers to be able to view a preview version of the video but in a lower quality while preventing them from seeing a full version. This for example can be used in a pay per view scheme where a lower quality preview version is available from the outset to attract the viewers interest, e.g., Li et al. [33]. The difference between sufficient and transparent is the fact that there is no minimum quality requirement for sufficient encryption. Encryption schemes which can do sufficient encryption cannot necessarily ensure a certain quality and are thus unable to provide transparent encryption.

Wavelet based codecs in general can deal with sufficient and transparent encryption due to the structure of the wavelet decomposition. In applications where confidentiality is required it is often better to utilize transport encryption except where other considerations, e.g., computational performance or key distribution, favor selective encryption. In section 2.1.2 a selective encryption method for the MC-EZBC is introduced and a comparison with transport encryption, in terms of scaling performance on MANEs, is introduced and evaluated.

1.2. Visual Evaluation of Security with Image Metrics

When it comes to image and video coding the desire is usually to either increase quality through better algorithms or likewise to decrease image/video size without impacting quality. The notion of quality is thus central when dealing with visual media. The problem is that quality is actually the perceived quality as observed by human subjects, which is not uniform. This means in order to get a proper quality estimate a number of human observers are required to reach a significant mean opinion score (MOS). The ITU recommendations for the "Methodology for the subjective assessment of the quality of television pictures" [26] and "Audiovisual quality in multimedia services" [29] specify testing methodology and environment as well as that the number of observers should be at least fifteen. Overall, this leads to a very high cost of quality assessment due to space and equipment requirements, i.e., laboratory space and viewing devices, as well as human observers. The time requirement is similarly high for even small sized test sets.

The solution to the high cost comes in the form of objective image metrics. Image metrics are designed to reflect the human judgement given certain impairments and can thus be used as a cheap and fast replacement for actual human observers for quality testing. This shifts the considerable time and cost effort towards the development of image metrics which is a considerable improvement. However, if each image metric would be judged on a different test set with different observers the inter metric comparability would suffer. To prevent this, and to further reduce the cost of engineering novel image metrics, fixed databases are used. The most prominent image databases are the LIVE Image Quality Assessment Database [49], Tampere Image Database (TID) [38, 39, 40], MICT Image Quality Evaluation Database [21] and Subjective quality assessment IRCCyN/IVC database [32].

Based on these databases a large number of image metrics have been published. The main difference between these metrics is the degree to which they model the human visual system (HVS). Usually it holds that a better simulation of the HVS increases the correlation with the MOS at the cost of computational speed. Improvement in this field is thus always possible, either through improvement of correlation with the MOS or by increasing the speed of computation. In section 2.2.1 we show how such an improvement can be done by introducing a new image metric which is effective, i.e., it has a high correlation with MOS, as well as efficient, i.e., fast to compute.

When it comes to selective encryption image metrics are also of importance. The target sce-

narios for sufficient and transparent encryption specifically target certain qualities. In the case of sufficient encryption we want a video/image to be below a certain quality threshold at all times. For transparent encryption the resulting quality is required to be in a quality range between the highest allowed quality and the lowest required quality, i.e., the quality required in order for the visual media to be usable as a preview. Depending on content and methodology this has to be evaluated for each type of media which is encrypted, as such a huge workload is required of the visual quality estimation. While this could be done by humans subjects during the design phase it is hardly praticial in an actual application. This leaves only image metrics to fill the role as visual security estimators.

Certain image metrics like the PSNR [34, 10] and SSIM [59, 61] are frequently used for this task since they are well understood, easy to implement and relatively fast. However, there are other security metrics specifically designed to fill the role of visual security estimators, e.g., the local entropy metric [52] and the local feature based visual security metric [54]. Usually the authors of security metrics claim generality for the introduced metric without evaluation outside the specific method for which they were designed for. This approach is problematic, since on the one hand image metrics are used without assessing their applicability to the given scenario. On the other hand, security metrics are introduced without proper testing, i.e., usually they are evaluated on less strict terms than regular image metrics, which is especially precarious given they are used in a security relevant task. However, the major problem when it comes to the evaluation of security metrics is the lack of clear guidelines or consideration on how to evaluate security metrics. In section 2.2.2 we evaluated regular image metrics and showed that they are not fit for the evaluation of low quality images, which is usually the case for sufficient encryption. We then extended the work to defining a methodology how to evaluate security metrics and assessed security and image metrics from literature.

The basic task of an image metrics, regardless of their proximity to the HVS, is to assess the difference between two images. Different image metrics utilize different image features to facilitate this comparison. This huge number of potential image feature comparison methods could potentially be used in other fields of science which require some kind of image comparison. In an attempt to transfer knowledge from one field to another we attempted to utilize image metrics as biometric comparators for iris recognition.

Basically, the task of iris recognition in biometrics is to use two images, one from an enrollment step which is linked to an identity, and another which is recorded during the authentication or identification process. For authentication the user claims an identity and the biometric feature is used, by comparing it with the stored and known identity, to estimate the veracity of the claim. For identification the user just presents his biometric feature and the system, by comparing the given feature with all features stored in the database, has to identify the user. So at a basic level the task of a biometric system is to identify a user by comparing two recordings of a biometric feature. This cleanly links the task of image metrics and biometric comparators, although surrounding parameters are somewhat different. An example of this difference would be the fact that for image comparison the average luminance of an image has a prominent role, which is present in almost all image metrics. Biometric systems on the other hand compensate difference in luminance as a preprocessing step since it is affected by recording conditions, e.g., recording of the iris during day or night. In section 2.2.3 we evaluated the use of image metrics, without compensating for the differences in biometric preprocessing, with high success. This shows that an exchange of domain knowledge from image metrics to biometric comparators is possible and should be investigated further.

2. Contribution

The work published in recent years can be roughly divided into two categories: the use of the motion compensated embedded zero block coder as a solution for universal multimedia access, and research in image quality based around, but not limited to, the requirements needed for selective encryption.

2.1. The Motion Compensated Embedded Zeroblock Coder as Solution to Universal Multimedia Access

2.1.1. In Network Adaptation

In-network adaptation is important since it reduces the load on the server. Specifically, there is the option of using multicast on the server even though clients receive different version of a video stream. This is done by utilizing in network adaptation, i.e. a full video is streamed from the server and adapted in the network when necessary. This has a number of restrictions. First, the video stream has to be capable of scaling without reencoding, since the network elements do not have the capacity to do reencoding. Secondly, the network elements need to know how to perform the scaling on the bitstream. And thirdly, network elements need access to the bitstream, i.e. they need the key for traditional encryption scenarios and thus create a potential point of attack.

Regarding scaling capability, both H.264/SVC and the MC-EZBC (or any other wavelet based codec) are capable of scaling. Both H.264/SVC and the MC-EZBC are about equal in quality, with the H.264/SVC being slightly higher quality for fewer scaling points and the MC-EZBC being slightly higher quality for a high number of scaling options [35, 14]. However, the main difference is that for the H.264/SVC the scaling options need to be specified during encoding while the MC-EZBC codec, through the wavelet structure, has natural scaling options. Thus, both codecs can be used for in network adaptation.

Regarding the signaling of bitstream layout to in network elements for scaling there are a number of options.

One is to use a codec agnostic description, the generic Bitstream Syntax Description (gBSD) [37] based on MPEG-21 Part 7 "Digital Item Adaptation" (DIA) [25]. The description framework we developed for the MC-EZBC is given in [15] along with an evaluation of overhead. The problem with this approach is the fact that we have to remodel the scalability options of the wavelet based bitstream in the gBSD data. This in turn leads to a rather large overhead since the gBSD data has to be transported alongside the original bitstream.

The other option is to use a standardized way for transporting multimedia data, the Real-time Transport Protocol (RTP) [44] and the Real Time Streaming Protocol (RTSP) [45] which in turn utilizes RTP. The interesting part is that the existing technology can already deal with H.264/SVC, e.g., [58] describes the H.264/AVC payload for RTP and multimedia aware network elements (MANE) and [31] extends this to H.264/SVC. The H.264/SVC is segmented into network abstraction layer units (NALUs) and scaling on MANEs is performed on a NALU basis. Thus if a mapping can be found from the MC-EZBC bitstream structure to the H.264/SVC NALU structure the existing hardware and software can be utilized to transport MC-EZBC bitstreams. We introduced such a mapping in [11], which also contains a comparison to the

adaptation method utilizing gBSD data. Overall we have shown that the utilization of MC-EZBC into NALU embedding generates less overhead on the network and on the MANEs and is preferable.

In [11] we also did a evaluation of encryption options. The options are to encrypt the bitstream, either prior to or after embedding into a NALU stream, or to use transport encryption, the Secure Real-time Transport Protocol (SRTP), defined in RFC-3711 [2], which is a profile of the RTP. It was found that bitstream encryption outperforms transport encryption from a computational standpoint. Furthermore, when utilizing bitstream encryption the MANEs do not need to know the key which reduces the number of potential attack points in the system. No difference was found when it comes to applying bitstream encryption prior to or after NALU embedding.

Publications (sorted chronologically)

[15] HOFBAUER, H., AND UHL, A. The cost of in-network adaption of the MC-EZBC for universal multimedia access. In *Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA '09)* (Salzburg, Austria, Sept. 2009)

[14] HOFBAUER, H., STÜTZ, T., AND UHL, A. Secure Scalable Video Compression for GVid. In *Proceedings of the 3rd Austrian Grid Symposium* (Linz, Austria, 2009), J. Volkert, T. Fahringer, D. Kranzlmüller, R. Kobler, and W. Schreiner, Eds., vol. 269 of *books@ocg.at*, Austrian Computer Society, pp. 88–102

[11] HELLWAGNER, H., HOFBAUER, H., KUSCHNIG, R., STÜTZ, T., AND UHL, A. Secure transport and adaptation of MC-EZBC video utilizing H.264-based transport protocols. *Elsevier Journal on Signal Processing: Image Communication 27*, 2 (2011), 192–207

2.1.2. Selective Encryption Schemes for Wavelet-based Codecs

Selective encryption in our case refers to two different selection methods. One is selective encryption in the sense that format compliance has to be achieved, i.e. the encrypted bitstream has to be a valid MC-EZBC bitstream and a decoder should be able to handle it. The other sense of selective encryption will be referred to as partial encryption here and in our publications to better distinguish between these two types. Partial encryption refers to the selection of data to actually encrypt from the subset which can be encrypted while still maintaining format compliance. This is done in order to target specific application scenarios. An example would be transparent encryption, where a certain quality should be maintained, which would lead to the selection of high frequency data for partial encryption.

We published our encryption method for the MC-EZBC in [16], which was the first encryption method for the MC-EZBC to appear in literature.

Regarding security Lookabaugh et al. [36] showed that selective encryption is sound and demonstrated its relation to Shannon's work, [47]. However, Said [43] showed that side information can compromise security. In [16] attacks on the encrypted visual data were investigated. In [17] we further investigated the security of the original approach and showed that confidential encryption is not possible while maintaining format compliance. Furthermore, we showed that motion vector information can also be used to compromise security. Consequently we also extended our encryption approach to encompass motion vector fields.

When it comes to selective encryption another stated benefit is reduction in computational cost and thus increased encryption speed and performance. This performance gain is often limited since parsing the bitstream can take up the same or more time than simply encrypting

it. In [11] we showed that in our application case, i.e. secure transportation of a video, the performance of selective encryption is measurably higher than transport encryption. Furthermore, through the use of format compliant end to end encryption the necessity for a secure key distribution system to MANEs in the network is removed.

Publications (sorted chronologically)

[16] HOFBAUER, H., AND UHL, A. Selective encryption of the MC EZBC bitstream for DRM scenarios. In *Proceedings of the 11th ACM Workshop on Multimedia and Security* (Princeton, New Jersey, USA, Sept. 2009), ACM, pp. 161–170

[17] HOFBAUER, H., AND UHL, A. Selective encryption of the MC-EZBC bitstream and residual information. In *18th European Signal Processing Conference, 2010 (EUSIPCO-2010)* (Aalborg, Denmark, Aug. 2010), pp. 2101–2105

[11] HELLWAGNER, H., HOFBAUER, H., KUSCHNIG, R., STÜTZ, T., AND UHL, A. Secure transport and adaptation of MC-EZBC video utilizing H.264-based transport protocols. *Elsevier Journal on Signal Processing: Image Communication 27*, 2 (2011), 192–207

2.2. Visual Security and Image Metrics

2.2.1. An Effective and Efficient Image Metric

During image and video encoding and manipulation the task is usually to achieve a high quality in respect to a human observer. Optimally, human observers would be used, but setting up a testing lab, introducing human observers to the task and performing the tests is both time consuming and expensive. Thus, image metrics are used instead of human observers in many tasks. Image metrics range from simple and technical distortion measures, such as MSE or PSNR to highly sophisticated image metrics, e.g., VIF [48] or CPA1 [4], which take into account the HVS to a great extent and closely simulate results as obtained from human observers.

The problem in practice is that simple image metrics like the PSNR are easy to understand and implement and fast, however the correlation to human judgement is lacking [23, 18]. Image metrics which model the human visual system (HVS) on the other hand generally produce good results but are difficult to understand and implement and oftentimes excruciatingly slow. There are metrics in between these two extremes, e.g., SSIM [57], NICE [7] or LFBVS [54], however they are significantly slower and more complicated to implement than the PSNR while not providing the high precision quality estimations like the VIF or CPA1. As such in practice inferior quality estimation is accepted when speed of calculation is a requirement, mostly in image and video coding tasks, while offline experimentations use higher precision image metrics.

Consequently, there is room for improvement by speeding up image metrics without sacrificing correlation with human observers. The SSIM is one of better examples of this, it is relatively fast and provides a good estimation of human judgement. In [19] we developed an image metric which is fast and shows a high correlation to human judgement, on average it outperforms even the CPA1.

Publications (sorted chronologically)

[19] HOFBAUER, H., AND UHL, A. An effective and efficient visual quality index based on local edge gradients. In *IEEE 3rd European Workshop on Visual Information Processing* (Paris, France, July 2011), p. 6pp

2.2.2. Applicability of Visual Security Metrics for Selective Encryption

When using selective encryption a part of the plain text remains by definition. When the selective encryption output is also format compliant, which is often the case since format compliance is one of the main reasons to use selective encryption, the cipher text can be decoded just like a normal video or image. The resulting quality of this decoding is usually poor since the cipher text introduces distortions. This leads to the relative obvious concept of visual security, i.e. the security of the scheme is based on the reconstruction of the video through an attack and regular decoding. However, since we by design retain some information in the visual domain the possible application scenarios can also be extended. Usual goals are sufficient encryption, where the output should be of very low quality in order to prevent unauthorized users from accessing the media, or transparent encryption, where a low quality version of the output should be guaranteed as a preview version.

Image metrics are routinely tested on databases [49, 21, 32, 40] which contain a fair number of image impairments as well as mean observer scores by a number of human observers. This is the ground truth with which image metrics are evaluated. The problem with these evaluation is that a given metric is usually tested over the whole range of image impairments ranging from high to low quality. And while most metrics perform admirably well for low impairments, i.e., high quality, and over the whole quality range, image metrics correlation to human observer scores frequently fall off for high impairments, i.e., low quality. In [18] we showed this behaviour for a large number of image metrics on widely used databases and suggest that this has to taken into account when selecting a image metric for a low quality application. Additionally, we showed that there is a lack of image metrics which perform as well on the low quality range as they do on the high quality range.

This poor metric behavior especially impacts the use of image metrics as security metrics. Sufficient encryption for example is a prime candidate for a low quality evaluation task. And this is specifically the case for which most image metrics fail. This is a twofold problem, on the one hand the desirable properties for security metrics have never been formalized and on the other hand image metrics are only evaluated on the whole quality range which, as shown in [18], misrepresents the low quality case. In [20] we attempt to alleviate this problem by specifying properties of image metrics which are required for a security analysis task. Furthermore, we evaluated a number image metrics which are frequently used in literature or which claim to be security metrics based on these properties. The unfortunate conclusion of the work is that none of the currently proposed security metrics are fit for the task.

Publications (sorted chronologically)

[18] HOFBAUER, H., AND UHL, A. Visual quality indices and low quality images. In *IEEE 2nd European Workshop on Visual Information Processing* (Paris, France, July 2010), pp. 171–176

[20] HOFBAUER, H., AND UHL, A. An evaluation of visual security metrics. *IEEE Transactions on Multimedia* (2013), 15 pages. submitted

2.2.3. Image Metrics as Comparators for Iris Recognition

The International Organization for Standardization (ISO) specifies iris biometric data to be recorded and stored in (raw) image form (ISO/IEC FDIS 19794-6), rather than in extracted templates (e.g. iris-codes) achieving more interoperability as well as vendor neutrality [9]. The storage of the raw iris image instead of iris templates allows for a more sophisticated approach during template matching. This triggered the idea of utilizing image metrics in the matching

process to assess whether a technology transfer from the field of image metrics to the field of biometric recognition would be possible. The processing chain of traditional iris recognition (and other biometric) systems has been left almost unchanged, following Daugman's approach [8] consisting of (1) *segmentation and preprocessing* normalizing the iris texture by unrolling into doubly-dimensionless coordinates, (2) *feature extraction* computing a binary representation of discriminative patterns of the rectified iris texture, and (3) *biometric comparison* in feature space involving the fractional HD as dissimilarity measure [41].

In [13] we applied image quality metrics to iris textures as well as iris templates. While image metrics do not outperform traditional feature vector-based techniques, results were better than expected when normalized input was used. This shows that a knowledge transfer is possible between image quality metrics and biometric comparators.

Given the promising results of [13] we extended the research to include fusion with other image metrics as well as biometric comparators in [12]. The results gained from this experiment shows an improvement of the total accuracy and justify the applicability of this approach. However, experiments also highlight that not every combination of comparators improve recognition, which was claimed by several authors [42]. Rather, the results suggest that the fusion of comparators utilizing complementary information is necessary to benefit from biometric fusion and increase recognition accuracy.

[13] HOFBAUER, H., RATHGEB, C., UHL, A., AND WILD, P. Iris recognition in image domain: Quality-metric based comparators. In *Proceedings of the 8th International Symposium on Visual Computing (ISVC'12)* (Crete, Greece, July 2012), pp. 1 – 10

[12] HOFBAUER, H., RATHGEB, C., UHL, A., AND WILD, P. Image metric-based biometric comparators: A supplement to feature vector-based hamming distance? In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'12)* (Darmstadt, Germany, Sept. 2012), pp. 1 – 5

3. Publications

This chapter presents publications as originally published, reprinted with permission from the corresponding publishers. The copyright of the original publications is held by the respective copyright holders, see the following copyright notices. In order to fit the paper dimension, reprinted publications may be scaled in size and/or cropped.

Selective Encryption of the MC EZBC Bitstream for DRM Scenarios

Heinz Hofbauer and Andreas Uhl
Department of Computer Sciences
University of Salzburg
Jakob-Haringer Str. 2, A-5020 Salzburg, Austria
{hhofbaue,uhl}@cosy.sbg.ac.at

ABSTRACT

Universal Multimedia Access (UMA) calls for solutions where content is created once and subsequently adapted to given requirements. With regard to UMA and scalability, which is required often due to a wide variety of end clients, the best suited codecs are wavelet based (like the MC-EZBC) due to their inherent high number of scaling options. However, we do not only want to adapt the content to given requirements but we want to do so in a secure way. Through DRM we can ensure that the actual content is safe and copyright is observed. However, traditional encryption removes the option of scalability in the encrypted domain which is opposed to what we want to achieve for UMA. The solution is selective encryption where only a part of the content is encrypted, enough to ensure safety but at the same time little enough to keep scalability intact. Towards this goal we discuss various methods of applying encryption to the bitstream produced by the MC-EZBC in order to keep scalability intact in the encrypted domain while also keeping security intact with regard to various DRM scenarios.

Categories and Subject Descriptors

K.4.4 [**Computer and Society**]: Electronic Commerce—*Security*; I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*

General Terms

Security

Keywords

Security, in-network adaption, wavelet, selective encryption, scalability,DRM

1. INTRODUCTION

The use of digital video in todays world is ubiquitous. Videos are viewed on a wide range of clients, ranging from hand held devices with QVGA resolution (320x240) over PAL (768x576) or NTSC (720x480) to HD 1080p (1920x1080) or higher. Furthermore, streaming servers should be able to broadcast over the internet with regard to a wide range of bandwidths, from fixed high bandwidth lines like ADSL2 to changing low bandwidths for mobile wireless devices. In such an environment it is simply not possible to encode a video for every application scenario. So content providers either have only a fixed number of options available or they use scaling video technology to adapt the video for bandwidth and resolution requirements of the client. The concept of creating the content once and adapting it to the current requirements is preferable and is better known as Universal Multimedia Access (UMA) [25].

One of the enabling technologies of UMA is the use of scalable video coding. This averts the need for transcoding on the server side and enables the server to scale the video. However, even scaling takes up computation time and reduces the number of connections the server can accept. Furthermore, variable bandwidth conditions, which happen frequently on mobile devices, further taxes the server with the need to adapt the video stream. The solution to this is usually in-network adaption, shifting the need to scale to the node in the network where a change in bandwidth is occurring. The core adaption with these restrictions takes place on the server and adaption due to actual channel capability is done in-network. For design options and comparisons of in network adaption of the H.264/SVC codec see Kuschnig et al. [10]. Wu et al. [26] give an overview of other aspects of streaming video ranging from server requirements to protocols, to QoS etc.

For video streaming in the UMA environment, i.e. a high number of possible bandwidths and target resolutions, wavelet based codecs should be considered. Wavelet based codes are naturally highly scalable and rate adaption as well as spatial and temporal scaling is easily achieved. Furthermore, wavelet based codecs achieve a coding performance similar to H.264/SVC, c.f. Lima et al. [13]. For an overview about wavelet based video codecs and a performance analysis as well as techniques used in those codecs see the overview paper by Adami et al. [1]. Under similar considerations Eeckhaut et al. [5] developed a complete server to client video delivery chain for scalable wavelet-based video. The main concern of research regarding UMA is usually performance with respect to scaling and in-network adaption. However, digital rights management and security is also a prime concern.

Shannon [22] in his work on security and communication

MM&Sec'09, September 7–8, 2009, Princeton, New Jersey, USA.

made it clear that the highest security is reached through a secure cipher operating on a redundancy free plain text. Current video codecs exploit redundancy for compression and we can consider the bitstream to be a redundancy free plain text in the sense of Shannon. Thus for maximum security we just need to encrypt the whole bitstream with an state of the art cipher, i.e. AES. But we also loose the flexibility of the scalable bitstream. If we want to continue scaling in the network we have to provide the key to every node in the network where we want to perform scaling. However, the required key management is another likely security risk since it generates more attack points, i.e. key transmission and the receiving network node could be targeted to gain access to the key. However, if we relax our security standard, i.e. we do not want perfect security, then it is possible to combine security and scalability. This is exactly what we will assess in this paper.

Selective encryption is the encryption of only a part of the bitstream we wish to protect, usually with the goal of keeping some information contained in the file accessible. While this lowers the security of the encrypted bitstream it also yields benefits. The first thing we should realize is that often we do not need full security, take television broadcasting for example. It is not necessary to prevent people from recognizing what movie is airing on an encrypted channel, we just want to reduce the viewing experience without the corresponding key. This is also a good example why we want to keep information intact: we do not want the receiver thinking it receives noise (and properly encrypted signals should look like a random signal) but we want it to recognize a valid signal, e.g. a video stream, we just do not want the receiver to be able to reconstruct the contents. Other goals could be to retain scalability, to generate preview versions from an encryption stream and so on.

Regarding security Lookabaugh et al. [14] showed that selective encryption is sound and demonstrated its relation to Shannon's work. However, in practice a bitstream is not always redundancy free, as required by Shannon. For example, Said [21] showed that side information can compromise security. And of course even the best video codec does not exploit all redundancies in the bitstream. As such, it is expedient to include an attack in the examination of a selective encryption scheme to be able to gauge the actual security. For an overview about prior selective encryption methods see the papers by Massoudi et al. [19] and Liu et al. [16].

So as stated our main goal is to keep scalability intact while providing security to some extent. The possible security goals we want to achieve with selective encryption in different DRM scenarios are as follows:

Confidentiality Encryption means complete security, except for the information we want to give away. This is not easily achieved, since headers and other information which are necessary to recognize a bitstream can contain information which can lead to an identification of the content, see [6] for an example of such an attack.

Sufficient Encryption means we do not require full security, just enough security to prevent abuse of the data. This is of course heavily dependent on what we want to achieve. In this case we want to prevent people without a key to be able to view the video sequence. This does not mean that we do not want them to recognize what is in the video sequence, we just want to reduce the visual quality to a level which is regarded as unviewable by the general public. Another goal of sufficient encryption is the reduction of computational complexity, e.g. less time or memory required as compared to traditional encryption.

Transparent Encryption means we want people to see a preview version of the video but in a lower quality while prevent them from seeing a full version. This is basically a pay per view scheme where a lower quality preview version is available from the outset to attract the viewers interest. The distinction is that for sufficient encryption we do not have a minimum quality requirement, and often encryption schemes which can do sufficient encryption cannot ensure a certain quality and are thus unable to provide transparent encryption. Also, computational complexity for transparent encryption is secondary, the main goal is to provide a preview version.

Regarding the standard H.264/AVC/SVC there has also been done research regarding selective encryption. For both AVC and SVC Magli et al. [17, 18] created a transparent encryption scheme. All the other works presented are regarding sufficient encryption of AVC only. The only bitstream oriented encryption schemes, i.e. encryption after compression, are done by Shi et al. [23] and Iqbal et al. [9] and are not format compliant, i.e. a standard coder would not be able to decode the encrypted bitstream. The methods proposed by Li et al. [12], Bergeron et al. [2] and Lee and Nam [11] are to our knowledge format compliant but also compression integrated. Especially the compression integrated algorithms are troublesome to use since a change of keys would require a new encoding of the bitstream.

We want to apply selective encryption to the bitstream produced by the MC-EZBC [8, 3, 4, **?**] which is a t+2D scalable video codec. This choice was made mainly because the source code is available[1], which enables our experiments. Scalability in a video codec means that after one encoding step we get a bitstream which can be scaled to different bit rates, spatial and temporal (i.e. frame rate) resolutions, without reencoding the video sequence. The MC-EZBC uses motion compensated temporal filtering, with 5/3 CDF wavelets, followed by regular spatial filtering, with 9/7 CDF filtering, see fig. 3 for a GOP size of 8. This method, temporal first and spatial later, is referred to as t+2D coding scheme. For temporal filtering a full decomposition is used and thus the GOP size is discernible by the number of temporal decomposition levels t, i.e. GOP size = 2^t. Both temporal and spatial filtering are done in a regular pyramidal fashion. Statistical dependencies are exploited by using a bit plane encoder, the name giving embedded zero bit coder (EZBC), and motion vectors are encoded with differential pulse code modulation followed by an arithmetic coding scheme. Also note that I frames lead each GOP and furthermore can appear later in a GOP in case of a scene change (the dashed outline in fig. 3, lower part, shows possible occurrences of further I frames).

The outline of the paper is as follows. Section 2 gives an overview of the goals we want to achieve with the selective encryption, the method we use and a performance

[1]The source code for the ENH-MC-EZBC is available from `http://www.cipr.rpi.edu/research/mcezbc/`.

analysis. Experimental results for sufficient and transparent encryption are given in section 3. A summary, conclusion and outlook to future work is given in section 4.

2. SELECTIVE ENCRYPTION

Our goal with selective encryption is to achieve sufficient and transparent encryption while conserving the scalability in the encrypted domain. If we were to use regular encryption we would have to decrypt the bitstream prior to scaling and reencrypt it afterwards, which of course also requires that we have the key at the node which does the scaling. With our proposed method we can directly perform scaling on the encrypted bitstream, which not only saves time (since we can skip the de- and encryption steps), but also simplifies key management since we now only need the key at the endpoints of the channel. However, assuming that the unencrypted bitstream is our plaintext and the selectively encrypted bitstream is the ciphertext, then some portions of the ciphertext are copies of the plaintext. This means that perfect security, as specified by Shannon, can not be achieved, as this would require a full traditional encryption with a state of the art cipher.

A preview is naturally a lower quality version of the original sequence, but so is a downscaled version for a device which has a limited resolution. For example, the preview sequence of a HD video might be even better than the normal quality of the sequence if it is viewed on a mobile phone. This dichotomy cannot be readily resolved since really low level end devices border the region to sufficient encryption, e.g. a preview for a video sequence on a cell phone may not be viewable at all. And versions which could be considered preview sequences on a hand held device might be regarded as unviewable when watched on HD ready devices, e.g. when upscaling a sqCIF version of the sequence to a HD resolution the occurring pixelation will effectively degrade quality.

2.1 Bitstream

A schematic overview of the MC-EZBC bitstream is given in fig. 1 and an illustration of the decomposition of a GOP is given in fig. 3. The main layout is a header followed by GOP sizes (this is the size of the image data in a GOP) followed by a sequential ordering of GOPs. Each GOP is lead by a header, giving scene change information, i.e. which frames are I frames, followed by the motion field and image data. For both motion field and image data the frames are kept separate, i.e. no interleaving of frames, and frames are ordered lowest to highest temporal resolution (which is equal to lowest to highest temporal frequency bands). Likewise for each frame the image data is stored from lowest to highest resolution (which is equal to lowest to highest spatial frequency bands). Each base layer and each enhancement layer is stored as chunk of data (not shown in the figure), meaning a leading header giving the length of the data block followed by the data block itself.

For a parsing of the bitstream the layout into chunks is beneficial since we do not have to search for marker sequences but can directly skip large parts of the file. Also when headers, including chunk headers, and GOP size information is kept intact the whole bitstream can subsequently be parsed correctly, which is important to be able to scale after the encryption. In our context the encryption of image data is called *selective encryption*, i.e. we do not encrypt headers, motion fields or chunk size. From the remaining

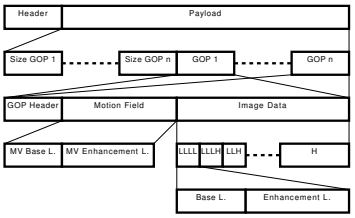


Figure 1: The layout of the MC-EZBC bitstream

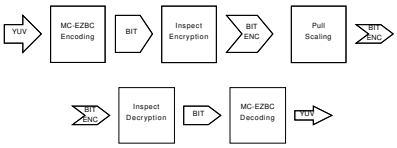


Figure 2: Workflow of the encoding, scaling, decoding process with encryption

data, which constitutes about 99% of the bitstream we can choose what to actually encrypt. If we choose to encrypt all we will denote that *full selective encryption*, if we choose to encrypt a subset we denote that as *partial selective encryption*. The size of the data in chunks is not aligned in any way and scaling happens in the image data chunks. As such we need an encryption scheme which can encrypt arbitrary block length and which does not reorder bytes, e.g. no ciphertext stealing. Given this information the choice of AES in OFB mode seems reasonable, since OFB mode has the desired properties of keeping the bitstream in order while AES is a well known state of the art cipher. Note that every cipher which does not rearrange bytes and can be cut of is useable here, e.g. basically every stream cipher. Since the visual data is easily accessible in the bitstream it seems to be a good choice to separate encryption and encoding, resulting in the work flow shown in fig. 2. The program `pull` was provided with the MC-EZBC source and does bitstream adaption, `inspect` is our tool to view the layout of the bitstream, encrypt and attack it.

2.2 Scaling Performance Analysis

The computational performance of selective encryption vs. traditional encryption is discussed controversially in literature. Basically, parsing and locating of what to encrypt generates an overhead and often a full traditional encryption is faster, especially with fast ciphers like AES. One can of course claim that the added advantage of keeping the ability to scale in the encrypted domain is worth the tradeoff of 'slow' encryption but it is still interesting to see how well we do.

2.2.1 Runtime Overview

Table 1 shows an overview of a full run through the work flow outlined above, and shown in fig. 2. The sequence en-

Table 1: Performance of the various steps in the work flow for the Flower sequence with a total of 128 frames and GOP size 128.

encoding	15m 47s 33ms	97.67%
encryption	148ms	0.02%
scaling	96ms	0.01%
decryption	50ms	0.01%
decoding	22s 344ms	2.30%
total	16m 9s 671ms	100.00%

coded was the well known flower sequence with a total of 128 frames and a temporal resolution of 7, resulting in a GOP size of 128. The highest quality version of the sequence is encrypted (all image data but no headers or motion vectors), then the sequence is downscaled to 128kbps (in the encrypted domain) and subsequently decrypted. What we see is that compared to encoding, and even decoding, the encryption and decryption process is extremely fast, and scaling is likewise. However, in terms of performance we should rather look at the absolute values, since if a bitstream is given (e.g. in retrieval scenarios like video on demand) encoding is not considered. For the highest quality version of the sequence we can encrypt, or decrypt, with a speed of roughly 1.15ms/frame and for the 128kbps version we have about 0.4ms/frame for full selective encryption. This translates to a throughput of about 870 frames per second for the full quality stream and 2500 frames per second for the downscaled version.

2.2.2 *Traditional vs. Full Selective Encryption*

While overall the performance is quite good the question remains how the full selective encryption process compares to full traditional encryption when scaling is applied. Taking the same high quality flower bitstream as above we perform full traditional encryption and full selective encryption, where the latter amounts to 99.41% of this bitstream. The encrypted bitstream is then downscaled. For traditional encryption we need to decrypt the bitstream prior to scaling and reencrypt it after scaling was performed. For full selective encryption we can directly scale the encrypted bitstream.

Full traditional encryption takes 114ms and full selective encryption takes 148ms, resulting in a speedup of 0.77. So if we do not scale the full traditional encryption is faster. Full selective encryption encrypts nearly the same amount of data as traditional encryption and also has a parsing overhead.

When we perform scaling however full selective encryption is faster since we can skip the decryption and encryption steps before and after scaling. Scaling takes 96ms for both encryption methods. With traditional encryption we have to decrypt before (114ms) and encrypt after (39ms) scaling. Thus, we get a total of 249ms for traditional encryption and 96ms for full selective encryption resulting in a speedup of 2.59.

The performance of partial selective encryption will be discussed in section 3.3.

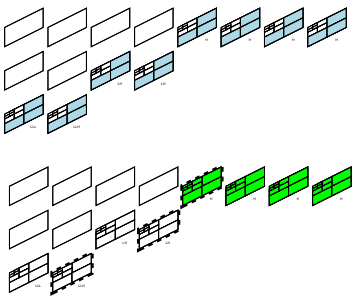


Figure 3: Overview of the decomposition of a GOP with GOP size 8 with marked high temporal layer (lower part), high spatial layer (upper part) and possible I frames as dashed outline on the lower part.

3. EXPERIMENTAL RESULTS

Since the various parts of the bitstream are basically wavelet decomposed signals we have a clear idea what to encrypt using partial selective encryption. For sufficient encryption we will target the low frequency bands, both temporally and spatially, as well as I frames. For transparent encryption we will encrypt the high bands, reducing the detail level of the sequence. To illustrate this, fig. 3 shows an overview of the decompositions, the marked frames in the lower part show the highest temporal band versus the highest spatial band in the upper part. The lower part of the figure also depicts possible I frames after the first frame in the GOP. Of course feeding a random signal into the arithmetic decoder will produce visual garbage in any case so it is expedient to consider an attack on the encrypted video sequence. This provides us not only with more insight into how well the sufficient encryption does but also gives us a method to remove the encrypted part of the sequence for the generation of the preview video for transparent encryption.

There are a number of possible attacks in literature. For an overview of selective encryption and attacks see Engel et al. [7] for JPEG2000, and Lookabaugh et al. [15] for MPEG-2. Specifically there are attacks which copy structurally similar symbols from one part of the bitstream to another or inject a forged version into the bitstream. This aims at removing the distortion introduced by decoding the encrypted bitstream or making decoding possible at all. These attacks also try to improve the resulting quality of the attack by forging the injected part of the bitstream in a way to minimize the decoding error. In literature such an attack is known as *error concealment attack* or *replacement attack*, a detailed description of such an attack can be found in Podesser et al. [20].

We will consider the error concealment attack of nulling out the encrypted part of the sequence. This basically exploits the fact that the arithmetic coder then maps the attacked part of the sequence to the most common output. While this also messes up the length of the bitstream segment with regard to the decoder we can still use it since the length is explicitly given. This allows the decoder to

properly reset after the attacked part of the sequence and continue the proper decoding. Also note that although in the still images presented here structural information may not, or only hardly be visible, the structure can often be seen better when the actual video sequence is seen in motion. So even if the attacked images sometimes give the impression that we have achieved confidential security, this is not the case. Also note that we will use the encryption only on selected parts of the image like the low temporal bands to get a better idea how this influences the video sequence, while in an actual application scenario one would probably mix these encryption schemes, e.g. encrypt low temporal and spatial bands at the same time.

The sequences used in this section will be Container and Waterfall, both with a length of 256 frames and a GOP size of 256 (leading to 8 temporal levels) with CIF resolution. No scaling was performed and a full quality sequence was used as base for the experiments.

3.1 Sufficient Encryption

For sufficient encryption we to target the parts of the bitstream which codes the visually most significant data. The codec exploits redundancy and inter frame dependencies and concentrates the high information content of the video in the lower frequency bands, both temporal and spatial. The low frequency frames effect all frames in their GOP through the wavelet synthesis and are thus prime targets for sufficient encryption. Likewise, the I-frame introduce information into the current GOP and effect frames in a pyramidal fashion (stemming from temporal decomposition). This makes I frames also good candidates for sufficient encryption. In the following we will look at the influence of I frames and low frequency frames for sufficient encryption. Each possibility will be evaluated on its own to better gauge the effect it has on the resulting video quality.

3.1.1 I Frame Encryption

To encrypt I frames is a good way to conceal a high amount of visual information. Figure 4 shows the PSNR per frame plot for the Container and Waterfall sequences for the baseline, encrypted and attacked version of the stream. Here the encrypted version is a decoding of the stream without prior attack or decryption, the attacked version has the encrypted parts of the bitstream nulled prior to decoding it. Depending on the sequence the attack can only obtain a limited amount of information: for Container which is a slow pan most information is stored into the motion field so naturally the refinement information has less energy. The Waterfall sequence on the other hand is a zoom which cannot be compensated as well by the motion estimation and this is clearly visible in the attacked version where we basically have a comparison of the refinement information with the original sequence. For a comparison of image quality between Container and Waterfall see fig. 5. In any case as can be seen from the PSNR plot the visual quality can be considered to be sufficiently degraded for our purpose, and even our attack hardly improves the visual quality.

3.1.2 Low Frequency Band Encryption

The next part of the bitstream which contains a high amount of information are the low frequency bands, temporal as well as spatial. Both are good candidates for encryption.

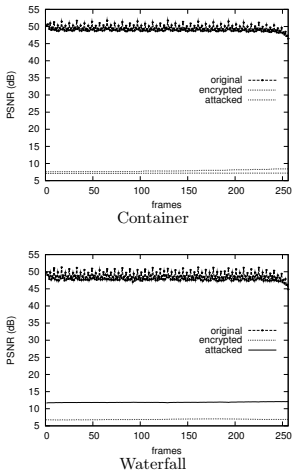


Figure 4: PSNR per frame plot for the Container and Waterfall sequences for encrypted and attacked I frames.

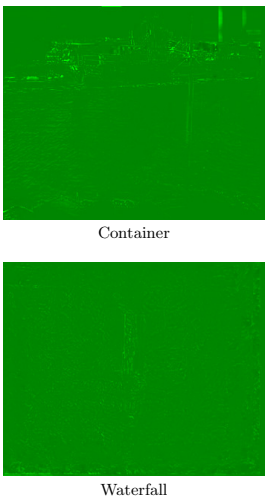


Figure 5: Frame 128 of the Container and Waterfall sequence with encrypted and attacked I frames.

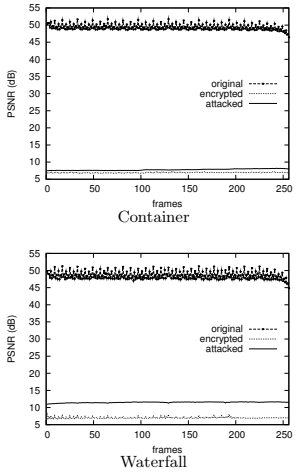


Figure 6: PSNR per frame plot for the Container and Waterfall sequence for encrypted and attacked low spatial frequencies.

The PSNR over frame plots for the encryption of low frequency spatial bands for both sequences, again original, attacked and encrypted versions, are given in fig. 6. The PSNR plot looks quite similar to the I frame case, as we actually did encrypt parts of the I frames as well. The advantage of encrypting the low frequency bands is of course that we also encrypt large parts of the temporal refinement information. To get a rough idea of how much information is left, fig. 7 shows frame 128 for the Container sequence in encrypted, decoded and attacked version. The encrypted version is a garbled output which stems from the fact that we actually input a random signal into the arithmetic decoder. The attacked image in this case looks rather inconspicuous but still gives of quite a bit of information when it is viewed as a motion sequence. This is also the main distinction between encrypted I frames and encrypted low spatial frames. The I frame version shows a much clearer attacked image where edges can be directly identified while the low spatial frequency version really needs motion to properly recognize structure. This can be easily seen when comparing the attacked Container sequence in fig. 7 (low spatial bands) and fig. 5 (I frames).

Encrypting the low temporal frequencies we expect something similar to the I frame version since GOPs in the MC-EZBC bitstream start with I frames, this coincides with the lowest temporal frequency. The PSNR plot for Container and Waterfall can be seen in fig. 9 and frame 128 of the decoded, attacked and encrypted version of the Waterfall sequence can be seen in fig. 8(a). What we can clearly see, and which was to be expected, is that for the Waterfall sequence, which contains a scene change, the PSNR rises after

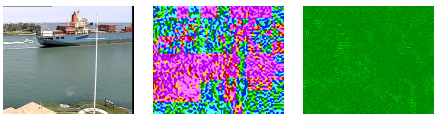
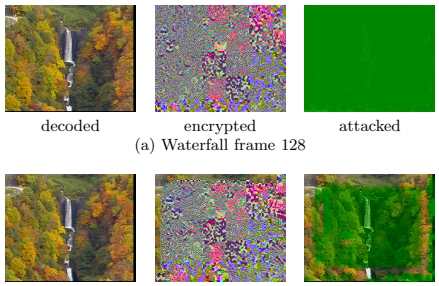


Figure 7: Comparison of encrypted, decrypted and attacked image to the original of frame 128 from the Container sequence (low spatial frequencies).



(a) Waterfall frame 128

(b) Waterfall frame 175

Figure 8: Comparison of encrypted, decrypted and attacked image to the original of frame 128 and 175 from the Waterfall sequence (low temporal frequencies).

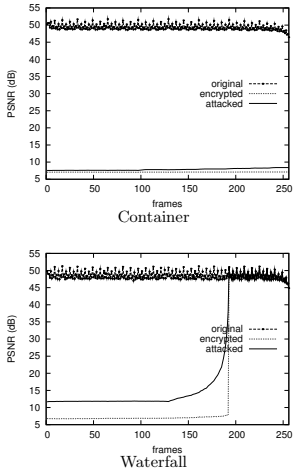


Figure 9: PSNR per frame plot for the Container and Waterfall sequence for encrypted and attacked low temporal frequencies.

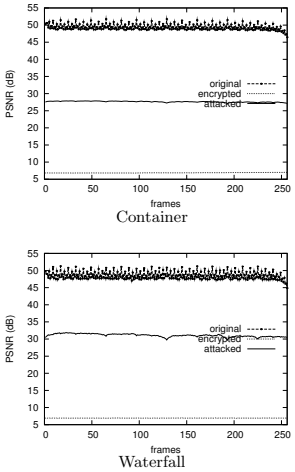


Figure 10: PSNR per frame plot for the Container and Waterfall sequence for encrypted and attacked high spatial frequencies.

the I frame in the later part of the sequence. Figure 8(b) shows frame 175 of the above versions for the Waterfall sequence, the influence of the I frame is clearly visible. The difference to the encrypted low spatial frequencies is rather obvious. Low temporal frequencies are full leading frames of the GOP, while low spatial frequencies are the low frequency information of all frames. Thus, low spatial frames include all I frames while low temporal frequencies only include leading I frames. Apart from the fact that we cannot ignore I frames when encrypting low temporal frames we can clearly see that encrypting low temporal frames also sufficiently destroys the visual quality. However, since we have to encrypt all I frames in addition to the low temporal frequencies it is usually sufficient to either encrypt I frames or low spatial frequencies (with or without full I frame encryption). Encryption of low spatial bands give a substantial gain vs. encryption of I frames only because they further destroy the visual quality of the difference frames. All versions however are sufficient to destroy the visual quality, while none gives confidential encryption.

3.2 Transparent Encryption

For transparent encryption the refinement information, residing in high frequency temporal and spatial bands, can be encrypted. The optimal solution would be to be able to completely choose a target PSNR for the preview image, this is not possible however since we only have a limited amount of steps, i.e. the decomposition depth of the sequence. However, adaption in this rough scale is possible and should be done.

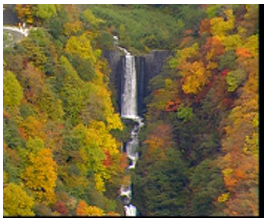
3.2.1 High Spatial Frequency Bands

Figure 10 gives the PSNR plot for both Container and Waterfall sequences, with attacked highest spatial frequency band. The drop in PSNR is clearly visible, the impact on the visual quality is not quite as obvious however, as illustrated in fig. 11(a), for the Container sequence. What we can see is that our attack (in this case rather preview image generation) is working well. However, the reduction in visual quality is not really as high as expected. To remedy this we will have to encrypt an additional layer of the decomposition. Figure 12 gives an overview what changes in this case for the Waterfall sequence. Now the degradation in visual quality is clearly visible, even though the PSNR dropped only an additional 5 dB. This also gives an impression of the scale on which we can adjust the visual quality with this method.

3.2.2 High Temporal Frequency Bands

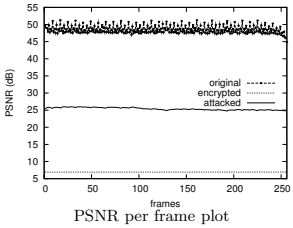
For high temporal bands the matter is a bit different. While spatial bands directly affect image quality, temporal bands do so to a lesser degree. They influence visual quality of course through blurring effects stemming from temporal filtering, but the main effect is a reduction in temporal resolution, i.e. frames per second. This can only partially be shown in a PSNR plot and still images, but nonetheless fig. 13 shows the PSNR plots for Container and Waterfall where the highest two (of eight total) temporal bands are encrypted. The visual impact can be seen in fig. 11(b), for the Waterfall sequence, the main effect being blurring which can be best seen at the waterfall itself. To show a stronger version of the blurring effect we also did a version where the

(a) Container Preview

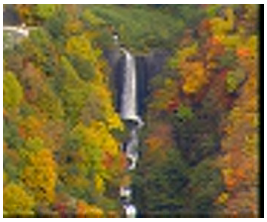


(b) Waterfall Preview

Figure 11: Preview image of the Container, high spatial frequencies, and Waterfall sequence, high temporal frequencies, frame 128.

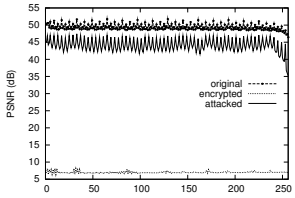


PSNR per frame plot

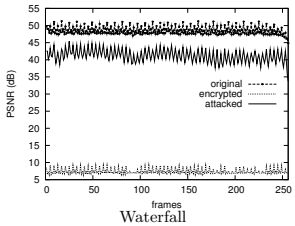


frame 128 of the preview sequences

Figure 12: PSNR per frame plot for the Waterfall sequence and frame 128 of the preview (two highest spatial frequency bands encrypted) sequence.



Container



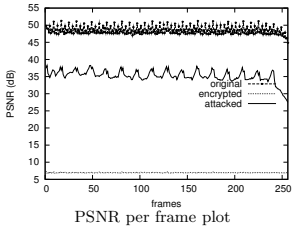
Waterfall

Figure 13: PSNR per frame plot for the Container and Waterfall sequence for encrypted and attacked high temporal frequencies.

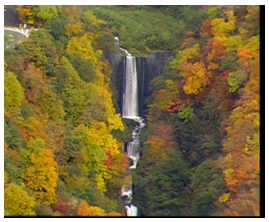
highest four (half total) temporal frames are encrypted, seen in fig. 14 (both PSNR and visuals). When looking at the PSNR plots we can see some spikes in the preview image. These are the images in the temporal sequence which best fit the original sequence, the degradation following these frames stems from the fact that the still frames are simply not changed but the original sequence continues and moves away from the good fitting frame. Overall the visual viewing quality is heavily impaired since a bucking effect with blurring is introduced when a sufficient number of temporal frames are encrypted. However, when only the highest temporal layer is encrypted the skipping of every second frame is actually nearly not noticeable since the merging of temporally adjacent frames partly conceals the missing frame. It should also be noted that the highest temporal band is actually a full half of all frames, leading to a high amount of data to be encrypted when comparing this to the spatial case.

3.3 Partial Selective Encryption Performance

Instead of giving the information about encryption time and amount individually in each section, we have collected the information for all tests in table 2 for easier comparison. The sequence used was waterfall with 256 frames and GOP size 256 as well. The information given is for encryption only, scaling is not taken into account here since it was already discussed in section 2.2. The interesting thing to note here is that as soon as we take a step away from full selective encryption we are actually faster than full traditional encryption. While this was not our main concern it was a definite side target of sufficient encryption. Transparent en-



PSNR per frame plot



Frame 128 of the preview sequence

Figure 14: PSNR per frame plot for the Waterfall sequence and frame 128 of the preview (four highest spatial frequency bands encrypted) sequence.

Table 2: Performance comparison of the various selective encryption methods and full traditional encryption.

What was encrypted	time	% of Bitstream
Sufficient Encryption		
I-frames only	28ms	5.52%
lowest spatial band	99ms	34.79%
lowest temporal band	21ms	5.47%
Transparent Encryption		
highest spatial band	181ms	91.58%
two highest temporal bands	148ms	72.53%
four highest temporal bands	181ms	89.97%
Full Encryption		
full selective encryption	217ms	99.76%
full traditional encryption	201ms	100%

cryption, while faster than full selective or even traditional encryption, is slower than sufficient encryption. This is not surprising since the refinement layers, i.e. higher temporal bands, are significantly larger than the lower frames. Since transparent encryption has to target those high bands the amount of data to be encrypted is increased.

For sufficient encryption we have seen that the encryption of the lowest spatial bands performs best in terms of destroying visual quality followed closely by I-frames only. In terms of speedup we have about 2 for lowest spatial bands and more than 7 for I-frames when compared to full traditional encryption. When considering that even the sequence with encrypted I-frames is practically unusable, the choice is obviously the I-frame version since it gives the higher speedup.

For transparent encryption the speedup for spatial and temporal bands is about the same when we want to achieve a similar quality. Given that encryption speed is not even an objective for transparent encryption we can easily state that both versions are quite applicable. The real choice which to use thus is not performance but rather target quality.

4. CONCLUSION

We have introduced different ways to selectively encrypt the MC-EZBC bitstream with regard to transparent as well as sufficient encryption while being able to scale the bitstream in the encrypted domain. The proposed encryption schemes are fast and computationally cheap. Furthermore, the proposed encryption schemes meet all requirements of UMA while keeping security intact.

Concerning sufficient encryption we have shown that the destruction of the visual quality can easily and efficiently be achieved, but one has to be aware that encrypting low temporal bands is not enough, i.e. I frames have to be included. Overall the best practice is to either use I frames, low spatial bands or both combined, since I frames contain all the base layer information and low spatial frames contain the highest amount of energy from base and enhancement layers. For sufficient encryption we also achieved a gain in computational performance, e.g. when using only low spatial bands we require less than half the time of full traditional encryption.

Concerning transparent encryption we have shown that it is possible to achieve a reduction in quality by encrypting high spatial and frequency bands. While both methods are rather limited when it comes to possible output qualities, when combining both we have a sufficient number of possible quality steps. Assuming three spatial and eight temporal bands we would have a total of 24 possible output qualities. One should note however that, while reduction of visual quality through spatial encryption can easily be quantified this is not so simple for temporal bands, mainly because we are lacking a proper metric to measure bucking and lagging behavior in video sequences, except from the blurring which can be clearly seen in the PSNR plots. Concerning computational performance we can only register a slight improvement over full traditional encryption.

In future work we will look at the encryption of the motion fields, and closer investigate if it is possible to achieve full security, i.e. confidentiality, by encrypting motion fields as well as visual data. Furthermore, the use of a technique similar to the sliding window approach Stütz et al. introduced for JPEG2000 [24] would be beneficial to reduce the computational performance of transparent encryption.

5. REFERENCES

[1] N. Adami, A. Signoroni, and R. Leonardi. State-of-the-art and trends in scalable video compression with wavelet-based approaches. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(9):1238–1255, Sept. 2007.

[2] C. Bergeron and C. Lamy-Bergor. Compliant selective encryption for H.264/AVC video streams. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing, MMSP'05*, pages 1–4, Oct. 2005.

[3] P. Chen, K. Hanke, T. Rusert, and J. W. Woods. Improvements to the MC-EZBC scalable video coder. In *Proceedings of the IEEE Int. Conf. Image Processing ICIP*, Barcelona, Spain, 2003.

[4] P. Chen and J. W. Woods. Bidirectional MC-EZBC with lifting implementation. In *IEEE Transactions on Circ. and Systems for Video Technology*, volume 14, pages 1183–1194, 2003.

[5] H. Eeckhaut, H. Devos, P. Lambert, D. De Schrijver, W. Van Lancker, V. Nollet, P. Avasare, T. Clerckx, F. Verdicchio, M. Christiaens, P. Schelkens, R. Van de Walle, and D. Stroobandt. Scalable, wavelet-based video: From server to hardware-accelerated client. *Multimedia, IEEE Transactions on*, 9(7):1508–1519, Nov. 2007.

[6] D. Engel, T. Stütz, and A. Uhl. Format-compliant JPEG2000 encryption in JPSEC: Security, applicability and the impact of compression parameters. *EURASIP Journal on Information Security*, (Article ID 94565):doi:10.1155/2007/94565, 20 pages, 2007.

[7] D. Engel, T. Stütz, and A. Uhl. A survey on JPEG2000 encryption. *Multimedia Systems*, 2009. to appear.

[8] S.-T. Hsiang and J. W. Woods. Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank. *Signal Processing: Image Communication*, 16(8):705–724, May 2001.

[9] R. Iqbal, S. Shirmohammadi, and A. E. Saddik. Compressed-domain encryption of adapted H.264 video. In *Proceedings of the 8th International Symposium on Multimedia, ISM'06*, pages 979–984, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[10] R. Kuschnig, I. Kofler, M. Ransburg, and H. Hellwagner. Design options and comparison of in-network H.264/SVC adaptation. *Journal of Visual Communication and Image Representation*, Sept. 2008.

[11] H.-J. Lee and J. Nam. Low complexity controllable scrambler/descrambler for H.264/AVC in compressed domain. In K. Nahrstedt, M. Turk, Y. Rui, W. Klas, and K. Mayer-Patel, editors, *Proceedings of ACM Multimedia 2006*, pages 93–96. ACM, 2006.

[12] Y. Li, L. Liang, Z. Su, and J. Jiang. A new video encryption algorithm for H.264. In *Proceedings of the Fifth International Conference on Information, Communications and Signal Processing, ICICS'05*, pages 1121– 1124. IEEE, Dec. 2005.

[13] L. Lima, F. Manerba, N. Adami, A. Signoroni, and R. Leonardi. Wavelet-based encoding for HD applications. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1351–1354, July 2007.

[14] T. D. Lookabaugh and D. C. Sicker. Selective encryption for consumer applications. *IEEE Communications Magazine*, 42(5):124–129, 2004.

[15] T. D. Lookabaugh, D. C. Sicker, D. M. Keaton, W. Y. Guo, and I. Vedula. Security analysis of selectiveley encrypted MPEG-2 streams. In *Multimedia Systems and Applications VI*, volume 5241 of *Proceedings of SPIE*, pages 10–21, Sept. 2003.

[16] X. Lu and A. M. Eskicioglu. Selective encryption of multimedia content in distribution networks: Challenges and new directions. In *Proceedings of the IASTED International Conference on on Communications, Internet and Information Technology (CIIT 2003)*, Scottsdale, AZ, USA, Nov. 2003.

[17] E. Magli, M. Grangetto, and G. Olmo. Conditional access to H.264/AVC video with drift control. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME'06*. IEEE, July 2006.

[18] E. Magli, M. Grangetto, and G. Olmo. Conditional access techniques for H.264/AVC and H.264/SVC compressed video. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008. to appear.

[19] A. Massoudi, F. Lefèbvre, C. D. Vleeschouwer, B. Macq, , and J.-J. Quisquater. Overview on selective encryption of image and video, challenges and perspectives. *EURASIP Journal on Information Security*, 2008(Article ID 179290):doi:10.1155/2008/179290, 18 pages, 2008.

[20] M. Podesser, H.-P. Schmidt, and A. Uhl. Selective bitplane encryption for secure transmission of image data in mobile environments. In *CD-ROM Proceedings of the 5th IEEE Nordic Signal Processing Symposium (NORSIG 2002)*, Tromso-Trondheim, Norway, Oct. 2002. IEEE Norway Section. file cr1037.pdf.

[21] A. Said. Measuring the strength of partial encryption schemes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'05)*, volume 2, Sept. 2005.

[22] C. E. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656–715, Oct. 1949.

[23] T. Shi, B. King, and P. Salama. Selective encryption for H.264/AVC video coding. In E. Delp and P. Wong, editors, *Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VIII*, volume 6072 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 461–469, Feb. 2006.

[24] T. Stütz and A. Uhl. On efficient transparent JPEG2000 encryption. In *Proceedings of ACM Multimedia and Security Workshop, MM-SEC '07*, pages 97–108, New York, NY, USA, Sept. 2007. ACM Press.

[25] A. Vetro, C. Christopoulos, and T. Ebrahimi. From the guest editors - Universal multimedia access. *IEEE Signal Processing Magazine*, 20(2):16 – 16, 2003.

[26] D. Wu, Y. T. Hou, W. Zhu, Y.-Q. Zhang, and J. M. Peha. Streaming video over the internet: approaches and directions. In *Circuits and Systems for Video Technology, IEEE Transactions on*, volume 11, pages 282–300, Mar 2001.

Cost of In-Network Adaption of MC-EZBC for Universal Multimedia Access

Heinz Hofbauer
Salzburg University
Department of Computer Sciences
hhofbaue@cosy.sbg.ac.at

Andreas Uhl
Salzburg University
Department of Computer Sciences
uhl@cosy.sbg.ac.at

Abstract

A core concept of universal multimedia access is the use of scalable content. The scalable content should be as flexible as possible to achieve the credo of creating the content once and adapting it to fulfill given requirements. As far as flexibility goes wavelet based codecs are superior. A problem which arises is that the adaptation does not necessarily happen on a device which is aware of the codec which was used in content creation. To rectify this non-awareness, MPEG-21 introduced digital item adaptation in part 7 to abstract the bitstream of a given video. This allows in-network adaptation on nodes which are DIA aware to adapt any video stream as long as a DIA description is given. The drawback is that the DIA description must be sent parallel to the original video sequence. In this paper we will look at how a DIA description for a t+2D scalable wavelet codec looks like. We will evaluate the possibilities we have with various description options and we will also look at the overhead generated by the DIA description.

1 Introduction

The use of digital video in todays world is ubiquitous. Videos are viewed on a wide range of clients, ranging from hand held devices with QVGA resolution (320x240) over PAL (768x576) or NTSC (720x480) to HD 1080p (1920x1080) or higher. Furthermore, streaming servers should be able to broadcast over the internet with regard to a wide range of bandwidths, from fixed high bandwidth lines like ADSL2 to changing low bandwidths for mobile wireless devices. In such an environment it is simply not possible to encode a video for every application scenario. So content providers either have only a fixed number of options available or they use scaling video technology to adapt the video for bandwidth and resolution requirements of the client. The concept of creating the content once and adapting it to the current requirements is preferable and is better known as Universal Multimedia Access (UMA) [10].

One of the enabling technologies of UMA is the use of scalable video coding. This averts the need for transcoding on the server side and enables the server to scale the video. However, even scaling takes up computation time and reduces the number of connections the server can accept. Furthermore, variable bandwidth conditions, which happen frequently on mobile devices, further taxes the server with the need to adapt the video stream. The solution to this is usually in-network adaptation, shifting the need to scale to the node in the network where a change in bandwidth is occurring. Figure 1 shows an example of this scenario, where a mobile device requests a video stream from the server which fits its capabilities. The core adaptation with these restrictions takes place on the server and on the fly adaptation due to actual channel capability is done in-network. Wu et al. [11] give an overview of other aspects of streaming video ranging from server requirements to protocols, to QoS etc.

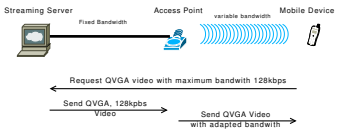


Figure 1. Example of video adaptation for a mobile device on the server and in the network.

For video streaming in this environment, i.e. a high number of possible bandwidths and target resolutions, wavelet based codecs can be considered. Wavelet based codes are naturally highly scalable and rate adaptation as well as resolution or temporal scaling is easily achieved. Furthermore, wavelet based codecs achieve a coding performance similar to H.264/SVC, c.f. Lima et al. [7].

For this reason we will consider the ENH-MC-EZBC wavelet based video codec for in-network adaptation. This choice was made mainly because the source code is available [1], which enables our experiments. The MC-EZBC codec [4, 12] is a scalable t-2D video codec which uses motion compensated temporal filtering, with 5/3 CDF wavelets, followed by regular spatial filtering, with 9/7 CDF filtering, see fig. 2 for a GOP size of 8. This method, temporal first and spatial later, is referred to as t+2D coding scheme. For temporal filtering a full decomposition is used and thus the GOP size is discernible by the number of tem-

[1] The source for the ENH-MC-EZBC is available from http://www.cipr.rpi.edu/research/mcezbc/.

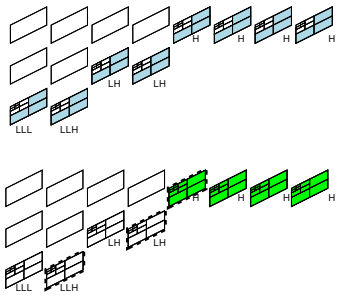


Figure 2. Overview of the decomposition of a GOP with GOP size 8 with marked (gray parts) high temporal layer (bottom), high spatial layer (top) and possible I-frames as dashed outline on the lower half.

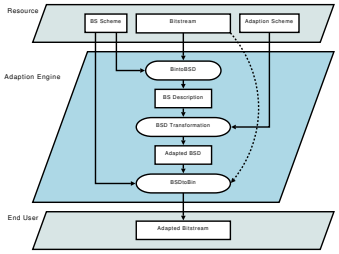


Figure 3. Overview of the adaptation process using the bitstream syntax description language.

poral decomposition levels. Both temporal and spatial filtering is done in a regular pyramidal fashion. Statistical dependencies are exploited by using a bit plane encoder, the name giving embedded zero bit coder. Motion vectors are encoded with DPCM followed by an arithmetic coding scheme.

For an overview about wavelet based video codecs and a performance analysis as well as techniques used in those codecs see the overview paper by Adami et al. [1].

The scalability of the video codec is important for UMA which means it is also necessary for servers and network nodes to be able to perform scaling. This can either be achieved by making them aware of the video codec, which would make upgrading to a different codec later quite troublesome, or by abstracting the actual bitstream. MPEG-21 gives a specification how such an abstraction has to look like. Part 7 of MPEG-21 [5] deals with Digital Item Adaption (DIA), more precisely it specifies a Bitstream Syntax Description Language (BSDL) which is based on the XML schema as specified by the W3C. The idea behind DIA with BSDL is that a syntax description of the bitstream is available and can be used to extract a XML description of the bitstream. On this abstraction of the bitstream the scaling is performed and mapped back to the original bitstream via the BSDL, see fig. 3 for an illustration of the process. As a result each node in the network only needs to be capable of understanding and handling DIA as per MPEG-21 part 7. For bitstreams which do not follow a marker based syntax, specifically if parsing the bitstream would be required to generate a description, the approach using BSDL does not work. For this cases a generic bitstream syntax description (gBSD) is available in MPEG-21 part 7 (see Panis et al. [8]) which can be used to directly describe the bitstream.

Usually when research is done on in-network adaptation the focus is on client and server layout as well as computational demand on the network node which performs scaling. That is, memory consumption on the network node and the scaling options as well as resulting video quality under those scaling options are evaluated, e.g. Eeckhaut et al. [3]. What is missing, especially when considering gBSD rather than BSDL, is that the transfer of the bitstream description also takes up bandwidth. Essentially for a fixed bandwidth the use of a bitstream description reduces the available bandwidth for the actual bitstream which in turn results in a reduction of video quality. For similar research regarding h.264/SVC see Kuschnig et. al. [6].

In section 2 we will describe the MC-EZBC bitstream in such detail as is necessary to map it to gBSD and provide possible gBSD descriptions of the bitstream. Section 3 will look at the overhead generated by gBSD and section 4 will give a conclusion and outlook.

2 Mapping an MC-EZBC Bitstream to gBSD

In the following the bitstream of the MC-EZBC will be described with regard to the gBSD mapping. Then we will give a brief description of the gBSD elements we use to map the bitstream to gBSD and give two possible mappings. The scaling option we want to maintain for in-network adaptation reflects which information we will need in the gBSD. We will focus on the two reasonable end points of the spectrum, i.e. full scalability in order to retain the advantage the wavelet based codec has vs. regular rate-distortion scaling with a limited number of scaling points reflecting the application scenario given in fig. 1.

2.1 MC-EZBC Bitstream

The basic layout of the MC-EZBC bitstream is depicted in the upper part of fig. 4 and a more detailed overview of the 'image data' required for fine grain scalability is given in lower part.The bitstream is lead by a general header giving resolution, frame rate, prediction options etc., most of which stay the same during scaling. The header however has three fields we need to adjust when scaling is performed: a `bitrate` field giving the bit rate to which the bitstream is scaled, `t_level` giving the number of temporal layers dropped and `s_level` giving the number of

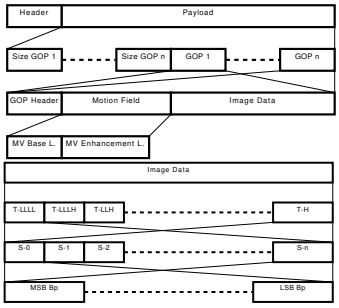


Figure 4. Layout of the MC-EZBC bitstream.

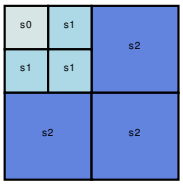


Figure 5. Grouping of decompositions for a frame with two spatial decomposition levels.

spatial layers dropped. The header is followed by a GOP size list giving the size of a GOP without GOP header size and motion field, i.e. only image data size. For any scaling done the GOP size list has to be adjusted to reflect the new size of image data.

Following this general information are the motion and image data ordered by GOPs in increasing order. Each GOP contains a GOP header, basically giving scene change information, i.e. which frames are encoded as I frames. Following the GOP header is the motion field for the current GOP. The GOP header and motion field are never changed during scaling, i.e. motion vectors are never scaled with the image data. Following the motion field is the image data in frame order of temporal decomposition, c.f. fig. 2 and fig. 4 lower part.

The layout of the image data consists of a number of data chunks consisting of size information and data. For each frame every spatial decomposition level is given as one chunk where color information and direction of decomposition are grouped together, fig. 5 illustrates this. The order of this chunks in the bitstream is from lowest subband to highest subband. For scaling, the size information of the chunks needs to be reset to the reduced data in the chunk, thus a description of the bitstream has to be at least down to the level of chunks. For a limited number of scaling options this would be enough since the chunk data can be subdivided into blocks which we can remove. However, if we want to retain the full scalability capability of the wavelet bitstream we have to go into more detail. In each chunk there is a three byte header which may never be removed for regular scaling, however when the whole resolution is dropped these three bytes can be dropped too. Then the data is ordered in terms of bitplanes, most significant to least significant. The reason we need the bitplane information is that the scaling algorithm performs quantization at a bitplane level, so for an implementation of the scaling algorithm the size information of the bitplanes is paramount.

2.2 gBSD Mapping

We use the gBSD from MPEG-21 DIA for describing the bitstream. While the gBSD allows more structural information to go into the description we will keep the bitstream description simple so as not to generate too much overhead.

The gBSD is prefaced with a `dia:DIA` root tag specifying namespaces followed by a `dia:Description` tag specifying the description type (`gBSDType`) followed by address information. Since the MC-EZBC bitstream is byte based we set it to `addressUnit="byte"` and `addressMode="Absolute"`. The address mode gives the method of accessing parts of the bitstream, this is reflected by the use of start and length attributes in subsequent tags. For the bitstream description we need two different types of tags. First we need a copy and paste descriptor stating that a part of the original bitstream should be carried over to the scaled version. The `gBSDUnit` tag is used for this purpose, we give start and length information to mark a part of the bitstream to be kept. Additionally we need access to the size information, in a scaling case this is not simply a copy from the original bitstream but needs to be adapted. The `Parameter` tag is used for this, which gives the length of the data block to insert into the bitstream. The actual information contained in the parameter is given by the required child `Value`. The attribute `xsi:type` gives the type of data and the content of the tag gives the actual value. By using parameter and value we can access the actual value and change it according to the adaptation, while the gBSDUnit tags let us copy parts of the actual bitstream. Both parameter and gBSDUnit also have an attribute `marker` which allows to give a handle to the tag to access it directly. For more information on the tags and attributes used see MPEG-21 part 7 [5].

First we will look at the description of the bitstream for a two case scenario to get lower bound for the limited case scenario. The temporal and spatial resolutions stay fixed and we give the option of scaling to 1024kbps and 512kbps. We only need to describe the bitstream down to the level of chunks of image data. Also, since we do not want resolution dropping the header description is simplified since only the bitrate field need to be changed. The GOP size list needs to be described by the parameter tag as it will change when scaling is done. Motion vectors and GOP headers however can be put together as one gBSDUnit since they are consecutive and will remain unchanged. Following the GOP headers are the chunks of image data, the header is described by a parameter tag and the data is described by using two gBSDUnits reflecting the two scaling options we want. One

```
...
<dia:Description xsi:type="gBSDType"
addressUnit="byte" addressMode="Absolute">
    <gBSDUnit start="0" length="14" marker="hdr1"/>
    <Parameter length="2" marker="bitrate Q0">
        <Value xsi:type="xsd:unsignedShort">1024</Value>
    </Parameter>
<gBSDUnit start="16" length="80" marker="hdr2"/>
...
    <Parameter length="2" marker="hdr Q0">
        <Value xsi:type="xsd:unsignedShort">118</Value>
    </Parameter>
    <gBSDUnit start="545775" length="18" marker="data"/>
    <gBSDUnit start="545793" length="100" marker="data Q0"/>
    <Parameter length="2" marker="hdr Q0">
        <Value xsi:type="xsd:unsignedShort">185</Value>
    </Parameter>
    <gBSDUnit start="545895" length="21" marker="data"/>
    <gBSDUnit start="545916" length="164" marker="data Q0"/>
</dia:Description>
</dia:DIA>
```

Figure 6. gBSD representation of the flower sequences scaled to 1024 kbps with marker for 512kbps.

```
...
<dia:Description xsi:type="gBSDType"
addressUnit="byte" addressMode="Absolute">
    <gBSDUnit start="0" length="14" marker="hdr1"/>
    <Parameter length="2" marker="bitrate Q1">
        <Value xsi:type="xsd:unsignedShort">512</Value>
    </Parameter>
<gBSDUnit start="16" length="80" marker="hdr2"/>
...
    <Parameter length="2" marker="hdr Q1">
        <Value xsi:type="xsd:unsignedShort">18</Value>
    </Parameter>
    <gBSDUnit start="545775" length="18" marker="data"/>
    <Parameter length="2" marker="hdr Q1">
        <Value xsi:type="xsd:unsignedShort">21</Value>
    </Parameter>
    <gBSDUnit start="545895" length="21" marker="data"/>
</dia:Description>
</dia:DIA>
```

Figure 7. gBSD representation of the flower sequence, downscaled to 512kbps.

gBSDUnit locates the data we need for the 512kbps version of the bitstream, the next describes the additional data for the 1024kbps case. The rest of the data, in case we start with a bitstream with bitrate greater than 1024kbps, does not need to be described since it will be cut out in case of scaling anyway. Figure 6 gives a part of the description of the bitstream which can be used to scale to 1024kbps. It also shows the description of the header where it can be seen that only the bitrate has to be described as parameter and that it needs to be set to 1024 to properly reflect the bitrate of the stream. The resulting description of the stream still consists of two gBSDUnit descriptions discerning between 512 and 1024 kbps. Compare this to fig. 7 which describes the bitstream for the 512kbps scaling case. The shown part of the description refers to the same section of the bitstream as the 1024kbps case. It is clear that the scaling of the bitstream also entails a scaling of the gBSD description. But, it also is clear that adding another scaling option for this fixed case requires the insertion of another gBSDUnit partition for each chunk.

The second scenario we want to look at is full grained scalability. For this case we have to render a finer description of the bitstream down to bitplane level. The overhead in the header is rather small, we just need to add the resolution drop fields as parameters. For the GOP size list, GOP headers and motion vectors nothing changes compared to the two case scenario. For the description of the image data chunks however we need a lot more detail and consequently a lot more gBSDUnit tags. First we need the subband header, which will not be changed in any case. This can either be described as an extra tag or can be contained in the first bitplane following the size information. We choose the latter version since the scaling algorithm must be aware of the header anyway when calculating the overhead. For the rest of the image data we have to model each bitplane as a separate gBSDUnit since the size of the bitplanes is required by the scaling algorithm. Despite the possibility that a bitplane is reduced in size we still can describe them with a gBSDUnit because the actual content does not change and a reduction in size can be achieved by resetting the length attribute. Figure 8 shows a part of the gBSD description for a downscaled version to 512kbps, the bitplane quantization can be clearly seen, i.e. the lowest bitplane in the figure is the 57th.

```
...
    <gBSDUnit start="272992" length="1" marker="data sp 59"/>
    <gBSDUnit start="272993" length="2" marker="data sp 58"/>
    <gBSDUnit start="272995" length="1" marker="data sp 57"/>
    <Parameter length="2" marker="hdr">
        <Value xsi:type="xsd:unsignedShort">19</Value>
    </Parameter>
    <gBSDUnit start="272998" length="6" marker="data sp 79"/>
    <gBSDUnit start="273004" length="3" marker="data sp 71"/>
    <gBSDUnit start="273007" length="1" marker="data sp 67"/>
    <gBSDUnit start="273008" length="3" marker="data sp 63"/>
    <gBSDUnit start="273011" length="5" marker="data sp 62"/>
    <gBSDUnit start="273016" length="1" marker="data sp 59"/>
...
```

Figure 8. Detailed gBSD representation of the flower sequence, downscaled to 512kbps.

In our examples the description is kept as simple as possible so as not to use up too much bandwidth. However, for an actual application it would be beneficial to retain some structure by nesting gBSDUnits. While this increases the size of the gBSD description it makes XSLT writing much easier and helps to avoid errors.

3 Evaluation of gBSD Overhead

In the two case scenario we can give a good approximation of the overhead. The framerate f in the video sequence and the encoded sequence stay the same but the number of GOPs is dependant on the temporal decompositions t. The number of spatial decompositions s depends on the resolution of the original and can change from sequence to sequence. We also have an approximate number of bytes each descriptive element of the gBSD requires. The number of bytes a Parameter p and gBSDUnit g require are 105 and 55 bytes respectively. This numbers are calculated with average variable length information (i.e. length value, start value), additionally we have a overhead for the DIA declaration which is 393 bytes. This means that the start and length fields as well as well as the value of parameters are only estimated since this information can vary widely. However, the use of a typical marker element is included since the marker will be a near constant in length. We can now calculate an approximate size of the gBSD. The main header consists of one changeable field with size p and two gBSDUnits of size g which stay constant. With a temporal resolution of t we have a GOP size of 2^t and the number of

Scenario	kbps	size	compressed
Bitstream	full	5.5M	
Bitstream	1024	540k	
Bitstream	512	272k	
Detailed gBSD	full	1.2M	112k
Detailed gBSD	1024	400k	32k
Detailed gBSD	512	268k	20k
Two case gBSD	full	84k	8k
Two case gBSD	1024	84k	8k
Two case gBSD	512	64k	4k

Table 1. Comparison of bitstream and gBSD file sizes for the flower sequence with 128 frames, GOP size of 128 and two spatial decompositions.

GOPs is $G = f/2^t$, for each GOP the header is followed by a GOP size entry as a parameter p. For each GOP we have a single gBSDUnit for the GOP header and motion vectors. Then for each frame we have a single chunk for each spatial decomposition level $(s+1)g$, i.e. if we do two spatial decompositions we have three subbands, see fig. 5. The chunks here have to be separated into the number of cases C we want to deal with. The resulting approximation in byte is thus size S:

$$S = 393 + \underbrace{p + 2g}_{\text{header}} + \underbrace{G * p}_{\text{GOP size list}} + G \underbrace{(g + 2^t(s+1)(p+Cg))}_{\text{single GOP}}$$

For a sequence with 128 frames, $t = 7$ and $s = 2$ this would estimate a gBSD file size of 81kb for the two case gBSD and 60kb for the downscaled version. This is compared experimentally to the actual file sizes of the flower sequence in table 1. The table gives the file size of the bitstream under the bitstream 'scenario', it also gives the two case gBSD file sizes scaled and unscaled. Note that the description for full is simply the complete gBSD containing both scenarios. When scaling to 1024kbps we still use the full description and for 512kbps the gBSD is reduced by the gBSDUnit sizes describing the 1024kbps parts of the bitstream. As can be seen the approximation given reflects the actual gBSD size quite accurately. Note that other sequences have a similar size for the limited case scenario, differing by less than 2%.

The transmission of the gBSD description will usually not be in plain text. XML which is the basis of gBSD can be compressed quite well, see Augeri et al. [2]. We used bzip2 to generate the compressed file sizes as given in table 1. While this may not be the best way to compress the data with regard to network nodes, where a XML aware compression scheme would be beneficial to save memory and time (see Timmerer et al. [9]), we will still use it as a baseline as it offers better compression.

For the detailed description case it is not possible to give a formula since the description, especially when scaling is performed, is heavily dependant on the layout of the bitplanes. The number and size of the bitplanes however vary widely depending on the content of the sequence. During scaling it is possible that a number of bitplanes with a low amount of data are removed which reduces the gBSD file size drastically. For a different sequence the same scaling operation could only reduce the number of bytes in the affected bitplanes which would leave the gBSD size nearly unchanged. Figure 9 gives a plot of gBSD file size over kbps for a number of sequences. What is interesting here is that the gBSD file size is higher for low motion videos like Akiyo. This is due to the fact that sequences like Akiyo can be predicted very well which results in a low file size. Consequently, when scaling to a certain bitrate is performed, fewer bitplanes have to be dropped and likewise fewer bitplane descriptions are deleted. This results in a larger gBSD file when compared to high motion sequences.

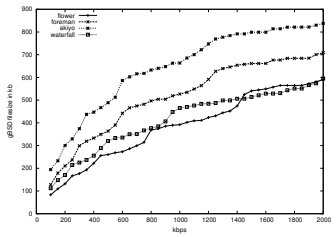


Figure 9. File size of the detailed gBSD description depending on the bitrate for a number of sequences, each with 128 frames and one GOP.

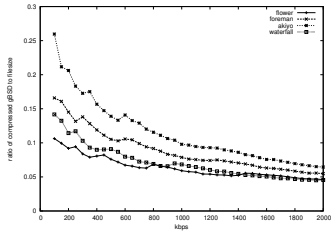


Figure 10. Ratio of compressed gBSD descriptions to bitstream size plotted over kbps for various sequences.

Figure 10 shows an overview over the ratio of compressed gBSD descriptions to bitstream size. What can be seen is that the ratio gets worse as the bitstream size is reduced. While the description of the bitplanes scales along with the bitstream there is also a fixed amount of data consisting of DIA overhead and bitstream headers. The effect of low reduction in bitplane description and fixed overhead is especially detrimental. This effect is especially observed for low motion sequences like Akiyo where the *compressed* gBSD size can reach up to 25% of the bitstream size.

4 Conclusion

We have seen that the overhead of the gBSD description can be quite high depending on the actual bitrate of the video stream. This is true for both a limited case scenario as well as for a detailed description. The difference is that for a limited case scenario we can precalculate the estimated size of the description and it is the same whether we use two cases for a 2048kbps or a 128kbps bitstream. For the detailed description we retain the full flexibility of the wavelet codec even for in-network adaptation. At the same time however the description can become quite large, this is especially true for video sequences which attain a high compression ratio with the codec. Thus the use of either limited cases or detailed description depends on the application scenario. The main problem with the detailed description is that the network node can not judge how much of the gBSD will remain after scaling. As such, it is hard to allocate an overhead bandwidth to calculate the target rate to which to scale the video sequence. It would be possible to do a number of iterations but doing so would result in delay and higher computational load on the node. As such, the detailed description is somewhat problematic to use when the gBSD actually has to be transfered over the network link. However, there are scenarios when this is not necessary, e.g. the example in fig 1. Here we have a high bandwidth link to the access point where we can transfer the video sequence without problem. The ability to do a fine grained scaling is beneficial here since we can optimally use the available bandwidth to the client. Furthermore, the end client does not need the gBSD anymore so the possible overhead is of no concern here. The limited case scenario is for applications where the gBSD needs to be sent too. Since we can approximate how much overhead the description will take the bitstream can be scaled with that in mind. This prevents bandwidth problems and still enables us to do scaling, the only drawback is that we loose the full flexibility of the wavelet codec once the gBSD is generated. However, for generation of the gBSD we still enjoy that same flexibility. Since the bitstream does not need to be altered we can tailor the gBSD specifically to any application scenario given.

Overall we have seen that there is a cost involved in using gBSD to enable in-network adaptation. On the other hand we can bring the flexibility of wavelet based codecs to the network. This brings us closer to the UMA idea of serving every possible end device in a flexible way without having to re-encode the video sequence when new application scenarios arise.

References

[1] N. Adami, A. Signoroni, and R. Leonardi. State-of-the-art and trends in scalable video compression with wavelet-based approaches. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(9):1238–1255, Sept. 2007.

[2] C. J. Augeri, D. A. Bulutoglu, B. E. Mullins, R. O. Baldwin, and I. L. C. Baird. An analysis of xml compression efficiency. In *ExpCS '07: Proceedings of the 2007 workshop on Experimental computer science*, page 7, New York, NY, USA, 2007. ACM.

[3] H. Eeckhaut, H. Devos, P. Lambert, D. De Schrijver, W. Van Lancker, V. Nollet, P. Avasare, T. Clerckx, F. Verdicchio, M. Christiaens, P. Schelkens, R. Van de Walle, and D. Stroobandt. Scalable, wavelet-based video: From server to hardware-accelerated client. *Multimedia, IEEE Transactions on*, 9(7):1508–1519, Nov. 2007.

[4] S.-T. Hsiang and J. W. Woods. Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank. *Signal Processing: Image Communication*, 16(8):705–724, May 2001.

[5] ISO/IEC 21000-7:2007. Information technology – Multimedia framework (MPEG-21) – Part 7: Digital Item Adaptation, Nov. 2007.

[6] R. Kuschnig, I. Kofler, M. Ransburg, and H. Hellwagner. Design options and comparison of in-network H.264/SVC adaptation. *Journal of Visual Communication and Image Representation*, Dec. 2008.

[7] L. Lima, F. Manerba, N. Adami, A. Signoroni, and R. Leonardi. Wavelet-based encoding for HD applications. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1351–1354, July 2007.

[8] G. Panis, A. Hutter, J. Heuer, H. Hellwagner, H. Kosch, C. Timmerer, S. Devillers, and M. Amielh. Bitstream syntax description: a tool for multimedia resource adaptation within MPEG-21. In *Special Issue on Multimedia Adaptation*, volume 18 of *Signal Processing: Image Communication*, pages 721–747, Sept. 2003.

[9] Timmerer, C., Kofler, I., Liegl, J., and Hellwagner, H. An Evaluation of Existing Metadata Compression and Encoding Technologies for MPEG-21 Applications. In *Proceedings of the 1st IEEE International Workshop on Multimedia Information Processing and Retrieval (IEEE-MIPR 2005)*, pages 534–539, Irvine, California, USA, December 2005.

[10] A. Vetro, C. Christopoulos, and T. Ebrahimi. From the guest editors - Universal multimedia access. *IEEE Signal Processing Magazine*, 20(2):16 – 16, 2003.

[11] D. Wu, Y. T. Hou, W. Zhu, Y.-Q. Zhang, and J. M. Peha. Streaming video over the internet: approaches and directions. In *Circuits and Systems for Video Technology, IEEE Transactions on*, volume 11, pages 282–300, Mar 2001.

[12] Y. Wu, A. Golwelkar, and J. W. Woods. MC-EZBC video proposal from Rensselaer Polytechnic Institute. *ISO/IEC JTC1/SC29/WG11, MPEG2004/M10569/S15*, Mar. 2004.

Secure Scalable Video Compression for GVid

Heinz Hofbauer and Thomas Stütz and Andreas Uhl *

Abstract. *GVid is a Grid service that enables the secure and transparent integration and development of graphical user interface applications in the Grid. It separates the potentially computationally complex task of data creation and visualization, e.g., scientific simulations, from the comparably computationally inexpensive task of transmission and display of the visual data. A Grid application produces visual data and GVid takes care of the encoding, the secure and efficient transmission and the display of the visual data. As the transmission parameters and grid node properties are highly variable, special compression schemes have to be chosen to cope with these requirements. Beneficial for such requirements is the application of scalable compression formats, such as H.264/SVC (Scalable Video Coding) and MC-EZBC (Motion-Compensated Embedded Zerotree Block Coding). As simulation data may be sensitive, e.g., in the case of medical simulations, the secure transmission and storage of the visual data has to be guaranteed. Format-specific encryption schemes offer improved functionality due to the preservation of scalability in the encrypted domain. In this work the compression performance of state-of-the-art scalable video compression systems is evaluated and format-specific encryption schemes are proposed and discussed.*

1. Introduction

The GVid framework and implementation has been introduced and discussed in previous work [6,10]. The GVid framework separates the task of data generation and visualization from the comparably computationally inexpensive task of transmission and display of the visual data. This separation is especially reasonable if the visual data is displayed on a computationally weak device. Mobile devices have become the most frequent computing platform for a majority of users, even if many of them are not even aware that there mobile device is essentially a general purpose computer with an extended set of hardware. Thus Andrew S. Tanenbaum's ironic statement "Computers are different from telephones. Computers do not ring." [12] has lost its context. A major difference between telephones and computers remain the different computational capabilities and further constraints of telephones, which are nowadays almost exclusively mobile devices. Mobile devices suffer from slower CPUs, less memory, lower resolution displays, and network connections with lower bandwidth, but with a higher probability of connection loss. Especially the restricted computational capabilities are a convincing argument for the separation of data generation and visualization from the comparably computationally inexpensive task of transmission and display. This topic is currently in the focus of research, e.g., Advanced Micro Devices (AMD) is currently working on a supercomputer for graphic rendering to enable 3D game playing for cellphones [9]; an approach rather similar to GVid. Additionally the varying network parameters paired with a higher probability of connection loss for mobile devices pushes the development of another line of research, namely scalable and error resilient for-

*Department of Computer Sciences, University Salzburg, Salzburg, Austria, email: {hhofbaue, tstuetz, uhl}@cosy.sbg.ac.at

mats and transmission systems for visual data. Scalable visual data formats enable simple and fast rate adaptation. In previous work the scalable still image standard JPEG2000 has been employed for intra-frame compression. At present the scalable extension of the video coding standard H.264 (SVC) has been finalized and thus an applicable scalable video compression system is now available. A different approach to implement a scalable video format compared to the traditional layered design of H.264/SVC is followed by the wavelet-based MC-EZBC codec. Both schemes offer state-of-the-art scalable video compression and are therefore evaluated for the suitability as compression codecs within the GVid framework (see section 2. for details on the GVid structure and section 3. for details on the codecs). Their compression performance is evaluated in section 3.3. In section 4. format-specific encryption approaches are discussed for the two schemes together with a motivation and introduction to format-specific encryption. A format-specific encryption scheme for MC-EZBC is proposed in this work. The major advantages of format-specific encryption schemes are the preservation of scalability in the encrypted domain, i.e. rate adaptation can still be conducted, and a potentially improved error robustness and resilience. A concluding comparison of the two compression systems and their corresponding format specific encryption schemes is given in section 5. Additionally an outline of future work is presented, discussing the potentials to improve the runtime performance of scalable compression systems via parallel and distributed compression within the Grid.

2. GVid: Secure Interactive Video Transmission

The GVid software is a result of a joint project of the Institute of Graphics and Parallel Processing (GUP) at the Joh. Kepler University Linz and the Department of Computer Sciences at the University of Salzburg, which included Thomas Köckerbauer, Dieter Kranzlmüller, Martin Polak, Herbert Rosmanith, Thomas Stütz and Andreas Uhl.

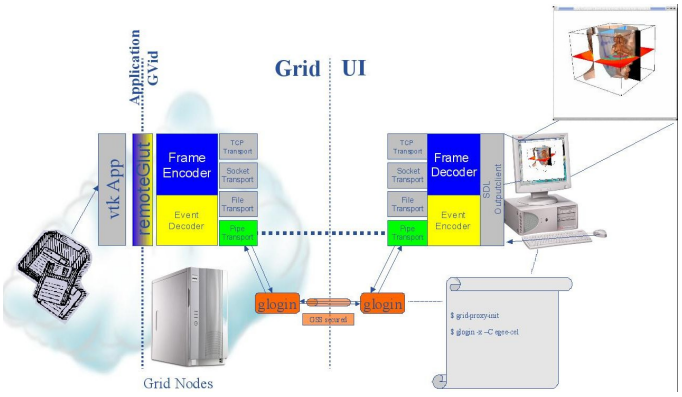


Figure 1. GVid Component Overview

2.1. The Structure of GVid

The aim of GVid software design was to support as many applications as easily as possible. Therefore, several input adapters exist that are responsible for acquiring the visual data of the application.

Currently a freeGLUT [1], a vtk [2] and a X11 based input adapter are implemented. The X11 input adapter enables every X11 application to be transmitted over the Grid.

Figure 1 illustrates an overview of the GVid design. An application provides the visual data through one of the several input adapters and GVid takes care of the encoding, the secure and efficient transmission and the display of the visual data. New compression and security schemes can be easily integrated. Currently a compression plug-in for MPEG-4 (Xvid) and JPEG2000 are integrated. Xvid does not provide a scalable format stream and JPEG2000 does not exploit inter frame redundancy. Thus for Xvid rate adaptation or delivery of streams at different rates can not be done efficiently and for JPEG2000 bandwidth could be saved by exploiting inter frame redundancy and yielding more efficient compression. A scalable video compression system would perfectly meet the requirements of efficient compression and scalability of the video format stream.

2.2. Confidentiality and Scalability in the GVid Framework

Scalability enables efficient rate adaptation, an important feature in an environment characterized by highly frequent network bandwidth changes. A scalable (video) format is the fundamental basis for efficient rate adaptation and enables advanced streaming and multicast scenarios, such as receiver driven layered multicast (RLM) [8]. RLM solves the adaptation to changing network conditions by receiver actions, i.e. join and leave of IP multicast groups, (receiver driven). However, IP multicast is not widely deployed and other implementations have to be considered for rate-adaptive streaming.

The idea of the application of scalable format streams for network adaptation has been extended to in-network adaptation systems, in which adaptation is dynamically performed in the network by a MANE (media aware network element). The basic setup is illustrated in figure 2. These in-network adaptation systems are assumed to offer rapid adaptation to changing network conditions as the delay for the propagation of changed network parameters is minimized. However, implementing such in-network systems within the scope of already existing and well-established transmission protocols, such as RTP, has been proved to contain certain pitfalls [7, 17]. Nonetheless, the idea of in-network adaptation can be considered sensible and as a potential candidate for the integration in the GVid framework. Integrating security services, i.e. confidentiality in in-network adaptation systems, is not straight-forward. The application of well-established security tools, e.g., SRTP, SSL or IPSEC, is not possible as the necessary information to perform rate-adaptation within the network is concealed and thus not available at the MANE. Thus if confidentiality and in-network adaptation are to be combined, format-specific encryption schemes, that preserve the information necessary for rate adaptation, are needed.

In multiple client scenarios (see figure 2) the application of scalable compression systems offers substantial advantages. In these scenarios the visual output of a Grid application is transmitted to multiple clients, each with its own preferences and parameters for the visual content ant its transmission (e.g., rate and resolution). If conventional compression systems (i.e., systems not delivering scalable format streams) are employed, a separate compression task for each client has to be performed. These separate compression tasks are, considering the computational complexity of state-of-the-art video compression, an enormous burden. The solution of separate compression tasks does not scale well with the number of clients, i.e., each new client with distinct preferences adds another separate compression task. Scalable compression systems can solve this issue, as only one single compression task generates a scalable format stream, that can efficiently be adapted to each client's preferences. This paradigm of a single encoding step with subsequent computationally efficient adaption steps is

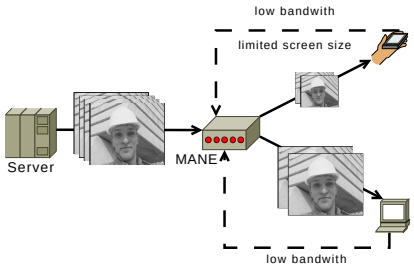


Figure 2. Example of a single video sequence from the server which is adapted to the given capabilities of two end devices.

referred to as Universal Multimedia Access (UMA) [14]. In case that confidential transmission has to be guaranteed, well-established security tools could be applied, but again these solutions do not scale well with the number of clients. In fact a separate encryption task has to be performed for each client (even if the clients share the same preferences). Format-specific encryption schemes offer a well-scaling solution. These schemes encrypt a scalable format stream in a specific scalability-preserving fashion. The still scalable but secured format stream can efficiently be adapted to the clients preferences and confidentially transmitted.

In conclusion we can state that the application of scalable compression systems and format specific encryption within GVid offers ample benefits and should thus be seriously considered.

Only recently SVC, the scalable extension of H.264, has been standardized [5] and therefore it is worth evaluating the suitability of this new compression system for the application within GVid. Additionally the wavelet-based scalable video codec MC-EZBC is evaluated.

3. State-of-the-Art Scalable Video Compression

In the following two scalable video compression systems are presented, which also represent two different approaches to implement scalable video coding.

SVC follows the traditional design of layered video coding [5], while MC-EZBC is a t+2D wavelet-based video codec with motion-compensated temporal filtering.

3.1. H.264/SVC

A major design requirement for SVC has been the backwards compatibility to the existing H.264/AVC. Thus SVC format streams are valid H.264/AVC format streams (format-compliant with respect to the non-scalable H.264/AVC format) and thus decodeable by H.264/AVC compliant decoders. Major parts of the H.264 AVC video coding system have been adopted, including most of the H.264 AVC syntax and semantics. An SVC format stream contains a base layer and one or more enhancement layers each may augment the user experience in one of three dimensions (temporal/spatial/quality).

3.1.1. Temporal Scalability

A format stream is temporally scalable if it contains sub streams with lower frame rates. Due to the flexible inter prediction in H.264/AVC, the implementation of temporal scalability within H.264/AVC/SVC has been straightforward by employing special prediction structures, e.g., dyadic temporal enhancement layers with hierarchical B-pictures. In figure 3 the dyadic hierarchical B-picture prediction structure is illustrated, but temporal scalability in H.264/AVC/SVC is not limited to dyadic prediction structures; SVC offers the syntax to easily extract a sub stream with a reduced frame rate by simply dropping parts of the format stream.

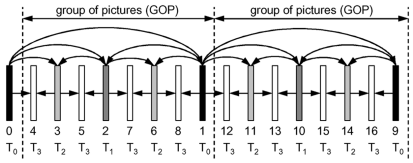


Figure 3. Prediction hierarchy of B-pictures in SVC

3.1.2. Spatial Scalability

A format stream is spatially scalable if it contains sub streams with different resolutions. SVC implements spatial scalability with a conventional multilayer approach. A base layer (lower resolution) is encoded in H.264/AVC compliant fashion, while the enhancement layers (containing higher resolutions) may apply inter layer prediction in order to exploit redundancies between the layers. Spatial scalability with arbitrary resolutions is supported.

3.1.3. Quality Scalability

A format stream is quality scalable if it contains substreams with different qualities, in a signal to noise ratio (SNR) sense, but same resolution. In SVC the so called key-picture concept, also known as medium grain scalability (MGS), is employed to enable quality scalability.

3.1.4. SVC NAL units

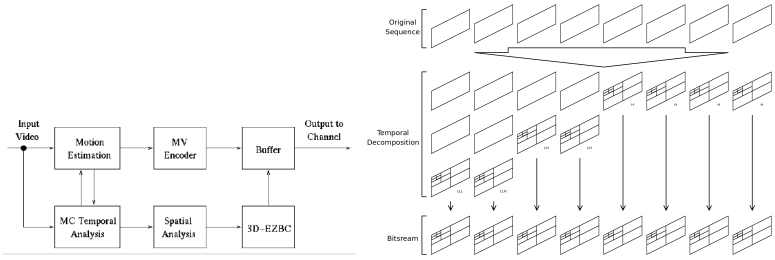
A network abstraction layer (NAL) unit in H.264 is preceded by an 1-byte NAL unit header, containing most importantly the NAL unit type. On the basis of the NAL unit type the NAL unit data is processed. For SVC the NAL header is extended, a three byte extension is added. This extension contains a dependency_id, which identifies the spatial layer to which the NAL unit data contributes, a temporal_id, which specifies the temporal layer of the NAL unit, and a quality_id, which specifies to which quality layer the NAL unit contributes.

3.2. MC-EZBC

The MC-EZBC [4, 19] coder is a t+2D wavelet coder, i.e., a wavelet transform is applied for temporal decomposition as well as for spatial decomposition. The abbreviation t+2D implies that the temporal decomposition combined with motion estimation is applied before the spatial decomposition (both apply pyramidal decomposition structures). The 9/7 CDF (Cohen-Daubechies-Feauveau) wavelet

filters are applied for spatial decomposition, while temporal decomposition is conducted with the CDF 5/3 wavelet filters. Furthermore adaptive prediction techniques are employed. The layout of the MC-EZBC coder is shown in figure 4(a)

Figure 4(b) illustrates the encoding process for a group of pictures (GOP). The frames of a raw video sequence are split into GOPs, which are independently coded. In a GOP frames are decomposed temporally and then spatially. Note that the ordering of the coded frames follows the temporal decomposition level, i.e., the deepest temporal low-pass frames are the first contributions in the final stream of a GOP.



(a) Layout of the MC-EZBC coder

(b) Encoding of a GOP showing spatial and temporal decomposition

Figure 4. MC-EZBC

3.2.1. Temporal Scalability

The MC-EZBC format stream automatically supports temporal scalability; this property is due to the temporal wavelet decomposition. If the GOP size is 2^t, then the number of temporal resolutions, i.e., different frame rates is t. Temporal scaling is done by dropping levels from the temporal decomposition, i.e., one step of temporal scaling reduces the frame rate by half. For example a GOP with $16 = 2^4$ frames could be reduced to $8, 4$ or 2 frames. Only dyadic temporal prediction structures are permitted.

3.2.2. Spatial Scalability

Spatial scalability is also automatically supported and like temporal scalability is done by dropping high frequency wavelet bands. Again scaling operations are discrete with steps of half the resolution of the previous step, e.g. a CIF (352×288) video could be scaled to qCIF (176×128) or sqCIF (88×64). Only dyadic spatial resolution changes are permitted.

3.2.3. Quality Scalability

Quality scalability unlike spatial or temporal scalability is more flexible. SNR scalability of the MC-EZBC achieves multiple bitrates within a single format stream. The coded wavelet coefficient data is arranged in an embedded bitstream, i.e., a truncated segment of the coded data is still decodeable and results in a quantized representation of the wavelet coefficient data.

3.3. Performance Evaluation

The following performance evaluation is intended to give an overview of the capabilities of the two scalable compression formats, SVC and MC-EZBC, and their suitability for the application within the GVid framework. For more extensive and exhaustive treatments on the compression performance of these codecs the reader is referred to [18]

In the assessment of the compression performance of scalable compression systems subtle pitfalls are hidden. The two scalable compression systems may contain substreams with different resolutions. However, the lower resolution versions of the original content contained in the format streams of SVC and MC-EZBC are different. In SVC the subsampling method at the encoder-side is not specified in the standard, however, the upsampling method is specified and it is therefore sensible to employ the corresponding subsampling method. In the MC-EZBC the subsampling method is defined by the low-pass filter of the spatial wavelet decompostion (9/7 CDF). Thus taking a common reference for quality assessment, peak signal to noise ratio (PSNR) calculation, for both schemes always and systematically favours one of the compression systems. Therefore the compression performance for lower resolution substreams is assessed for each compression system individually with the correct reference, i.e., the lower resolution reference sequences for the MC-EZBC are generated with the low-pass filters of the 9/7 CDF and the lower resolution reference sequences for SVC are generated with the subsampling filters fitting to the normative upsampling filters.

The quality for lower frame rate substreams is assessed with reference to the original sequence where frames have been dropped, i.e., every second frame is dropped if the frame rate is halved.

In this evaluation the well-known foreman sequence in the CIF format (352x288) with 96 frames at a frame rate of 30 fps is employed.

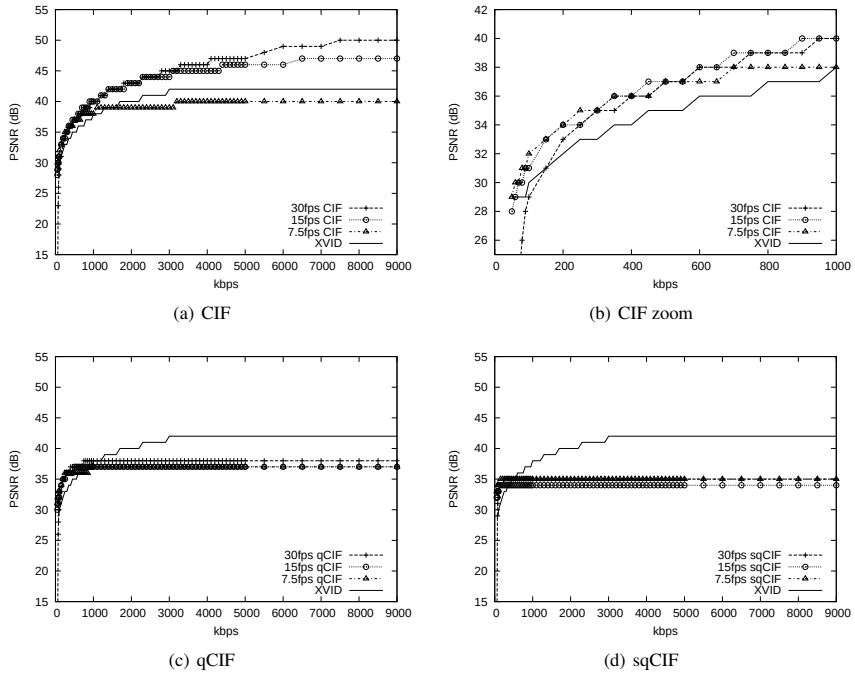
3.3.1. Performance of the MC-EZBC

The compression performance of the MC-EZBC is summarized in figure 5. Most notably are the multiple bitrates contained within the single scalable MC-EZBC stream, illustrated by dots in the figure. It is also noteworthy that for a regular CIF version with full framerate the MC-EZBC performs better than the widely used XVID codec, fig. 5(a) and 5(b).

3.3.2. Performance of the H.264/SVC

The main issue for the performance evaluation is the definition of suitable encoder configurations. The encoder configuration is decisive for the compression performance and it also defines extraction points (i.e., bitrates at which reconstruction is possible). In general, it can be summarized that temporal scalability comes for free and even improves the compression performance, while the other types of scalability decrease the compression performance, but increase the number of extraction points. H.264/SVC is a layered video codec allowing only a discrete number of extraction points. This is a major difference to the MC-EZBC codec, which allows the extraction of arbitrary bitrates from the stream.

The following figures illustrate the extraction points and their respective PSNR for different encoder configurations. A point in the figure represents an extraction point.



(a) CIF (b) CIF zoom

(c) qCIF (d) sqCIF

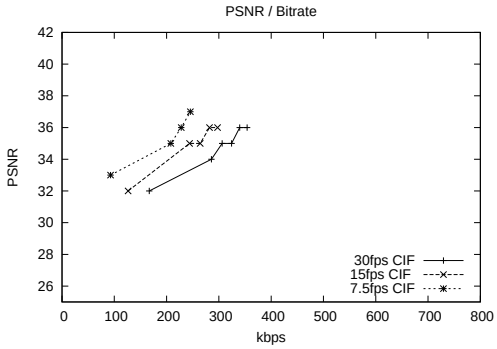
Figure 5. Rate-distortion plots for the foreman sequence and different resolutions (CIF, qCIF and sqCIF) as well as a zoomed version for the low bitrates of the CIF plot.

First we discuss two configurations that implement temporal and quality scalability. These configurations are suitable for computationally strong devices such as home PCs, therefore smaller resolutions and a simple base layer (e.g., suitable for mobile devices) are omitted.

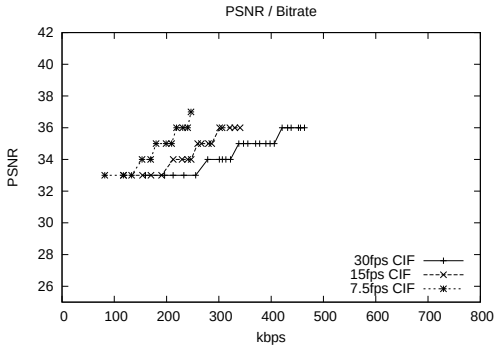
Figure 6(a) shows a simple configuration with only one spatial resolution and one MGS enhancement layer.

Figure 6(b) shows a simple configuration with only one spatial resolution and 8 MGS enhancement layers. MGS is a mode very similar to progressive JPEG, namely the spectral selection mode of operation. In this configuration the 16 transform coefficients of the 4x4 transform are grouped into 8 partitions each containing exactly two transform coefficients.

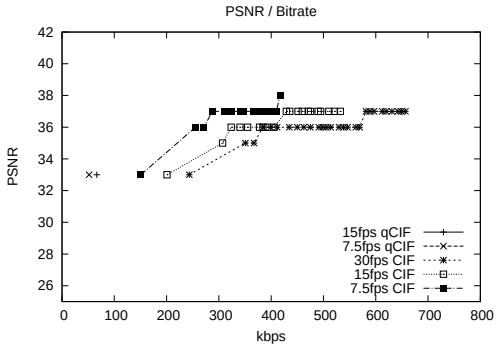
Additionally lower resolution substreams can be defined. For the fine configuration a QCIF resolution is contained in the substream. This substream is encoded within the limits and constraints of the H.264/AVC baseline profile (CAVLC). The bitstream may be used to serve both a computationally weak device such as a mobile phone and a PC. The number of reference frames is set to 1. In figure 6(c) the extraction points for this configuration are illustrated.



(a) Results for the coarse configuration.



(b) Results for the coarse2 configuration.



(c) Results for the fine configuration.

Figure 6. The foreman sequence with 30 fps under different coder configurations.

3.3.3. Comparison between H.264/SVC and MC-EZBC

MC-EZBC's compression performance (at least of the encoder configurations we have tested) is at least equal to H.264/SVC (see figure 7). It has to be noted, that the results for H.264/SVC have been obtained with the reference software JSVM and that other implementations of the H.264/SVC standard may offer better compression performance. If both codecs are compared to state-of-the-art MPEG-4 / H.263 encoders (Xvid), the clear resume is that both perform significantly better for a broad range of bitrates (see figure 8).

The advantage of the MC-EZBC is its higher flexibility in terms of possible extraction points; beneficial if fine grained rate adaptation is to be performed.

There are several arguments for H.264/SVC: It is backwards-compatible to H.264, which allows the base layer to be decoded with a compliant H.264 decoder, e.g., special hardware chips. It is scalable in terms of computational complexity. The base layer can be encoded such that decoding has a very low computational complexity, e.g., arithmetic coding can be omitted.

Another advantage of H.264/SVC is related to the interactive usage possible in the GVid framework. It allows zero structural delay, i.e., the inter-prediction process can be configured to allow only forward prediction. Thus every frame can immediately be coded and transmitted. In case of MC-EZBC this is not possible as frames have to be processed on a GOP-basis, i.e., a number of frames (the GOP size) have to be buffered and delayed until the coding and transmission can be conducted. The introduced delay is adverse to interactive usage, where low delay is preferred.

In conclusion, MC-EZBC offers better rate adaptation, but H.264/SVC provides other important features MC-EZBC lacks.

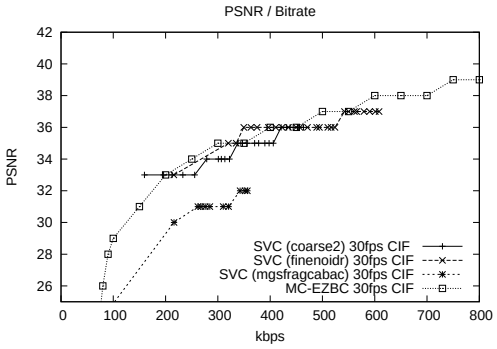


Figure 7. MC-EZBC compared to H.264/SVC

4. Format-Specific Encryption Schemes

Format-specific scalability-preserving encryption schemes are necessary in order to combine efficient transmission and confidentiality.In the following format-specific encryption schemes are discussed for H.264/SVC and MC-EZBC.

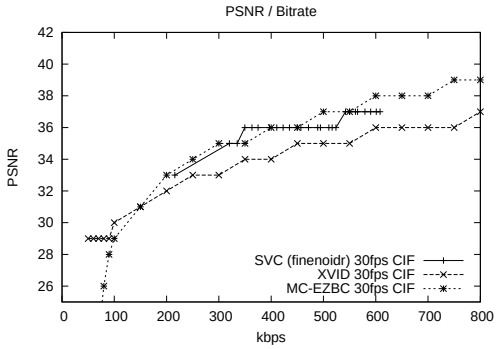


Figure 8. Comparison of Xvid, MC-EZBC and H.264/SVC

4.1. H.264/SVC-Specific Encryption

In order to preserve the scalability of the H.264/SVC stream the information to which dependency layer, temporal layer and quality layer a NAL unit contributes, which is part of the SVC extended NAL unit header, has to be preserved. Thus only encrypting the NAL unit body preserves scalability. However, straight-forward conventional encryption of the NAL unit body is problematic, as NAL unit bodies obey certain syntax rules. Namely marker sequences, that e.g., signal the beginning and the end of a NAL unit, are forbidden. Thus conventional encryption of NAL unit bodies is likely to break the system at some point, e.g., if the H.264/SVC byte-stream format is used and a marker sequence is accidentally generated in a NAL unit body, the entire synchronization is lost.

A way to prevent such behaviour is to ensure that format-specific encryption produces a format-compliant encrypted stream (format-compliant encryption). As a result it can be guaranteed that a decoder does not crash decoding such a stream.

In [11], the H.264/SVC header is preserved and unspecified NAL unit types are employed to signal encrypted data. For the most frequent NAL unit types (NUTs 1, 5, 14, 20) a direct mapping to unspecified NAL unit type values is defined. For all other NAL unit types, the original NAL unit header is preserved as the first payload byte and a certain unspecified NAL unit type is used to signal these encrypted NAL units. However, if packaging is applied as specified in the RFC 3984 [15] and the draft RFC defining the RTP payload for H.264/SVC video [16], all but one (NUT 0) of the unspecified NUT values are already assigned a specific meaning. Hence, the only possibility to employ unspecified NAL unit types to signal encrypted data is NUT 0 [3]. A NAL unit selected for encryption is prefixed by a NAL unit header with NUT 0, and the original NAL unit header and the H.264/SVC header are the first bytes of the encrypted NAL unit payload, and the remaining NAL unit payload is encrypted. However, special care must be taken to avoid marker sequences (H.264 marker sequences are prefixed by at least two zero bytes). This is a problem if encryption is applied more than once, i.e., encrypted NAL units are encrypted. A straight-forward solution is to set the NRI field in the NAL unit header to a value not equal to 0.

4.1.1. Format Compliance and Encryption

Although the encrypted NAL unit has to be ignored by a compliant decoder, certain syntax requirements have to be met by the encrypted NAL unit. These requirements are given in [5]; namely that within the NAL unit, the following three-byte sequences shall not occur at any byte-aligned position: 0x000000, 0x000001, 0x000002, and 0x000003.

Additionally, within the NAL unit, any four byte sequence that starts with 0x000003 other than the following sequences shall not occur at any byte-aligned position: 0x00000300, 0x00000301, 0x00000302, and 0x00000303. Additionally, the last byte of a NAL unit shall not be 0x00.

The encryption scheme has to ensure that these requirements are met. Therefore, after encryption the procedure for the encapsulation of an SODB (string of data bits) within an RBSP (raw byte sequence payload) [5] has to be applied. For the case of two consecutive 0x00 bytes, this procedure ensures that the NAL unit does not end with a 0x00 byte. If a NAL unit ends with a 0x00 byte, it has to end with two consecutive 0x00 bytes for all currently specified RBSP types.

Encrypted NAL unit payloads may not have this property and thus special care has to be taken for the encryption of the last byte of a NAL unit. In our approach we use AES in Counter Mode and treat the last byte with special care.

Every cipher byte, except the last one, is the plaintext byte XORed with a keystream byte. The last cipher byte c is derived from the plaintext byte p and a keystream byte k (optimally in the range [0x00,0xfe], which can be ensured by ignoring 0xff bytes from the keystream) in the following way:

$$c = (p - 1 + k) \texttt{mod 0xfe} + 1$$

For decryption the following procedure is applied:

$$p = (c - 1 - k) \texttt{mod 0xfe} + 1$$

In order to ensure format compliance and decodability by any conformant decoder, an appropriate set of NAL units has to be selected for encryption.

4.2. MC-EZBC-Specific Encryption

As format-specific encryption for MC-EZBC heavily relies on its bitstream format, we start the with a thorough discussion of the MC-EZBC format. A schematic overview of the MC-EZBC format stream is given in figure 9, the organization of GOP data is outlined in figure 4(b). The main header followed by GOP sizes (this is the size of the image data in a GOP) followed by coded data of sequential GOPs. In the following the coded data of a GOP is referred to as GOP as well. Each GOP is lead by a header, giving scene change information, i.e. which frames are I frames, followed by the motion field and coded image data. Both motion field and image data are ordered by frame; frames are ordered lowest to highest temporal resolution. The image data of a frame is also arranged from lowest to highest resolution and a spatial decomposition of a frame is grouped together as a basic image data unit (we will call them chunks from now on). Each chunk is preceded by a leading header defining the length of the chunk and groups all chroma information of a given decomposition level. The image data in a chunk is ordered by importance regarding SNR scalability and is the result of a bitplane coder. This enables SNR scaling through truncation of image data (and adjustments of GOP size information and chunk length).

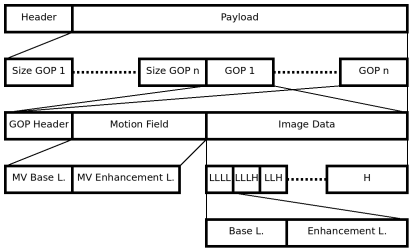


Figure 9. The layout of the MC-EZBC bitstream

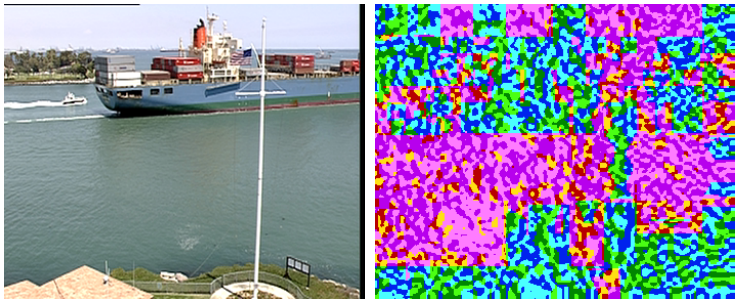
Only when all headers, including chunk headers, and GOP size information are kept intact the whole bitstream can subsequently be parsed correctly, which is of ultimate importance for the preservation of scalability in the encrypted domain. Additionally the motion vector data is coded differently from the image data; the length of the motion vector data is not explicitly signalled, but it has to be determined by arithmetic decoding (until a termination marker is encountered). Thus headers and motion vector data will not be encrypted in our encryption scheme, but solely the coded image data. Since the coded image data is byte aligned we need an encryption scheme which can encrypt blocks of arbitrary length, e.g., AES in OFB mode.

Our MC-EZBC-specific encryption scheme is format-compliant in the sense that the decoder can decode the encrypted format streams (and does neither crash nor complain). This is because the arithmetic decoder has to deal with SNR scalability and thus utilizes the chunk length information to prevent misalignment. We exploit this decoder property with our encryption. In case the arithmetic decoder tries to decode to much, as would be the case when regular scaling is done, the chunk length prevents the decoder from reading data of the next chunk. Additionally, when the decoder finishes early the rest of the chunk is skipped and the decoder is properly realigned for the next chunk. This is part of the error correction of the decoder which prevents misalignment when bit flips occur in the image data during transmission.

To increase the speed of the encryption and decryption processes it is possible to encrypt only a fraction of the image data. In order to minimize the amount of data to be encrypted, while maximizing its impact on the degradation of image quality we need to encrypt the parts of the bitstream which carries the most important visual information, e.g., I-frames of low frequency bands of the wavelet decomposition. Figure 10 illustrates this by comparing frame 128 of the Container sequence to the decoding of the encrypted sequence. In this figure only the low spatial frequencies have been encrypted.

4.3. Comparison of the Format-Specific Encryption Schemes

Both format-specific encryption schemes offer efficient encryption of the coded video data, while preserving the scalability. Both schemes preserve format-compliance (i.e., the decoder does not crash). Thus both schemes are well-suited for the integration in the GVid framework.



(a) decoded (b) encrypted

Figure 10. Comparison of encrypted image to the original of frame 128 from the Container sequence (low spatial frequencies)

5. Conclusion and Future Work

We have evaluated two state-of-the-art scalable video compression systems, namely H.264/SVC and MC-EZBC, for their suitability as compression codecs in the GVid framework. Their compression performance is competitive to conventional video compression systems, such as the MPEG-4 implementation Xvid. Their scalable format streams offer improved performance for multiple application scenarios. As the application of conventional security tools for confidentiality circumvent the advantages of scalable compression systems, format-specific encryption tools are necessary. For both H.264/SVC and MC-EZBC format-specific and even format-compliant encryption schemes have been proposed and discussed. Both encryption schemes meet the requirements well and can be recommended for integration in the GVid framework.

Future work will focus on the parallelization and optimization of the scalable video compression systems, as the current implementations are still not capable of real-time compression.

References

[1] FreeGLUT - The Free OpenGL Toolkit. online presentation: http://freeglut.sourceforge.net, December 2005.

[2] VTK - The Visualization Toolkit. online presentation: http://www.vtk.org, January 2006.

[3] Hermann Hellwagner, Robert Kuschnig, Thomas Stütz, and Andreas Uhl. Efficient in-network adaptation of encrypted H.264/SVC content. *Elsevier Journal on Signal Processing: Image Communication*, 24(9):740 – 758, July 2009.

[4] Shih-Ta Hsiang and J. W. Woods. Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank. *Signal Processing: Image Communication*, 16(8):705–724, May 2001.

[5] ITU-T H.264. Advanced video coding for generic audivisual services, November 2007.

[6] T. Köckerbauer, M. Polak, T. Stütz, and A. Uhl. GVid - video coding and encryption for advanced Grid visualization. In J. Volkert, T. Fahringer, D. Kranzlmüller, and W. Schreiner, editors,

Proceedings of the 1st Austrian Grid Symposium, volume 210 of *books@ocg.at*, pages 204–218, Schloss Hagenberg, Austria, 2006. Austrian Computer Society.

[7] R. Kuschnig, I. Kofler, M. Ransburg, and H. Hellwagner. Design options and comparison of in-network H.264/SVC adaptation. *Journal of Visual Communication and Image Representation*, September 2008.

[8] Steven McCanne, Van Jacobson, and Martin Vetterli. Receiver-driven layered multicast. In *SIGCOMM '96: Conference proceedings on Applications, technologies, architectures, and protocols for computer communications*, pages 117–130, New York, NY, USA, August 1996. ACM.

[9] Philip Ross. Cloud computing's killer app: Gaming. *IEEE Spectrum*, March 2009.

[10] Thomas Stütz and Andreas Uhl. Evaluation of compression codecs and selective encryption schemes for GVid. In J. Volkert, T. Fahringer, D. Kranzlmüller, and W. Schreiner, editors, *Proceedings of the 2nd Austrian Grid Symposium*, volume 221 of *books@ocg.at*, pages 28–41, Innsbruck, Austria, 2007. Austrian Computer Society.

[11] Thomas Stütz and Andreas Uhl. Format-compliant encryption of H.264/AVC and SVC. In *Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'08)*, Berkeley, CA, USA, December 2008. IEEE Computer Society.

[12] Andrew S. Tanenbaum. *Computer networks*. Prentice Hall, 3rd edition, 1996.

[13] A. Uhl and A. Pommer. *Image and Video Encryption. From Digital Rights Management to Secured Personal Communication*, volume 15 of *Advances in Information Security*. Springer-Verlag, 2005.

[14] A. Vetro, C. Christopoulos, and T. Ebrahimi. From the guest editors - Universal multimedia access. *IEEE Signal Processing Magazine*, 20(2):16 – 16, 2003.

[15] S. Wenger, M.M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer. RTP Payload Format for H.264 Video. RFC 3984, February 2005.

[16] S. Wenger, Y. Wang, T. Schierl, and A. Eleftheriadis. RTP Payload Format for SVC Video. Internet Draft draft-ietf-avt-rtp-svc-14, September 2008.

[17] S. Wenger, Y. Wang, T. Schierl, and A. Eleftheriadis. RTP Payload Format for SVC Video. Internet Draft draft-ietf-avt-rtp-svc-17, February 2009.

[18] M. Wien, H. Schwarz, and T. Oelbaum. Performance analysis of SVC. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1194–1203, September 2007.

[19] Y. Wu, A. Golwelkar, and J. W. Woods. MC-EZBC video proposal from Rensselaer Polytechnic Institute. *ISO/IEC JTC1/SC29/WG11, MPEG2004/M10569/S15*, March 2004.

VISUAL QUALITY INDICES AND LOW QUALITY IMAGES

Heinz Hofbauer and Andreas Uhl

Department of Computer Sciences
University of Salzburg
{hhofbaue, uhl}@cosy.sbg.ac.at

ABSTRACT

Visual quality indices are frequently used instead of human evaluation for the quality assessment of impaired images (or video material). These visual quality indices are in turn evaluated on databases containing impaired images in conjunction with a score given by evaluation with human observers. The fitness of these indices are judged on the entire quality scale of the respective database. However, this leads to the incorrect assumption that these quality indices perform well over the whole range possible qualities. This is unfortunately not true, especially towards the low quality range of images these quality indices often show little actual correlation to human judgement. In this paper a number of visual quality indices will be evaluated with regard to the lower quality spectrum of impairments and it will be shown that the overall fitness of a quality index is not generally related to its performance regarding high impairment.

Index Terms— Image analysis, Quality control

1. INTRODUCTION

The assessment of image and video quality is important whenever image and videos are transmitted (gauging of transmission errors), encoded (compression vs. quality) etc. Optimally a jury of humans would judge the impact of the impairment, however the high time and cost required to do this are prohibitive. Thus, visual quality indices(VQI) are used to simulate the assessment that should be made by humans. In order to judge the correlation of VQI to the average human judgement a number of databases have been created containing distorted images along with a mean opinion score (MOS) of human observers, for example LIVE [1] or TID [2]. These databases typically contain different testsets which correspond to typical application scenarios, e.g. JPEG or JPEG2000 compression, distortion scenarios, e.g. transmission errors or denoising, and operations on images, e.g. gaussian blur or masking.

The evaluation of a VQI is usually done over the whole range of impairments in a given database. This is reasonable to estimate the overall fitness of VQIs but there are certain shortcomings in this approach. For example in high compression scenarios a VQI which does well overall is less useful than one which performs best for the given low quality scale (and the same is essentially true for a high quality range). The underlying problem is that VQIs overall performance does not correlate to performance for low quality scenarios or even, though less frequently, for high quality scenarios. Typical low quality scenarios are low bitrate videos [3, 4], video streaming [5, 6], assessment of transmission errors [7] or quality control for transparent encryption [8].

While there is some previous work regarding low quality video sequences [9], there is, to the extent of the authors knowledge, no information available regarding low quality image assessment over the range of recent VQIs. To rectify this shortcoming, state of the art VQIs will be evaluated on known databases where the focus is on low and high quality subsets rather than the whole database. This will also show that the databases which are already in existence are sufficient to properly evaluate the performance of VQIs and point out that it would be good practice to evaluating performance in a more discerning way.

In order to facilitate the reproducibility of the research we restricted ourself to publicly available data and implementations.

In section 2 a brief overview of the evaluated VQIs will be given, in section 3 the evaluation based on the LIVE and selected subsets of the TID database will be given.

2. OVERVIEW OF VISUAL QUALITY INDICES

Modern VQIs often use a sophisticated approach on quality which heavily relies on knowledge about the human visual system (HVS). In the following we will give a short review of the VQIs which will be evaluated.

The most widely used VQI today is the peak signal-to-noise ratio (PSNR) since it is easy to implement and fast to compute. It is also well known that the PSNR does not reflect human judgement very well. In [10] Huynh-Thu and Ghanbari showed that as long as the content is unchanged the PSNR reasonably well reflects the human observer.

The luminance and edge similarity score (LSS and ESS) was introduced by Mao and Wu [11]. They used the informa-

tion on the HVS to find criteria how observers judge images. The edge information reflects the assessment of humans regarding the shape or contour of objects and the luminance score reflects changes in the color space. Both algorithm use 8×8 windows to assess the edge direction and mean luminance of a region in the image.

The visual signal-to-noise ratio (VSNR) [12] uses a two stage method of quality assessment. In the first stage contrast detection thresholds are calculated by using wavelet (DWT) based models of visual masking and summation to assess if the errors are perceivable by the HVS. If the errors are judged to be below the detection threshold the image is considered pristine. When the errors are above the threshold of detection a score is calculated by using the ratio of the RMS contrast to the weighted values of the perceived contrast and global precedence.

The structural similarity index measure SSIM [13] extracts three separate scores from the image and combines them into the final score. First the visual influence is calculated locally then luminance, contrast and structural scores are calculated globally. These separate scores are then combined with equal weight to form the SSIM score.

The multi-scale structural similarity index measure MS-SSIM [14] is an extension of the SSIM to take into account that the perceivability of image impairments is different depending on the sampling density of the image signal, e.g. as influenced by viewing distance. To take this into account the similarity scores are calculated at different spatial scales. The core operation is similar to SSIM, contrast and structural scores are calculated at each scale and the luminance score is calculated at the lowest scale. The factors for combining these scores where found by experiments with human observers.

Criterion v4.0 C4 [15] uses a detailed model of the HVS, and information regarding the score is extracted from a transformation of the image in the perceptual space. The transformation include compensation for display device gamma, perceptual colorspace, luminance normalization, contrast sensitivity functions, subband decomposition and modelling of masking effects. From this perceptual model the contrast orientation, length and width as well as the subband amplitude and average luminance, red-green chroma and yellow-blue chroma channels are extracted from characteristic points in the model. The local scores are generated as averaging of the extracted features and the overall score is generated by averaging the local scores.

For the visual information fidelity criterion VIF [16] a more refined model is used which starts with the modeling of the reference image using natural scene statistics (NSS). Furthermore, the possible distortion is modeled as signal gain and additive noise in the wavelet domain and parts of the HVS which have not been covered by the NSS are modeled, i.e. internal neural noise is modeled by using a additive white Gaussian noise model. Using this model the VIF score reflects the fraction of the reference image information which can be extracted from the impaired image.

The Weighted Signal to Noise Ration (WSNR) [17] is defined as the ratio of the average weighted signal power to the average weighted noise power. The weight function used is a contrast sensitivity function (CSF) which is gained by using the frequency response of HVS. A measure of the non-linear HVS response to a single frequency, the contrast threshold function (CTF), is used which is measured over the visible radial spatial frequencies. The CTF is the minimum amplitude necessary to detect a sine wave of a given angular spatial frequency. The CSF is the frequency response obtained by inverting the CTF.

With respect to implementations we used our own code[1] for PSNR, SSIM, LSS, ESS. For C4 the implementation from Carnec et al. was used and for all other VQIs the "MeTriX MuX Visual Quality Assessment Package[2] was used in version 1.1. Also note that UQI, IFC and NQM from Metrix Mux were not included in the evaluation since they are predecessors of other VQIs which were evaluated.

3. EVALUATION OF VISUAL QUALITY INDICES

In the following evaluations the Spearman rank order correlation (SROC) is used to compensate for non linearity. The VQIs were evaluated on two different databases for two reasons. First, we want to show that the shortcomings of the VQIs are not based on the distortion and image types of a single database. Secondly, we want to show that the problems of VQI with lower quality images are not a result of the evaluation method employed in the database assembly. The DMOS value of these two databases was derived differently, the LIVE database uses a linear scale of perceived impairment where observers judge each image individually while the TID database uses a direct comparison of two impaired images where the observer selects the higher quality image and thus creates a ranking in order of perceived impairment.

3.1. Evaluation on the LIVE Database

The first comparison will be on base of the LIVE database[3]. The comparison is between the full range of the database as would be used for regular VQI evaluation, table 1, the low quality part of the database with a DMOS of greater than 80 (70 for gblur in order to keep the number of distortions high enough), table 3, and the high quality range with a DMOS lower than 40, table 2. In these tables value with SROC lower than 0.5 are underlined to show low correlation and the best score per testset is printed in a bold font.

To get a better overview fig. 1 illustrates the relation of the SROC from tables 1 through 6. Figure 2 shows the same

[1] http://www.wavelab.at
[2] http://foulard.ece.cornell.edu/gaubatz/metrix_mux/
[3] http://live.ece.utexas.edu/research/quality/

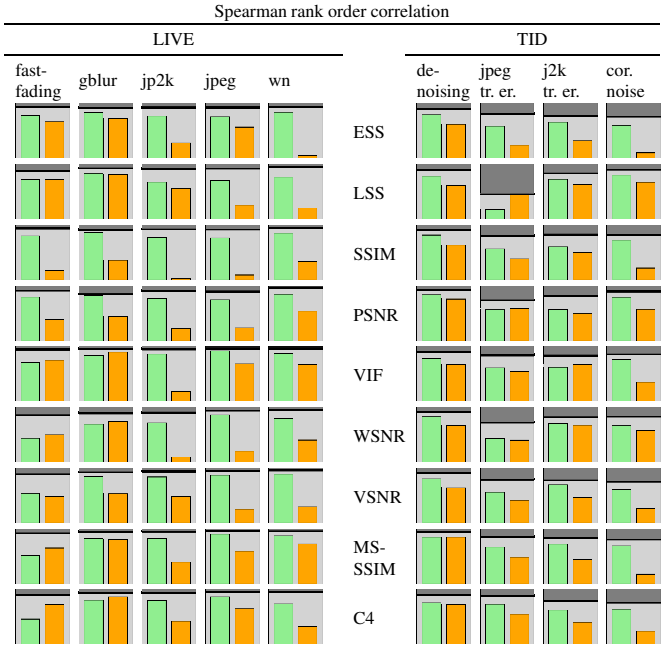


Fig. 1. For each entry the SROC for the full range of qualities is indicated by the line separating the light and dark background ranging from SROC $= 0.0$ at the bottom to SROC $= 1.0$ at the top. Superimposed are two bar charts showing the SROC for the high quality range on the left side and the low quality range on the right side.

Table 1. LIVE Image Quality Assessment Database

	fastfading	gblur	jp2k	jpeg	wn
ESS	0.956	0.939	0.944	0.945	0.958
LSS	0.898	0.941	0.928	0.939	0.967
SSIM	0.957	0.935	0.940	0.940	0.968
PSNR	0.927	0.865	0.923	0.913	0.982
VIF	**0.965**	**0.972**	**0.968**	**0.984**	**0.985**
WSNR	0.873	0.909	0.920	0.958	0.973
VSNR	0.903	0.941	0.955	0.966	0.978
MS-SSIM	0.932	0.958	0.965	0.979	0.973
C4	0.919	0.956	0.959	0.975	0.970

Table 2. LIVE Image Quality Assessment Database, high quality (DMOS ≤ 40)

	fastfading	gblur	jp2k	jpeg	wn
ESS	0.799	0.848	0.790	0.775	0.854
LSS	0.738	0.853	0.688	0.720	0.779
SSIM	**0.821**	**0.879**	0.789	0.780	0.868
PSNR	0.809	0.838	0.780	0.764	0.865
VIF	0.722	0.853	**0.880**	**0.942**	0.889
WSNR	0.444	0.707	0.732	0.881	0.810
VSNR	0.553	0.861	0.856	0.887	**0.908**
MS-SSIM	0.529	0.839	0.841	0.918	0.894
C4	0.473	0.826	0.818	0.883	0.753

information for the Kendall τ_b rank order correlation, for reasons of brevity we did not give tables of τ_b values since overall the behavior is the same as for SROC which is nicely illustrated by the figures given.

It can be directly read from the comparison of low and high quality ranges that the VQIs, with a few exceptions, perform worse for the lower quality range than the higher quality range. Furthermore, the reduced performance for the lower quality range can not be reduced to a lower number of

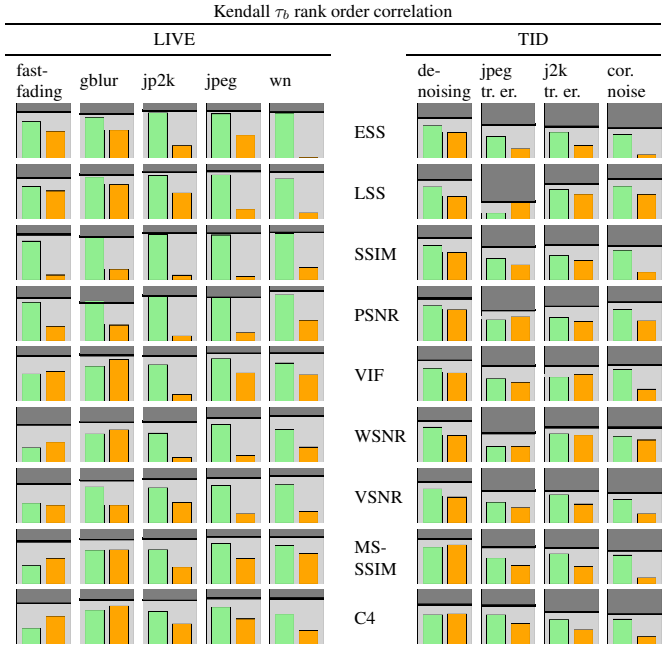


Fig. 2. For each entry the τ_b for the full range of qualities is indicated by the line separating the light and dark background ranging from $\tau_b = 0.0$ at the bottom to $\tau_b = 1.0$ at the top. Superimposed are two bar charts showing the τ_b for the high quality range on the left side and the low quality range on the right side.

Table 3. LIVE Image Quality Assessment Database, low quality (DMOS $>$ 80, *70)

	fastfading	gblur*	jp2k	jpeg	wn
ESS	0.686	0.74	0.291	0.583	0.062
LSS	0.738	0.828	**0.573**	0.268	0.210
SSIM	0.179	0.365	0.027	0.095	0.348
PSNR	0.398	0.451	0.227	0.243	0.549
VIF	**0.767**	**0.919**	0.191	**0.703**	0.692
WSNR	0.514	0.76	0.100	0.209	0.414
VSNR	0.492	0.547	0.491	0.257	0.303
MS-SSIM	0.665	0.819	0.409	0.607	**0.741**
C4	0.741	0.882	0.436	0.666	0.342

comparison images since the higher quality range used is the same distance from the mean of the DMOS values and thus, roughly, the same number of comparison images are used. A notable VQI is the VIF which displays good performance for all cases except high compression rates for JPEG2000 compression where it is among the worst. All VQIs however show certain deficiencies regarding low quality images, even though the actual deficiency is dependant on the distortion introduced. Furthermore, even if the distortion is known beforehand, the evaluation over the full database can be misleading. As an example the best VQI to evaluate highly compressed JPEG2000 images would be the LSS, but the overall performance of LSS regarding JPEG2000 compression is among the worst.

Furthermore, while the reduction in performance is usually in the lower end of the quality spectrum this is not always so. Compare for example the performance of C4 and WSNR for the high and low quality range of the fastfading and gblur testsets. For both VQIs and both testsets the performance on highly impaired images is better than for high quality version.

Table 4. TID2008 Tampere Image Database 2008 (v1.0)

	denoising	jpeg tr. er.	j2k tr. er.	cor. noise
ESS	0.922	0.827	0.789	0.777
LSS	0.912	0.460	0.850	**0.917**
SSIM	0.930	0.818	0.840	0.833
PSNR	0.942	0.752	0.831	0.916
VIF	0.919	0.858	0.851	0.870
WSNR	0.934	0.738	0.834	0.848
VSNR	0.929	0.806	0.791	0.766
MS-SSIM	**0.957**	0.874	**0.853**	0.819
C4	0.918	**0.901**	0.808	0.777

Table 5. TID2008 Tampere Image Database 2008 (v1.0), high quality (DMOS > 3.5)

	denoising	jpeg tr. er.	j2k tr. er.	cor. noise
ESS	0.815	0.596	0.676	0.613
LSS	0.798	0.179	**0.744**	**0.819**
SSIM	0.834	0.582	0.615	0.738
PSNR	0.856	0.577	0.578	0.805
VIF	0.801	0.623	0.634	0.781
WSNR	0.846	0.438	0.720	0.677
VSNR	0.823	0.575	0.715	0.619
MS-SSIM	**0.868**	0.680	0.739	0.703
C4	0.774	**0.745**	0.640	0.655

3.2. Evaluation on the TID Database

The second comparison will be based on the TID Tampere Image Database 2008[4]. Due to reasons of space we only present a subset of the 17 testsets contained in the database, these are "Image denoising", "JPEG transmission errors", "JPEG2000 transmission errors" and "Spatially correlated noise" abbreviated as "denoising", "jpeg tr. er.", "j2k tr. er." and "cor. noise" respectively in the tables. The comparison is again between the full quality range, Table 4, the low quality part DMOS of lower than 3.5, table 6, and the high quality range greater than 3.5, table 5. Again, in these tables values with SROC lower than 0.5 are underlined to show low correlation and the best score per testset is printed in a bold font.

Overall the comparison again shows that the VQIs perform worse for a low quality subset than a high quality subset, even if the performance over the full range is high. However, there are some VQIs which perform better for lower qualities. The LSS for example performs better on the low quality version of the JPEG transmission error testset than on the full quality range.

Furthermore, like with the LIVE database, a high performance of an VQI over the whole quality range can not be taken as indicator that the VQI will perform optimally in either a low or high quality range.

[4] http://www.ponomarenko.info/tid2008.htm

Table 6. TID2008 Tampere Image Database 2008 (v1.0), low quality (DMOS < 3.5)

	denoising	jpeg tr. er.	j2k tr. er.	cor. noise
ESS	0.634	0.248	0.335	0.109
LSS	0.628	0.471	0.649	**0.685**
SSIM	0.650	0.393	0.507	0.224
PSNR	0.773	**0.600**	0.513	0.584
VIF	0.690	0.562	**0.691**	0.357
WSNR	0.682	0.411	0.686	0.590
VSNR	0.652	0.425	0.480	0.269
MS-SSIM	**0.868**	0.497	0.451	0.180
C4	0.743	0.567	0.415	0.247

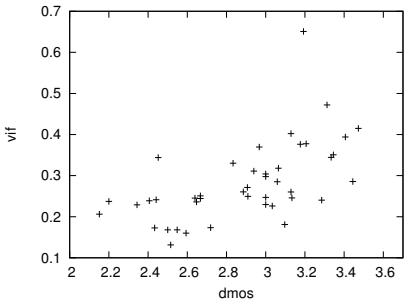


Fig. 3. Scatter plot of VIF over DMOS for the JPEG 2000 transmission errors testset from the TID 2008 database for low quality images (DMOS < 3.5)

To examine in more detail which effects lead to this problem fig. 3 gives a scatter plot of VIF ratings over DMOS for the low quality range of the J2K transmission errors testset, containing 44 of the 100 image in the testset. It can clearly be seen that the overall tendency of higher VIF for higher DMOS ratings holds. It is also clear that locally a high variance in VIF ratings can be observed leading to large mismatches. To illustrate the problem observe that the lowest quality according to VIF would be at about DMOS 2.5, resulting in 8 images being rated higher quality than the corresponding observers would, this is more than 18% of the image in the low quality testset. The same holds regarding the highest quality image according to the VIF.

4. CONCLUSION

It was shown that even seemingly well performing VQIs actually have flaws which can be seen under close scrutiny. When performance is measured only using the full quality range

provided by image evaluation databases these flaws tend to be concealed since the overall correlation between DMOS and VQI overrides the large variance when taking into account a subset of qualities. Two statements can be made regarding the availability of VQIs and the corresponding testing of VQIs. One, there is a lack of VQIs which target the low quality images and performs well over a wide range of distortion types. There are VQIs which are well suited for evaluation of such images when the distortion type is known in advance and a proper VQI can be chosen, however, this becomes less clear when a mix of two or more distortion types can be expected. Second, the evaluation process of VQIs should be done in a more elaborate way, specifically it should differentiate between overall fitness and fitness on low and high quality ranges to better identify shortcomings of certain VQIs. While at least a split into low and high qualities should be done, it might be expedient to differentiate between low, medium and high quality ranges where the database allows, i.e. when enough levels of distortion in the database exist to keep the significance high.

5. REFERENCES

[1] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," http://live.ece.utexas.edu/research/quality.

[2] N. Ponomarenko, F. Battisti, K. Egizarian, J. Astola, and V. Lukin, "Metrics performance comparison for color image database," in *Fourth international workshop on video processing and quality metrics for consumer electronics*, Arizona, USA, Jan. 2009, p. 6 p.

[3] Mark A. Masry and Sheila S. Hemami, "CVQE: A metric for continuous video quality evaluation at low bit rates," in *Human Vision and Electronic Imaging*, Bernice E. Rogowitz and Thrasyvoulos N. Pappas, Eds., Santa Clara CA, Jan. 2003, vol. 5007 of *SPIE Proceedings*, pp. 116–127.

[4] E. P. Ong, W. Lin, Zhongkang Lu, S. Yao, and M. H. Loke, "Perceptual quality metric for h.264 low bit rate videos," in *IEEE International Conference on Multimedia and Expo*, Toronto, Ont., July 2006, pp. 677–680.

[5] L. Superiori, O. Nemethove, W. Karner, and M. Rupp, "Cross-layer detection of visual impairments in h.264/avc video sequences streamed over umts networks," in *Proceedings of IEEE 1st International Workshop on Cross Layer Design*, Jinan, Shandong, China, Sept. 2007, pp. 96–99.

[6] M. Ries, O. Nemethova, and M. Rupp, "Video quality estimation for mobile h.264/AVC video streaming," *Journal of Communications*, vol. 3, no. 1, pp. 41–50, 2008.

[7] Michael Gschwandtner, Andreas Uhl, and Peter Wild, "Transmission error and compression robustness of 2D Chaotic Map image encryption schemes," *EURASIP Journal on Information Security*, vol. 2007, no. Article ID 48179, pp. doi:10.1155/2007/48179, 16 pages, 2007.

[8] Thomas Stütz and Andreas Uhl, "On efficient transparent JPEG2000 encryption," in *Proceedings of ACM Multimedia and Security Workshop, MM-SEC '07*, New York, NY, USA, Sept. 2007, pp. 97–108, ACM Press.

[9] E. P. Ong, M. H. Loke, W. Lin, Zhongkang Lu, and S. Yao, "Perceptual quality metric for h.264 low bit rate videos," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, HI, 2007, vol. 1, pp. 889–892.

[10] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, June 2008.

[11] Y. Mao and M. Wu, "Security evaluation for communication-friendly encryption of multimedia," in *Proceedings of the IEEE International Conference on Image Processing (ICIP'04)*, Singapore, Oct. 2004, IEEE Signal Processing Society.

[12] Damon Chandler and Sheile Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, Sept. 2007.

[13] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[14] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th IEEE Asilomar Conference on Signals, Systems and Computers*, Nov. 2003, pp. 1398–1402.

[15] Mathieu Carnec, Patrick Le Callet, and Dominique Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing: Image Communication*, vol. 23, no. 4, pp. 239–256, Apr. 2008.

[16] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, May 2006.

[17] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

SELECTIVE ENCRYPTION OF THE MC-EZBC BITSTREAM AND RESIDUAL INFORMATION

Heinz Hofbauer and Andreas Uhl

Department of Computer Sciences
University of Salzburg
{hhofbaue, uhl}@cosy.sbg.ac.at

ABSTRACT

When selective encryption is used for security in DRM schemes some information of the original bitstream is intentionally left in plain text. This can have various reasons, e.g. generating preview versions for try and buy scenarios. In the case of the MC-EZBC there is also the goal of retaining the scaling capability in the encrypted domain. However, since parts of the bitstream remain in plaintext this information is available to a potential attacker at all times. In this paper we will assess which attacks can be done with this residual information. Consequently we will extend a prior version of selective encryption for the MC-EZBC to include motion vectors.

1. INTRODUCTION

The use of digital video in todays world is ubiquitous. Videos are viewed on a wide range of clients, ranging from hand held devices with QVGA resolution (320x240) over PAL (768x576) or NTSC (720x480) to HD 1080p (1920x1080) or higher. Furthermore, streaming servers should be able to broadcast over the internet with regard to a wide range of bandwidths, from fixed high bandwidth lines like ADSL2 to various low bandwidths for mobile wireless devices. In such an environment it is simply not possible to encode a video for each application scenario. So content providers either have only a fixed number of options available or they use scaling video technology to adapt the video for bandwidth and resolution requirements of the client. The concept of creating the content once and adapting it to the current requirements is preferable and is better known as Universal Multimedia Access (UMA) [10].

One of the enabling technologies of UMA is the use of scalable video coding. This averts the need for transcoding on the server side and enables the server to scale the video. However, even scaling takes up computation time and reduces the number of connections the server can accept. Furthermore, variable bandwidth conditions, which happen frequently on mobile devices, further taxes the server with the need to adapt the video stream. The solution to this is usually in-network adaption, shifting the need to scale to a node in the network where a change in bandwidth is occurring. The core adaption with these restrictions takes place on the server and adaption due to varying channel capability is done in-network. For design options and comparisons of in network adaption of the H.264/SVC codec see Kuschnig et al. [8]. Wu et al. [11] give an overview of other aspects of streaming video ranging from server requirements to protocols, to QoS etc.

For video streaming in the UMA environment, i.e. a high number of possible bandwidths and target resolutions, wavelet based codecs should be considered. Wavelet based codes are intrinsically highly scalable and rate adaption as well as spatial and temporal scaling can easily be done. Furthermore, wavelet based codecs achieve a coding performance similar to H.264/SVC, c.f. Lima et al. [9]. For an overview about wavelet based video codecs and a performance analysis as well as techniques used in those codecs see the overview paper by Adami et al. [1]. Under similar considerations Eeckhaut et al. [4] developed a complete server to client video delivery chain for scalable wavelet-based video. The main concern of research regarding UMA is usually performance with respect to scaling and in-network adaption. However, digital rights management and security is also a prime concern.

These considerations on network streaming and the inherent scaling capability of wavelet based codecs lead to the development of a selective encryption approach [6] for the MC-EZBC(motion compensated embedded zeroblock coder) [7, 3] video codec. In this approach information was left in plain text in order to be format compliant, meaning that even the encrypted bitstream is decodeable by a standard decoder. Additionally, this approach allows scalability in the encrypted domain.

In section 1.1 an overview will be given about security, selective encryption and objectives of an attack. In order to facilitate the understanding of the encryption method and attacks a short overview of the MC-EZBC bitstream will be given in section 1.2.

In section 2 we will investigate the information which was intentionally left in plain text, namely motion fields and header information in order to mount attacks on the video sequence. While we will specifically look at the MC-EZBC video codec similar attacks are possible on other video and image codecs, e.g. [5] for a header information attack on JPEG2000.

In section 3 the selective encryption method will be extended to include motion vectors and section 4 will give a summary over the attacks and the extended encryption approach.

1.1 Overview Over Selective Encryption

Selective encryption refers to encrypting, carefully selected, parts of a plaintext. Two common reasons for this approach are reduction in resources, usually time saved when only a part of a plaintext is encrypted, and maintaining properties of the plaintext in the encrypted domain. The discussed selective encryption approach for the MC-EZBC is of the second kind where the objective is to retain the ability to scale the encrypted bitstream.

Furthermore selective encryption can be utilized to pro-

tect only parts of the bitstream for digital rights management (DRM) scenarios, e.g. a freely decodeable preview version with embedded but encrypted high quality version.
The possible security goals we want to achieve with selective encryption in different DRM scenarios are as follows:

Confidentiality Encryption means MP security (message privacy). The formal notion is that if a system is MP-secure an attacker can not efficiently compute any property of the plaintext from the ciphertext [2].

Sufficient Encryption means we do not require full security, just enough security to prevent abuse of the data. Regarding video this could for example refer to destroying visual quality to a degree which prevents a pleasant viewing experience.

Transparent Encryption means we want people to be able to view a preview version of the video but in a lower quality while prevent them from seeing a full version. This is basically a pay per view scheme where a lower quality preview version is available from the outset to attract the viewers interest. The distinction is that for sufficient encryption we do not have a minimum quality requirement, and often encryption schemes which can do sufficient encryption cannot ensure a certain quality and are thus unable to provide transparent encryption.

Regarding attacks the focus will be to breach message privacy under the assumption that the visual data is fully encrypted. We will look at header and motion field information and determine what information can be produced regarding the content of the video sequence.

1.2 The MC-EZBC Bitstream

A schematic overview of the MC-EZBC bitstream is given in fig. 1 and an illustration of the decomposition of a GOP is given in fig. 2. The main layout is a header followed by GOP sizes (this is the size of the image data in a GOP) followed by a sequential ordering of GOPs. Each GOP is lead by a header, giving scene change information, i.e. which frames are I frames, followed by the motion field and image data. Motion field and image data are kept separate. For image data the frames are ordered lowest to highest temporal resolution (which is equal to lowest to highest temporal frequency bands). Likewise for each frame the image data is stored from lowest to highest resolution (which is equal to lowest to highest spatial frequency bands). Motion vector fields are stored lowest to highest temporal resolution and in order of frame for each temporal band, in case a given frame is stored as an I-frame the motion vector field for this frame is omitted. Each base layer and each enhancement layer is stored as chunk of data (not shown in the figure), meaning a leading header giving the length of the data block followed by the data block itself.

For a parsing of the bitstream the layout into chunks is beneficial since we do not have to search for marker sequences but can directly skip large parts of the file. Also when headers, including chunk headers, and GOP size information is kept intact the whole bitstream can subsequently be parsed correctly, which is important to be able to scale after the encryption.

2. RESIDUAL INFORMATION

The original approach to selective encryption of the MC-EZBC [6] leaves the header and motion information unencrypted. The motion information is left unencrypted in order to be able to decode the bitstream with the original MC-EZBC implementation. The header information is used for scaling and has to be changed when scaling is performed, so an encryption is not possible. In the following this residual information from the selective encryption approach will be used to gain information about the encrypted video sequence. The akiyo, bus, coastguard, container, flower, foreman, mobile, news, silent, tempete and waterfall sequences are used to perform these tests. In the following subsections we refer to full selective encryption which means the format compliant encryption of image data, cf. fig. 1 and [6], leaving header information and motion fields in plain text.

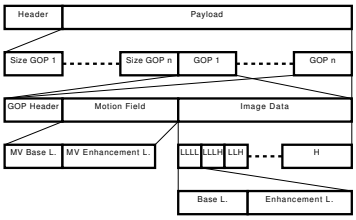


Figure 1: The layout of the MC-EZBC bitstream

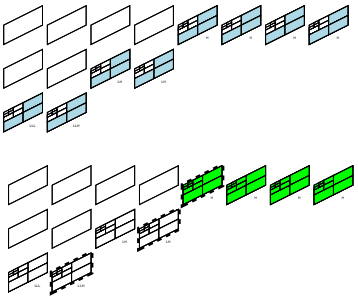


Figure 2: Overview of the decomposition of a GOP with GOP size 8 with marked high temporal layer (lower part), high spatial layer (upper part) and possible I frames as dashed outline on the lower part.

2.1 Header Information

Assuming an attacker intercepts a video stream which is encrypted using full selective encryption. Assuming further that we do have a catalog of available videos from the source of the stream. If this information is present can we identify the video sequence which was intercepted? If this is possible message privacy would be breached since an attacker is able to identify the video sequence.

Since the header information is available a video stream with the same scaling parameters (bitrate and resolution) can be requested. The size of the motion field and visual data is

a part of the plain text headers in the encrypted stream. Using this information we can identify whether the requested stream matches the intercepted stream. Also note that this can be done even if the new stream is sent encrypted and we have no possibility of decrypting it. For each stream requested a similarity score S will be calculated in comparison to the intercepted stream as follows,

$$S = \sum_{i \in MV} (o_i - c_i)^2$$

where MV is the set of indices of motion vector chunk lengths, o_i and c_i are the length of the ith motion vector chunk of the original and comparison sequence respectively.

In the following experiment the sequences were split into subsequences, each 8 frames in size, in order to simulate a larger catalog of video sequences as well as to show that even for this low number of frames the similarity score identifies the source sequence with precision.

In fig. 3 a plot is shown where the waterfall32 subsequence, starting at frame 32, is compared to other subsequences, including waterfall. The dashed line shows subsequences not connected to the waterfall sequence, the solid line show subsequences from the waterfall sequence and the mark at the abscissa shows the waterfall32 subsequence.

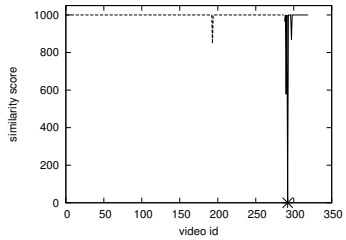


Figure 3: Similarity calculation for various sequences compared to waterfall32.

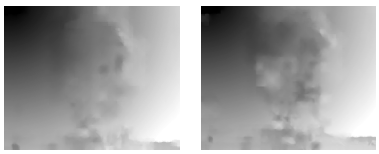
The plot is artificially capped at 1000 in order to show the more interesting lower range of similarities. From the illustration it can be seen that other subsequences originating from the same sequence can also give a similarity response, overall this is because the type of motion throughout subsequences are quite similar. The only other subsequence with a decent similarity response is news8, sequence id 193, which is a near static image with a downward motion from dancers in the center similar to the motion in the waterfall sequence. This similarity of motion over 8 frames is the reason for the response, the longer the subsequences the lower the similarity response outside a video sequence will be.

2.2 Motion Vector Information

Assuming an attacker intercepts a full selective encrypted sequence it is possible to inject image data into the bitstream in order to gain information about the content, which again breaches message privacy.

A visual object can be injected into the video sequence by encoding a still image sequence of the injected object and merging the two sequences using the motion information from the original sequence and the visual data from the still image sequence. The main header can be kept since it is the same for both sequences resulting from using the same parameters for encoding the still image sequence. Motion header and image header information is taken from the respective sequence. This leaves only the GOP length information to be adjusted which is a trivial task. Regarding which object to inject there are two possible courses, one is to analyze the motion field in order to gain information about the sequence. The other is to identify the sequence by using the header information as described in the previous section and utilize side channel information.

By analyzing the motion field it is relatively easy to determine in which parts of the image actual motion is happening as opposed to general movement like panning or zooming. A simple way of doing this is injecting a gradient image and watch the resulting sequence. In the example of the foreman sequence it is easily discernible that the sequence is of the head and shoulders type, see fig. 4.



Frame 50 Frame 100

Figure 4: Frames 50 and 100 of the foreman sequence with injected gradient image.

The related attack is given in fig. 5 where a head is inserted into the foreman sequence. For encoding a GOP length of 128 was used and just the first GOP of the sequence will be used here. A head which roughly fits the proportions of the moving object in the center of the image was inserted, the inserted head has not the exact right size nor the right proportions. Note that the background in the inserted image was left blank since there is nearly no background motion in this GOP to work with.
While only two frames of the sequence are compared in fig. 5 it can be seen that the inserted head goes through the same motion as the original foreman head. In the actual video sequence even the movements of the mouth are perceivable. In any case the quality is a dramatic improvement over a direct decoding of the encrypted sequence, frame 15 and 62 are shown in fig. 6.

Under the assumption that the video sequence can be identified through the header information a search can be done for still images from the actual sequence. Given that such a still image can be found, either a preview version or a screenshot of the video sequence, a much better approximation can be done. In the example given in fig. 7 we used frame 20 of the foreman sequence to inject. The steps for injecting the image data are the same as for the more general case, but the result is much better. This is mostly due to the pictures being more similar and thus the artefacts introduced by motion compensation are less visible.

This second attack using motion vectors also makes the identification of the video sequence through header informa-

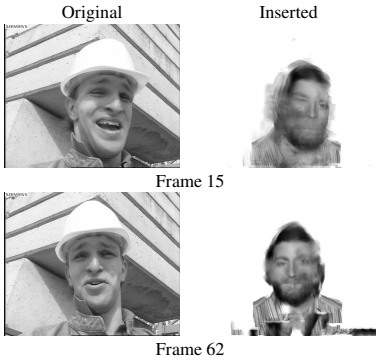


Figure 5: Foreman frames 15 and 62 compared with an injected image of a head.

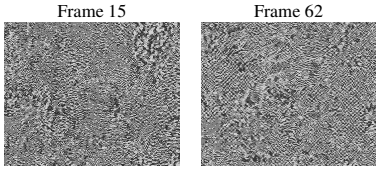


Figure 6: Frame 15 and 62 of a direct decoding of the encrypted foreman sequence.

tion much more dangerous. Not only do we gain knowledge about the video sequence but we can mount a more effective attack on the sequence.

3. SELECTIVE ENCRYPTION WITH MOTION VECTORS

The current MC-EZBC video codec supports scalable motion vectors [12]. Motion vectors are available for each temporal resolution and thus are structured in order of temporal resolution first and frame order in the given resolution second. In terms of the bitstream the motion data is, like the image data, given in chunks, i.e. a leading header gives the length information of the following block of data. The amount of motion data in relation to the whole bitstream, depending on the bitrate of the sequence, ranges from 0.5% (full bitrate) to nearly 40% (128kbps) under full temporal resolution.

The primary goal of adding encryption to motion vectors is still to keep the scalability intact in the encrypted domain. However unlike with corrupt image data the decoder is far less resistant to errors in the motion vectors. This results in format compliance only on a bitstream level, i.e. the bitstream can still be parsed and scaled, but the standard decoder will most likely be unable to deal with the random input of the encrypted motion vectors.

The encryption of the data in motion field chunks is not

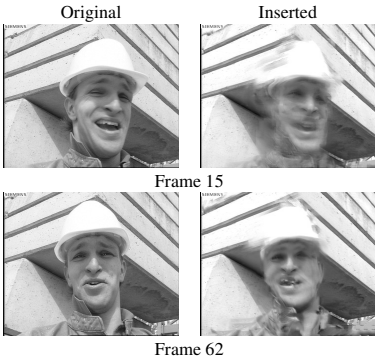


Figure 7: Foreman frames 15 and 62 compared with an injected image of foreman frame 20.

block aligned so a stream cipher has to be used. Furthermore, scaling away higher temporal resolution can disrupt ciphers in feedback mode, like AES in OFB, when the feedback is used over all chunks. Consequently it is best like with the original version of the encryption algorithm to use feedback only in a given chunk. The motion data encryption alone can not be used for sufficient or transparent encryption.

In order to assess the encryption of motion vectors only two attacks are used. One is the injection of a zero motion field into the bitstream similar to what is described in section 2.2. The other is to fix up the decoder to prevent it from crashing during motion field decoding. In the case where motion data is required beyond the bound of a chunk we introduce a one bit spike to prevent the decoder from locking up in a loop waiting for a symbol. Furthermore, the referencing to image data outside the boundaries of a given frame is prevented. The fix of the decoder will in the following be referred to as "mvfix" attack.

For sufficient encryption, depending on the video sequence, the quality can be too high. Figure 8 shows the PSNR of the tempete sequence, high global motion, and silent sequence, a head and shoulder sequence with low global motion, for injection and mvfix attacks. In this attacks all motion fields were encrypted. For sequences with distinct global motion the mvfix attack does better because the residuals are distributed throughout the image while for a zero motion field the residual information is accumulated which leads to severe color bleeding. Figure 9 illustrates the color bleeding effect frames 62 and 250 of the tempete sequence. This effect becomes less distinct when the GOP size decreases. Additionally, the mvfix attack introduces more jitter resulting in a lower viewing quality.

Regarding transparent encryption the problem is how to control a target quality. The way to use motion vector encryption for transparent encryption would be to force the receiver to downscale on the temporal resolution, i.e. reducing the frame rate. Since the downsampling is done with wavelets the difference from the original frames are somewhat hard to measure since video quality indices (VQI)

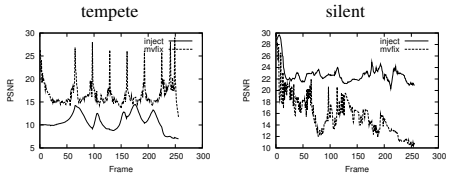


Figure 8: PSNR plot showing the comparison of the injection and mvfix attacks on the tempete and silent sequence with GOP size 256

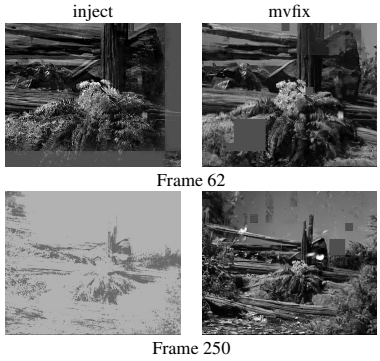


Figure 9: Comparison of injection and mvfix attacks based on frames 62 and 250 of the tempete sequence with GOP size 256.

like PSNR would rate the blurring effects introduced by the downsampling as severe degradation even if the content is still viewable. Furthermore, the impact of zero motion injection or mvfix attacks are hard to evaluate purely on the basis of a VQI.

4. CONCLUSION

It was shown that confidentiality can not be reached with selective encryption for the MC-EZBC, header data alone can be used to identify a video sequence. Motion fields if left unencrypted have been shown to compromise content, i.e. an approximation of the content can be created using only motion vectors.

An enhancement of a selective encryption scheme to include motion vectors has been introduced and discussed in detail. The encryption of motion vectors alone has been shown to be insufficient for transparent or sufficient encryption schemes. However, the encryption of motion vectors can prevent reconstruction attacks as presented in this paper and should be used in conjunction with the selective encryption of image data.

Furthermore, since header data has to be left in plain text in order to allow scalability in the encrypted domain the identification attack is always possible. This shows that full cryptographic security can only be achieved with traditional methods, e.g. AES encryption over the whole bitstream.

REFERENCES

[1] N. Adami, A. Signoroni, and R. Leonardi. State-of-the-art and trends in scalable video compression with wavelet-based approaches. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(9):1238–1255, September 2007.

[2] M. Bellare, T. Ristenpart, P. Rogaway, and T. Stegers. Format-preserving encryption. In *Proceedings of Selected Areas in Cryptography, SAC '09*, volume 5867, pages 295–312, Calgary, Canada, August 2009. Springer-Verlag.

[3] P. Chen, K. Hanke, T. Rusert, and J. W. Woods. Improvements to the MC-EZBC scalable video coder. In *Proceedings of the IEEE Int. Conf. Image Processing ICIP*, volume 2, pages 81–84, Barcelona, Spain, 2003.

[4] H. Eeckhaut, H. Devos, P. Lambert, D. De Schrijver, W. Van Lancker, V. Nollet, P. Avasare, T. Clerckx, F. Verdicchio, M. Christiaens, P. Schelkens, R. Van de Walle, and D. Stroobandt. Scalable, wavelet-based video: From server to hardware-accelerated client. *Multimedia, IEEE Transactions on*, 9(7):1508–1519, November 2007.

[5] Dominik Engel, Thomas Stütz, and Andreas Uhl. Format-compliant JPEG2000 encryption in JPSEC: Security, applicability and the impact of compression parameters. *EURASIP Journal on Information Security*, 2007(Article ID 94565):20 pages, 2007.

[6] Heinz Hofbauer and Andreas Uhl. Selective encryption of the MC EZBC bitstream for DRM scenarios. In *Proceedings of the 11th ACM Workshop on Multimedia and Security*, pages 161–170, Princeton, New Jersey, USA, September 2009. ACM.

[7] Shih-Ta Hsiang and J. W. Woods. Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank. *Signal Processing: Image Communication*, 16(8):705–724, May 2001.

[8] R. Kuschnig, I. Kofler, M. Ransburg, and H. Hellwagner. Design options and comparison of in-network H.264/SVC adaptation. *Journal of Visual Communication and Image Representation*, pages 529–542, September 2008.

[9] L. Lima, F. Manerba, N. Adami, A. Signoroni, and R. Leonardi. Wavelet-based encoding for HD applications. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1351–1354, July 2007.

[10] A. Vetro, C. Christopoulos, and T. Ebrahimi. From the guest editors - Universal multimedia access. *IEEE Signal Processing Magazine*, 20(2):16 – 16, 2003.

[11] Dapeng Wu, Yiwei Thomas Hou, Wenwu Zhu, Ya-Qin Zhang, and Jon M. Peha. Streaming video over the internet: approaches and directions. In *Circuits and Systems for Video Technology, IEEE Transactions on*, volume 11, pages 282–300, Mar 2001.

[12] Yongjun Wu, Konstantin Hanke, Thomas Rusert, and John W. Woods. Enhanced MC-EZBC scalable video coder. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(10):1432–1436, October 2008.

AN EFFECTIVE AND EFFICIENT VISUAL QUALITY INDEX BASED ON LOCAL EDGE GRADIENTS

Heinz Hofbauer and Andreas Uhl

Department of Computer Sciences
University of Salzburg
{hhofbaue, uhl}@cosy.sbg.ac.at

ABSTRACT

The structural similarity index measure is a well known and widely used full reference visual quality index. In this paper we introduce a new full reference visual quality index based on local edges and edge gradients in the wavelet domain. The proposed metric corresponds better to human judgement and is more efficient, in terms of computational complexity, than the structural similarity index measure. Furthermore, the proposed metric is more efficient than other state of the art metrics and surpasses them for certain visual impairment classes.

Index Terms— Image analysis, Quality control

1. INTRODUCTION

The assessment of image and video quality is important for transmission (assessment of transmission errors) and compression (compression vs. quality). Optimally an evaluation where human observers judge the perceived quality or impairment should be performed. However, the time and cost requirements to perform such tests are high. Thus, algorithms to assess image quality automatically are employed, which are referred to as visual quality indices (VQI). These VQIs are in turn compared to human assessment over a number of distortion and compression types via established databases. It is important for VQIs to strongly correlate to human judgement. Furthermore, ease of use and low computational complexity are desired traits.

Still widely used the peak signal-to-noise ratio (PSNR) is unrivaled in speed and ease of use. However, it is also well known that the correlation to human judgement is somewhat lacking [1]. With this problem in mind newer VQIs were developed which take the human visual system (HVS) into account in order to increase the correlation with human judgement.

These VQIs utilize the knowledge of the HVS to a lesser or higher extent. However, the trend over all VQIs is the more information about the HVS is included in the generation of a quality score the more complex and time consuming the VQI becomes. This ranges from the fast luminance and edge similarity score (LSS and ESS) as introduced by Mao and Wu [2] to the refined but slow visual information fidelity criterion (VIF) by Sheikh and Bovik [3] and CPA1 by Carosi et al. [4]. In terms of HVS LSS and ESS uses the basic knowledge that edge information reflects the assessment of humans regarding the shape or contour of objects and the luminance score reflects changes in the color space. The VIF on the other hand uses a more refined model which starts with the modeling of the reference image using natural scene statistics (NSS). Furthermore, the possible distortion is modeled as signal gain and additive noise in the wavelet domain and parts of the HVS which have not been covered by the NSS are modeled, i.e. internal neural noise is modeled by using a additive white Gaussian noise model.

However, the most widely used VQI is the structural similarity index measure (SSIM) by Wang et al. [5] because it offers an excellent tradeoff between performance and quality. The SSIM extracts three separate scores from the image and combines them into the final score. First the visual influence is calculated locally then luminance, contrast and structural scores are calculated globally. These separate scores are then combined with equal weight to form the SSIM score.

In this paper we propose a VQI which is more efficient as well as more effective than the SSIM and for certain cases is even better than VQIs which are heavily modelled on the HVS. The proposed metric is a full reference metric, i.e. both original and impaired image are used, based on local edge direction and gradient which exploits properties of the HVS to some extent. Based on the utilization of local edge gradients we will refer to this metric as LEG for the rest of the paper.

Local edge information is a good feature when it comes to image comparison, classification or retrieval. Local binary patterns (LBP) have been widely employed to assess local edge information. LBPs compare a center pixel to its neighborhood to gain an edge description, and have been successfully used from face recognition, e.g. [6], to medical image classification, e.g. [7]. Usually these LBPs are used in the form of histograms in the comparison of images. In our case image based histograms do not provide enough error localization, thus a different approach of direct comparison is used.

Furthermore, we use a one-step wavelet decomposition

which allows the utilization of HVS based knowledge to increase the accuracy of error assessment. The decomposition serves a number of purposes. First the peripheral vision is taken into account by using the low frequency subband. The eye of an observer focuses on one point but also takes in surrounding information. Furthermore, the HVS is sensitive to different spatial and temporal frequencies which are present in a stimulus and this is modelled by using the high frequency subbands. Additionally, error information is perceived stronger among large edge structures in the image and the high frequency bands can be used for structure detection. In the proposed algorithm this structural information is used, by means of local edge gradients, as a weighting function for the local error features generated form the low frequency bands. Furthermore, the difference of overall luminance will be used to model the light sensitivity of the eye allowing to detect impairments which do not change the structure of the image.

In order to assess how the LEG performs in comparison to other metrics based on edge features the local feature based visual security metric (LFBVS) by Tong et al. [8] as well as the natural image contour evaluation (NICE) quality index by Rouse et al. [9] will be evaluated and compared to LEG. The LFBVS uses basic color features as well as edge amplitude and direction on local blocks. The NICE uses gradient maps on different scales, adjusts for possible image shift by using a morphological dilation with a plus shaped structuring element. The actual score is computed by doing a thresholding on the image and calculating differences. For the implementation of the NICE we choose the version with only one scale, Sobel edge detector and morphological dilation since it is significantly faster than using steerable wavelet decomposition but shows similar performance, c.f. [9]. For weighting the hamming distance and normalization as given in [10] was used.

The implementations of NICE, LFBVS, LEG and SSIM which are used in the following sections are available online at `http://www.wavelab.at/sources/VQI/`.

In section 2 the algorithm for the LEG VQI will be given. In section 3 we give the evaluation and analysis of the VQI with respect to LIVE, MICT and IVC image database as well as a comparison to other metrics. Section 4 concludes the paper.

2. ALGORITHM

Let I and O denote the impaired and original (gray scale) image of size $W \times H$ with maximum pixel value $M = 2^b$ with b bits per pixel. The following steps are performed to calculate the LEG index.

Step 1: The following steps only use edge difference, consequently a change in the image which does not influence the structural influence would go unnoticed. To compensate for this the difference in luminance between I and O is calculated:

$$\mathrm{lum}(I,O) = 1 - \sqrt{\frac{|\mu(O) - \mu(I)|}{M}},$$

$$\mu(X) = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H} X(x,y),$$

where $X(x,y)$ is the pixel value of image X at position x, y.

Step 2: One step wavelet decomposition with Haar wavelets resulting in four sub images for each image X denoted as X_0 for the LL-subband, and X_1, X_2, X_3 for LH, HH and HL subband respectively. Figure 1 illustrates the decomposition of a sample image, upper left is X_0 and the numbering continues clockwise.

Fig. 1: Wavelet decomposition of the lighthouse2 image with the Haar wavelet.

Step 3:A local edge map is calculated for each position x, y in the image, reflecting the change in coarse structure of the image. Through the Haar wavelet decomposition and the comparison of each center pixel with its neighborhood the actual area of influence of the original image is a 6×6 window. The local edge map is used to prevent faulty information gained from gradients on parts of the image where the edges are off by a certain degree.

$$\mathrm{le}(I,O,x,y) = \begin{cases} 1 & \text{if } EDC(I,O,x,y) = 8, \\ 0.5 & \text{if } EDC(I,O,x,y) = 7, \\ 0 & \text{otherwise.} \end{cases}$$

$$EDC(I,O,x,y) = \sum_{p \in N(x,y)} ED(I,O,x,y,p)$$

$$ED(I,O,x,y,p) = \begin{cases} 1 & \text{if } I(x,y) < I(p) \text{ and } O(x,y) < O(p), \\ 1 & \text{if } I(x,y) > I(p) \text{ and } O(x,y) > O(p), \\ 0 & \text{otherwise.} \end{cases}$$

where $N(x,y)$ is the eight neighborhood of the pixel x, y. Edge extension is done by copying the last edge value, e.g. $I(-1,0) := I(0,0)$.

Step 4: In order to assess the contrast changes a difference of gradients in a neighborhood will be calculated. The effect of a contrast change will be more pronounced along large edges. Thus, this step serves as a contrast sensitivity function as well as a detector for large structures in the image.

$$\mathrm{led}(I,O,x,y) = \frac{1}{8}\sum_{p\in N(x,y)}\left(1-\sqrt{\frac{|LD(I,O,x,y,p)|}{M}}\right)^2$$

$$LD(I,O,x,y,p) = (O(x,y)-O(p)) - (I(x,y)-I(p))$$

Step 5: The edge score is calculated by using local edge conformity (le) and local edge difference (led).

$$\mathrm{es}(I,O) = \frac{4}{WH}\sum_{x=1}^{\frac{W}{2}}\sum_{y=1}^{\frac{H}{2}}\Big(\mathrm{le}(I_0,O_0,x,y) * \frac{1}{3}\sum_{i=1}^{3}\mathrm{led}(I_i,O_i,x,y)\Big).$$

Step 6: The LEG visual quality index is calculated by combining es and lum:

$$\mathrm{LEG}(I,O) = \mathrm{lum}(I,O)\,\mathrm{es}(I,O).$$

An example of the generated edge score and edge conformity is given in fig. 2, where the original image O (lower left) as well as an impaired image I (top right) is given in low and high quality. The top left illustrates the values of $\mathrm{le}(I,O)$ and the lower right gives the average of the led values as used in **step 5**, i.e. $\frac{1}{3}\sum_{i=1}^{3}\mathrm{led}(I_i,O_i,x,y)$. The values have been mapped onto a continuous gray scale, $[0,1]\mapsto[0,255]$, where black indicates a high amount of errors and white represents no distortion between I and O.

2.1. Analysis of Weight Functions

The features used in the metric are average luminance (μ), edge direction conformity (EDC) and contrast change difference (LD) which are combined with weight functions in order to model their impact on the HVS. For example, should a small difference in μ have a huge impact, super linear weight, or a small impact, sub linear weight. In order to find the best weight function for each feature the individual weight functions were tested on the IVC database.

It turned out that super linear weight performed better, indicating that the HVS is able to detect even small aberrations from the original. For this reason we tested various functions, as depicted in fig. 3, with varying degrees of superlinearity: different slope for linear with capping at the boundaries, as used in le for EDC; logarithmic or square root, again with different forms of superlinearity, e.g., normal square root as used in lum for μ or squared as used in led for LD.

(a) High Quality

(b) Low Quality

Fig. 2: Comparison of high and low quality error maps. The original image is in the lower left corner, the impaired version is in the top right corner, the edge error map is top left and the difference of gradient error map is lower right.

Table 1 shows results for this test, for reasons of brevity only some results are presented. The table gives the Spearman rank order correlation (SROC) for a given function on each testset of the IVC database as well as the summation of the SROC scores for overall comparison. The test uses the final weight functions for all features except for the feature indicated in the table for which the given weight function is evaluated. The weight functions indicated in the table are those illustrated in fig. 3.

The trends towards the selected weight functions as used in the final version of LEG is clearly visible.

3. EXPERIMENTAL RESULT AND ANALYSIS

In order to evaluate the LEG VQI we compare it to the widely used SSIM. Furthermore, a comparison to slow VQIs which highly exploit the HVS, such as VIF and CPA1, is included. In addition a comparison to fast VQIs, PSNR and LSS, which rely less on the HVS are performed. Furthermore, a com-

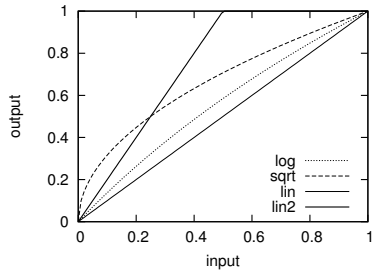


Fig. 3: Different forms of weight functions, log is actually calculated as $log = \log(1+x)/\log(2)$ to compensate for the singularity at 0 and lin2 is $\min(2x, 1)$.

		Testset					
feature	weight	lar	j2000	flou	jpeg	lumichr	sum
LEG		0.8504	0.8837	0.9692	0.8962	0.9392	4.5386
μ	lin	0.8535	0.8708	0.9669	0.8942	0.9473	4.5327
μ	log	0.8535	0.8708	0.9669	0.8942	0.9473	4.5327
LE	sqrt	0.7448	0.8361	0.9263	0.8015	0.6985	4.0072
LE	log	0.8038	0.8724	0.9429	0.8375	0.8073	4.2639
LE	lin	0.8087	0.8871	0.9519	0.8551	0.8615	4.3643
LE	lin2	0.8390	0.8918	0.9692	0.8909	0.9338	4.5248

Table 1: Result of the test of a selected number of weight functions and features on the IVC database.

parison with the NICE and LFBVS, which use similar image features as LEG, is included.

The comparison is done based of the LIVE (release 2) [11], MICT [12] and IVC [13] databases. Three different image databases were used because a single database can be biased due to different evaluation methodology, number of participants, demography, expertise of participants etc. An example of this are the JPEG and JPEG 2000 compression testsets which are contained in each of the three databases. Throughout this section the absolute value of the Spearman rank order correlation (SROC) is given. We also used the Kendall tau (τ) measure for evaluation and both measures coincide.

The SROC for LIVE is given in table 2a, for MICT table 2b lists the results and for IVC the results are given in table 2c, where each table shows separate entries per testset. Likewise the results for τ are given in table 4. The main focus of the comparison is SSIM versus LEG and the higher correlation per testset is given in bold for both SROC and τ for easier reference. Table 3 shows the SROC for NICE and LFBVS in comparison to LEG and SSIM, for brevity reasons no Kendall τ results are given for this comparison.

Overall the LEG outperforms the SSIM except for two

(a) LIVE

testset	LSS	PSNR	SSIM	LEG	VIF	CPA1
fastfading	0.843	0.891	0.942	**0.971**	0.965	0.881
gblur	0.916	0.782	0.903	**0.966**	0.972	0.927
jp2k	0.953	0.895	0.936	**0.945**	0.968	0.958
jpeg	0.970	0.881	0.946	**0.960**	0.984	0.962
wn	0.965	0.985	**0.962**	0.960	0.985	0.984

(b) MICT

testset	LSS	PSNR	SSIM	LEG	VIF	CPA1
jpeg	0.839	0.285	0.631	**0.938**	0.907	0.725
jpeg2000	0.908	0.860	**0.915**	0.914	0.956	0.923

(c) IVC

testset	LSS	PSNR	SSIM	LEG	VIF	CPA1
flou	0.889	0.805	0.869	**0.969**	0.973	0.902
j2000	0.938	0.850	0.851	**0.884**	0.936	0.927
jpeg	0.952	0.674	0.805	**0.896**	0.924	0.906
lumichr	0.914	0.563	0.749	**0.939**	0.878	0.834
lar	0.863	0.699	0.711	**0.850**	0.888	0.881

Table 2: Absolute Spearman rank order correlation (SROC) for LSS, PSNR, SSIM, LEG, VIF, CPA1 on the IVC, MICT and LIVE databases.

cases. For the white noise testset on the LIVE database the SSIM reaches a SROC of 0.96151 while the LEG only scores 0.96000, a negligible difference given that the performance over the other testsets is significantly higher. The other exception is the JPEG 2000 testset on the MICT database where again the SSIM outperforms the LEG by a small margin. However, for the JPEG 2000 testsets on both LIVE and IVC the LEG performs better. This fluctuation is most likely due to the, relatively low, number of human observations which are used as ground truth.

When it comes to class of similar metrics we see that LEG outperforms NICE on all three databases and all testsets. LFBVS on the other hand is slightly better than LEG on the IVC lar testset and significantly better than LEG, and all other metrics including VIF and CPA1, on the IVC lumichr testset.

For some testsets the LEG even surpasses the VIF and CPA1 VQIs. On average the CPA1 is surpassed by LEG although only by a small margin, the LEG is on average surpassed by the VIF with an equally small margin. For metrics which use similar features the LEG on average outperforms both LFBVS and NICE. The average difference of a given VQI to the LEG is shown in table 5, and table 6 for NICE and LFBVS, for the SROC on a per database basis. Positive results indicate a VQI which is on average better than the LEG on the given database.

(a) LIVE

testset	SSIM	LFBVS	NICE	LEG
fastfading	0.942	0.932	0.959	**0.971**
gblur	0.903	0.937	0.949	**0.966**
jp2k	0.936	0.853	0.925	**0.945**
jpeg	0.946	0.920	0.950	**0.960**
wn	**0.962**	0.891	0.910	0.960

(b) MICT

testset	SSIM	LFBVS	NICE	LEG
jpeg	0.631	0.814	0.861	**0.938**
jpeg2000	**0.915**	0.584	0.881	0.914

(c) IVC

testset	SSIM	LFBVS	NICE	LEG
flou	0.869	0.955	0.924	**0.969**
j2000	0.851	0.825	0.882	**0.884**
jpeg	0.805	0.889	0.860	**0.896**
lumichr	0.749	**0.969**	0.809	0.939
lar	0.711	**0.853**	0.795	0.850

Table 3: Absolute Spearman rank order correlation for SSIM, LFBVS, NICE, LEG on IVC, MICT and LIVE databases.

(a) LIVE

testset	LSS	PSNR	SSIM	LEG	VIF	CPA1
fastfading	0.668	0.707	0.784	**0.848**	0.841	0.698
gblur	0.751	0.584	0.726	**0.841**	0.858	0.761
jp2k	0.811	0.711	0.771	**0.785**	0.843	0.817
jpeg	0.849	0.691	0.795	**0.822**	0.892	0.827
wn	0.835	0.894	**0.832**	0.824	0.894	0.888

(b) MICT

testset	LSS	PSNR	SSIM	LEG	VIF	CPA1
jpeg	0.644	0.199	0.446	**0.791**	0.737	0.524
jpeg2000	0.742	0.682	**0.752**	0.747	0.821	0.764

(c) IVC

testset	LSS	PSNR	SSIM	LEG	VIF	CPA1
flou	0.741	0.667	0.709	**0.878**	0.899	0.741
j2000	0.803	0.726	0.693	**0.705**	0.790	0.784
jpeg	0.822	0.519	0.627	**0.719**	0.791	0.750
lumichr	0.752	0.442	0.563	**0.813**	0.718	0.631
lar	0.659	0.571	0.571	**0.656**	0.713	0.705

Table 4: Absolute Kendall tau (τ) for LSS, PSNR, SSIM, LEG, VIF, CPA1 on the IVC, MICT and LIVE databases.

database	LSS	PSNR	SSIM	VIF	CPA1
LIVE	−0.031	−0.073	−0.023	0.015	−0.018
MICT	−0.053	−0.354	−0.153	0.005	−0.102
IVC	0.003	−0.189	−0.111	0.012	−0.018

Table 5: Comparison of average SROC per database in relation to LEG.

database	SSIM	LFBVS	NICE
LIVE	−0.023	−0.053	−0.022
MICT	−0.153	−0.227	−0.055
IVC	−0.111	−0.010	−0.054

Table 6: Comparison of average SROC per database in relation to LEG.

3.1. Runtime Efficiency Analysis

For an efficiency analysis the LEG metric is compared to the SSIM, LSS, PSNR, NICE and LFBVS. The CPA1 and VIF are not included since they are only available as matlab implementations and the matlab overhead would unfairly influence the results. However, the CPA1 uses a full frame Fourier transformation which is of complexity $\mathcal{O}(N \log N)$ while the LEG is $\mathcal{O}(N)$. And for the VIF, Sheikh et al. [3] evaluated that the VIF is 6.5 times slower than the MS-SSIM. The MS-SSIM is a multi scale variant of the SSIM utilizing wavelet decompositions and repeat calculations of the SSIM for different subbands, thus the VIF is at least 6.5 times slower than the SSIM.

As testset the 779 impaired images from the LIVE database were used. The results of the comparison are given in table 7 where it is shown that LEG is faster than the SSIM for best, worst and average case but can not match the efficient simplicity of LSS or PSNR. For the average case LEG is 6 times faster than the SSIM.

For the edge based metrics, NICE, LFBVS and LEG, the performance is similar with the LEG outperforming LFBVS and NICE by a factor of 1.46 and 1.22 respectively.

metric	$\bar{t}$	t_{min}	t_{max}
SSIM	764 ms	588 ms	882 ms
LFBVS	186 ms	140 ms	247 ms
NICE	155 ms	107 ms	193 ms
LEG	127 ms	94 ms	198 ms
LSS	52 ms	40 ms	83 ms
PSNR	55 ms	43 ms	68 ms

Table 7: Runtime performance of the given metrics over 779 impaired images of the LIVE database.

4. CONCLUSION

We have proposed a novel visual quality index based on local edge features augmented by knowledge about the human visual system. We have given the algorithmic details of a new visual quality index, LEG, and have evaluated it in terms of runtime efficiency and correlation with human judgement.

It was shown that the proposed VQI is more effective, i.e., corresponds better to human judgement, than the SSIM and is superior in runtime efficiency. Furthermore, on average the proposed metric is more effective than the CPA1 and has a far superior runtime efficiency than either CPA1 or VIF.

The LEG also has a higher runtime efficiency than NICE and LFBVS, which use image features similar to those used in LEG. Furthermore, LEG is more effective than NICE and on average more effective than LFBVS.

5. REFERENCES

[1] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, June 2008.

[2] Y. Mao and M. Wu, "Security evaluation for communication-friendly encryption of multimedia," in *Proceedings of the IEEE International Conference on Image Processing (ICIP'04)*, Singapore, Oct. 2004, IEEE Signal Processing Society.

[3] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, May 2006.

[4] Maurizio Carosi, Vinod Pankajakshan, and Florent Autrusseau, "Towards a simplified perceptual quality metric for watermarking applications," in *Proceedings of SPIE, Multimedia on Mobile Devices*, San Jose, CA, USA, Jan. 2010, vol. 7542, SPIE.

[5] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[6] Shengcai Liao, Xiangxin Zhu, Zhen Lei, Lun Zhang, and Stan Li, "Learning multi-scale block local binary patterns for face recognition," in *Advances in Biometrics*, pp. 828–837. Springer, 2007.

[7] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, "Pit pattern classification using extended local binary patterns," in *Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine (ITAB'09)*, Larnaca, Cyprus, Nov. 2009.

[8] Lingling Tong, Feng Dai, Yongdong Zhang, and Jintao Li, "Visual security evaluation for video encryption," in *Proceedings of the international conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 835–838, ACM.

[9] S. S. Hemami D. Rouse, "Natural image utility assessment using image contours," in *IEEE International Conference on Image Processing (ICIP'09)*, Cairo, Egypt, Nov. 2009, pp. 2217–2220.

[10] S. S. Hemami D. Rouse, "The role of edge information to estimate the perceived utility of natural images," in *Western New York Image Processing Workshop (WNYIP)*, Rochester, NY, Sept. 2009, p. 4 pp.

[11] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," http://live.ece.utexas.edu/research/quality.

[12] Yuukou Horita, Keiji Shibata, and Yoshikazu Kawayoke, "MICT image quality evaluation database," http://mict.eng.u-toyama.ac.jp/mictdb.html.

[13] Patrick Le Callet and Florent Autrusseau, "Subjective quality assessment IRCCyN/IVC database," 2005, http://www.irccyn.ec-nantes.fr/ivcdb/.

Secure Transport and Adaptation of MC-EZBC Video Utilizing H.264-based Transport Protocols☆

Hermann Hellwagner[b], Heinz Hofbauer[a,*], Robert Kuschnig[b], Thomas Stütz[a], Andreas Uhl[a]

[a] *Department of Computer Sciences, University of Salzburg*
Jakob-Haringer-Straße 2, 5020 Salzburg, Austria
[b] *Institute of Information Technology, Klagenfurt University*
Universitätsstraße 65–67, 9020 Klagenfurt, Austria

Abstract

Universal Multimedia Access (UMA) calls for solutions where content is created once and subsequently adapted to given requirements. With regard to UMA and scalability, which is required often due to a wide variety of end clients, the best suited codecs are wavelet based (like the MC-EZBC) due to their inherent high number of scaling options. However, most transport technologies for delivering videos to end clients are targeted toward the H.264/AVC standard or, if scalability is required, the H.264/SVC. In this paper we will introduce a mapping of the MC-EZBC bitstream to existing H.264/SVC based streaming and scaling protocols. This enables the use of highly scalable wavelet based codecs on the one hand and the utilization of already existing network technologies without accruing high implementation costs on the other hand. Furthermore, we will evaluate different scaling options in order to choose the best option for given requirements. Additionally, we will evaluate different encryption options based on transport and bitstream encryption for use cases where digital rights management is required.

Keywords: Scalable Video Coding (MC-EZBC), In-network Adaptation, RTP/SRTP MANE, generic Bitstream Syntax Description (gBSD), Video Encryption, Selective Encryption, Format Compliance

1. Introduction

The use of digital video in today's world is ubiquitous. Content consumers desire to retrieve content through a multitude of networks, from 3G to broad-

☆Supported by Austrian Science Fund (FWF) project P19159-N13.
*Corresponding author
Email addresses: hermann.hellwagner@uni-klu.ac.at (Hermann Hellwagner), hhofbaue@cosy.sbg.ac.at (Heinz Hofbauer), robert.kuschnig@uni-klu.ac.at (Robert Kuschnig), tstuetz@cosy.sbg.ac.at (Thomas Stütz), uhl@cosy.sbg.ac.at (Andreas Uhl)

band Internet, on a broad range of consumer devices, from cell phones to high performance PCs. However, consumers do not care about the technicality necessary to provide the content over this wide range of networks but rather about their quality of experience (QoE), i.e., they want to consume the best possible quality in a timely manner. This creates a problem for content providers since it is costly, in both time and storage space consumption, to provide content for every conceivable end device and network link. Re-encoding on the other hand is expensive in the way that it requires significant time which reduces the QoE for end users.

The solution to this problem is called Universal Multimedia Access (UMA) [1]. The goal of UMA is to encode content once and adapt it in a timely manner to current end user requirements. One of the enabling technologies of UMA is the use of scalable video coding. This averts the need for transcoding on the server side and enables the server to scale the video. However, even scaling requires computation time and reduces the number of connections the server can accept. Furthermore, variable bandwidth conditions, which happen frequently on mobile devices, further tax the server with the need to adapt the video stream. The solution to this is usually in-network adaptation, shifting the need to scale to the node in the network where a change in bandwidth is occurring. The core adaptation with these restrictions takes place on the server and adaptation due to actual channel capability is done in-network.

For video streaming in the UMA environment, i.e., a high number of possible bandwidths and target resolutions, wavelet based codecs should be considered. Wavelet based codecs are naturally highly scalable and rate adaptation as well as spatial and temporal scaling is easily achieved. Furthermore, wavelet based codecs achieve a coding performance similar to H.264/SVC, c.f. Lima et al. [2]. Under similar considerations Eeckhaut et al. [3] developed a complete server to client video delivery chain for scalable wavelet-based video. However, there are already standardized ways of transporting multimedia data, namely the Real-time Transport Protocol (RTP) [4]. Similarly, there is a protocol for handling a single or several time-synchronized stream of continuous media, e.g., audio and video, the Real Time Streaming Protocol (RTSP) [5] which can use RTP as its mode of transportation. Besides RTP and RTSP the MPEG-21 Part 7 "Digital Item Adaptation" (DIA) [6] can be used to provide content related metadata. A codec agnostic description, the generic Bitstream Syntax Description (gBSD) [7], can also be used as a basis for an informed adaptation process.

In order to use existing technology, i.e., RTP streaming and in network adaptation, modules for handling the motion compensated embedded zero bit codec (MC-EZBC) have to be created to facilitate packetization for RTP and media awareness for adaptation nodes. However, the existing technology can already deal with H.264/SVC, e.g., [8] describes the H.264/AVC payload for RTP and multimedia aware network elements (MANE) and [9] extends this to H.264/SVC. Since the H.264/* bitstream is build from network abstraction layer units (NALUs), the fastest route to utilize the existing infrastructure is to encapsulate the MC-EZBC into a NALU bitstream which presents itself as H.264/SVC to those components. Following this route it is, apart from the

MC-EZBC to NALU conversion, trivial to use the existing infrastructure. Also note that, while we only take a look at MC-EZBC to NALU conversion, such a conversion can be constructed for other scalable video codecs and the theoretical and experimental analysis will by and large also hold for those conversions.

In this paper we will provide a method of encapsulating the MC-EZBC into a NALU bitstream. Additionally, we will investigate how this encapsulated bitstream can be transported, encrypted, and scaled, and at what cost in terms of payload overhead and network delay. Furthermore, we will look at surrounding issues which have to be taken into account, e.g., initial vectors for encryption.

In section 1.1 we will describe the basics of the chosen wavelet based video codec, the MC-EZBC, in section 2.1 a description of the layout of the bitstream will be given and the adaptation to the RTP packetization scheme will be given in section 2.2. An overview of the MPEG-21 DIA generic Bitstream Syntax Description (gBSD) will be given in section 2.3. Section 2.4 describes additional requirements for the RTP streaming process for the MC-EZBC and presents the outline of the encapsulation process.

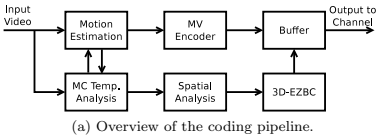
The main concern of research regarding UMA is usually performance with respect to scaling and in-network adaptation. However, digital rights management and security is also a prime concern for providers of commercial videos. Furthermore there are a range of other aspects of video streaming, ranging from server requirements to protocols, to QoS etc., Wu et al. [10] give a good overview of these aspects. General principles and possible goals of digital rights management (DRM) will be explained in section 1.2 and application of encryption to the MC-EZBC codec will be discussed in section 3.

In section 4 we will compare the different aspects and options of the adaptation and streaming process theoretically and experimentally.

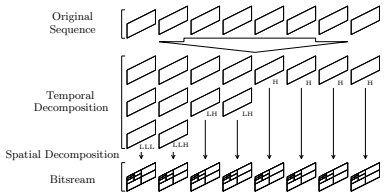
1.1. The Motion Compensated Embedded Zero Bit Codec (MC-EZBC)

For reasons of scalability which fit the UMA principle we use the enhanced MC-EZBC wavelet based video codec for in-network adaptation. This choice was made mainly because the source code is available [1], which enables our experiments. The MC-EZBC codec [11, 12, 13, 14] is a scalable t-2D video codec which uses motion compensated temporal filtering, with 5/3 CDF wavelets, followed by regular spatial filtering, with 9/7 CDF filtering, an overview of the encoding pipeline is given in fig. 1a. This method, temporal first and spatial later, is referred to as t+2D coding scheme, see fig. 1b for an example of this decomposition for a group of picture (GOP) size of 8. For temporal filtering a full decomposition is used and thus the GOP size is discernible by the number of temporal decomposition levels. Both temporal and spatial filtering is done in a regular pyramidal fashion. Statistical dependencies are exploited by using a bit plane encoder, the name giving embedded zero bit coder. Motion vectors are encoded with DPCM followed by an arithmetic coding scheme.

[1] The source for the ENH-MC-EZBC is available from `http://www.cipr.rpi.edu/research/mcezbc/`.



(a) Overview of the coding pipeline.



(b) Decomposition of a GOP of size 8 showing the arrangement of temporally and spatially decomposed frames in the bitstream.

Figure 1: MC-EZBC encoding overview.

For an overview of wavelet based video codecs and a performance analysis as well as techniques used in those codecs see the overview paper by Adami et al. [15]. Again, while we concentrate only on the MC-EZBC in this paper the encapsulation process described later can in a modified version still be applied to other scalable video codecs. Likewise the analysis performed will also be indicative for other scalable video codecs.

1.2. Overview of Encryption and Digital Rights Management

Shannon's work [16] on security and communication shows that the highest security is reached through a secure cipher operating on almost redundancy free plain text. Current video codecs exploit redundancy for compression and we can consider the bitstream to be a redundancy free plain text in the sense of Shannon. Thus for maximum security we just need to encrypt the whole bitstream with an state of the art cipher, i.e., the Advanced Encryption Standard (AES) [17]. However, the choice was made to keep information in plain text in order to facilitate scalability in the encrypted sequence. Regarding security, Lookabaugh et al. [18] showed that such a selective encryption is sound and demonstrated its relation to Shannon's work. However, Said [19] showed that side information can compromise security.

Thus, we can differentiate between:

Traditional Encryption or full encryption where the full range of the plaintext is encrypted and security in the sense of Shannon is achieved.

Selective Encryption or partial encryption where, carefully selected, parts of the plaintext are left unencrypted. Two common reasons for this approach are reduction in resources, usually time saved when only a part of a plaintext is encrypted, or maintaining properties of the plaintext in the encrypted domain.

The encryption approach used for the MC-EZBC is of the second kind where the objective is to retain the ability to scale the encrypted bitstream, which is not possible when using traditional encryption.

Furthermore selective encryption can be utilized to protect only parts of the bitstream for digital rights management (DRM) scenarios, e.g., a freely decodeable preview version with embedded but encrypted high quality version. The possible security goals we want to achieve with selective encryption in different DRM scenarios are as follows:

Confidentiality Encryption means MP security (message privacy). The formal notion is that if a system is MP-secure an attacker can not efficiently compute any property of the plaintext from the ciphertext [20].

Sufficient Encryption means we do not require full security, just enough security to prevent abuse of the data. Regarding video this could for example refer to destroying visual quality to a degree which prevents a pleasant viewing experience.

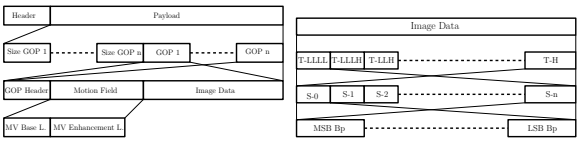
Transparent Encryption means we want consumers to be able to view a preview version of the video but in a lower quality while preventing them from seeing a full version. This is basically a pay per view scheme where a lower quality preview version is available from the outset to attract the viewer's interest. The distinction is that for sufficient encryption we do not have a minimum quality requirement, and often encryption schemes which can do sufficient encryption cannot ensure a certain quality and are thus unable to provide transparent encryption.

2. Particulars of the Protocols

In this section we will describe the details of the MC-EZBC bitstream which are required to perform scaling. Furthermore we will describe the NALU bitstream requirements related to the encapsulation of the MC-EZBC bitstream in order to provide scalability on the transport layer. Likewise, the subset of gBSD syntax elements related to describing the MC-EZBC bitstream are discussed. The requirements introduced by utilizing the RTP are explained and an overview of the process which encapsualtes the MC-EZBC bitstream into gBSD and NALU with respect to RTP are presented.

2.1. MC-EZBC Bitstream

The basic layout of the MC-EZBC bitstream is depicted in fig. 2a and a more detailed overview of the 'image data' required for fine grain scalability is shown



(a) General bitstream layout. (b) Layout of image data in the bitstream.

Figure 2: Layout of the MC-EZBC bitstream.

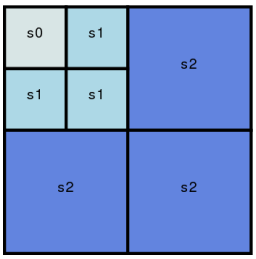


Figure 3: Grouping of decompositions for a frame with two spatial decomposition levels.

in fig. 2b. The bitstream is lead by a general header giving resolution, frame rate, prediction options etc., most of which stay the same during scaling. The header however has three fields we need to adjust when scaling is performed: a `bitrate` field giving the bit rate to which the bitstream is scaled, `t_level` giving the number of temporal layers dropped and `s_level` giving the number of spatial layers dropped. The header is followed by a GOP size list giving the size of a GOP without GOP header size and motion field, i.e., only specifying the image data size. For any scaling done the GOP size list has to be adjusted to reflect the new size of image data.

Following this general information are the motion and image data ordered by GOP, i.e.: Header, motion vectors of GOP 1, image data of GOP 1, motions vectors of GOP 2, and so on. Each GOP contains a GOP header, containing scene change information, i.e., which frames are encoded as I frames. Following the GOP header is the motion field for the current GOP. The GOP header and motion field are not changed during scaling, i.e., motion vectors are not scaled with the image data. Following the motion field is the image data in frame order of temporal decomposition, c.f. fig. 1b and fig. 2 lower part.

The layout of the image data consists of a number of data chunks consisting

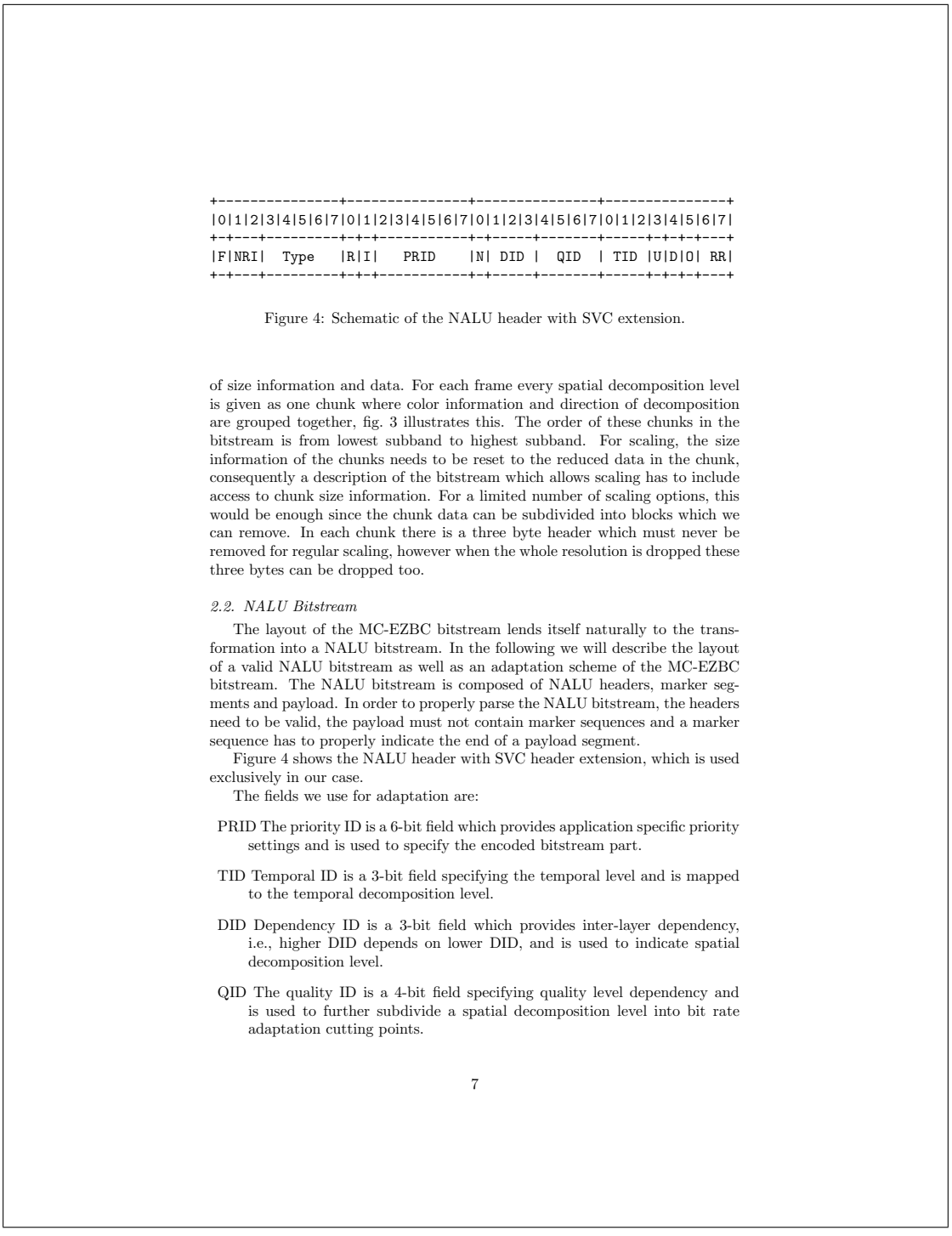


Figure 4: Schematic of the NALU header with SVC extension.

of size information and data. For each frame every spatial decomposition level is given as one chunk where color information and direction of decomposition are grouped together, fig. 3 illustrates this. The order of these chunks in the bitstream is from lowest subband to highest subband. For scaling, the size information of the chunks needs to be reset to the reduced data in the chunk, consequently a description of the bitstream which allows scaling has to include access to chunk size information. For a limited number of scaling options, this would be enough since the chunk data can be subdivided into blocks which we can remove. In each chunk there is a three byte header which must never be removed for regular scaling, however when the whole resolution is dropped these three bytes can be dropped too.

2.2. NALU Bitstream

The layout of the MC-EZBC bitstream lends itself naturally to the transformation into a NALU bitstream. In the following we will describe the layout of a valid NALU bitstream as well as an adaptation scheme of the MC-EZBC bitstream. The NALU bitstream is composed of NALU headers, marker segments and payload. In order to properly parse the NALU bitstream, the headers need to be valid, the payload must not contain marker sequences and a marker sequence has to properly indicate the end of a payload segment.

Figure 4 shows the NALU header with SVC header extension, which is used exclusively in our case.

The fields we use for adaptation are:

PRID The priority ID is a 6-bit field which provides application specific priority settings and is used to specify the encoded bitstream part.

TID Temporal ID is a 3-bit field specifying the temporal level and is mapped to the temporal decomposition level.

DID Dependency ID is a 3-bit field which provides inter-layer dependency, i.e., higher DID depends on lower DID, and is used to indicate spatial decomposition level.

QID The quality ID is a 4-bit field specifying quality level dependency and is used to further subdivide a spatial decomposition level into bit rate adaptation cutting points.

More specifically, since we always use the SVC extension header, the header type (denoted by `Type`) is always set to 20. The priority ID reflects the type of data from the original MC-EZBC bitstream: header information (PID 0), GOP header information (PID 1), motion field (PID 2) and image data (PID 3). The GOP length information of the original MC-EZBC bitstream is dropped.

In order to ensure that no marker sequences appear in the bitstream, an escape sequence can be used to escape such marker information. The following table shows the transforms:

`0x000000` → `0x00000300`
`0x000001` → `0x00000301`
`0x000002` → `0x00000302`
`0x000003` → `0x00000303`

Also note that the escaped sequences are not allowed to appear in the bitstream but since this is done by inserting `0x03` and the fact that `0x000003` is also in the marker sequence list this problem solves itself.

Another problem with transforming the bitstream is that the NALU header is prefixed with a marker sequence which is of the form `0x0000 (00)* 01`. Usually three byte sequences are used, except for the the first header which uses a four byte sequence as a synchronization marker. The problem is the arbitrary number of zero bytes in the marker sequence. The specification was done with H.264/SVC in mind where an encoded slice can not end in `0x00`. For the MC-EZBC however this is not the case and thus a trailing zero byte would be counted as belonging to a marker and be lost. To fix this, we append `0x03` to the end of every payload.

The transformation from the NALU bitstream to a MC-EZBC bitstream is a bit more complicated. The data in the NALU bitstream follows the same order as the bitstream representation of the MC-EZBC, i.e., no reordering has to be performed. But since we need to reconstruct the header information for the MC-EZBC bitstream in case scaling occurred, we need to put the information in a treelike structure representing the temporal and spatial decompositions of the MC-EZBC. This is done by monitoring drops in NALU header fields, i.e., a drop in a field refers to parsing a lower *ID value than the previous parsed *ID value. For example, if a drop in the QID occurs we move to a different spatial decomposition or a different frame, depending if a drop in DID is also detected, or to a different GOP and so on. After this is done, we need to restructure the whole bitstream in order to find the maximum decomposition levels, e.g., if there is a resolution drop in one GOP, the other GOPs need to be adjusted to reflect this, in order to properly determine a resolution for the overall header. When this is done the overall header information is calculated and corrected and the GOP length information which was dropped in the transformation to the NALU bitstream is reconstructed.

2.3. gBSD

The gBSD is part of the MPEG-21 part 7 "Digital Item Adaptation" and is used to describe a bitstream in a format agnostic way. This enables devices to

understand a single high level interface (gBSD) and thus perform operations on a bitstream, e.g., scaling, without knowledge about the actual bitstream. While the gBSD allows more structural information to go into the description, we will keep the bitstream description simple so as not to generate too much overhead. For more information on the tags and attributes used see MPEG-21 part 7 [6].

The gBSD is prefaced with a `dia:DIA` root tag specifying namespaces followed by a `dia:Description` tag specifying the description type (`gBSDType`) followed by address information. Since the MC-EZBC bitstream is byte based, we set it to `addressUnit="byte"` and `addressMode="Absolute"`. The address mode gives the method of accessing parts of the bitstream, this is reflected by the use of start and length attributes in subsequent tags. For the bitstream description we need two different types of tags.

First we need a copy descriptor specifying that a part of the original bitstream should be retained in the scaled version. The `gBSDUnit` tag is used for this purpose, it takes start and length information to mark a part of the bitstream to be kept.

Additionally we need access to the bitstream in positions where the header has to be adapted, e.g., size information in a scaling case. Such information can not be copied over from the original bitstream but has to be adapted depending on the target resolution or bitrate. The `Parameter` tag is used for this purpose and gives the length of the data block to insert into the bitstream. The actual information contained in the parameter is given by the required child `Value`. The attribute `xsi:type` gives the type of data and the content of the tag gives the actual value.

By using `Parameter` and `Value` we can access the actual value and change it according to the adaptation, while the `gBSDUnit` tags let us copy parts of the actual bitstream. Both `Parameter` and `gBSDUnit` also have an attribute `marker` which allows to give a handle to the tag to access it directly.

Figure 5 shows a part of the description of the bitstream for the flower sequence which can be used to scale to 1024kbps and 512kbps. It also shows the description of the header where it can be seen that only the bitrate has to be described as `Parameter` and that it needs to be set to 1024 to properly reflect the bitrate of the stream. The resulting description of the bitstream consists of two `gBSDUnit` descriptions discerning between 512 and 1024 kbps.

In order to perform repeated adaptations in the network, the gBSD has to encompass all adaptation possibilities and has to be kept accurate. In order to do this, the gBSD has to be adapted via extensible stylesheet language transformations (XSLT) which is done on the network adaptation node. However, the more fine grained the adaptation choices should be, the more fine grained the gBSD has to be which results in a bigger gBSD file and a more complicated XSLT script. The gBSD together with the XSLT script produce an overhead which limits the size of the actual bitstream, so it is best to keep them as simple as possible. Furthermore, if no more adaptation steps are necessary, the gBSD file can be dropped, i.e., from the last node in the network to the end device the full channel bandwidth can be used.

Figure 6 illustrates how gBSD is used for adaptation, fig. 6a shows the overall

```
...
<dia:Description xsi:type="gBSDType"
        addressUnit="byte" addressMode="Absolute">
    <gBSDUnit start="0" length="14" marker="hdr1"/>
    <Parameter length="2" marker="bitrate Q0">
        <Value xsi:type="xsd:unsignedShort">1024</Value>
    </Parameter>
    <gBSDUnit start="16" length="80" marker="hdr2"/>
...
    <Parameter length="2" marker="hdr Q0">
        <Value xsi:type="xsd:unsignedShort">118</Value>
    </Parameter>
    <gBSDUnit start="545775" length="18" marker="data"/>
    <gBSDUnit start="545793" length="100" marker="data Q0"/>
    <Parameter length="2" marker="hdr Q0">
        <Value xsi:type="xsd:unsignedShort">185</Value>
    </Parameter>
    <gBSDUnit start="545895" length="21" marker="data"/>
    <gBSDUnit start="545916" length="164" marker="data Q0"/>
</dia:Description>
</dia:DIA>
```

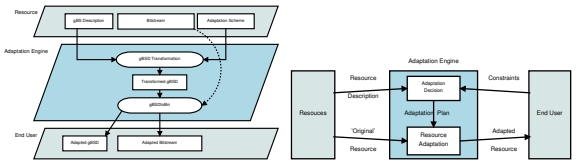
Figure 5: gBSD representation of the flower sequences quality scaling options for 1024 kbps and 512kbps.

layout of an adaptation process, a bitstream and a corresponding gBSD are sent together. According to an adaptation scheme the adaptation engine can scale the bitstream, and adapt the gBSD to fit the scaled bitstream. The adaptation scheme can be fixed, i.e., only certain fixed scaling options are included, or it can be generated based on user preference or requirement, this part of the adaptation engine process is illustrated in fig. 6b. The adaptation based on user preference, especially if more than one user is involved, however increases the size of the gBSD since more options have to be taken into account. Furthermore, either the overhead is increased by creating a more complex adaptation scheme, which anticipates possible user preferences, or the delay is increased by having the adaptation engine request a custom adaptation scheme from the server. A more detailed information about the gBSD adaptation of the MC-EZBC is available in [21]. The paper also shows that there are problems with the gBSD for different types of sequence, like the increase in relative gBSD description size in low motion sequences.

2.4. RTP

Apart from the NALU encapsulation the RTP streaming requires timing information for the packetization, cf. [8]. Furthermore, in order to stream the gBSD with along the same channel utilizing RTP it has to be embedded in the NALU bitstream. This is done by adding supplemental enhancement information (SEI) messages, cf. [22], to the NALU bitstream.

In order to produce timing information for the RTP server, the conversion from MC-EZBC to NALU will also produce an XML output which describes



(a) Overview of a gBSD based adaptation. (b) Overview of the gBSD adaptation process.

Figure 6: Overview of the gBSD adaptation planning process.

the resulting NALU bitstream, including timing information which can be calculated from the framerate given in the original MC-EZBC header and the frame number. This XML description can be used not only as a source of timing information but also as a basis to generate interleaved gBSD descriptions. Should a gBSD be required the produced XML description can be annotated to create the basis of an SEI embedded gBSD description. The annotated XML file can then be broken up to conform to the desired access units (AU) of the bitstream, i.e., the interleaving granularity. This AU gBSD fragments are then compressed and wrapped in an SEI message and inserted into the NALU bitstream in such a way that they precede the AU which they describe. Figure. 7 gives a schematic overview of the transformation process.

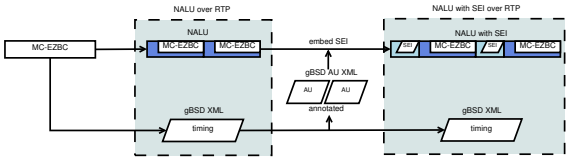


Figure 7: Scheme for MC-EZBC to NALU encapsulation with SEI embedding.

3. Encryption

In order to encrypt the content and still retain the ability to scale there are two options, content encryption, i.e., encrypt the bitstream either on a MC-EZBC or NALU level, or transport encryption. Both methods have advantages and disadvantages regarding computational requirements and security provided.

3.1. Transport Encryption

Transport encryption can be done by using the Secure Real-time Transport Protocol (SRTP), defined in RFC3711 [23]. SRTP is a profile to the Real-

Time Transport Protocol (RTP), defined in RFC3550 [4], providing encryption, message authentication and protection against replay attacks for both uni- and multicast.

The drawback of using the SRTP is the need to decrypt the whole communication on any MANE, where potential scaling takes place. The decryption on each MANE is required, whether scaling is performed or not, in order to inspect the bitstream to determine if scaling has to be performed. This puts a high computational strain on the MANE, which has to decrypt as well as encrypt, compared to encryption only on the server and decryption only on the client. Furthermore, since the key for decryption has to be known on any MANE where scaling can take place each MANE introduces a potential attack point to the system.

On the other hand, we gain security against replay attacks since the whole of the communication is encrypted. Furthermore, the delay for delivery to the consumer is reduced in comparison to prior encryption. Since no prior encryption is employed the streaming can start sooner and the overhead of encryption is distributed in time over the whole streaming process.

However, this option still does not provide confidential security akin to traditional encryption. Due to the headers of the encapsulating SRTP packages remaining in plain text, the length information can be used combined with side channel attacks to compromise security, see Hellwagner et al. [24].

3.2. Bitstream Encryption

For bitstream encryption the choices are either to encrypt the MC-EZBC prior to encapsulating it into a NALU bitstream, or to directly encrypt the NALU bitstream. However, the use of the NALU bitstream for encryption is somewhat problematic. A cipher should optimally produce output resembling a uniform distribution, and thus can create marker sequences which have a special meaning, cf. section 2.2. However, the creation of marker sequences can be prevented or remedied, for a more detailed discussion of NALU encryption see Hellwagner et al. [24], in this paper we will not look into encryption on a NALU level. Encrypting the MC-EZBC on the other hand is easier in technical terms. Through the length information in data chunks, no marker sequences are needed and the content of a chunk can be directly encrypted. Furthermore, the transformation to NALU automatically takes care of possible NALU marker sequences as described in section 2.2. When utilizing UMA, a highest quality source video is used; thus in order to reduce computational cost, encryption should be performed after defining quality levels, i.e., only a part of the source video is used. Furthermore to reduce parsing cost the best option is to include encryption into the NALU encapsulation process. When the encryption is applied just prior to NALU encapsulation, the occurrence of possible marker sequences is automatically taken care of by the encapsulation process.

There are a number of options on how to encrypt the MC-EZBC bitstream depending on the desired results in terms of DRM, i.e., transparent or sufficient encryption, discussed in more detail in [25]. However, in order to allow scaling in the encrypted domain, information about the bitstream has to be kept in

plaintext, i.e., headers. This information can be used in side channel attacks as shown in [26]. In these side channel attacks, the fact that the lengths of encoded video sequence parts correlate to the contained video material is exploited. This can be combined with the information of possible streaming content (the side channel) to identify which video is streamed. While this does not allow an attacker to reconstruct the visual material, the confidentiality is broken. In this section we will mainly look initial vectors for encryption, how the encryption schemes presented in [25] can be used in the NALU encapsulation scenario and how they compare to transport encryption (SRTP).

3.2.1. Considerations for the Initial Vectors

The high scalability of the MC-EZBC bitstream introduces some requirements for a potential encryption method. First and foremost is the ability to perform quality scalability which enables the bitstream to be cut at byte aligned positions. This enforces the use of stream ciphers or block ciphers in streaming mode, e.g., AES in CBC, CFB, OFB or counter mode, [27]. Additionally, due to the scalability in temporal and spatial resolutions as well as scalability in quality, a cipher needs to be restarted for each new chunk. Ciphertext feedback (CBC, CFB) is obviously not able to bridge the resulting gap of data, since ciphertext feedback uses prior ciphertext information to generate a key for following ciphertext. In case of missing data, the keystream for following ciphertext can no longer be constructed. But since information about the original length of a chunk is not kept, pre-ciphertext feedback (OFB, CTR) are also unable to continue over this gap of data. In this case the ciphertext itself is not needed but an iteration is performed in order to construct the keystream and the number of iterations is tied to the length of the missing data. This requires some form of providing an initial vector (IV) for each chunk of data. The data of a chunk has no fixed minimal length and can be scaled down to arbitrary small size. This prohibits the use of plaintext or ciphertext for crafting new IVs for the next chunk in the bitstream.

The solution is to send IVs separately or generate them from a separate source. A separate source could be a single IV which is encrypted to generate a different IV and thus iteratively generate the IVs of the chunks as they appear in the bitstream. This can however lead to synchronization errors, i.e., when a whole GOP is dropped, the next chunk in the bitstream and all subsequent chunks would receive faulty IVs. This happens because the GOPs are not numbered and synchronization can not be restored. Something similar can happen when a whole frame is dropped, then from this frame forward the rest of the GOP would receive a faulty IV. However the next GOP can be properly synchronized because the number of frames in a GOP is known. Similarly, a dropped spatial resolution level would result in the faulty IVs for the rest of the frame and be synchronized at the beginning of the next frame.

This leads to the following options:

- Send a limited number of IVs and generate subsequent IVs by iterated encryption:

  - A single IV is sent at the beginning, resulting in the lowest overhead but can result in synchronization loss for the whole bitstream when a whole GOP is dropped during transport.
  - An IV is sent for each GOP, leading to synchronization at the GOP borders but frame drops can destroy the rest of the current GOP.
  - An IV is sent for each frame, synchronization is now per frame but a resolution drop can destroy the rest of the current frame.
- Send a single IV for each chunk of data in the bitstream. This has the highest overhead but desynchronization can not occur.

Regarding overhead we can give a simple upper bound by looking at the number of spatial resolutions. The number of frames per GOP remains the same since full temporal decomposition is used. Assuming a framerate of f with s spatial decomposition levels we can simply give the overhead as:

$$o_{IV} = f * (s + 1) * b,$$

for a block size of b. For AES of a PAL video, a resolution of 768x576, 6 decomposition steps and 25 frames per second, this would result in an overhead of $o_{IV} = 21.875$kbps. To put this into relation, consider streaming over an old, low bandwidth IEEE 802.11 WLAN with a channel capacity of 2Mbps this would be $\approx 0.01\%$ of the channel capacity. In essence, the overhead of sending frequent IVs is negligible and does hardly impact channel bandwidth. For newer WLAN standards featuring higher bandwidth the overhead of sending frequent IVs becomes even less of a problem.

4. Comparison and Evaluation

In this section we will compare the overhead introduced by encapsulation in NALU and gBSD respectively. Since RTP is used as transport protocol for both NALU and gBSD, we will not take into account the RTP overhead since it is the same for both formats.

4.1. Protocol Overhead

In [28] it is shown that seven quality levels are usually enough to support almost all required target applications. While this is a reasonable goal for comparison we will look at the overhead in a more general fashion. This is done mainly because if a request for a certain bandwidth, framerate or bitrate is issued the MC-EZBC source can be transformed on the fly to support the requested target scalability which can lead to an actual lower number of scaling options. As a result of a lower number of scaling options the encapsulating protocols generate less overhead. The encoding overhead is important since the actual bitrate of the bitstream can only be the channel bitstream minus the required overhead.

In the following we will denote the number of frames as f, the number of temporal decompositions as t and the number of spatial decompositions as s. Consequently we have a GOP size of 2^t and the number of GOPs is $G = f/2^t$, for simplicity we assume that the framerate is a multiple of the GOP size, and in total we have $s+1$ spatial decomposition bands. Furthermore we will denote the number of quality levels by q.

4.1.1. Evaluation of gBSD Overhead

For the number of bytes each descriptive element of the gBSD requires, we use an approximation obtained from empirical analysis of the used bitstreams and resulting gBSD descriptions. While most of the markers have a fixed structure, like element and attribute names, the value of the attributes change depending on the encoded sequence, see fig. 5 as an example containing the `gBSDUnit` element. The average size in bytes a `Parameter` and `gBSDUnit` require are $p = 105$ and $g = 55$ bytes respectively. These numbers are calculated with average variable length information (i.e., length value, start value) but excluding the marker attribute since the value is essentially user defined. Additionally, we have an overhead for the DIA declaration which is 393 bytes, which is the length of the fixed header, see. fig. 5. This means that the start and length fields as well as the value of parameters are only estimated since this information can vary widely. However, the use of a typical marker element is included since the marker will be a near constant in length. We can now calculate an approximate size of the gBSD. The main header consists of three changeable fields, bitrate, spatial and temporal scaling level, with size p and five `gBSDUnit`s of size g which stay constant. The main header is followed by a list of GOP sizes, with one entry per GOP, each entry in the list is given as a `Parameter` with size p. For each GOP we have a single `gBSDUnit` for the GOP header and motion vectors. Then for each frame we have a single chunk for each spatial decomposition level. The chunks here have to be separated into the number of quality levels we want to deal with. The resulting approximation in byte is thus size S:

$$S = 393 + \underbrace{3p + 5g}_{\text{header}} + \underbrace{G\,p}_{\text{GOP size list}} + G\underbrace{(g + 2^t(s+1)(p+qg))}_{\text{single GOP}}$$

For a sequence with 128 frames, $t = 7$ and $s = 2$ this would estimate a gBSD file size of 81kB for two quality levels and 60kB for the downscaled version. However, this assumes that the gBSD is transferred in plaintext which is unusual. A gBSD description is text based and can be compressed quite well, see Augeri et al. [29] for an overview. Furthermore, there are XML aware compression schemes which are designed for ease of access on network nodes and alleviate the need to decompress the description of the whole bitstream, see Timmerer et al. [30] for an overview. For the rest of this paper we will use bzip2 as compressor for gBSD which will compress by an order of magnitude. For a more detailed overview of gBSD regarding MC-EZBC and compression see Hofbauer et al. [21].

Size increase when including gBSD			
	Filesize in byte for		
sequence	NALU	NALU+SEI	increase
bbbunny	12868259	12946658	0.61%
sintel	13424680	13546838	0.91%
football	1139309	1167002	2.43%
harbour	901722	928638	2.98%
crew	676359	702960	3.93%
foreman	449419	474891	5.66%

Table 1: Overhead of bzip2 compressed gBSD SEI inclusion into bitstreams of different quality levels.

While this is an overhead calculation for the whole bitstream it can be used as approximation for the AU based description as well. In order to create a well formed gBSD document the header has to be replicated which increases the overall size, but simultaneously the relative length information from the start of the description is shorter, resulting in lower p and g values. As such the given equation can be either used directly for overhead calculation, or in parts if a better fitting calculation is desired. As an example we can consider the overhead calculation for the whole sequence with GOP based gBSD descriptions. This can be easily done by extracting the single GOP part of the given equation and adding the cost of the header; the resulting overhead has to be taken into account for each GOP. The resulting overhead is

$$S_{AU_{GOP}} = G * (393 + (g + 2^t(s+1)(p+qg)).$$

What is problematic about this overhead is the fact that the overhead size is only dependent on the scaling options but not the quality of the contained bitstream. This means that for a given gBSD description the overhead relative to the size of the bitstream increases with decreasing quality. Table 1 shows an example for this increase in size when including bzip2 compressed SEI messages, as described in fig. 7, for various bitrates. The sequences used in the table are of CIF resolution with GOP size 16, 6 spatial levels and with bitrates 1045kbps(football), 822kbps (harbour), 611kbps (crew) and 398kbps (foreman), and 720p resolution with GOP size 16, 4 spatial levels and bitrates 3072kbps (bbbunny) and 2048kbps (sintel). The CIF sequences have a runtime of 10.24 sec while bbbunny and sintel have a runtime of 33sec and 52sec respectively.

4.1.2. Evaluation of NALU Overhead

For every piece of payload we have to take into account the marker sequence leading up to it (3 bytes), the NALU SVC header (4 bytes) as well as the payload end marker (1 byte). We denote the fixed overhead value as $o_f = 8$. Furthermore we have an overhead of 1 byte since the first NALU marker is a

4 byte synchronization marker, and we have a reduction in size resulting from the drop of the GOP size table of the original MC-EZBC bitstream which gives the overall overhead adjustment $o_o = 1 - G\,4$, since every size entry in the GOP table is a long integer 4 bytes in size. Thus, we can give the overhead as

$$O = o_o + o_f + o_f G(2 + q(s+1)2^t)$$

A NALU is created for the global header and for every GOP q NALUs are created per temporal and spatial resolution. This only reflects the fixed overhead, a further overhead occurs when marker sequences appear in the original bitstream and have to be escaped. However this can not be given in a deterministic fashion. Assuming uniform distribution of byte values we can calculate the chance P of a marker appearing at any given byte position as:

$$P = \underbrace{\frac{1}{2^8}}_{\text{0x00}} * \underbrace{\frac{1}{2^8}}_{\text{0x00}} * \underbrace{\frac{2^2}{2^8}}_{\text{0x}\{00,01,02,03\}} = \frac{1}{2^{22}}.$$

In this unlikely case a single byte is inserted into the three byte sequence, extending it by 4/3, thus on average the size of the bitstream will increase by a factor $F = 1 + \frac{1}{3*2^{20}} \approx 1.00000032$. The increase in size due to this factor is practically negligible. Furthermore, unlike the gBSD overhead this size increase is multiplicative instead of additive, i.e., dependent on the size of the original bitstream. Thus, while the overhead of the gBSD description stays the same for reduced quality versions the overhead due to this factor is reduced together with the bitstream size.

4.2. Encryption Performance

A direct comparison of bitstream encryption and transport encryption is not really possible. Bitstream based encryption is done only on the server and client and introduces a constant delay until streaming can start. Transport encryption (SRTP) on the other hand encrypts while streaming and thus the load on the server and client are distributed over the time it takes to stream the video sequence, but additional load is produced on the MANE where decryption and encryption also has to take place. With SRTP the delay to start streaming is basically shifted to frame delays during transport. As such we will, and can, not provide a direct comparison, rather both methods are looked at differently. Transport encryption will be looked at during the evaluation of adaptation performance since both are tied together.

For bitstream based encryption it is most important to get a notion of how long the delay to start streaming is since this has a direct influence on consumer satisfaction (QoE). In order to evaluate the time requirement for encryption for different DRM scenarios, a number of selective encryption types are used. As a baseline we will use the same cipher used for selective encryption and encrypt the whole bitstream. In order to better gauge the influence of the parsing overhead generated when using selective encryption, the same video sequence is

used but with different quality levels. Table 2a gives the encryption performance for a full quality version of the foreman sequence, the full quality version has a bitrate of about 9.5Mbps. For comparison we use a reduced quality version with a bitrate of 398kbps, which is later also used in the analysis of streaming performance, given in table 2b. For each bitrate version we performed different types of encryption which correlate to possible DRM applications. For more information about the encryption process and resulting quality see [25].

Full selective encryption refers to the encryption of all image data, i.e., excluding headers. Due to the plaintext headers, scaling is still possible with full selective encryption. This method is put in direct comparison with full traditional encryption, i.e., encryption of the whole bitstream including headers and motion fields. This option generates no parsing overhead but does not allow scalability in the encrypted domain. The parsing overhead for both bitrate versions is the same, since the layout of the bitstream is unchanged. This leads to an actual reduction in encryption time for very high quality bitstreams even for full selective encryption. For low bitrates however the overhead is quite significant, in the 398kbps test case the parsing overhead nearly doubles the time required for encryption.

Sufficient encryption refers to a significant reduction in visual quality. This is typically done by encrypting the base layer and leaving the enhancement layers intact. This leads to a significant reduction in encryption time in relation to full selective encryption. The time reduction is more pronounced for higher quality versions of the bitstream because more refinement information is contained in the bitstream and thus the reduction in the amount of data to be encrypted is more pronounced. Table 2 shows the two extremes, on one hand we have a high reduction in quality, and consequently the amount of data which needs encryption. For this case the parsing overhead renders any selective encryption slower than full traditional encryption. On the other hand, the high quality case shows that the parsing overhead becomes negligible in comparison to the amount of data which need encryption. Thus, the higher the quality the more time reduction can be gained from selective encryption.

Transparent encryption usually targets enhancement layer information in order to allow a decoding of a decent base layer quality as preview version. This version normally, except for low quality versions of a bitstream, encrypts an amount of data between sufficient and full encryption. Likewise the amount of time required for encryption is between full selective and sufficient.

Regarding which kind of encryption to use we can distinguish between application scenarios. Since we want to keep scalability intact, full traditional encryption can not be used. When the goal is to produce sufficient encryption, the best option usually is to encrypt I-frames only. I-frames have to be included even when encrypting only lowest spatial bands, in order to prevent the introduction of higher quality content in case of a scene change. Thus, the I-frames only option is in any case faster than the encryption of lowest spatial bands, since this would necessarily include I-frames. When transparent encryption is desired, the options are highest spatial or highest temporal bands. Which option to choose depends strongly on the video sequence, i.e., when encrypting

What was encrypted	Time	% of Bitstream
Sufficient Encryption		
I-frames only	49ms	21.34%
lowest spatial band	80ms	35.54%
lowest temporal band	84ms	39.85%
Transparent Encryption		
highest spatial band	179ms	88.96%
two highest temporal bands	156ms	75.74%
Full Encryption		
full selective encryption	201ms	99.50%
full traditional encryption	207ms	100%

(a) Full quality ($\approx$ 9.5Mbps)

What was encrypted	Time	% of Bitstream
Sufficient Encryption		
I-frames only	16ms	63.42%
lowest spatial band	15ms	65.98%
lowest temporal band	14ms	77.70%
Transparent Encryption		
highest spatial band	13ms	50.02%
two highest temporal bands	12ms	22.00%
Full Encryption		
full selective encryption	18ms	87.60%
full traditional encryption	10ms	100%

(b) Reduced quality (398kbps)

Table 2: The time required for selective encryption and the amount of the bitstream actually encrypted for the foreman sequence with CIF resolution, 256 frames and GOP size of 16.

a scene which contains little motion the drop in framerate from encryption of high temporal bands will hardly be noticeable. Otherwise, encryption of highest temporal bands usually contains less information and consequently is faster. For a more in depth discussion of encryption types and application scenarios see [25].

4.3. Adaptation Performance

As discussed in previous sections there are certain options for scaling and encryption. However, depending on the method chosen, the computational load for adaptation is increased. If SRTP is utilized for encryption, the stream has to be de- and encrypted on the MANE. Likewise, gBSD description allows a more fine grained scalability but introduces an overhead in data sent as well as computational load on the MANE. While effects other than computational requirements have already been discussed, the question of computational load is still open. In this section we will compare gBSD and direct NALU scaling over RTP as well as SRTP to gauge the effects on server, client and MANE.

4.3.1. Evaluation Setup

As test setup we use a loop to measure timing information accurately, i.e., both server and client are on the same machine. The server is connected to the client via a MANE running on a second machine. Both machines have the same hardware, a DELL Optiplex 960 with Intel® Core™2 Quad Q9650 (3GHz, 1333MHz, 2x6MB L2 Cache) CPU with 4GB of DDR2 RAM. The machines are connected via Gigabit LAN using an Intel PRO/1000 GT Network Adapter and run the same software with Ubuntu Linux 10.04 as OS. A schematic drawing of the setup is given in fig. 8.

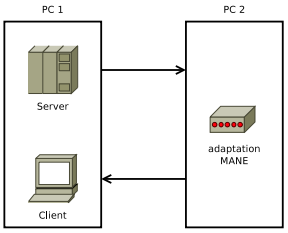


Figure 8: Schematic of test setup

As test sequences the well known crew, football, foreman and harbour sequences are used in CIF resolution with a running length of 10.24sec. The CIF sequences use a GOP size of 16 with a total of 256 frames and an fps of 25. Furthermore, two test sequences are chosen from an application point of view,

the trailers for the Sintel[2] and Big Buck Bunny[3] (abbreviated to bbbunny in tables and figures) movies in 720p resolution with a length of 52sec and 33sec respectively. The two trailers are encoded with a GOP size of 16. For the test each sequence was set to two quality levels. The quality levels and number of possible scaling points for temporal and spatial resolution for all sequences are given in table 3.

Sequence	resolution	T	S	Q1	Q0
bbbunny	720p	4	4	3072	2048
sintel	720p	4	4	2048	1045
football	CIF	4	6	1045	822
harbour	CIF	4	6	822	611
crew	CIF	4	6	611	398
foreman	CIF	4	6	398	256

Table 3: Overview of the video sequences in the testset. The quality levels Q0 and Q1 are given in kbps, the scaling options for temporal (T) and spatial (S) resolution are equal to the number of wavelet decompositions in the respective domain.

Scaling Test	Passed levels					
	CIF Resolution			720p Resolution		
	Temporal	Spatial	Quality	Temporal	Spatial	Quality
None	4	6	2	4	4	2
Temporal	**3**	6	2	**3**	4	2
Spatial	4	**4**	2	4	**3**	2
Quality	4	6	**1**	4	4	**1**

Table 4: Overview of scaling tests, showing which temporal, spatial and quality levels are passed through, bold numbers indicate scaling.

For evaluation, four tests were performed per video sequence, unscaled transport, quality scaling, temporal (framerate) scaling and spatial (resolution) scaling. For each test, 20 streams were simultaneously sent from server to client with adaptation on the MANE. Table 4 gives an overview of which levels are passed during which test. A temporal level of 4 represents the original 16 frames per GOP while a temporal level of 3 indicates a GOP size of 8, and consequently half the original framerate. For each sequence there are two quality levels, the quality levels differ for each sequence and are given in table 3 as Q0 and Q1 respectively. For spatial scaling of CIF sequences, 6 refinement levels reproduce

[2] `http://www.sintel.org`
[3] `http://www.bigbuckbunny.org`

the original CIF resolution while passing only the first 4 levels results in a reduction of resolution to SQCIF 88×72. For 720p sequences, 4 refinement levels reproduce the original sequence at 720p (1280×720) while passing only 3 levels results in a reduction of resolution to 640×360.

4.3.2. In-Network Performance Evaluation

For the performance evaluation we use the testset as described above with both RTP and SRTP. The difference in memory, CPU and frame delay when using NALU and gBSD for adaptation will be investigated. The gBSD is used to describe an underlying NALU bitstream. The NALU bitstream can easily be used to scale spatial and temporal resolution as provided by the MC-EZBC bitstream. Furthermore, during encapsulation of MC-EZBC into a NALU bitstream the number and range of quality scaling points can be freely chosen. However, it is not possible to scale according to higher semantics, e.g., marking certain frames or GOPs as less important. To enable such scaling options, gBSD can be used but this incurs an overhead in the bitstream and, through XML parsing and processing, in computational load. To facilitate a fair comparison, the scaling options for the NALU bitstream as given in table 4 are also used for gBSD testing.

What we expect to see is that the use of gBSD results in a distinct impact on memory and CPU usage on the MANE due to decompression and processing of the XML description. Likewise the use of SRTP is assumed to incur a higher CPU usage on client, server and MANE due to encryption and decryption. Regarding delay in delivery time, both gBSD and SRTP are expected to negatively impact frame delay due to processing cost.

Figure 9 shows the average memory and CPU consumption for the 20 parallel streams on the server, MANE and client for transport via RTP. For each stream 30000 frames were sent. In the figure, NALU refers to scaling based on NALU and SEI refers to scaling with a gBSD description, which is compressed and embedded in the bitstream as SEI messages.

Likewise figure 10 shows the CPU and memory consumption for the same test when using SRTP. This produces an overhead on server, MANE and client due to the encryption and decryption of the bitstream in order to process it. The ordinate for the MANE is different from server and client in order to see the difference for client and server memory and CPU consumption. However, to facilitate comparison between RTP and SRTP the ordinate scales are the same for each case. What is evident from these figures is that encryption for SRTP incurs a significant overhead, especially on the MANE which needs to decrypt as well as encrypt, leading to an almost fourfold increase in CPU consumption. Furthermore, the use of gBSD for scaling results in increased memory consumption on the MANE. This increased memory consumption is more pronounced when the relative size of the gBSD compared the NALU is higher, compare table 1. Additionally the decompression and processing of the SEI gBSD messages increases CPU consumption on the MANE.

In addition to the CPU and memory consumption for scaling, the processing on the MANE incurs a frame delay. Figure 11 plots the cumulative distribution

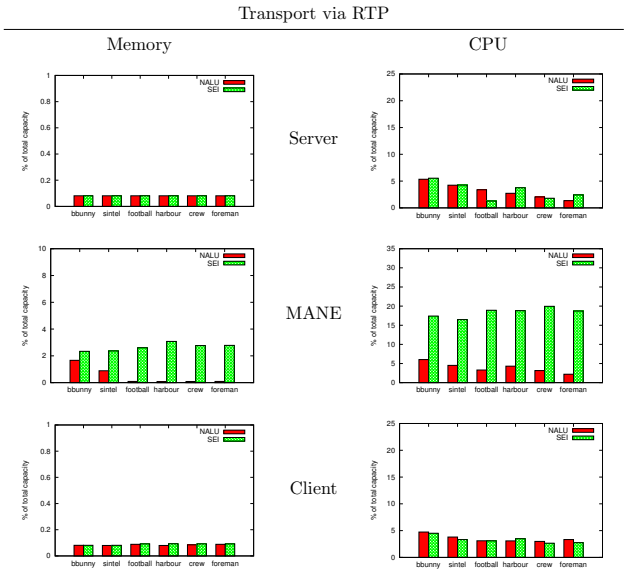


Figure 9: Average of CPU and memory consumption in percent over 20 simultaneous RTP streams and four scaling tests for Server, Client and MANE.

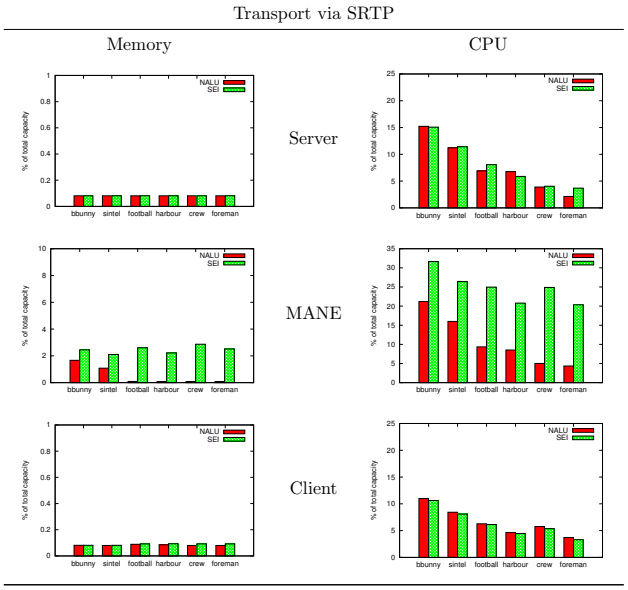


Figure 10: Average of CPU and memory consumption in percent over 20 simultaneous SRTP streams and four scaling tests for Server, Client and MANE.

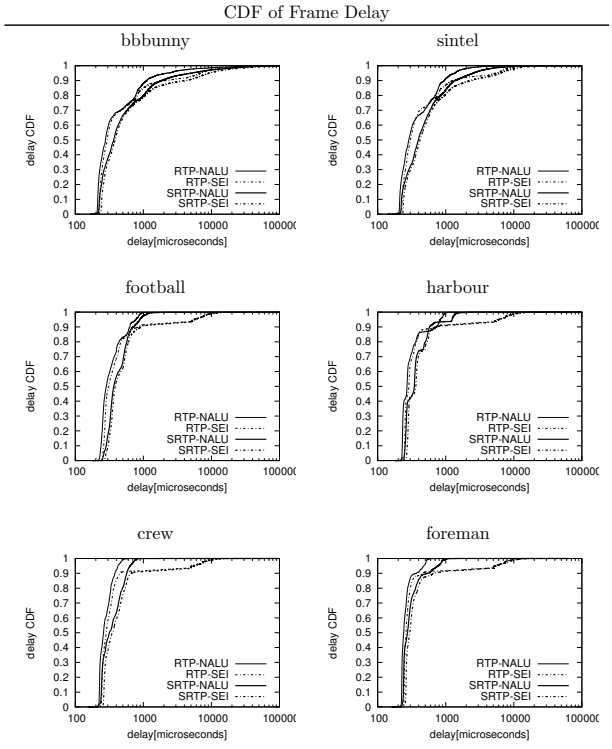


Figure 11: Comparison of cumulative frame delay (CDF) for NALU and SEI adaptation using (S)RTP as transport.

function (CDF) for the frame delay over the actual delay, given in microseconds on a logarithmic scale. This test is done for unscaled contents only since this is the worst case. Any scaling of contents results in less information the MANE has to send and thus smaller outgoing buffers and consequently lower delay. The delay is given based on transport protocol, SRTP and RTP, as well as encapsulation type, NALU and gBSD with SEI. What can be seen is that SRTP causes more delay than RTP since the required decryption and encryption steps have to be performed prior to adaptation checking and sending. Likewise SEI messages incur a higher delay than pure NALU based adaptation. This is due to the fact that the gBSD has to be decompressed and inspected before passing the adapted bitstream along to the client. Furthermore, the higher the bitrate the higher the frame delay, this stems from additional computational demand and fuller outgoing buffers. However, even for higher bitrate sequences the overall relation of SRTP, RTP, NALU and SEI holds.

sequence	RTP		SRTP	
	NALU	SEI	NALU	SEI
bbunny	13490	18521	28478	40524
sintel	2685	9934	6704	13688
football	925	10072	1165	10471
harbour	962	10428	1410	11136
crew	490	9682	756	10102
foreman	529	9501	938	10327

Table 5: Frame delay in μs for CDF= 0.99.

The CDF plot shows that the overall behavior is as expected, both SEI and SRTP incur a delay in delivery time. To better assess the actual impact rather than the general notion, we will take a closer look at the time delay for CDF= 0.99. Table 5 gives the average time, over 30000 frames, to deliver 99% of the image sequence to the end user, i.e., only 1% of the image sequence will take longer to deliver to the client. What can be seen is that the impact of SEI over NALU is tremendous: for RTP SEI is slower than NALU, but the slowdown becomes less sever the higher the overall processing cost. For SRTP the behavior is similar but overall less pronounces since the encryption and decryption overhead slows down both scaling methods. For NALU the switch from RTP to SRTP incurs a significant slowdown while for SEI the impact of SRTP over RTP is less pronounced. This is due to the decryption and encryption being faster by an order of magnitude than the decoding and parsing of SEI. Table 6 gives the factors of slowdown for all cases.

Overall, the expected impact in delivery time due to SRTP and SEI messages over RTP and NALU can clearly be seen.

NALU → SEI		
sequence	RTP	SRTP
bbbunny	1.37	1.42
sintel	3.70	2.04
football	10.89	8.99
harbour	10.84	7.90
crew	19.76	13.36
foreman	17.96	11.01

(a) Slowdown for RTP and SRTP when switching scaling method from NALU to SEI

RTP → SRTP		
sequence	NALU	SEI
bbbunny	2.11	2.19
sintel	2.50	1.38
football	1.26	1.04
harbour	1.47	1.07
crew	1.54	1.04
foreman	1.77	1.09

(b) Slowdown for NALU and SEI when switching from bitstream encryption to SRTP

Table 6: Frame delay slowdown factor for the different scaling and encryption options.

5. Conclusion

We have introduced a mapping of a wavelet based video coding format, the MC-EZBC format, to an H.264/SVC compatible bitstream in order to utilize existing transport and scaling protocols and technologies, i.e., RTP. Furthermore, we compared the bitstream based encryption to transport encryption, i.e SRTP, and evaluated different scaling technologies, i.e., NALU based adaptation versus MPEG-21 Part 7 'Digital Item Adaptation' with gBSD. In addition we have also provided an overhead estimation which is introduced by the mapping of MC-EZBC to a NALU based bitstream as well as the overhead introduced by the inclusion of gBSD in the bitstream.

When it comes to scaling, it is clear that a NALU based approach is better since it generates less overhead in terms of bitstream size. Furthermore, when compared to gBSD, the memory and CPU consumption on network scaling nodes is lower by a significant amount and consequently NALU based adaptation has a lower frame delay. Consequently, even though the NALU based approach is less flexible than gBSD based adaptation, NALU based adaptation should be the baseline and only in those cases where scalability beyond NALU capabilities is desired a gBSD based description should be used.

Regarding encryption, the available options are transport encryption via SRTP and bitstream based encryption on either NALU or MC-EZBC level. It is clear that encryption of the NALU bitstream provides no benefit over encryption of MC-EZBC bitstream prior to the mapping process. When comparing MC-EZBC based encryption to transport encryption, it was shown that the computational load on scaling network elements is much higher for transport encryption and an additional frame delay is introduced. Furthermore, the encryption and decryption of the streamed video content required on every MANE poses a security risk. However, transport encryption has less overall delay to start streaming than bitstream encryption. Any form of DRM, e.g., transparent encryption multicast with sufficient encryption, required a bitstream based en-

cryption since SRTP can not handle those cases. Confidential encryption is not possible since header attacks to leak information about the streamed content are always possible, whether they operate on the plain text information used to scale the bitstream in-network or on the packetization headers of the streaming protocol.

References

[1] A. Vetro, C. Christopoulos, and T. Ebrahimi. From the guest editors - Universal multimedia access. *IEEE Signal Processing Magazine*, 20(2):16 – 16, 2003.

[2] L. Lima, F. Manerba, N. Adami, A. Signoroni, and R. Leonardi. Wavelet-based encoding for HD applications. In *2007 IEEE International Conference on Multimedia and Expo*, pages 1351–1354, July 2007.

[3] H. Eeckhaut, H. Devos, P. Lambert, D. De Schrijver, W. Van Lancker, V. Nollet, P. Avasare, T. Clerckx, F. Verdicchio, M. Christiaens, P. Schelkens, R. Van de Walle, and D. Stroobandt. Scalable, wavelet-based video: From server to hardware-accelerated client. *IEEE Transactions on Multimedia*, 9(7):1508–1519, November 2007.

[4] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. RFC 3550, July 2003.

[5] H. Schulzrinne, A. Rao, and R. Lanphier. Real Time Streaming Protocol (RTSP). RFC 2326, April 1998.

[6] ISO/IEC 21000-7:2007. Information technology – Multimedia framework (MPEG-21) – Part 7: Digital Item Adaptation, November 2007.

[7] Gabriel Panis, Andreas Hutter, Jörg Heuer, Herman Hellwagner, Harald Kosch, Christian Timmerer, Sylvain Devillers, and Myriam Amielh. Bitstream syntax description: a tool for multimedia resource adaptation within MPEG-21. In *Special Issue on Multimedia Adaptation*, volume 18 of *Signal Processing: Image Communication*, pages 721–747, Sept. 2003.

[8] S. Wenger, M.M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer. RTP Payload Format for H.264 Video. RFC 3984, February 2005.

[9] R. Kuschnig, I. Kofler, M. Ransburg, and H. Hellwagner. Design options and comparison of in-network H.264/SVC adaptation. *Journal of Visual Communication and Image Representation*, 19(8):529–542, September 2008.

[10] Dapeng Wu, Yiwei Thomas Hou, Wenwu Zhu, Ya-Qin Zhang, and Jon M. Peha. Streaming video over the Internet: approaches and directions. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(3):282–300, Mar 2001.

[11] Shih-Ta Hsiang and J. W. Woods. Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank. *Signal Processing: Image Communication*, 16(8):705–724, May 2001.

[12] Y. Wu, A. Golwelkar, and J. W. Woods. MC-EZBC video proposal from Rensselaer Polytechnic Institute. *ISO/IEC JTC1/SC29/WG11, MPEG2004/M10569/S15*, March 2004.

[13] P. Chen, K. Hanke, T. Rusert, and J. W. Woods. Improvements to the MC-EZBC scalable video coder. In *Proceedings of the IEEE Int. Conf. Image Processing ICIP*, volume 2, pages 81–84, Barcelona, Spain, 2003.

[14] Peisong Chen and John W. Woods. Bidirectional MC-EZBC with lifting implementation. *IEEE Transactions on Circ. and Systems for Video Technology*, 14(10):1183–1194, 2004.

[15] N. Adami, A. Signoroni, and R. Leonardi. State-of-the-art and trends in scalable video compression with wavelet-based approaches. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(9):1238–1255, September 2007.

[16] Claude E. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656–715, October 1949.

[17] National Institute of Standards and Technology. FIPS-197 - advanced encryption standard (AES), November 2001.

[18] T. D. Lookabaugh and D. C. Sicker. Selective encryption for consumer applications. *IEEE Communications Magazine*, 42(5):124–129, 2004.

[19] A. Said. Measuring the strength of partial encryption schemes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'05)*, volume 2, pages 1126–1129, September 2005.

[20] M. Bellare, T. Ristenpart, P. Rogaway, and T. Stegers. Format-preserving encryption. In *Proceedings of Selected Areas in Cryptography, SAC '09*, volume 5867, pages 295–312, Calgary, Canada, August 2009. Springer-Verlag.

[21] Heinz Hofbauer and Andreas Uhl. The cost of in-network adaption of the MC-EZBC for universal multimedia access. In *Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA '09)*, Salzburg, Austria, September 2009.

[22] ITU-T H.264. Advanced video coding for generic audiovisual services, November 2007.

[23] M. Baugher, D. McGrew, M. Naslund, E. Carrara, and K. Norrman. The Secure Real-time Transport Protocol (SRTP). RFC 3711 (Proposed Standard), March 2004.

[24] Hermann Hellwagner, Robert Kuschnig, Thomas Stütz, and Andreas Uhl. Efficient in-network adaptation of encrypted H.264/SVC content. *Elsevier Journal on Signal Processing: Image Communication*, 24(9):740 – 758, July 2009.

[25] Heinz Hofbauer and Andreas Uhl. Selective encryption of the MC EZBC bitstream for DRM scenarios. In *Proceedings of the 11th ACM Workshop on Multimedia and Security*, pages 161–170, Princeton, New Jersey, USA, September 2009. ACM.

[26] Heinz Hofbauer and Andreas Uhl. Selective encryption of the MC-EZBC bitstream and residual information. In *18th European Signal Processing Conference, 2010 (EUSIPCO-2010)*, pages 2101–2105, Aalborg, Denmark, August 2010.

[27] National Institute of Standards and Technology. FIPS-81 - DES modes of operation, November 1981.

[28] T. Rusert, K. Hanke, P. Chen, and J. W. Woods. Recent improvements to MC-EZBC. *ISO/IEC JTC1/SC29/WG11 MPEG, M9232*, page 14pp., December 2002.

[29] Christopher J. Augeri, Dursun A. Bulutoglu, Barry E. Mullins, Rusty O. Baldwin, and III Leemon C. Baird. An analysis of XML compression efficiency. In *ExpCS '07: Proceedings of the 2007 Workshop on Experimental Computer Science*, page 7, New York, NY, USA, 2007. ACM.

[30] Timmerer, C., Kofler, I., Liegl, J., and Hellwagner, H. An Evaluation of Existing Metadata Compression and Encoding Technologies for MPEG-21 Applications. In *Proceedings of the 1st IEEE International Workshop on Multimedia Information Processing and Retrieval (IEEE-MIPR 2005)*, pages 534–539, Irvine, California, USA, December 2005.

Iris Recognition in Image Domain: Quality-metric based Comparators*

Heinz Hofbauer, Christian Rathgeb, Andreas Uhl, and Peter Wild

Multimedia Signal Processing and Security Lab
Department of Computer Sciences, University of Salzburg, Austria
{hhofbaue,crathgeb,uhl,pwild}@cosy.sbg.ac.at

Abstract. Traditional iris recognition is based on computing efficiently coded representations of discriminative features of the human iris and employing Hamming Distance (HD) as fast and simple metric for biometric comparison in feature space. However, the mapping into feature space is likely to cause loss of information, which is crucial especially in the case of unconstrained acquisition. In this paper we propose the application of quality-metric based comparators operating directly on iris textures, i.e. without transformation into feature space. For this task, the Structural Similarity Index measure (SSIM), Local Edge Gradients metric (LEG), Natural Image Contour Evaluation (NICE), Edge Similarity Score (ESS) and Peak Signal to Noise ratio (PSNR) is evaluated. Obtained results on the CASIA-v3 iris database confirm the applicability of this type of iris comparison technique.

Keywords: Iris reconition, biometric comparators, image quality-metrics, image domain

1 Introduction

Iris recognition is considered one of the most robust and reliable biometric technologies obtaining recognition rates above 99% and equal error rates of less than 1% on several data sets. Compared to other modalities, the iris offers the advantages of being extractible at-a-distance and on-the-move [12]. Taking into account the ever-increasing demand on biometric systems operating in less constrained environments, new iris feature extraction methods have been proposed continously over the past decade [2]. Still, the processing chain of traditional iris recognition (and other biometric) systems has been left almost unchanged, following Daugman's approach [3] consisting of (1) *segmentation and preprocessing* normalizing the iris texture by unrolling into doubly-dimensionless coordinates, (2) *feature extraction* computing a binary representation of discriminative patterns of the rectified iris texture, and (3) *biometric comparison* in feature space involving the fractional HD as dissimilarity measure, see Fig. 1.

* This work has been supported by the Austrian Science Fund, project no. L554-N15 and the Austrian FIT-IT Trust in IT-Systems, project no. 819382.

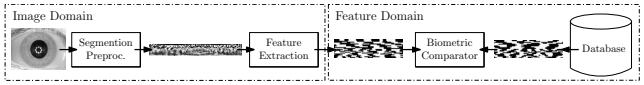


Fig. 1. Common processing chain: images are preprocessed and adaquate feature extractors generate (mostly binary) feature vectors, stored as biometric templates.

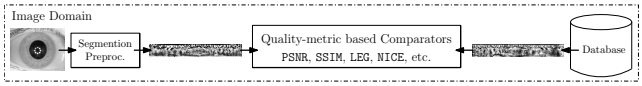


Fig. 2. Proposed processing chain: images are preprocessed and quality-metric based comparators (operating in image domain) estimate similarities between pairs of images.

Recently, several improvements with respect to the employed matching technique as alternatives to the fractional HD have been proposed. In [6], [15] feature extraction is left unchanged, aiming to exploit even more information from the stored biometric template. However, it is clear, that the mapping of original iris texture to short feature vectors, often less than a few hundred bytes in size, induces a strong loss of information. Therefore in this work, we target the application of comparison techniques able to operate in the image domain of normalized iris textures, see Fig. 2. This architecture involves several benefits: (1) The problem of iris recognition can be mapped to a standard image processing problem, benefiting of results in this domain. (2) Features and comparators can be easily replaced without the necessity of re-enrollment, as the entire iris image is stored for comparison and available as reference for future comparators. (3) The approach allows for easier continuous updates, e.g. by averaging iris textures each time of successsful authentication. (4) Quality-based metrics in the image domain may be combined with other image domain methods, such as SIFT-based [1] or Phase-based [9] methods. Of course, the proposed technique may also be combined with traditional feature-scale methods (in which case feature extraction has to be incorporated into the comparison module), since global features used by image metrics complement the mostly localized biometric features. (5) Finally, new techniques like [6], [14] have shown, that an incremental refinement of comparison decisions saves precious computation time and can target the drawback of traditional image quality metrics being considered slow compared to trivial metrics, such as fractional HD. Regarding the security of the stored templates it is suggested to apply standard encryption algorithms (e.g. AES) in order to protect user privacy.

The following sections are organized as follows: related work is reviewed in Section 2. The proposed approach and quality metrics are introduced in Section 3. Experiments are outlined in Section 4 using an open iris database and comparing both original as well as normalized iris images. Finally, Section 5 summarizes the paper.

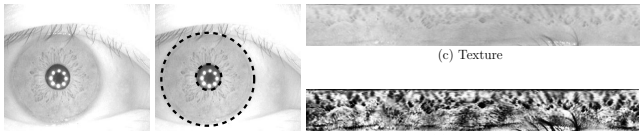
2 Related Work

In the context of iris biometrics, image quality metrics are largely understood as domain-specific indicators, e.g. focus assessment or measurement of pupil/iris diameter ratio, to be considered for quality checks rejecting samples if insufficiently suited for comparison [18]. Such metrics have also been applied for dynamic matcher selection in biometric fusion scenarios [20], i.e. quality is employed to predict matching performance and to select the comparator or adjust weighting of the fusion rule. Our approach is different in employing general purpose image quality metrics and their ability to measure the degree of similarity of image pairs if one of both images is subjected to a (more or less severe) degradation in quality. In our model, the degradation of a sample to be compared is not caused by compression, but by biometric noise factors (time, illumination, etc.), and the stored biometric gallery template represents the (updated) ideal representation of the biometric property of an individual.

Pursuing the idea of employing iris comparison in the image domain, the following works need to be acknowledged: Miyazawa *et al.* [13] identify the problem of feature-based iris recognition being highly dependent on the feature extraction process varying based on environmental factors, which can be avoided by computing features in the image domain. The authors suggest to apply 2D Fourier Phase components of iris images. This scheme is extended by Krichen *et al.* [9], who propose to combine global and local Gabor (i.e. wavelet instead of Fourier coefficients) phase-correlation-based iris matching directly on enhanced (using adaptive histogram equalization) iris textures for unconstrained acquisition procedures. They employ normalized cross-correlation and a Peak to Slob Ratio (PSR) as comparator, which uses mean and standard deviation of the correlation matrix. As Local correlation-based method they correlated sub-images of fixed size using correlation peak in terms of PSR and peak position of each window computing a score out of means and standard deviation. Alonso-Fernandez *et al.* [1] propose the application of Scale Invariant Feature Transformation (SIFT) for recognition, as a means of processing without transformation to polar coordinates, thus permitting less constrained image acquisition conditions. SIFT features can be extracted from original templates in scale space and matched using texture information around the feature points. Kekre *et al.* [7], [4] use the image feature set extracted from Haar Wavelets at various levels of decomposition and from walshlet pyramid for recognition. Simple Euclidean distance on the feature set is applied as the similarity measure. Furthermore, numerous advanced iris biometric comparators have been proposed [15].

3 Iris Recognition in the Image Domain

Given an image of the human eye as shown in Fig. 3 (a), the first task is the transformation into Daugman's rubbersheet model. While any accurate segmentation technique may be applied for this task, we employ the preprocessing chain in [19]. This method applies (1) reflection removal with image inpainting, (2) assessment of edge magnitude and orientation by a Weighted Adaptive Hough



(a) Image (b) Detection (c) Texture (d) Enhanced

Fig. 3. Preprocessing: (a) image of eye (b) detection of pupil and iris (c) unrolled iris texture (d) preprocessed iris texture.

Transform for initial center detection, followed by (3) polar and ellipsopolar transforms to detect boundary candidates, which are evaluated to (4) select the most reliable ones to be used for un-wrapping the image to a rectangular texture of 512×64 pixels. Since image-based methods are largely affected by different illumination [9], we further enhance the iris texture applying CLAHE (Contrast Limited Adaptive Histogram Equalization) [22], see Fig. 3 (b)-(d).

In image-domain iris processing, we store one full reference iris texture O per user. While template-updates can easily be handled in such a scenario, for evaluations we employ enrollment using the first eye image per user only. In order to score an authentication attempt given a claimed identity, the corresponding template image O is compared with the current sample image I. Both images of $W \times H$ pixels are compared by employing one of the following quality metrics $Q(s(I, m), O)$, where $s(I, m)$ denote a shifting of m pixels to the left or right in order to obtain a rotation invariant technique. For I and O the b bits per pixel are used with a maximum pixel value of $M = 2^b$.

All of the following image metrics[1] are full reference metrics, meaning they utilize information from the original and comparison image to calculate an assessment of the visual similarity. The following subsections describe details of applied image metrics and show, which features are used in the calculation of the quality assessment.

3.1 Peak Signal to Noise Ratio (PSNR)

The PSNR is still widely used because it is unrivaled in speed and ease of use.

The following steps are performed to calculate the PSNR.

Step 1: Calculate the mean squared error $\mathrm{MSE} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=i}^{H} (I(i,j) - O(i,j))^2$

Step 2: The PSNR is calculated:

$$\mathrm{PSNR} = 10 \log_{10} \Big(\frac{M^2}{\sqrt{\mathrm{MSE}}} \Big). \tag{1}$$

[1] The implementation is available online at `http://www.wavelab.at/sources/VQI/`

3.2 Structural Similarity Index Measure (SSIM)

The structural similarity index measure (SSIM) by Wang et al. [21] uses the local luminance as well as global contrast and a structural feature to calculate a score as follows.

Step 1: Each image is transformed by convolution with a 11×11 Gaussian filter.

Step 2: The luminance, contrast and structural scores can be calculated and combined in one step as follows.

$$\mathrm{SSIM}(I,O) = \frac{(2\mu_I\mu_O + c_1)(2\sigma_{IO} + c_2)}{(\mu_I^2 + \mu_O^2 + c_1)(\sigma_I^2 + \sigma_O^2 + c_2)}, \tag{2}$$

where μ_I is the average pixel value of image I, σ_I^2 is the variance of pixel values of image I and σ_{IO} is the covariance of I and O. The variables $c_1 = (k_1 M)^2$ and $c_2 = (k_2 M)^2$, with $k_1 = 0.01$ and $k_2 = 0.03$, are used to stabilize the division.

3.3 Local Edge Gradients Metric (LEG)

The image metric based on local edge gradients was introduced by Hofbauer and Uhl [5] and uses luminance and localized edge information from different frequency domains.

Step 1: First the global luminance difference between I and O id calculated as $\mathrm{LUM}(I,O) = 1 - \sqrt{\frac{|\mu(O)-\mu(I)|}{M}}$, where $\mu(X) = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H} X(x,y)$, and $X(x,y)$ is the pixel value of image X at position x, y.

Step 2: One step wavelet decomposition with Haar wavelets resulting in four sub images for each image X denoted as X_0 for the LL-subband, and X_1, X_2, X_3 for LH, HH and HL subbands, respectively.

Step 3:A local edge map is calculated for each position x, y in the image, reflecting the change in coarse structure of the image.

$$\mathrm{LE}(I,O,x,y) = \begin{cases} 1 & \text{if } \mathrm{EDC}(I,O,x,y) = 8, \\ 0.5 & \text{if } \mathrm{EDC}(I,O,x,y) = 7, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathrm{EDC}(I,O,x,y) = \sum_{p \in N(x,y)} \mathrm{ED}(I,O,x,y,p)$$

$$\mathrm{ED}(I,O,x,y,p) = \begin{cases} 1 & \text{if } I(x,y) < I(p) \text{ and } O(x,y) < O(p), \\ 1 & \text{if } I(x,y) > I(p) \text{ and } O(x,y) > O(p), \\ 0 & \text{otherwise.} \end{cases}$$

where $N(x,y)$ is the eight neighborhood of the pixel x, y.

Step 4: In order to assess the contrast changes a difference of gradients in a neighborhood is calculated

$$\mathrm{LED}(I,O,x,y) = \frac{1}{8}\sum_{p \in N(x,y)} \left(1 - \sqrt{\frac{|LD(I,O,x,y,p)|}{M}}\right)^2,$$

where $\mathrm{LD}(I, O, x, y, p) = (O(x, y) - O(p)) - (I(x, y) - I(p))$.

Step 5: The edge score is calculated by combining local edge conformity (LE) and local edge difference (LED) into

$$\mathrm{ES}(I, O) = \frac{4}{WH} \sum_{x=1}^{\frac{W}{2}} \sum_{y=1}^{\frac{H}{2}} \Big(\mathrm{LE}(I_0, O_0, x, y) \frac{1}{3} \sum_{i=1}^{3} \mathrm{LED}(I_i, O_i, x, y) \Big).$$

Step 6: The LEG visual quality index is calculated by combining ES and LUM.

$$\mathrm{LEG}(I, O) = \mathrm{LUM}(I, O)\, \mathrm{ES}(I, O). \tag{3}$$

3.4 Natural Image Contour Evaluation (NICE)

The NICE quality index by Rouse and Hemami [17, 16] uses gradient maps, adjusted for possible image shift by using a morphological dilation with a plus shaped structuring element. The actual score is computed by doing a thresholding on the image and calculating differences. The following steps are used to calculate the NICE score.

Step 1: Gradient amplitude image $\hat{I}$ is generated from I such that for $i \in [1, \ldots, W]$ and $j \in [1, \ldots, H]$ $\hat{I}$ is defined as $\hat{I}(i, j) = \sqrt{S_x(I, i, j)^2 + S_x(I, i, j)^2}$, where $\hat{I}(i, j)$ is the pixel value at location i, j and $S_x(I, i, j)$ and $S_y(I, i, j)$ are the results of a Sobol filter at position i, j in image I in direction x and y, respectively. Likewise $\hat{O}$ is generated from O.

Step 2: A binary image $B_{\hat{I}}$ is generated by thresholding with the average gradient amplitude value. That is, $B_{\hat{I}}(i, j) = 1$ if $\hat{I}(i, j) > T_{\hat{I}}$ and $B_{\hat{I}}(i, j) = 0$ otherwise, where $T_{\hat{I}} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \hat{I}(i, j)$.

Step 3: The binary image $B_{\hat{I}}$ is transformed into $B^+_{\hat{I}}$ by applying a morphological dilation with a plus shaped structuring element. That is, each pixel $B^+_{\hat{I}}(i, j)$ is set to 1 if at least one of the 4−connected neighbours of $B_{\hat{I}}(i, j)$ or $B_{\hat{I}}(i, j)$ is 1, otherwise $B^+_{\hat{I}}(i, j) = 0$.

Step 4: The NICE score is calculated based on the normalized Hamming distance as

$$\mathrm{NICE}(O, I) = \frac{\sum_{i=1}^{W} \sum_{j=1}^{H} (B^+_{\hat{O}}(i, j) - B^+_{\hat{I}}(i, j))^2}{\sum_{i=1}^{W} \sum_{j=1}^{H} B^+_{\hat{O}}(i, j)} \tag{4}$$

3.5 Edge Similarity Score (ESS)

The ESS was introduced by Mao and Wu [11] and uses localized edge information to compare two images.

Step 1: Each image is separate into N blocks of size 8×8.

Step 2: For each image I a Sobel edge detection filter is used on each block i to find the most prominent edge direction $e^i_I$and quantized into one of eight

directions (each corresponding to 22.5°). Edge direction 0 is used if no edge was found in the block.

Step 3: Calculate the ESS based on the prominent edges of each block:

$$\mathrm{ESS} = \frac{\sum_{i=1}^{N} w(e_I^i, e_O^i)}{\sum_{i=1}^{N} c(e_I^i, e_O^i)}, \tag{5}$$

where $w(e_1, e_2)$ is a weighting function defined as

$$w(e_1, e_2) = \begin{cases} 0 & \text{if } e_1 = 0 \text{ or } e_2 = 0 \\ |cos(\phi(e_1) - \phi(e_2))| & \text{otherwise,} \end{cases}$$

where $\phi(e)$ is the representative edge angle for an index e, and $c(e_1, e_2)$ is an indicator function defined as $c(e_1, e_2) = 0$ if $e_1 = e_2 = 0$ and $c(e_1, e_2) = 1$ otherwise. In cases where $\sum_{i=1}^{N} c(e_I^i, e_O^i) = 0$ the ESS is set to 0.5.

4 Experiments

Experiments are carried out on the CASIA-v3-Interval iris database[2] using left-eye images only. The database consists of good quality 320×280 pixel NIR illuminated indoor images where the applied test set consists of 1307 instances, a sample is shown in Fig. 3 (a).

Recognition accuracy is evaluated in terms of false none match rate (FNMR) at a certain false match rate (FMR). The FNMR defines the proportion of verification transactions with truthful claims of identity that are incorrectly rejected, and the FMR defines the proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed (ISO/IEC FDIS 19795-1), in particular, ZeroFMR defines the FNMR at a FMR of 0.1%. As score distributions overlap the Equal Error Rate (EER) of the system is defined (FNMR = FMR). At all authentication attempts 7 circular texture-shifts are performed in each direction for all comparators. A summary of obtained EERs and ZeroFMR rates at the corresponding decision thresholds for the underlying image quality metrics is given in Table 1. Receiver operating characteristics, which illustrate the tradeoff between FMR and FNMR, are plotted in Fig. 4 for experiments evaluating (a) metrics, as well as (b) impact of the used image type: original image, texture after segmentation, and enhanced texture after CLAHE normalization. Score distributions for each metric (normalized to $[0, 1]$) with respect to genuine (intra-) and impostor (inter-personal) comparisons are illustrated in Fig. 5.

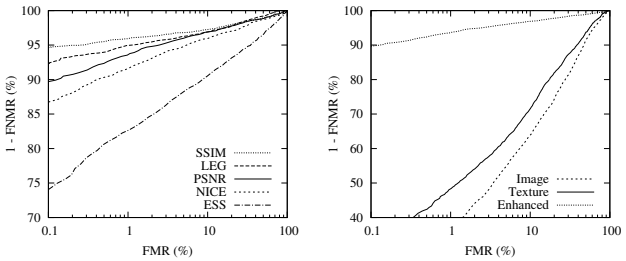
4.1 Which quality metrics are useful iris biometric comparators?

With respect to accuracy, the ranking of metrics is as follows: SSIM, LEG, PSNR, NICE, and ESS, with the first three metrics exhibiting EERs of less

[2] The Center of Biometrics and Security Research, CASIA Iris Image Database, http://www.idealtest.org

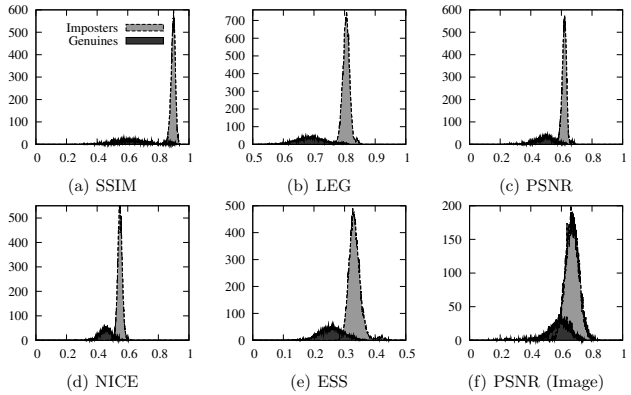
Table 1. Recognition performance of Quality Metrics

Algorithm	Type	EER	ZeroFMR	Threshold
SSIM	Enhanced	3.40%	5.34%	0.868
LEG	Enhanced	3.99%	7.72%	0.785
NICE	Enhanced	5.14%	13.32%	0.526
ESS	Enhanced	9.61%	25.97%	0.311
PSNR	Enhanced	4.21%	10.33%	0.592
PSNR	Texture	18.88%	65.37%	0.478
PSNR	Image	23.01%	80.67%	0.638
Ma *et al.*	Iris-Code	1.83%	2.02%	–
Ko *et al.*	Iris-Code	4.36%	18.45%	–



(a) Quality metrics (b) Image type (PSNR)

Fig. 4. Receiver operating characteristics by (a) quality metric, and (b) image type.



(a) SSIM (b) LEG (c) PSNR

(d) NICE (e) ESS (f) PSNR (Image)

Fig. 5. Genuine and impostor score distributions for (a) SSIM, (b) LEG, (c) PSNR, (d) NICE, (e) ESS for enhanced textures and (f) PSNR on original images.

than 5%. It is interesting to see, that PSNR with 4.21% EER performs quite well on the enhanced textures although it is the most simple metric. However, for high security applications with requested low FMR, SSIM with 5.34% ZeroFMR compared to 10.33% for PSNR is clearly the better alternative. Considering recognition accuracy image metrics do not outperform feature-based techniques [2]. For instance, on the same dataset the approaches of Ma *et al.* [10] and Ko *et al.* [8], which extract binary iris-codes obtain EERs of 1.83% and 4.36%, respectively (see Table 1). However, image metrics are rather useful as additional features in fusion scenarios.

4.2 How useful is texture enhancement and preprocessing?

In a second experiment, we tested the effect of texture enhancement and segmentation on iris recognition accuracy of quality metrics using PSNR as reference metric. Obtained results indicate a high degradation in case texture enhancement steps are skipped (18.88% EER instead of 4.21%). Recognition from the original eye images (without segmentation) further degraded results (23.01% EER), thus normalizaton and enhancement steps accounting for different illumination enriching the texture in the image (see Fig. 3) are extremely useful.

5 Summary

This paper applies quality metrics in image domain to the problem of iris recogniton. As opposed to the view that original iris textures exhibit too much noisy information to be used directly for comparison, we found that some metrics (SSIM, LEG, PSNR) provide quite reasonable accuracy (3.4%, 3.99% and 4.21% EER, respectively). The proposed architecture alleviates continuous template updates and enables a transparent replacement of comparators without re-enrollment. Iris texture enhancement is found to be essential to the accuracy of iris recognition in the image domain. Future work is targeted at a sophisticated analysis of fusion approaches of image-domain methods and a combination with serial comparison techniques to accelerate processing time.

References

1. F. Alonso-Fernandez, P. Tome-Gonzalez, V. Ruiz-Albacete, and J. Ortega-Garcia. Iris recognition based on sift features. In *Int'l Conf. on Biometrics, Ident. and Sec. (BIdS)*, pages 1–8, 2009.
2. K. W. Bowyer, K. Hollingsworth, and P. J. Flynn. Image understanding for iris biometrics: A survey. *Comp. Vis. Image Underst.*, 110(2):281 – 307, 2008.
3. J. Daugman. How iris recognition works. *IEEE Trans. Circ. and Syst. for Video Techn.*, 14(1):21–30, 2004.
4. H.B.Kekre, Sudeep D. Thepade, Juhi Jain, and Naman Agrawal. Iris recognition using texture features extracted from haarlet pyramid. *Int'l J. of Comp. App.*, 11(12):1–5, 2010. Found. Comp. Sc.

5. H. Hofbauer and A. Uhl. An effective and efficient visual quality index based on local edge gradients. In *IEEE 3rd Europ. Workshop on Visual Inf. Proc.*, page 6pp., 2011.
6. K. P. Hollingsworth, K. W. Bowyer, and P. J. Flynn. The best bits in an iris code. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 31(6):964–973, 2009.
7. H. B. Kekre, S. D. Thepade, J. Jain, and N. Agrawal. Iris recognition using texture features extracted from walshlet pyramid. In *Prof. Int'l Conf. & Workshop on Emerging Trends in Techn. (ICWET)*, pages 76–81. ACM, 2011.
8. J.-G. Ko, Y.-H. Gil, J.-H. Yoo, and K.-I. Chung. A novel and efficient feature extraction method for iris recognition. *ETRI Journal*, 29(3):399 – 401, 2007.
9. E. Krichen, S. Garcia-Salicetti, and B. Dorizzi. A new phase-correlation-based iris matching for degraded images. *IEEE Trans. on Systems, Man, and Cyb., Part B*, 39(4):924–934, 2009.
10. L. Ma, T. Tan, Y. Wang, and D. Zhang. Efficient iris recognition by characterizing key local variations. *IEEE Trans. on Image Processing*, 13(6):739–750, 2004.
11. Y. Mao and M. Wu. Security evaluation for communication-friendly encryption of multimedia. In *IEEE Int'l Conf. on Image Proc. (ICIP)*, 2004.
12. J. Matey, O. Naroditsky, K. Hanna, R. Kolczynski, D. LoIacono, S. Mangru, M. Tinker, T. Zappia, and W. Y. Zhao. Iris on the move: Acquisition of images for iris recognition in less constrained environments. *Proc. IEEE*, 94:1936–1947, 2006.
13. K. Miyazawa, K. Ito, T. Aoki, K. Kobayashi, and H. Nakajima. An efficient iris recognition algorithm using phase-based image matching. In *IEEE Int'l Conf. on Image Proc. (ICIP)*, pages 49–52, 2005.
14. C. Rathgeb, A. Uhl, and P. Wild. Incremental iris recognition: A single-algorithm serial fusion strategy to optimize time complexity. In *Proc. Int'l Conf. on Biometrics: Theory, App., and Syst. (BTAS)*, pages 1–6, 2010.
15. C. Rathgeb, A. Uhl, and P. Wild. Iris-biometric comparators: Minimizing trade-offs costs between computational performance and recognition accuracy. In *Proc. Int'l Conf. on Imaging for Crime Det. and Prev. (ICDP)*, pages 1–7, 2011.
16. D. Rouse and S. S. Hemami. Natural image utility assessment using image contours. In *IEEE Int'l Conf. on Image Proc. (ICIP)*, pages 2217–2220, 2009.
17. D. Rouse and S. S. Hemami. The role of edge information to estimate the perceived utility of natural images. In *Western New York Image Proc. Workshop (WNYIP)*, page 4 pp., 2009.
18. I. Tomeo-Reyes, J. Liu-Jimenez, I. Rubio-Polo, and B. Fernandez-Saavedra. Quality metrics influence on iris recognition systems performance. In *IEEE Int'l Carnahan Conf. on Security Technology (ICCST)*, pages 1–7, 2011.
19. A. Uhl and P. Wild. Weighted adaptive hough and ellipsopolar transforms for real-time iris segmentation. In *Proc. Int'l Conf. on Biometrics (ICB)*, 2012. to appear.
20. M. Vatsa, R. Singh, A. Noore, and A. Ross. On the dynamic selection of biometric fusion algorithms. *IEEE Trans. on Inf. Forensics and Sec.*, 10(3):470 – 479, 2010.
21. Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Proc.*, 13(4):600–612, 2004.
22. K. Zuiderveld. Contrast limited adaptive histogram equalization. In Paul S. Heckbert, editor, *Graphics Gems IV*, pages 474–485. Morgan Kaufmann, 1994.

IMAGE METRIC-BASED BIOMETRIC COMPARATORS: A SUPPLEMENT TO FEATURE VECTOR-BASED HAMMING DISTANCE?

H. Hofbauer, C. Rathgeb, A. Uhl, and P. Wild

Multimedia Signal Processing and Security Lab
Department of Computer Sciences, University of Salzburg, Austria
{hhofbaue,crathgeb,uhl,pwild}@cosy.sbg.ac.at

ABSTRACT

In accordance with the ISO/IEC FDIS 19794-6 standard an iris-biometric fusion of image metric-based and Hamming distance (HD) comparison scores is presented. In order to demonstrate the applicability of a knowledge transfer from image quality assessment to iris recognition, Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Local Edge Gradients metric (LEG), Edge Similarity Score (ESS), Local Feature Based Visual Security (LFBVS), and Visual Information Fidelity (VIF) are applied to iris textures, i.e. query textures are interpreted as noisy representations of registered ones. Obtained scores are fused with traditional HD scores obtained from iris-codes generated by different feature extraction algorithms. Experimental evaluations on the CASIA-v3 iris database confirm the soundness of the proposed approach.

Index Terms— Biometrics, image quality metrics, iris recognition, biometric fusion

1. INTRODUCTION

Iris recognition takes advantage of random variations in the iris. The details of each iris are phenotypically unique yielding recognition rates above 99% and equal error rates of less than 1% on diverse data sets. In past years the ever-increasing demand on biometric systems operating in less constrained environments entails continuous proposals of new iris feature extraction methods [1]. Still, the processing chain of traditional iris recognition (and other biometric) systems has been left almost unaltered, following Daugman's approach [2] consisting of (1) segmentation and preprocessing, (2) feature extraction, and (3) biometric comparison.

The International Organization for Standardization (ISO) specifies iris biometric data to be recorded and stored in (raw) image form (ISO/IEC FDIS 19794-6), rather than in extracted templates (e.g. iris-codes) achieving more interoperability as well as vendor neutrality [3]. Biometric databases, which store raw biometric data, enable the incorporation of future improvements (e.g. in segmentation stage) without re-enrollment of registered users. While the extraction of rather short (a few hundred bytes) binary feature vectors provides a compact storage and rapid comparison of biometric templates, information loss is inevitable. This motivates a fusion of comparators operating in image domain (e.g. image metrics) and traditional HD-based comparators requiring binary feature vectors. The contribution of this work is the proposal of a fusion scenario combining image metrics and traditional HD-based approaches. In contrast to common believe that original iris textures exhibit too much variation to be used directly for recognition we proof that (1) quality metric, interpreting iris textures as a noisy reproduction of the reference sample, can be employed for recognition, and (2) global features extracted by image metrics tend to complement localized features encoded by traditional feature extraction methods.

This paper is organized as follows: related work is reviewed in Section 2. Subsequently, the proposed fusion scenario is described in detail in Section 3. Experimental results are presented in Section 4. Section 5 concludes the paper.

2. RELATED WORK

In the context of iris biometrics, image quality metrics are largely understood as domain-specific indicators, e.g. focus assessment or measurement of pupil/iris diameter ratio, to be considered for quality checks rejecting samples if insufficiently suited for comparison [4]. Such metrics have also been applied for dynamic matcher selection in biometric fusion scenarios [5], i.e. quality is employed to predict matching performance and to select the comparator or adjust weighting of the fusion rule. Our approach is different in employing general purpose image quality metrics and their ability to measure the degree of similarity of image pairs if one of both images is subjected to a (more or less severe) degradation in quality. In our model, the degradation of a sample to be compared is not caused by compression, but by biometric noise factors (time, illumination, etc.).

This work has been supported by the Austrian Science Fund, project no. L554-N15 and the Austrian FIT-IT Trust in IT-Systems, project no. 819382.

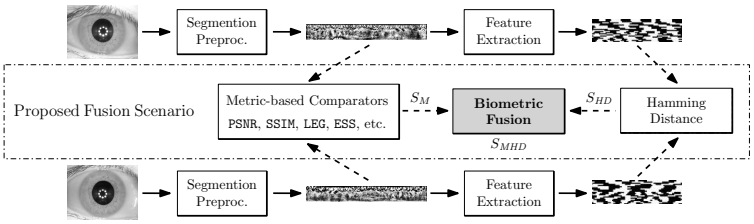


Fig. 1. Proposed Fusion Scenario: image quality metric-based scores are combined with Hamming distance-based feature-level scores in order to obtain a final comparison score.

Information fusion in biometrics is an efficient means to enhance the accuracy of a biometric system by employing multiple modalities, sensors, or comparators [6]. Compared to other types of fusion, score level fusion enables transparent enhancement of biometric systems by combining the matching scores of multiple comparators yielding a score vector $S = (s_1, \ldots, s_m)$, which is combined using a fusion rule, such as e.g. sum rule $s = \sum_{i=1}^{m} s_i$ or product rule $s = \prod_{i=1}^{m} s_i$ [7]. Park *et al.*[8] investigate this fusion type for local and global Gabor feature-vector based algorithms and found their proposed SVM-based fusion of HD scores to outperform each single Gabor filter when restricting the features to reliable regions. In previous work [9], we have investigated score level fusion for combining best with worst HD-based alignment of iris codes for enhanced iris matching. If comparators are weakly dependent and still contain rich discriminative information, the combined score can be expected to provide better discrimination between genuine and imposter comparisons. An essential step before employing such fusion rules is a normalization of scores [6].

3. FUSION OF IRIS RECOGNITION ALGORITHMS AND IMAGE QUALITY METRICS

The proposed fusion scenario is shown in Fig. 1. At authentication, segmentation and pre-processing is performed on a given pair of iris images. Subsequently, resulting iris textures are compared applying a distinct image metric. The preliminary comparison score, denoted by S_M, is normalized and fused with the according HD-based score, denoted by S_{HD}, after feature extraction has been applied to both iris textures, in order to obtain the final score S_{MHD}. The biometric fusion is performed applying sum-rule fusion [6], i.e. S_{MHD} is defined as,

$$S_{MHD} = \frac{1}{2}(S_M + S_{HD}). \quad (1)$$

In the following subsections modules of the proposed system, which comprise (1) segmentation and pre-processing, (2) iris-biometric feature extractors, and (3) image metrics, are described in detail. All of the applied image metrics[1] are full reference metrics, meaning they utilize information from the original and comparison image to calculate an assessment of the visual similarity.

3.1. Pre-processing and Feature Extraction Algorithms

We apply multi-stage iris segmentation using a weighted version of adaptive Hough transform for iterative iris center detection at the first stage and pupillary and limbic boundary detection by applying an ellipsopolar transform and assessing gradient information for finding the second boundary based on the outcome of the first [10]. After having obtained a parametrization of inner and outer iris boundaries, the iris texture is unwrapped using Daugman's doubly dimensionless representation [2] and enhanced using contrast-limited adaptive histogram equalization [11].

In the feature extraction stage we employ custom implementations of two different algorithms used to extract binary iris-codes. The first one was proposed by Ma *et al.*[12]. Within this approach the texture is divided into 10 stripes to obtain 5 one-dimensional signals, each one averaged from the pixels of 5 adjacent rows, hence, the upper 512×50 pixel of preprocessed iris textures are analyzed. A dyadic wavelet transform is then performed on each of the resulting 10 signals, and two fixed subbands are selected from each transform. In each subband all local minima and maxima above a adequate threshold are located, and a bit-code alternating between 0 and 1 at each extreme point is extracted. Using 512 bits per signal, the final code is then $512 \times 20 = 10240$ bit. The second feature extraction method follows an implementation by Masek[2] applying filters obtained from a Log-Gabor function. Here, a row-wise convolution with a complex Log-Gabor filter is performed on the texture pixels. We use the same texture size and row-averaging into 10 signals prior to applying the one-dimensional Log-Gabor filter. The 2 bits of phase information are used to generate a binary code, which therefore is again $512 \times 20 = 10240$ bit.

[1] Implementation available at `www.wavelab.at/sources/VQI/`, except for VIF for which we used MetriX MuX from `foulard.ece.cornell.edu/gaubatz/metrix_mux`.

[2] L. Masek: Recognition of Human Iris Patterns for Biometric Identification, Master's thesis, University of Western Australia, 2003

3.2. Peak Signal to Noise Ratio (PSNR)

The PSNR is still widely used because it is unrivaled in speed and ease of use.

The following steps are performed to calculate the PSNR.

Step 1: Calculate the mean squared error $\mathrm{MSE} = \frac{1}{WH} * \sum_{i=1}^{W}\sum_{j=i}^{H}(I(i,j) - O(i,j))^2$

Step 2: The PSNR is calculated:

$$\mathrm{PSNR} = 10\log_{10}\Big(\frac{M^2}{\sqrt{\mathrm{MSE}}}\Big). \tag{2}$$

3.3. Structural Similarity Index Measure (SSIM)

The SSIM by Wang *et al.*[13] uses the local luminance as well as global contrast and a structural feature.

Step 1: Each image is transformed by convolution with a 11×11 Gaussian filter.

Step 2: The luminance, contrast and structural scores can be calculated and combined in one step as follows.

$$\mathrm{SSIM}(I,O) = \frac{(2\mu_I\mu_O + c_1)(2\sigma_{IO} + c_2)}{(\mu_I^2 + \mu_O^2 + c_1)(\sigma_I^2 + \sigma_O^2 + c_2)}, \tag{3}$$

where μ_I is the average pixel value of image I, σ_I^2 is the variance of pixel values of image I and σ_{IO} is the covariance of I and O. The variables $c_1 = (k_1 M)^2$ and $c_2 = (k_2 M)^2$, with $k_1 = 0.01$ and $k_2 = 0.03$, are used to stabilize the division.

3.4. Local Edge Gradients Metric (LEG)

The image metric based on local edge gradients was introduced by Hofbauer and Uhl [14] and uses luminance and localized edge information from different frequency domains.

Step 1: First the global luminance difference between I and O id calculated as $\mathrm{LUM}(I,O) = 1 - \sqrt{\frac{|\mu(O)-\mu(I)|}{M}}$, where $\mu(X) = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}X(x,y)$, and $X(x,y)$ is the pixel value of image X at position x, y.

Step 2: One step wavelet decomposition with Haar wavelets resulting in four sub images for each image X denoted as X_0 for the LL-subband, and X_1, X_2, X_3 for LH, HH and HL subbands, respectively.

Step 3:A local edge map is calculated for each position x, y in the image, reflecting the change in coarse structure. $\mathrm{LE}(I,O,x,y) = \max(0, \mathrm{EDC}(I,O,x,y) - 6)/2$, i.e. $\mathrm{LE} = 1$ if $\mathrm{EDC} = 8$, $\mathrm{LE} = 0.5$ if $\mathrm{EDC} = 7$ and 0 otherwise. Here $\mathrm{EDC}(I,O,x,y) = \sum_{p\in N(x,y)}\mathrm{ED}(I,O,x,y,p)$, where $N(x,y)$ is the eight neighborhood of the pixel x, y, with $\mathrm{ED}(I,O,x,y,p) = 1$ if edge directions for I and O match, i.e. if $I(x,y) < I(p)$ and $O(x,y) < O(p)$ or $I(x,y) > I(p)$ and $O(x,y) > O(p)$, otherwise $\mathrm{ED}(I,O,x,y,p) = 0$.

Step 4: In order to assess the contrast changes a difference of gradients in a neighborhood is calculated by $\mathrm{LED}(I,O,x,y) = \frac{1}{8}\sum_{p\in N(x,y)}\left(1 - \sqrt{\frac{|LD(I,O,x,y,p)|}{M}}\right)^2$, with $\mathrm{LD}(I,O,x,y,p) = (O(x,y) - O(p)) - (I(x,y) - I(p))$.

Step 5: The edge score is calculated by combining local edge conformity (LE) and local edge difference (LED) into

$$\mathrm{ES}(I,O) = \frac{4}{WH}\sum_{x=1}^{\frac{W}{2}}\sum_{y=1}^{\frac{H}{2}}\Big(\mathrm{LE}(I_0,O_0,x,y) * \frac{1}{3}\sum_{i=1}^{3}\mathrm{LED}(I_i,O_i,x,y)\Big).$$

Step 6: The LEG visual quality index is calculated by combining ES and LUM.

$$\mathrm{LEG}(I,O) = \mathrm{LUM}(I,O)\,\mathrm{ES}(I,O). \tag{4}$$

3.5. Edge Similarity Score (ESS)

The ESS was introduced by Mao and Wu [15] and uses localized edge information to compare two images.

Step 1: Each image is separate into N blocks of size 8×8.

Step 2: For each image I a Sobel edge detection filter is used on each block i to find the most prominent edge direction $e_I^i$and quantized into one of eight directions (each corresponding to $22.5°$). Edge direction 0 is used if no edge was found in the block.

Step 3: Calculate the ESS based on the prominent edges of each block:

$$\mathrm{ESS} = \frac{\sum_{i=1}^{N} w(e_I^i, e_O^i)}{\sum_{i=1}^{N} c(e_I^i, e_O^i)}, \tag{5}$$

where $w(e_1, e_2)$ is a weighting function defined as

$$w(e_1,e_2) = \begin{cases} 0 & \text{if } e_1 = 0 \text{ or } e_2 = 0 \\ |cos(\phi(e_1) - \phi(e_2))| & \text{otherwise,} \end{cases}$$

where $\phi(e)$ is the representative edge angle for an index e, and $c(e_1, e_2)$ is an indicator function defined as $c(e_1, e_2) = 0$ if $e_1 = e_2 = 0$ and $c(e_1, e_2) = 1$ otherwise. In cases where $\sum_{i=1}^{N} c(e_I^i, e_O^i) = 0$ the ESS is set to 0.5.

3.6. Local Feature Based Visual Security (LFBVS)

The LFBVS was introduced by Tong et al. [16] utilizes localized edge and luminance features which are combined and weighted according to error magnitude, i.e. error pooling.

Step 1: Separate an image I into N blocks B_i^I of size 16×16.

Step 2: Calculate the average $\mu(B_i^I)$ and standard deviation $\sigma(B_i^I)$ of the pixel luminance values in the given block. Calculate the local luminance feature $\mathrm{LUM}(I,O,i) = (|\mu(B_i^O) - \mu(B_i^I)| + |\sigma(B_i^O) - \sigma(B_i^I)|)\,/\,2L_{\max}$.

Step 3: For each pixel in the macroblock (excluding borders) calculate the (luminance) edge directions $\delta_x(x,y) = L(x+1,y) - L(x-1,y)$, $\delta_y(x,y) = L(x,y+1) - L(x,y-1)$. Generate a histogram $H_i^I[d] = A$ of cumulative edge amplitude strength $a = \sqrt{\delta_x(x,y)^2 + \delta_y(x,y)^2}$ over edge directions d(8-bins for 360) for each block. And using the histogram calculate the local edge density feature $\mathrm{ED}(I,O,i) = \sum_{d=1}^{8}|H_i^O[d] - H_i^I[d]|\,/\,\sum_{d=1}^{8}\max(H_i^O[d], H_i^I[d])$

Table 1. Obtained results for the proposed fusion scenario.

	EER (%)							
	Ma *et al.*	Masek	PSNR	SSIM	LEG	ESS	LFBVS	VIF
Ma *et al.*	1.43	1.46	1.56	1.53	**1.32**	2.51	2.01	1.65
Masek		1.77	1.97	**1.72**	**1.58**	2.43	2.12	1.78
PSNR			4.21	**3.08**	**3.34**	4.69	**3.60**	2.11
SSIM				3.40	3.40	4.51	**2.71**	2.18
LEG					3.99	5.76	**3.46**	2.10
ESS						9.61	**4.90**	2.20
LFBVS							5.54	**1.86**
VIF								2.06

Step 4: Calculate a local visual score incorporating local luminance and edge density $LVS(I,O,i) = 0.2LUM(I,O,i) + 0.8ED(I,O,i)$. Order the local visual features $\mathrm{OLVS}(I,O,j) = \mathrm{LVS}(I,O,i_j)$ such that $\forall x < i_j\, \mathrm{LVS}(I,O,x) \leq \mathrm{LVS}(I,O,i_j)$ and $\forall x > i_j\, \mathrm{LVS}(I,O,x) \geq \mathrm{LVS}(I,O,i_j)$.

Step 5: Weigh the ordered local visual feature scores to further increase the prominent errors,

$$\mathrm{LFBVS}(I,O) = \sum_{i=1}^{N} \exp^{i/N-0.5} \mathrm{OLVS}(O,I,i) \,/\, \sum_{i=1}^{N} \exp^{i/N-0.5}.$$

3.7. Visual Information Fidelity (VIF)

The VIF by Sheikh and Bovik [17] uses a refined model which starts with the modeling of the reference image using natural scene statistics (NSS). Furthermore, the possible distortion is modeled as signal gain and additive noise in the wavelet domain and parts of the HVS which have not been covered by the NSS are modeled, i.e. internal neural noise is modeled by using a additive white Gaussian noise model. While the VIF can not be described in the available space the calculation roughly consists of the following steps.

Step 1: NSSs are calculated based on gaussian scale mixture (GSM) model based on the wavelet domain.

Step 2: Calculate a model for the distorted image based on the GSM model from the original image combined with signal gain and additive noise in the wavelet domain (this compensates for white noise and image blur in the image domain).

Step 3: Extend the model to include information from HVS, i.e. optical point spread, contrast sensitivity and internal neural noise, which is not covered by the NSS model.

Step 4: Calculate the amount of the original signal, taking into account different wavelet subbands, which can be reconstructed from the distorted signal given the NSS and the HVS model, this reconstructible fraction of the original signal is termed VIF.

4. EXPERIMENTAL STUDY

Experiments are carried out on the CASIA-v3-Interval iris database[3] using left-eye images only. The database consists of good quality 320×280 pixel NIR illuminated indoor images where the applied test set consists of 1307 instances.

Recognition accuracy is evaluated in terms of false none match rate (FNMR) and false match rate (FMR). The FNMR defines the proportion of verification transactions with truthful claims of identity that are incorrectly rejected, and the FMR defines the proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed (ISO/IEC FDIS 19795-1). As score distributions overlap the EER of the system is defined (FNMR = FMR). At all authentication attempts 7 circular texture-shifts and according bit-shift are performed in each direction for all comparators. Image metric scores are normalized in a way that mean impostor scores are 0.5 and low scores indicate high similarity. Obtained performance rates in terms of EERs for single and paired combination of comparators are summarized in Table 1. According ROC curves of individual image metrics, feature extraction algorithms as well as selected fusion scenarios are plotted in Fig. 2. It is important to note, that all combinations (IrisCode-Metric and Metric-Metric) represent a challenging single-sensor multi-algorithm fusion scenario.

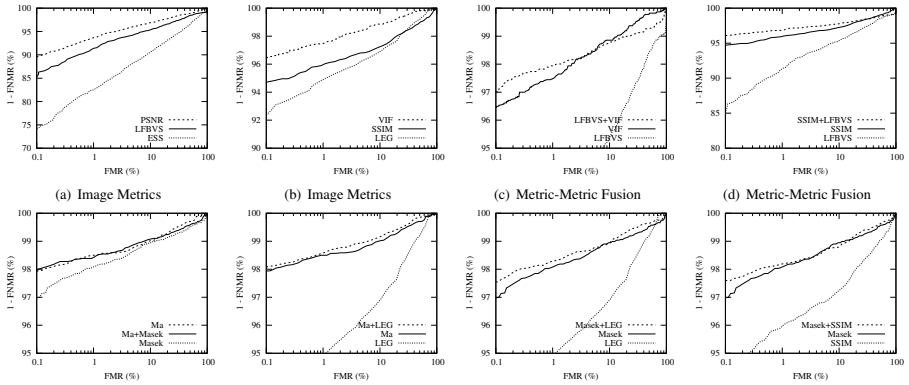
4.1. Combination of Image Metrics

Focusing on obtained EERs most individual image metrics do not represent an alternative to traditional iris-based feature extraction algorithms, see Table 1. While an exclusive application of best image metrics yield EERs $> 2\%$ (see Fig. 2 (a)-(b)) traditional feature extraction algorithms obtain EERs $< 1.5\%$ (see Fig. 2 (e)). However, as shown in Fig. 2 (c)-(d) distinct combinations of image metrics yield significant improvement in accuracy, e.g. a fusion of LFBVS and VIF yields an EER of 1.86%.

4.2. Combination of Metrics and Traditional Algorithms

For the applied simple sum-rule, a combination of applied feature extraction algorithms does not yield improvement with respect to recognition performance, see Fig. 2 (e). In addition, image metrics do not supplement traditional iris recognition algorithms in general. While the incorporation of most image metrics (e.g. PSNR, ESS and LFBVS) decreases performance distinct image metrics represent adequate complements (e.g. SSIM and LEG), see Table 1 and Fig. 2 (f)-(h). In particular, combinations of the LEG metric and applied feature extractors show significant improvements achieving EERs of 1.32% and 1.58%, respectively. Obtained results appear promising since image metrics are applied without any adaption using the most simple fusion rule to the proposed application scenario, i.e. adjusted implementations of image metrics are expected to further improve recognition accuracy.

[3]The Center of Biometrics and Security Research, CASIA Iris Image Database, http://www.idealtest.org



(a) Image Metrics (b) Image Metrics (c) Metric-Metric Fusion (d) Metric-Metric Fusion

(e) Trad. Algorithms (f) IrisCode-Metric Fusion (g) IrisCode-Metric Fusion (h) IrisCode-Metric Fusion

Fig. 2. Receiver Operation Characteristic (ROC) curves for image metric, traditional algorithms, and selected fusion scenarios.

5. CONCLUSION AND FUTURE WORK

In this paper a fusion of image metrics and traditional HD-based comparators is presented. It is demonstrated that the incorporation of distinct image metrics in a fusion scenario is able to significantly improve recognition accuracy of iris biometric systems.

Future work will comprise biometric fusions of several image metrics and traditional biometric comparators as well as an adaption of image metrics to biometric systems.Regarding security issues, image metrics will be assessed for comparing iris images in encrypted domain.

6. REFERENCES

[1] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn, "Image understanding for iris biometrics: A survey," *Comp. Vis. Image Underst.*, vol. 110, no. 2, pp. 281 – 307, 2008.

[2] J. Daugman, "How iris recognition works," *IEEE Trans. Circ. and Syst. for Video Techn.*, vol. 14, no. 1, pp. 21–30, 2004.

[3] J. Daugman and C. Downing, "Effect of severe image compression on iris recognition performance," *IEEE Trans. Inf. Forensics and Sec.*, vol. 3, pp. 52–61, 2008.

[4] I. Tomeo-Reyes, J. Liu-Jimenez, I. Rubio-Polo, and B. Fernandez-Saavedra, "Quality metrics influence on iris recognition systems performance," in *IEEE Int'l Carnahan Conf. Sec. Techn. (ICCST)*, 2011, pp. 1–7.

[5] M. Vatsa, R. Singh, A. Noore, and A. Ross, "On the dynamic selection of biometric fusion algorithms," *IEEE Trans. Inf. Forensics and Sec.*, vol. 10, no. 3, pp. 470 – 479, 2010.

[6] A. Ross and A. K. Jain, "Information fusion in biometrics," *Pattern Recogn. Lett.*, vol. 24, no. 13, pp. 2115–2125, 2003.

[7] A. Uhl and P. Wild, "Single-sensor multi-instance fingerprint and eigenfinger recognition using (weighted) score combination methods," *Int'l J. on Biometrics*, vol. 1, no. 4, pp. 442–462, 2009.

[8] H.-A. Park and K.R. Park, "Iris recognition based on score level fusion by using svm," *Pattern Recogn. Lett.*, vol. 28, pp. 2019–2028, 2007.

[9] C. Rathgeb, A. Uhl, and P. Wild, "Shifting score fusion: On exploiting shifting variation in iris recognition," in *Proc. 26th ACM Symp. Appl. Comp. (SAC'11)*, 2011, pp. 1–5.

[10] A. Uhl and P. Wild, "Weighted adaptive hough and ellipsopolar transforms for real-time iris segmentation," in *Proc. Int'l Conf. on Biometrics (ICB)*, 2012, to appear.

[11] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement," *J. VLSI Signal Process. Syst.*, vol. 38, no. 1, pp. 35–44, 2004.

[12] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Efficient iris recognition by characterizing key local variations," *IEEE Trans. Image Proc.*, vol. 13, no. 6, pp. 739–750, 2004.

[13] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[14] H. Hofbauer and A. Uhl, "An effective and efficient visual quality index based on local edge gradients," in *IEEE 3rd Europ. Workshop on Visual Inf. Proc.*, 2011, p. 6pp.

[15] Y. Mao and M. Wu, "Security evaluation for communication-friendly encryption of multimedia," in *IEEE Int'l Conf. on Image Proc. (ICIP)*, 2004.

[16] L. Tong, F. Dai, Y. Zhang, and J. Li, "Visual security evaluation for video encryption," in *Proc. Int'l Conf. on Multimedia*, 2010, MM '10, pp. 835–838.

[17] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Proc.*, vol. 15, no. 2, pp. 430–444, May 2006.

IEEE TRANSACTIONS ON MULTIMEDIA, VOL. X, NO. X, JUNE 201X

An Evaluation of Visual Security Metrics

Heinz Hofbauer, Andreas Uhl

***Abstract*—Visual security metrics are deterministic measures with the (claimed) ability to assess whether an encryption method for visual data is secure or not. These metrics are usually developed together with a particular encryption method in order to provide an evaluation of the encryption method based on the visual output of the encryption scheme. However, visual security metrics themselves are seldom evaluated and the claim to perform as visual security metric is not tied to the specific encryption method for which they were developed. In this paper we will systematically evaluate these visual security metrics (along with conventional image metrics used for the same task) for distinct media encryption application scenarios and show that they are not generally fit to perform their claimed task.**

***Index Terms*—security metrics, image metrics, selective encryption, confidence, partial encryption, sufficient encryption, transparent encryption.**

I. Introduction

The claim of visual security metrics (security metrics for brevity) is usually the ability to assess the security of an encryption method based on the output of the encryption of visual data. In particular, that is the evaluation of an encryption method based only on the visual output (the ciphertext), which is either an image or video. While such metrics are often created in conjunction with a specific encryption method and tested, if at all, only for this encryption method, the claim to perform as a security metric is usually universal. Furthermore, regular image quality metrics, most frequently PSNR and SSIM, are also utilized in literature to evaluate encryption methods, e.g. [1], [2].

Regarding cryptographic security, Shannon's work [3] on security and communication shows that the highest security is reached by applying a secure cipher to almost redundancy free plain text. Current image/video codecs exploit redundancy for compression and thus we can consider a bit stream to be a redundancy free plain text in the sense of Shannon. Thus, for maximal security, the encryption of the entire bit stream with an state of the art cipher, i.e. AES, would suffice ("conventional encryption"). However, there are well-founded reasons not to stick to this approach, but to apply specifically designed encryption routines:

I: The implementation of advanced application scenarios where visual data has to be retained, such as transparent/perceptual encryption and privacy preserving encryption. The goal of such applications can be different but directly enforces certain quality restraints. Transparent encryption aims at degrading the overall image quality to a certain extent. Privacy preserving encryption aims at high quality overall except where privacy is concerned, i.e. surveillance cameras should provide an operator with a crisp image but should encrypt faces to preserve privacy, e.g. Dufaux and Ebrahimi [4].

II: The preservation of properties and functionalities of the bitstream, such as format compliance, scalability, streaming/packetization, secure adaptation, fast forward, extraction of subsequences, transcodability, watermarking, and error resilience.

III: The reduction of computational complexity (especially in the context of mobile computing).

In many of those specifically designed encryption routines, techniques like Lightweight / Soft / Partial / Selective Encryption are employed, which achieve their respective advantages with a loss in security / secrecy as compared to conventional encryption. For example, in selective / partial encryption the choice is made to keep information in plain text. Lookabaugh et al. [5] showed that selective encryption is sound and demonstrated its relation to Shannon's work. However, Said [6] showed that side information can compromise security.

In order to be able to discuss the exact notion of security in such non-conventional encryption schemes, we need to distinguish distinct application scenarios of encryption schemes for visual data:

a) Confidentiality Encryption: Means MP security (message privacy). The formal notion is that if a system is MP-secure an attacker cannot efficiently compute any property of the plain text from the cipher text [7]. This can only be achieved by the conventional encryption approach.

b) Content Confidentiality: Is a relaxation of confidential encryption. Side channel information may be reconstructed or left in plaintext, e.g. header information, packet length, but the actual visual content must be secure in the sense that the image content must not be intelligible / discernible [8].

c) Sufficient Encryption: Means we do not require full security, just enough security to prevent abuse of the data. The content must not be consumable due to high distortion (e.g. for DRM systems) by destroying visual quality to a degree which prevents a pleasant viewing experience or destroys the commercial value. This implicitly refers to message quality security (MQ), which requires that an adversary cannot reconstruct a higher quality version of the encrypted material than specified for the application scenario [9].

d) Perceptual / Transparent Encryption: Means we want consumers to be able to view a preview version of the video but in a lower quality while preventing them from seeing a full version. This for example can be used in a pay per view scheme where a lower quality preview version is available from the outset to attract the viewers interest, e.g., Li et al. [10]. The difference between sufficient and transparent is the fact that there is no minimum quality requirement for sufficient encryption. Encryption schemes which can do sufficient encryption cannot necessarily ensure a certain quality and are

The authors are with the Department of Computer Sciences, University of Salzburg, e-mail: {hhofbaue, uhl}@cosy.sbg.ac.at

thus unable to provide transparent encryption.

Given these different application scenarios it is clear that depending on the goal, a security metric has to fulfill different roles. For example, under the assumption of sufficient encryption a given security metric would have to evaluate which quality is low enough to prevent a pleasant viewing experience. In contrast, for the transparent encryption case a metric not only has to judge whether the quality of an image or video is low enough, but also has to grade if the quality is high enough to be useful to attract interest. When it comes to content confidentiality the question of quality is no longer applicable. Content confidentiality requires that image content must not be identified by human or automated recongnition. This requirement also has to be maintained for any part of the image. Image metrics, in general, do not deal with such questions but rate the overall image quality, the question of intelligibility is usually not covered at all. A drastic example would be an image where only a small part of the image is partly visible. Classical metrics would judge the whole image and consequently would attribute a high security, even though a part of the image is still recognizable which contradicts content confidentiality. Still, it has to be pointed out that also content confidentiality can have different formings. To prevent a face recognition scheme from working properly it is sufficient to protect any facial information in a surveillance video, while humans could still be identified in such a video by using gait recognition. Furthermore, if the appearace of a person has to be concealed entirely, a much stronger extent of protection (i.e. higher security) is required. Finally, confidential encryption cannot be solely assessed with security metrics since the scope goes beyond assessing security based on the visual appearance only. Furthermore, we should note that the application of security metrics on video is performed at a frame by frame basis in literature. We will adopt this model but should note that for the discussion of confidential encryption motion data is of importance, e.g. Hofbauer and Uhl showed in [11] that a replacement attack combined with motion information can reveal the content of a scene even though the visual content of every frame is encrypted.

Consequently, depending on a given application scenario different properties are required from a security metric and different approaches to construct such a metric might perform better or worse for some applications scenarios. This dependence on the evaluation goal of a security metric is hardly ever discussed in the papers introducing a metric. Sufficient and transparent encryption scenarios have a clear and distinct link to the traditional notion of (low) visual quality, while it is highly questionable or at least doubtful if content confidentiality can be assessed by the classical quality notion. While the lack of relation to spatial areas of most security metrics could be compensated in the design to provide locally varying results, the lack of relation to intelligibility in general can probably not be easily resolved.

For both, security metrics and regular image metrics, in literature we do not find any evaluation whether a given metric can perform the claimed function or how such an evaluation correlates to actual security. However, for regular image metrics it is well known that the correlation with human observations over the full range of possible quality (from high to low quality) does not imply a good performance on a given subset. More specifically, it was pointed out recently by Hofbauer and Uhl [12] that most image metrics perform very poorly for the low quality range. For security metrics, not even this question has been covered so far.

In this paper we will try to remedy this situation and give an overview of requirements regarding security goals as well as evaluate the various metrics in relation to these goals. However, we will not deal with every application scenario equally explicitely. We will only make a first step to cover the content confidentiality scenario. The main reason for this is a lack of ground truth. It is not obvious how to generate gound truth for this scenario since there is a disparity between how an image metric works and what is necessary to evaluate content confidentiality. Image metrics, and as an extension security metrics, measure the quality of an image respective to human judgement. This works well for high quality images but suffers for low quality images where human observers can have difficulties differentiating between the severity of an impairment. Thus the methodology to systematically generate ground truth based on human observation needs to be changed for content confidentiality which is not in the scope of this paper. On the other hand, for the image quality-related scenarios (sufficient and transparent encryption), ground truth data is available, in the form of image impairment databases with mean opinion scores (MOS) based on a number of human observations.

In the following we will evaluate whether security metrics can actually be used to perform security assessment of encrypted visual data. In order to do this we will give an overview of security metrics used in literature in section II. In section III we will describe what is expected of security metrics and the methodology how security metrics can be evaluated. In section IV we will present the actual evaluation and section V will conclude the paper.

II. Overview Of Security Metrics

In this section a brief overview of security metrics will be given. The metrics discussed are taken from recent literature and are specifically designed to ascertain whether the image quality after encryption is sufficiently reduced. The metrics given in this section are discussed as general security metrics, i.e., not limited to the specific method with which they were designed together. References to the original work will be given for each security metric as well as some examples where the metric fails to assess given example images as would be expected from a general security metric. The SSIM and PSNR are also included in this overview. Even though they were not designed to be security metrics, they are frequently used as such, e.g., [13], [14] (SSIM),[15], [16] (PSNR) and [1], [2] (SSIM and PSNR).

All the following image metrics, with the exception of the local entropy metric (LE), are full reference metrics, meaning they utilize information from the original and comparison (encrypted) image to calculate an assessment of the visual similarity. The local entropy metric by Sun et al. is a no reference metric, i.e. it utilizes only the impaired image to

judge the resulting quality. By measuring entropy, LE can also be interpreted to assess the encrypted image compared to random noise (which exhibits maximal LE). Since all of the given security metrics are proposed to be general we will not differentiate between full- and non-reference metrics in the following but compare them solely on the task they are supposed to solve.

Considering the metrics' design, the LE seems to be the only one suited to cover content confidentiality. An image consisting entirely of noise obviously satisfies the requirements of content confidentiality. Thus, determining the difference to noise by measuring entropy can be interpreted as measuring the extent of security. In contrast, all other metrics determine the difference to the original plaintext image, thus rather correspond to the classical notion of quality.

Peak Signal to Noise Ratio (PSNR)

The peak signal-to-noise ratio (PSNR) is still widely used because it is unrivaled in speed and ease of use. However, it is also well known that the correlation to human judgement is somewhat lacking even for high and medium quality [17]. Figure 1 illustrates the performance of the PSNR metric on samples from the IVC-SelectEncrypt [18], [19] database (see section III).

(a) 24.4566 (b) 24.4767 (c) 33.2437

Fig. 1. PSNR metric scores for images from the IVC-SelectEncrypt database. According to PSNR images (a) and (b) are of the same quality and (c) is of much higher quality, i.e. less secure than (a) and (b).

Structural Similarity Index Measure (SSIM)

The structural similarity index measure (SSIM) by Wang et al. [20] extracts three separate scores from the image and combines them into the final score. First the visual influence is calculated locally then luminance, contrast and structural scores are calculated globally. These separate scores are then combined with equal weight to form the SSIM score. Figure 2 illustrates the performance of the SSIM metric on samples from the IVC-SelectEncrypt [18], [19] database.

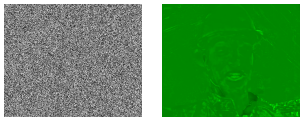
(a) 0.568381 (b) 0.561515 (c) 0.293894

Fig. 2. SSIM metric scores for images from the IVC-SelectEncrypt database. According to SSIM images (a) and (b) are of the same quality and (c) is of much lower quality, i.e. more secure than (a) and (b).

Edge Similarity Score (ESS)

The edge similarity score (ESS) was introduced by Mao and Wu [21] and uses localized edge direction information to compare two images. Figure 3 illustrates the performance of the ESS metric on the foreman sequence when encryption according to [22] is applied in comparison to white noise.

(a) Original (b) Processed



(c) White noise (0.273534) (d) Extracted (0.053214)

Fig. 3. ESS comparison for frame 80 of the foreman sequence (a). ESS judges the white noise (c) to be of higher quality than the residual information from the encrypted frame (d). In order to show the amount of information actually retained in the encrypted frame a post processed version is also shown (b).

Luminance Similarity Score (LSS)

The luminance similarity score (LSS) was introduced by Mao and Wu [21] and uses localized luminosity information to compare two images. Figure 4 illustrates the performance of the LSS metric on the foreman sequence when encryption according to [23] is applied in comparison to noise.

(a) Noise (-1.3080) (b) Extracted (-1.6388) (c) Original

Fig. 4. LSS comparison for a frame of the foreman sequence (c). LSS judges the noise (a) to be of higher quality than the residual information from the encrypted frame (b).

Neighborhood Similarity Degree (NSD)

The neighborhood similarity degree metric (NSD), introduced by Yao et al. [24], uses local pixel similarity correlation between original and impaired image. The NSD depends on two parameters, one to define the region for pixel similarity correlation (d) and one to define the similarity threshold (m). The parameters m and d were set to the same values as in the experiments in [24], i.e., $m = 5$, $d = 3$, border extension is done by repeating the last border pixel. Figure 5 illustrates the performance of the NSD metric on samples from the IVC-SelectEncrypt [18], [19] database.

Local Entropy (LE)

The local entropy metric was introduced by Sun et al. [25] (LE), it is a no reference metric operating only on an impaired image. The LE metric uses the average of normalized localized entropy scores, on 8×8 blocks, as image quality predictor. Figure 6 illustrates the performance of the LE metric on samples from the IVC-SelectEncrypt [18], [19] database.

Local Feature Based Visual Security (LFBVS)

The local feature based visual security metric (LFBVS) was introduced by Tong et al. [26] and utilizes localized edge and luminance features which are combined and weighted according to error magnitude, i.e. error pooling. Figure 7 illustrates the performance of the LFBVS metric on the silent sequence when encryption according to [22] is applied in comparison to white noise.

From the description of the various security metrics it can be seen that a wide range of approaches exist, from metrics targeting signal properties, e.g. PSNR which targets noise, to LE which targets local entropy, metrics which use higher level information, e.g. NSD or ESS which use a form of object detection (mostly based on edges), to metrics which use information about he HVS to improve their performance, e.g. SSIM or LFBS which use simulation of the fovea centralis and error pooling respectively. However, for every security metric the accompanying figure demonstrates a fault in the performance of the given metric. Such examples can be found for every metric of course, the question we will try to answer in the following sections is whether the demonstrated fault is singular or systemic.

(a) 0.066527 (b) 0.068766 (c) 0.111781

Fig. 5. NSD metric scores for images from the IVC-SelectEncrypt database. According to NSD images (a) and (b) are of the same quality and (c) is of much lower quality, i.e. more secure than (a) and (b).

(a) 0.039396 (b) 0.039555 (c) 0.045168

Fig. 6. LE metric scores for images from the IVC-SelectEncrypt database. According to LE images (a) and (b) are of the same quality and (c) is of much lower quality, i.e. more secure than (a) and (b).

(a) Original (b) Processed

(c) White noise (0.730991) (d) Extracted (0.757152)

Fig. 7. LFBVS comparison for frame 80 of the silent sequence (a). LFBVS judges the residual information from the encrypted frame (d) to be about the same as the information contained in white noise (c). In order to show the amount of information actually retained in the encrypted frame a post processed version is also shown (b).

III. Evaluation Methodology

In this section we will outline the evaluation methodology, the reason to use this methodology, and the application scenario which can be assessed employing a certain methodology. A discussion of desired outcome from these test for security metrics will also be given. This section is the guideline of how the security metrics and image metrics are evaluated for the use as security metrics in section IV.

A. Comparison to Regular Quality Metrics

In order to gauge the effectiveness of a dedicated security metric it is useful to compare them to regular metrics. If the security metrics improve on the regular image metrics in some aspect regarding the security considerations then the security metric is worthwhile even if a regular image metric outperforms it in quality control tasks. A second reason to include those metrics is to gauge whether they can be used as security metrics. In order to facilitate a fair comparison, three additional recent metrics are chosen (in addition to SSIM and PSNR which are often used as security metrics as well). The local edge gradient image metric (LEG) by Hofbauer and Uhl [27] shows a good correlation with human judgement and is reasonably fast to compute such that it can be used even under time constraints. The visual information fidelity (VIF) by Sheikh and Bovik [28] and CPA1 by Carosi et al. [29] outperform the SSIM and LEG in regard to correlation with human judgement but are a lot slower to compute [27].

B. Application Domain

Frequently security metrics are applied on a direct reconstruction of the encrypted bitstream. This can have adverse effects since encryption introduces noise which can hide plain text data and consequently a security metric might judge that an encryption method is more secure then it actually is. There are a number of options how a security metric can be applied as illustrated in fig. 8. All security metrics under evaluation are applied on a direct reconstruction (decoding) of the encrypted bitstream (which we will denote as *encrypted domain*) by the authors in the respective original papers. The other options would be to attack the encryption method in a way which does not break it but reduces the obfuscation of the plain text data in the encrypted domain (denoted *extracted* in the figure). Such attacks usually utilize knowledge about

the bitstream rather than the encryption method (other than location), typical attacks would be error concealment and replacement attacks against selective encryption schemes [30], [31]. Another possibility would be to utilize post processing to further help the metric detect residual information (denoted *processed* in the figure).

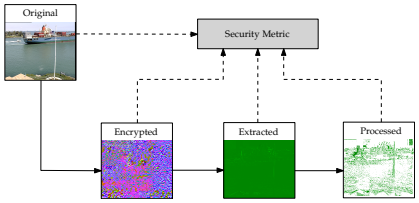


Fig. 8. The possible domains which can be used by a metric to compare to an original image. Either the direct output of an encrypted bit stream for format compliant encryption or an extraction of the plain text data which minimizes the disruptive effect of the encrypted data on the resulting bit stream. Another possibility would be post processing to further accent the residual plain text data contained in an image.

In the evaluation we only handle the difference between security metrics applied directly in the encrypted domain versus application in the extracted domain. The post processing step is only provided to better highlight the information remaining in an image. It cannot be directly used as an application domain since the post processing step is either specific to the encryption, in which case it should be included in the attack, or specific to the metric, in which case it should be included in the security metric. Post processing in general can influence different metrics in different ways, an example of this is given in fig. 9 where the post processing increases ESS and at the same time decreases the SSIM.

Fig. 9. An example of how post processing influences different metrics.

In order to test whether a security metric can operate in the encrypted domain a number of well known video sequences have been encrypted for three target qualities, utilizing ECBZ encryption methods described in [22]. To generate the different target qualities the selective encryption was applied to either all I-frames (low quality), low frequency bands of all frames (medium quality), and high frequency bands of all frames (high quality). Figure 10 illustrates the quality targets of the encryption process. The targets where chosen to contain high, medium and low residual information. Under the assumption that a security metric can operate in the encrypted domain it should be able to reliably order the encrypted frames of each sequence for every sequence from highest to lowest quality.

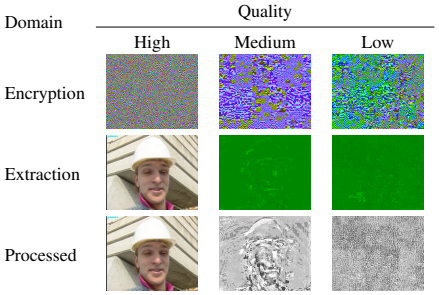


Fig. 10. A sample of the quality ordering test set based from the foreman sequence. Samples from the high, medium and low quality sequences are shown in the encryption, extraction and processed domain.

This evaluation consists of two comparisons: For each frame of each sequence the security metric must do two comparisons, high versus medium and medium versus low quality. Results of this ordering can be averaged over each sequence in order to get the number of correct orderings. Results around 50% are akin to random decisions while results close to 100% and 0% show a strong ability to order the encrypted images correctly and give the direction of the ordering. The reason why both 100% and 0% are valid orderings is because image metrics can either measure similarity of images, i.e., quality metrics, or the difference between images, i.e., impairment metrics.

Furthermore, since the low quality range chosen is in (or at least close to) the domain of content confidentiality this setting also serves as an indication whether an image metric might be useful for content confidentiality. While a good performance on this evaluation does not necessarily mean a image metric is qualified for content confidentiality, a low performance is a strong indicator that the metric is unfit for this task. Based on the information which parts of the data have been encrypted and the entirely evident differences in visual appearance, ground truth is out of question here.

C. Correspondence to HVS: Confidence

Besides the encryption application scenarios where a certain quality is required (sufficient and transparent encryption), further examples for the importance of the quality notion are watermarking where the resulting quality should not be below a certain threshold, and of course, lossy compression. However, the notion of quality in this cases is not as straightforward as it seems. On the one hand we use the term quality in the context of the human visual system (HVS), i.e. how a person consuming the content would judge the quality. On the other hand, the term quality can refer to the score returned by a (security) metric which is tied to the quality in the HVS sense. This relation is not exact and it is not inherently clear how to choose a metric which correlates to the HVS quality which is targeted, although in practice algorithm 1 is usually applied.

Algorithm 1 Method for finding a target metric score based on a target HVS quality.

1: Chose a source image.
2: Alter the image until it fits the perceived target quality.
3: Apply the security metric on the altered and original image, the resulting metric score is the target quality.

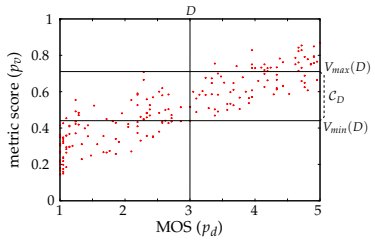


Fig. 11. Illustrations of zero false negative $V_{min}(D)$, zero false positives $V_{max}(D)$ and confidence $\mathcal{C}_D$ for a MOS value of $D = 3$ is shown based on the IVC-SelectEncrypt database and the LEG metric.

While this results in a target quality which can be used, we know nothing about how well this score actually reflects the human judgement, since it is well known that the correlation between human judgement and image metrics is not perfect. In other words, how confident can we be in the choice of image metric score in relation to the perceived quality?

In order to evaluate this, well known databases which contain impaired and encrypted images and the perceived quality, in the form of mean opinion scores (MOS), will be used. The databases contain a set of points p representing impaired images with associates values p_v for metric value and p_d for MOS value, ordered from lowest to highest quality. Based on a target MOS quality score D two values can be calculated, fig. 11 illustrates this.

Zero false negative: $V_{min}(D)$ refers to the metric value for which the following holds:

$$p_d > D \implies p_v > V_{min}(D). \tag{1}$$

That is if the metric score is below $V_{min}(D)$ we are sure that the perceived quality is below the MOS quality score (D).

Zero false positives: $V_{max}(D)$ refers to the metric value for which the following holds:

$$p_v > V_{max}(D) \implies p_d > D. \tag{2}$$

That is if the metric score is above $V_{max}(D)$ we are sure that the perceived quality is above the MOS quality score (D).

This also means that if a target metric quality score p_v^t is obtained as given by algorithm 1 we are assured that

$$V_{min}(D) \geq p_v^t \geq V_{max}(D). \tag{3}$$

Thus we can define the confidence $\mathcal{C}_D$ for a metric score based on a given perceived quality D as $\mathcal{C}_D := |V_{max}(D) - V_{min}(D)|$. A confidence score over the full perceived quality range can be given as

$$\mathcal{C} = \frac{1}{\#S} \sum_{D \in S} \mathcal{C}_D, \tag{4}$$

where S is the set of distinct MOS samples from the database.

Also note that we can interpret $\mathcal{C}$ as $\mu_{D \in S}(\mathcal{C}_D)$ and can also calculate $\sigma_{D \in S}(\mathcal{C}_D)$. The reason for calculating σ is to estimate how stable the confidence range is over the whole range of visual quality. This has to be taken into account since it is well known that image metrics exhibit different correlation to human judgement depending on the quality range, see e.g. Hofbauer and Uhl [12].

For security metrics, and image metrics in general, the lower $\mu(\mathcal{C}_D)$ and $\sigma(\mathcal{C}_D)$ the better algorithm 1 can be used to estimate a target image quality metric score.

Furthermore, since the signal is reduced to statistical components it is also of interest which shape the signal takes in conjunction with $\mu(\mathcal{C}_D)$ and $\sigma(\mathcal{C}_D)$. The shape, together with the monotonicity (see subsection III-D), can be used to indicate a possible application scenario for a security metric, essentially whether the security metric can be used for all quality ranges or only on high/low quality applications.

By shape of the signal we mean the distribution of outliers, where we define outlier based on the z score[1] of a datapoint D as

$$z_D = \frac{\mathcal{C}_D - \mu(\mathcal{C}_D)}{\sigma(\mathcal{C}_D)}. \tag{5}$$

We will define *high outliers* as outliers with $z_D < -1$ and likewise *low outliers* as outliers with $z_D > 1$, indicating a higher and lower confidence respectively. Based on the distribution of high and low outliers we can specify the shape of the signal as follows.

- A signal is **stable** if there are no outliers, i.e. $-1 \geq z_D \geq 1$ holds for all $D \in S$.
- A signal is **biased** if there exists a D_t such that $z_D < -1 \implies D < D_t$ and $z_D > 1 \implies D > D_t$ or $z_D < -1 \implies D > D_t$ and $z_D > 1 \implies D < D_t$. If a low D indicates a high quality we specify the shape to be **biased towards high quality** if $z_D < -1 \implies D < D_t$ or **biased towards low quality** if $z_D > 1 \implies D < D_t$. If a low D indicates low quality the definition is switched accordingly.
- A signal which is neither stable nor biased is considered **unstable**.

D. Correspondence to HVS: Monotonicity

What is required from image metrics in general is monotonicity with regard to human observations. That is, if an image metric decides that image A is of better quality than image B a human observer should also prefer image A over image B. This is akin to correlation but since the human visual system is not a linear system regular linear correlation is meaningless. Thus in order to ascertain the correlation of an image metric and human observations the notion of monotonicity is utilized. Rank order correlation, which essentially judges the

[1]NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, April 2012

monotonicity of the signals, is most often used, usually in the form of Spearman's rank oder coefficient (SROC) [32] or Kendall Tau (τ) [33].

Hofbauer and Uhl [12] pointed out that the correlation of an image metric over the full quality range does not imply that a high correlation is achieved for the low quality range. This is especially important for security metrics since certain application scenarios specifically target the low quality range of images, e.g. sufficient encryption. We cannot confine the evaluation to the low quality range since there are also applications for higher quality, i.e. transparent encryption. Also note that this is a dual property to the confidence in the sense that for the confidence we evaluate the relation of choosing a MOS value and evaluating the range of metric scores which can potentially fall onto this MOS value. Monotonicity is evaluated on specific sets of impairment and looks at how well an increase in metrics score reflects an increase in the MOS.

To properly evaluate security metrics for all encryption scenarios we will evaluate them using a high quality, a low quality and a full quality range dataset. The reason to also include the high quality range is to be inclusive in terms of possible application scenarios. For example the upgrade to high definition quality from PAL/NTSC quality is just as valid in terms of application scenarios as from hand held quality to PAL/NTSC quality. As basis for the evaluation we will use well known databases which contain mean opinion score of human judgement over different impairments. The SROC will be used for evaluation purpose, as is current best practice for metric evaluation.

From security metrics we would expect a high correlation with human judgement for the low quality range. While the low quality range is often the target of encryption some transparent encryption schemes could target a higher quality, consequently, a good correlation with human judgement on the high quality range is also desirable.

IV. Evaluation

In this section we will present the results of the evaluation process as detailed in section III. Each evaluation will be lead with a short description of the test data, contain the actual data from the evaluation and a discussion.

With respect to implementations we used our own code[2] for LSS, ESS, LE, NSD, LFBVS, SSIM and PSNR. For the VIF we used the implementation from the "MeTriX MuX Visual Quality Assessment Package"[3], version 1.1. For the CPA1 we used the matlab implementation provided by Florent Autrusseau[4].

A. Evaluation of the Application Domain

In order to evaluate the extraction versus encrypted domain applicability of metrics we used a number of standard sequences: akiyo, bus, coastguard, container, flower, foreman, mobile, news, silent, tempete and waterfall[5]. The ordering as discussed above is performed on a frame by frame basis and averaged over all the frames in a given sequence. Additionally we provide the average over all sequences in order to simplify the comparison. Table I shows the results for the encryption and extraction domain. The optimal result would be for a metric to perform equally well independent of the application domain.

From the overall averages it can be clearly seen that the performance in the encrypted domain is worse than in the extraction domain with the exception of LE. While the LE performs better in the encrypted domain than in the extraction domain the performance is still very low. In the extraction domain most metrics still perform poorly, only the LEG, SSIM, VIF, CPA1 and to a lesser degree the ESS exhibit good performance. Aside from the ESS all other security metrics perform extremely poorly. From this we can easily see that the application in the encrypted domain should not be performed (although it is routinely done in all corresponding papers).

When looking closer at the detail information from the extraction domain some interesting effects can be observed. While the performance of the LEG, SSIM and VIF are persistently good we can also observe that there are cases where even a metric which shows good performance can have problems. In the actual case the performance of the ESS is significantly reduced for the waterfall sequence and the CPA1 shows poor performance on the bus sequence. It stands to reason that there are other, untested, sequences which would lead to similar reduced performance for the LEG, SSIM and VIF. Another noteworthy fact is the appearance of 50% scores, in all such cases one of the comparisons always yielded the correct result and one always the incorrect result. In the case of the LE for example the high quality was consistently and correctly rated higher than the medium quality, but the low quality was also consistently and incorrectly rated higher than the medium quality resulting in the overall bad score. A somewhat similar case can also be observed for LSS and PSNR. The PSNR rated the foreman sequence consistently and correctly in order of high, medium, low quality but reversed the order for the container sequence, i.e. the ordering was low, medium, high.

B. Evaluation of Confidence

In order to evaluate the confidence, databases with either pertinent content, i.e. encrypted images, or a large dataset of distortions are optimal. The IVC-SelectEncrypt database [18], [19] contains various examples of JPEG 2000 transparent encryption and is an obviously useful tool for the evaluation of confidence regarding encrypted images. It is the only database available containing encrypted visual data and corresponding MOS. The test sets contained in the IVC-SelectEncrypt database (and their abbreviation) are traditional encryption (trad), truncation of the code stream (trunc), window encryption without error concealment (iwind_nec), window encryption with error concealment (iwind_ec), and wavelet packet encryption (res), for detailed information see Stütz et al. [19].

[2] http://www.wavelab.at/sources/VQI
[3] http://foulard.ece.cornell.edu/gaubatz/metrix_mux/
[4] http://www.irccyn.ec-nantes.fr/~autrusse/Softwares.html
[5] Available for example at http://media.xiph.org/video/derf

TABLE I
RESULTS FOR THE ORDERING OF HIGH VERSUS MEDIUM AND MEDIUM VERSUS LOW QUALITY SEQUENCES IN THE ENCRYPTION AND EXTRACTION DOMAIN. THE AVERAGES OVER ALL THE FRAMES IN A SEQUENCE AS WELL AS THE AVERAGE OVER ALL SEQUENCES IS SHOWN PER IMAGE METRIC.

Results for the Encryption Domain

	LEG	SSIM	LSS	ESS	PSNR	LFBVS	LE	NSD	VIF	CPA1
akiyo	16.80 %	50.00 %	49.61 %	44.14 %	52.73 %	50.00 %	60.94 %	48.05 %	44.92 %	49.61 %
bus	39.84 %	38.28 %	40.23 %	33.20 %	55.47 %	51.56 %	89.45 %	42.58 %	38.28 %	25.00 %
coastguard	50.00 %	39.06 %	33.98 %	43.36 %	58.20 %	58.98 %	75.39 %	42.19 %	50.78 %	25.39 %
container	24.61 %	48.44 %	20.70 %	39.06 %	34.77 %	50.00 %	77.73 %	56.25 %	58.59 %	3.52 %
flower	19.14 %	42.97 %	52.73 %	41.02 %	44.53 %	48.83 %	99.61 %	39.84 %	48.83 %	39.45 %
foreman	28.91 %	48.05 %	49.22 %	41.02 %	50.78 %	50.39 %	54.69 %	45.31 %	50.00 %	48.05 %
mobile	33.98 %	48.05 %	50.00 %	30.47 %	45.70 %	51.17 %	68.75 %	49.22 %	45.31 %	47.27 %
news	42.97 %	50.00 %	39.45 %	41.80 %	74.22 %	50.00 %	50.00 %	45.70 %	36.33 %	49.61 %
silent	26.95 %	33.20 %	24.22 %	10.16 %	25.00 %	60.16 %	99.22 %	37.50 %	32.03 %	14.84 %
tempete	41.41 %	48.05 %	50.00 %	39.45 %	50.00 %	50.00 %	74.22 %	48.05 %	44.92 %	46.48 %
waterfall	42.19 %	49.22 %	48.83 %	46.09 %	71.48 %	50.78 %	53.91 %	51.17 %	49.22 %	49.61 %
average	33.35 %	45.03 %	41.73 %	37.25 %	51.17 %	51.99 %	73.08 %	45.99 %	45.38 %	36.26 %

Results for the Extraction Domain

	LEG	SSIM	LSS	ESS	PSNR	LFBVS	LE	NSD	VIF	CPA1
akiyo	100.00 %	100.00 %	51.56 %	94.14 %	100.00 %	50.00 %	50.00 %	50.00 %	94.53 %	9.38 %
bus	99.61 %	100.00 %	51.17 %	96.88 %	99.22 %	50.00 %	50.00 %	46.09 %	98.05 %	23.44 %
coastguard	100.00 %	100.00 %	50.78 %	99.22 %	99.61 %	45.70 %	50.00 %	24.61 %	100.00 %	0.78 %
container	100.00 %	100.00 %	100.00 %	94.14 %	0.00 %	50.00 %	50.00 %	50.00 %	100.00 %	0.39 %
flower	100.00 %	100.00 %	95.31 %	92.58 %	6.25 %	41.80 %	50.00 %	47.27 %	98.44 %	9.38 %
foreman	100.00 %	100.00 %	53.12 %	99.61 %	50.00 %	50.00 %	50.00 %	50.00 %	99.61 %	0.00 %
mobile	98.44 %	100.00 %	92.19 %	97.66 %	43.36 %	50.00 %	50.00 %	41.02 %	98.44 %	1.95 %
news	100.00 %	100.00 %	50.00 %	99.61 %	100.00 %	50.00 %	50.00 %	50.00 %	99.61 %	0.00 %
silent	100.00 %	100.00 %	90.23 %	98.05 %	57.81 %	50.00 %	50.00 %	50.00 %	100.00 %	0.00 %
tempete	100.00 %	100.00 %	57.03 %	99.22 %	100.00 %	50.00 %	50.00 %	48.05 %	100.00 %	0.00 %
waterfall	95.70 %	99.22 %	69.53 %	82.81 %	92.97 %	50.00 %	50.00 %	24.61 %	98.05 %	0.78 %
average	99.43 %	99.93 %	69.18 %	95.81 %	68.11 %	48.86 %	50.00 %	43.79 %	98.79 %	4.19 %

However, the IVC-SelectEncrypt database has a rather small set of impairments, five per test set, and is focused on JPEG 2000 only.

In order to get a more diverse view on the confidence of metrics we utilize the LIVE database [34] to supplement the IVC-SelectEncrypt database of encrypted images. While the LIVE database does not contain encrypted images the quality range of images contained in the LIVE database reaches from high to low quality which makes it at least relevant for transparent encryption where a certain target quality is required. Furthermore, the low quality range of the LIVE database displays strong distortions an can be equated to encrypted images in the sense that strong distortions mask a lot of the information contained in an image. Consequently, the distorted images can be used to assess how well a metric can identify information contained in a distorted/encrypted image, which is exactly the property wanted from security metrics. An example of these strong distortions are given in fig 12, which contains an encrypted image from the IVC-SelectEncrypt database as well as heavily distorted versions of images from the LIVE database. These examples show that the LIVE database contains not only images which are similar to the IVC-SelectEncrypt database in terms of content masking but also image which are clearly in the quality realm of sufficient encryption. The test sets contained in the LIVE database (and their abbreviation in plots and figures) are JPEG 2000 compression (jp2k), JPEG compression (jpeg), white noise (wn), Gaussian blur (gblur), and bit errors in JPEG2000

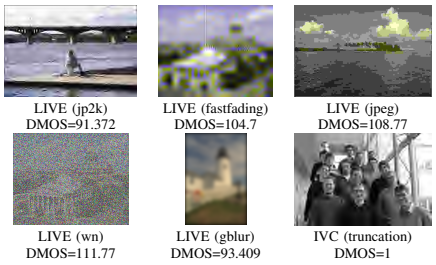


Fig. 12. The lowest quality images of each test set in the LIVE database as well as one of the lowest quality images from the IVC-SelectEncrypt database.

bit stream transmission over a simulated fast fading Rayleigh Channel (fastfading), for detailed information see Sheikh et al. [35].

While the TID database [36]also has a huge number of observer scores and distortion types it also has a severe drawback. Due to the evaluation process a given distortion type does not necessarily span the whole range of MOS values which would result in a non-uniform evaluation of confidence over the MOS range. Thus only the LIVE and IVC-SelectEncrypt databases are used for the evaluation of confidence.

Figure 13 shows the detailed evaluation of confidence on the LIVE database. For each metric the figure presents a scatter

plot of MOS and metric values, the bounding curves $V_{\min}(D)$ and $V_{\max}(D)$ as well as a plot of the local confidence value $\mathcal{C}_D$. Table II gives the confidence score values $\mu(\mathcal{C}_D)$ and $\sigma(\mathcal{C}_D)$ for each metric on the LIVE database. In order to make the confidence scores comparable a pseudo normalization was used. Bounded metrics are normalized and unbounded metrics, e.g., PSNR, are normalized by mapping the range of metric score actually occurring on the database into unity range. For the calculation of the signal shape the leading and trailing 10% of the DMOS range where not taken into account because at the high and low ends of the DMOS the difference in metric scores is limited due the boundary of the metric range, see figure 13, which would result in false outliers. The shape was calculated, like $\mu(\mathcal{C}_D)$ and $\sigma(\mathcal{C}_D)$, on pseudo normalized local confidence value $\mathcal{C}_D$. In figure 13 the outliers are indicated by down and up arrows for high and low quality outliers, respectively.

For the IVC-SelectEncrypt database the accordant plot and confidence scores are given in fig. 14 and table III, respectively. Notice that while a MOS score of 0 is high quality on the LIVE database a MOS score of 0 represents low quality on the IVC-SelectEncrypt database.

When it comes to confidence we can safely state that none of the metrics show overall good performance, i.e. none of the metrics are stable with a low $\mu(\mathcal{C}_D)$ and $\sigma(\mathcal{C}_D)$. However for certain test sets there are metrics with good performance, the CPA1 on the IVC-SelectEncrypt test set shows exceptionally high confidence and is stable. The CPA1 metric on the LIVE database however shows extremely poor performance.

Furthermore, $\mu(\mathcal{C}_D)$ and $\sigma(\mathcal{C}_D)$ alone cannot properly predict the performance of a metric unless it is stable. But $\mu(\mathcal{C}_D)$ is a good overall estimation and $\sigma(\mathcal{C}_D)$ together with signal shape is an indicator for the magnitude of the bias of the signal shape. As an example for this compare LEG and SSIM: While the LEG shows better overall $\mu(\mathcal{C}_D)$ and $\sigma(\mathcal{C}_D)$ the SSIM is biased much more than the LEG, i.e. the SSIM outperforms the LEG where its bias is, while the LEG outperforms the SSIM outside of the bias. This behaviour is clearly reflected when the $\sigma(\mathcal{C}_D)$ is considered in conjunction with the shape, i.e. even though the LEG is biased due to the small $\sigma(\mathcal{C}_D)$ we can deduct that the bias is far smaller than the bias of the SSIM which shows a high $\sigma(\mathcal{C}_D)$.

A similar conclusion can be drawn for the unstable shape. A metric with an unstable shape and a high $\sigma(\mathcal{C}_D)$ will have much more severe outliers than one with a low $\sigma(\mathcal{C}_D)$. This behaviour is nicely illustrated by the PSNR, although unstable, the magnitude of the outliers should be relatively small since the PSNR shows a low $\sigma(\mathcal{C}_D)$, which is exactly the behaviour shown in the plots of figure 13 and 14.

For a stable shape the $\mu(\mathcal{C}_D)$ becomes more important since it shows where the stable part of the confidence lies. The LE metric on the IVC-SelectEncrypt database is a prime example of this: While it is stable, the actual confidence score shows that it is stable in the sense that it exhibits exceedingly poor performance over the whole quality range.

Regarding confidence values and shape it can be seen from tables II and III that metrics perform differently on different test sets. If this is the case the worse value should be taken into account when it comes to overall performance. What is also noticeable from the two tables is the fact that the evaluated image and security metrics are more often biased towards the high quality range. Indeed on the IVC-SelectEncrypt database, which is the actual encryption database, not a single metric is biased towards the low quality range. Furthermore, the metrics biased towards the low quality metric range on the LIVE database, i.e. LEG, VIF and LFBVS, are all biased towards high quality on the IVC-SelectEncrypt database, and should thus be considered unstable overall.

To sum up the findings regarding the confidence of the metrics we can state the following: First, the LE, LSS and CPA1 show extremely poor performance overall. While the CPA1 performs exceptionally well on the IVC-SelectEncrypt database it performs very poorly on the LIVE database, so overall the performance of the CPA1 is not good.

Second, the LEG, VIF, ESS, LFBVS and PSNR, while not stable, exhibit a low $\sigma(\mathcal{C}_D)$ and are thus closest for being considered good metrics on the whole quality range. However, in each case the $\mu(\mathcal{C}_D)$ is overall relatively high, at least from a security standpoint, and thus could use some improvement, or replacement.

Third, the SSIM and NSD show a strong bias towards the high quality range. While the confidence for these metrics is overall not good the confidence on the high quality range is actually quite good. Consequently, when the application scenario is known to target the higher quality range, e.g. transparent encryption, these metrics should be considered.

C. Monotonicity for Low Quality Images

The monotonicity of an image metric over the MOS for the full quality range is what defines the quality of an image metric. However, for transparent encryption and evaluation of image distortion due to encryption it is important that a given metric has good monotonicity properties on the lower quality range rather than the overall quality range. It is well known, c.f. Hofbauer et al. [12], that this can be a problem so it is necessary to study this in more detail.

The absolute SROC values for the test sets of the LIVE database are given in table IV(a) for the full quality range, the low quality range and the high quality range. In the table, high, i.e. SROC > 0.9, (bold) and unsatisfactory, i.e. SROC < 0.5, (underlined) SROC scores are marked for each test set. For the IVC-SelectEncrypt database the same information is given in table IV(b).

In order to better compare the difference in high, low and overall rank order correlation for a given test set and database a graphical representation of the relation is given in table IV for the LIVE and IVC-SelectEncrypt databases. For each combination of test set and metric the graphical entry displays the range of possible SROC as a dark gray background, SROC$= 0$ at the bottom and SROC$= 1$ at the top. The light gray background bar shows the SROC value for the full quality range while the smaller bars give the SROC for high quality, green bar on the left, and low quality range, orange bar on the right.

What can be directly seen is that the overall performance of a metric does not imply a good performance for either

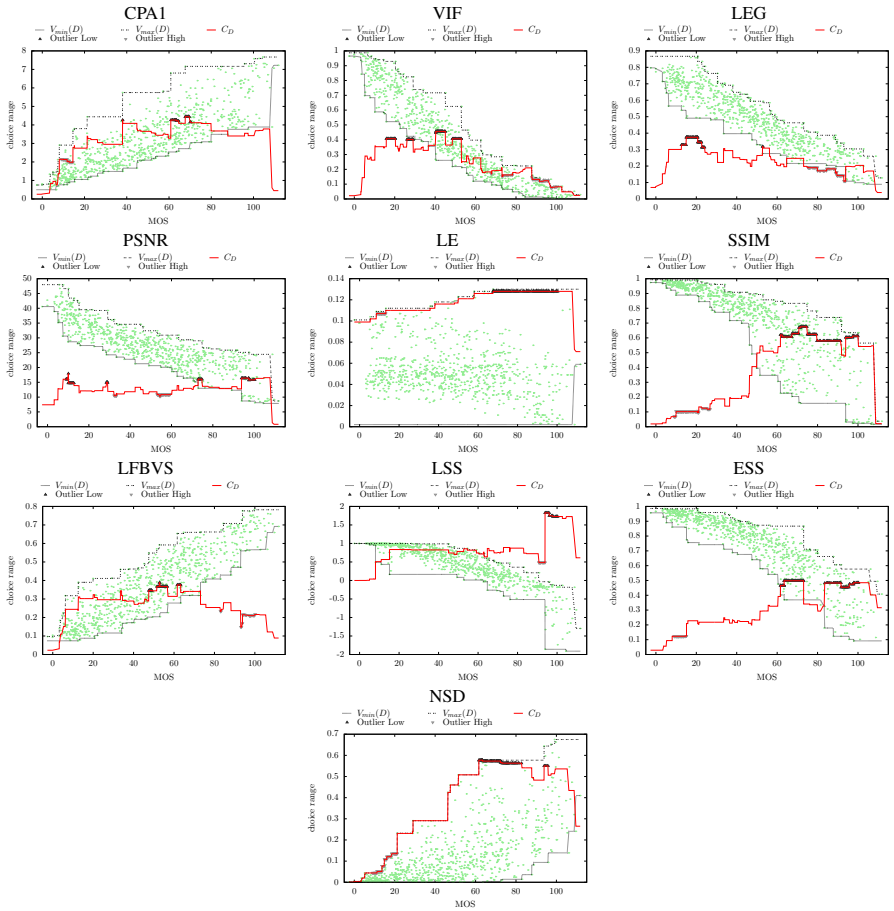


Fig. 13. Confidence Plots for LIVE database for the different image metrics. The plot shows the scatter plot for the MOS and metric score pairs, the plot of $V_{min}(D)$, $V_{max}(D)$ and $\mathcal{C}_D$.

TABLE II
AVERAGE AND STANDARD DEVIATION OF NORMALIZED CONFIDENCE AND SIGNAL SHAPE ON THE LIVE DATABASE.

	SSIM	LEG	VIF	CPA1	LSS	ESS	LFBVS	LE	NSD	PSNR
$\mu(\mathcal{C}_D)$	0.357	0.291	0.285	0.431	0.415	0.300	0.370	0.906	0.537	0.265
$\sigma(\mathcal{C}_D)$	0.225	0.070	0.110	0.109	0.159	0.133	0.071	0.069	0.277	0.038
Signal Shape	Bias High	Bias Low	Bias Low	Bias High	Bias High	Bias High	Bias Low	Bias High	Bias High	Unstable

TABLE III
AVERAGE AND STANDARD DEVIATION OF NORMALIZED CONFIDENCE AND SIGNAL SHAPE ON THE IVC-SELECTENCRYPT DATABASE.

	SSIM	LEG	VIF	CPA1	LSS	ESS	LFBVS	LE	NSD	PSNR
$\mu(\mathcal{C}_D)$	0.319	0.268	0.277	0.119	0.374	0.168	0.273	0.540	0.394	0.196
$\sigma(\mathcal{C}_D)$	0.226	0.077	0.098	0.066	0.173	0.090	0.139	0.107	0.274	0.063
Signal Shape	Bias High	Bias High	Bias High	Stable	Bias High	Bias High	Bias High	Stable	Bias High	Unstable

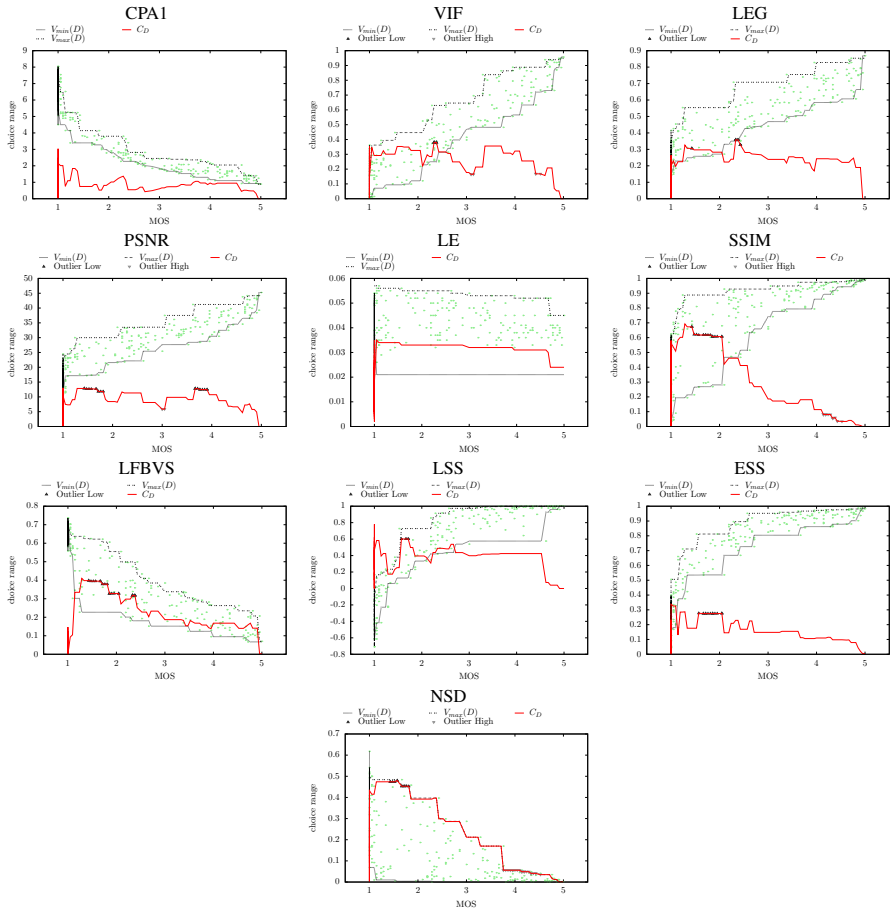


Fig. 14. Confidence Plots for IVC-SelectEncrypt database for the different image metrics. The plot shows the scatter plot for the MOS and metric score pairs, the plot of $V_{min}(D)$, $V_{max}(D)$ and C_D.

the high or low quality range, although most metrics tend to perform better on the high quality range. In some cases a metric can even perform better on a limited quality range than for the whole range of quality, e.g. the VIF performs better for the low quality end of resolution encryption on the IVC-SelectEncrypt database than for the whole range of resolution. In other cases the performance of full range case is drastically reduced for a reduced quality range, e.g. the VIF performs poorly on the high quality range for resolution encryption on the IVC-SelectEncrypt database. For other test sets the impact of either low or high quality is slight and most likely due to the reduced number of samples from the database, e.g. the VIF performs well for the white noise distortion from the LIVE database no matter which quality range. This in essence shows that the actual performance of an image metric is dependent on the distortion type as well as the quality range. For most test set and image metric combinations the SROC over the full quality range can at most be used as an upper limit for the limited quality cases, although there are exceptions, e.g. VIF for the low quality range of the resolution test set performs better than on the overall quality range for the same test set.

With regard to security metrics and their performance on the low quality range there is no noticeable difference in behavior to regular image metrics. Rather for each test set the best

TABLE IV
SPEARMAN RANK ORDER CORRELATION (SROC) FOR FULL, HIGH AND LOW QUALITY RANGE

Full Quality Range

	fastfading	gblur	jp2k	jpeg	wn
SSIM	**0.942**	**0.903**	**0.936**	**0.946**	**0.962**
LEG	**0.971**	**0.966**	**0.945**	**0.960**	**0.960**
VIF	**0.965**	**0.972**	**0.968**	**0.984**	**0.985**
CPA1	0.881	**0.927**	**0.958**	**0.962**	**0.984**
LSS	0.843	**0.916**	**0.953**	**0.970**	**0.965**
ESS	**0.933**	0.888	**0.929**	**0.949**	0.886
PSNR	0.891	0.782	0.895	0.881	**0.985**
LFBVS	**0.932**	**0.937**	0.853	**0.920**	0.891
LE	0.241	0.236	0.076	0.766	0.228
NSD	0.815	0.803	0.741	0.806	0.657

Full Quality Range

	iwind ec	iwind nec	resolution	trad	truncation
SSIM	**0.925**	**0.954**	0.887	**0.968**	0.879
LEG	0.869	**0.956**	0.876	**0.966**	0.863
VIF	**0.937**	**0.969**	0.767	**0.982**	**0.954**
CPA1	**0.925**	**0.953**	**0.906**	**0.967**	**0.959**
LSS	0.888	**0.907**	**0.955**	**0.948**	**0.948**
ESS	0.868	0.894	**0.933**	**0.906**	0.860
PSNR	**0.909**	**0.910**	**0.916**	**0.965**	0.889
LFBVS	0.756	**0.930**	**0.916**	**0.943**	0.878
LE	0.136	0.579	0.562	0.550	0.201
NSD	0.698	**0.922**	0.790	0.757	0.523

High Quality Range

	fastfading	gblur	jp2k	jpeg	wn
SSIM	0.517	0.815	0.710	0.879	0.875
LEG	0.652	0.792	0.638	0.761	0.734
VIF	0.632	0.782	0.797	**0.920**	0.891
CPA1	0.681	0.689	0.761	0.889	0.840
LSS	0.504	0.762	0.754	0.825	0.700
ESS	0.473	0.618	0.694	0.744	0.827
PSNR	0.517	0.662	0.676	0.801	0.869
LFBVS	0.289	0.707	0.393	0.651	0.572
LE	0.295	0.127	0.228	0.136	0.143
NSD	0.028	0.760	0.275	0.370	0.674

High Quality Range

	iwind ec	iwind nec	resolution	trad	truncation
SSIM	0.652	0.863	0.714	**0.975**	0.835
LEG	0.713	0.852	0.643	**0.920**	0.610
VIF	0.729	**0.907**	0.143	0.885	0.662
CPA1	0.683	0.890	0.000	0.868	**0.934**
LSS	0.705	**0.945**	0.393	0.865	**0.975**
ESS	0.809	0.846	0.571	0.679	0.794
PSNR	0.636	0.857	0.500	0.879	0.830
LFBVS	0.521	0.863	0.250	0.698	0.723
LE	0.095	0.489	0.464	0.431	0.019
NSD	0.251	0.736	0.179	0.236	0.602

Low Quality Range

	fastfading	gblur	jp2k	jpeg	wn
SSIM	0.662	0.487	0.635	0.339	0.802
LEG	0.893	0.872	0.617	0.699	0.804
VIF	**0.937**	**0.920**	0.646	0.829	**0.911**
CPA1	0.614	0.669	0.657	0.402	0.897
LSS	0.788	0.732	0.647	0.595	0.817
ESS	0.877	0.847	0.397	0.735	0.570
PSNR	0.611	0.408	0.510	0.046	0.897
LFBVS	0.863	0.849	0.391	0.702	0.659
LE	0.268	0.338	0.070	0.720	0.290
NSD	0.731	0.582	0.427	0.668	0.231

(a) SROC for the LIVE Image Quality Assessment Database

Low Quality Range

	iwind ec	iwind nec	resolution	trad	truncation
SSIM	0.191	0.694	0.644	0.680	0.695
LEG	0.141	0.823	0.490	0.652	0.181
VIF	0.518	0.732	0.823	**0.913**	0.832
CPA1	0.400	0.628	0.897	0.592	0.706
LSS	0.523	0.240	0.875	0.386	0.679
ESS	0.422	0.411	0.551	0.609	0.549
PSNR	0.251	0.474	0.688	0.663	0.541
LFBVS	0.188	0.562	0.507	0.416	0.022
LE	0.386	0.004	0.152	0.166	0.272
NSD	0.100	0.655	0.290	0.589	0.291

(b) SROC for the IVC-SelectEncrypt Image Quality Assessment Database

performance over the low quality range can be found among the traditional image metrics.

Comparing the high and low quality ranges we can see that the low quality range has a far higher number of unsatisfactory SROC scores. This shows that overall image metrics tend to perform better at differentiating the different strength in distortion for high quality images. A high SROC over the whole quality range indicates that the metric can differentiate between high and low quality images. In essence, image metrics which perform well on the overall quality range can still be utilized to identify sufficient encryption, even though a metric which exhibits good performance in the range of the quality threshold should be preferred. For transparent encryption, where the goal is to give the best image below a certain threshold, the monotonicity in the chosen range of quality becomes more important. In this case the target quality is a lot closer to the threshold so a high monotonicity, expressed by a high SROC, is required.

Another interesting aspect of the high versus low quality test is the fact that not a single metric among those tested has overall better performance on either the high or low quality range. Consequently, the metrics cannot be reduced to an overall SROC score and a bias towards either high or low. There are cases where the performance of both high and low quality is far lower than the overall quality, e.g., LFBVS on the jp2k test set, and no metric shows an overall preference for high or low quality range, e.g. for the SSIM the low quality range of the fastfading test set performs better while for the gblur test set the high quality range performs better. Thus, in order to evaluate whether an image metric is fit to be used as a security metric, tests regarding low and high quality performance have to be conducted.

Summing up the monotonicity tests we can state the following: Regarding the overall quality the VIF, CPA1, SSIM, LSS and LEG perform best, LFBVS, PSNR and ESS also show a good behaviour while NSD and especially LE perform poorly.

For the high quality range most quality metrics still show a decent performance, however, only SSIM, LEG, and PSNR exhibit no unsatisfactory performance in a single test set.

On the low quality side only the VIF exhibits good performance over all test sets. All other metrics have at least two test sets where their performance is unsatisfactory.

V. Conclusion and Future Work

We have outlined an evaluation method for security metrics based on practical application scenarios and considerations. These methods were used to evaluate state of the art security metrics, image metrics which are used as security metrics as well as state of the art image metrics. A summary of the evaluation and a basic evaluation score is given in table VI. From the summary it is clear that none of the security metrics are fit to perform as general purpose security metrics.

Regarding transparent and sufficient encryption, the LE and NSD metrics especially show that metrics engineered to fit a certain application scenario cannot claim generality. Furthermore, the SSIM and PSNR which were used as security metrics in literature also perform poorly. Most state of the art image metrics hardly perform the security metric task adequately, only the VIF, apart from a borderline confidence score and stability, demonstrates good performance.

Regarding content confidentiality we cannot make a strong statement, due to lack of ground truth for recognizability tests. However, the performance in the encrypted domain during the evaluation of the application domain gives a strong indication that no image metric, among those tested, can perform the task of evaluating the content confidentiality.

The inability to properly evaluate image metrics in regard to content confidentiality naturally leads to the conclusion that more data is required. This also holds true for the lower quality ranges in regard to regular metrics which would undoubtedly benefit from a dataset specially designed for high impairment cases. In future work we will gather ground truth data for content confidentiality (and will also design protocols how to properly capture human assessment for these data sets) and extend the work presented in this paper to properly encompass content confidentiality. Similarly, we will gather more data on the low quality range to better evaluate security metrics for transparent and sufficient encryption as well as aid in their development.

References

[1] S.-K. Au Yeung, S. Zhu, and B. Zeng, "Quality assessment for a perceptual video encryption system," in *Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE Int. Conf. on*, Jun. 2010, pp. 102–106.

[2] M. Khan, V. Jeoti, and A. Malik, "Perceptual encryption of JPEG compressed images using DCT coefficients and splitting of DC coefficients into bitplanes," in *2010 Int. Conf. on Intelligent and Advanced Systems (ICIAS)*, Jun. 2010, pp. 1–6.

[3] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, pp. 656–715, Oct. 1949.

[4] F. Dufaux and T. Ebrahimi, "Scrambling for privacy protection in video surveillance systems," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1168–1174, 2008.

[5] T. D. Lookabaugh and D. C. Sicker, "Selective encryption for consumer applications," *IEEE Communications Magazine*, vol. 42, no. 5, pp. 124–129, 2004.

[6] A. Said, "Measuring the strength of partial encryption schemes," in *Proc. of the IEEE Int. Conf. on Image Processing (ICIP'05)*, vol. 2, Sep. 2005, pp. 1126–1129.

[7] M. Bellare, T. Ristenpart, P. Rogaway, and T. Stegers, "Format-preserving encryption," in *Proc. of Selected Areas in Cryptography, SAC '09*, vol. 5867. Calgary, Canada: Springer-Verlag, Aug. 2009, pp. 295–312.

[8] T. Stütz and A. Uhl, "A survey of H.264 AVC/SVC encryption," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 3, pp. 325–339, 2012.

[9] ——, "Efficient format-compliant encryption of regular languages: Block-based cycle-walking," in *Proc. of the 11th Joint IFIP TC6 and TC11 Conf. on Communications and Multimedia Security, CMS '10*, vol. 6109, May 2010, pp. 81–92.

[10] Q. Li and I. J. Cox, "Using perceptual models to improve fidelity and provide resistance to valumetric scaling for quantization index modulation watermarking," *IEEE Trans. on Information Forensics and Security*, vol. 2, no. 2, pp. 127–139, Jun. 2007.

[11] H. Hofbauer and A. Uhl, "Selective encryption of the MC-EZBC bitstream and residual information," in *18th European Signal Processing Conf., 2010 (EUSIPCO-2010)*, Aalborg, Denmark, Aug. 2010, pp. 2101–2105.

[12] ——, "Visual quality indices and low quality images," in *IEEE 2nd European Workshop on Visual Information Processing*, Paris, France, Jul. 2010, pp. 171–176.

[13] E. J. Wharton, K. Panetta, and S. Agaian, "Simultaneous encryption/compression of images using alpha rooting," in *2008 Data Compression Conf. (DCC 2008)*, 2008, p. 551ff.

[14] Y. Zhou, K. Panetta, and S. Aagaian, "Partial multimedia encryption with different security levels," in *IEEE Conf. on Technologies for Homeland Security 2008*, May 2008.

[15] S. Lian, "Efficient image or video encryption based on spatiotemporal chaos system," *Chaos, Solitons & Fractals*, vol. 40, no. 5, pp. 2509 – 2519, 2009.

[16] L. Dubois, W. Puech, and J. Blanc-Talon, "Smart selective encryption of cavlc for h.264/avc video," in *Information Forensics and Security (WIFS), 2011 IEEE Int. Workshop on*, dec 2011, pp. 1 –6.

[17] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, Jun. 2008.

[18] F. Autrusseau, T. Stuetz, and V. Pankajakshan, "Subjective quality assessment of selective encryption techniques," 2010, http://www.irccyn.ec-nantes.fr/~autrusse/Databases/.

[19] T. Stütz, V. Pankajakshan, F. Autrusseau, A. Uhl, and H. Hofbauer, "Subjective and objective quality assessment of transparently encrypted JPEG2000 images," in *Proc. of the ACM Multimedia and Security Workshop (MMSEC '10)*. Rome, Italy: ACM, Sep. 2010, pp. 247–252.

[20] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[21] Y. Mao and M. Wu, "Security evaluation for communication-friendly encryption of multimedia," in *Proc. of the IEEE Int. Conf. on Image Processing (ICIP'04)*. Singapore: IEEE Signal Processing Society, Oct. 2004.

[22] H. Hofbauer and A. Uhl, "Selective encryption of the MC EZBC bitstream for DRM scenarios," in *Proc. of the 11th ACM Workshop on Multimedia and Security*, Princeton, New Jersey, USA, Sep. 2009, pp. 161–170.

[23] H. Hofbauer, T. Stütz, and A. Uhl, "Selective encryption for hierarchical MPEG," in *Communications and Multimedia Security, Proc. of the 10th IFIP Int. CMS 2006 Conf.*, vol. 4237, Oct. 2006, pp. 151–160.

[24] Y. Yao, Z. Xu, and J. Sun, "Visual security assessment for cipher-images based on neighborhood similarity," *Informatica*, vol. 33, pp. 69–76, 2009.

[25] J. Sun, Z. Xu, J. Liu, and Y. Yao, "An objective visual security assessment for cipher-images based on local entropy," *Multimedia Tools and Applications*, Mar. 2010.

[26] L. Tong, F. Dai, Y. Zhang, and J. Li, "Visual security evaluation for video encryption," in *Proc. of the Int. Conf. on Multimedia*, New York, NY, USA, 2010, pp. 835–838.

[27] H. Hofbauer and A. Uhl, "An effective and efficient visual quality index based on local edge gradients," in *IEEE 3rd European Workshop on Visual Information Processing*, Paris, France, Jul. 2011, p. 6.

[28] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing*, vol. 15, no. 2, pp. 430–444, May 2006.

[29] M. Carosi, V. Pankajakshan, and F. Autrusseau, "Towards a simplified perceptual quality metric for watermarking applications," in *Proc. of SPIE, Multimedia on Mobile Devices*, vol. 7542, San Jose, CA, USA, Jan. 2010.

[30] A. Uhl and A. Pommer, *Image and Video Encryption. From Digital Rights Management to Secured Personal Communication*. Springer-Verlag, 2005, vol. 15.

[31] T. Stütz and A. Uhl, "On JPEG2000 error concealment attacks," in *Advantages in Image and Video Technology: Proc. of the 3rd Pacific-Rim Symposium on Image and Video Technology, PSIVT '09*, Tokyo, Japan, Jan. 2009, pp. 851–861.

[32] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 100, no. 3/4, pp. 441–471, 1904.
[33] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun. 1938.
[34] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," http://live.ece.utexas.edu/research/quality.
[35] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[36] N. Ponomarenko, F. Battisti, K. Egizarian, J. Astola, and V. Lukin, "Color image database for evaluation of image quality metrics," in *Proc. of Int. Workshop on Multimedia Signal Processing*, Australia, Oct. 2008, pp. 403–408.

TABLE V
VISUAL REPRESENTATION OF THE SPEARMAN RANK ORDER CORRELATION (SROC) FOR THE LIVE IMAGE QUALITY ASSESSMENT AND IVC-SELECTENCRYPT DATABASES FOR FULL QUALITY RANGE (LIGHT GRAY), LOW QUALITY RANGE (ORANGE BAR ON THE RIGHT) AND HIGH QUALITY RANGE (GREEN BAR ON THE LEFT SIDE).

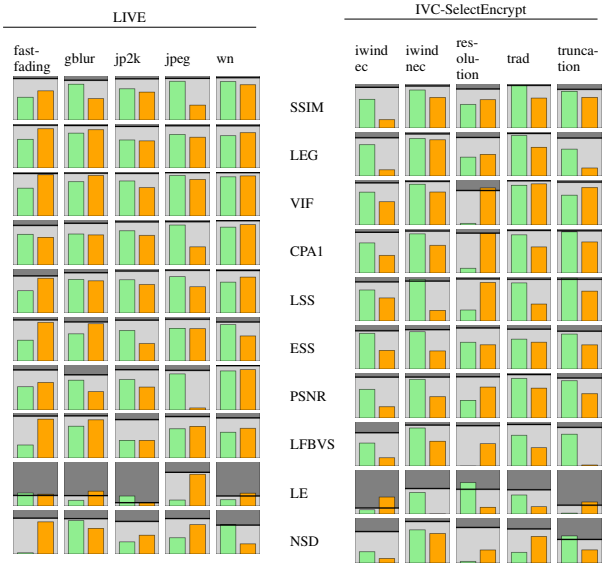


TABLE VI
SUMMARY OF EVALUATION

	SSIM	LEG	VIF	CPA1	LSS	ESS	PSNR	LFBVS	LE	NSD
Application Domain										
Encryption	*45.03 %*	33.35 %	*45.38 %*	36.26 %	*41.73 %*	37.25 %	*51.17 %*	*51.99 %*	73.08 %	*45.99 %*
Extraction	**99.93 %**	**99.43 %**	**98.79 %**	**4.19 %**	69.18 %	**95.81 %**	68.11 %	*48.86 %*	*50.00 %*	*43.79 %*
Confidence on the LIVE database										
$\mu(\mathcal{C}_D)$	0.357	**0.291**	**0.285**	0.431	0.415	0.300	**0.265**	0.370	*0.906*	*0.537*
$\sigma(\mathcal{C}_D)$	0.225	**0.070**	0.110	0.109	0.159	0.133	**0.038**	**0.071**	**0.069**	*0.277*
Signal Shape	Bias High	Bias Low	Bias Low	Bias High	Bias High	Bias High	Bias Low	Bias High	Bias High	*Unstable*
Confidence on the IVC-SelectEncrypt database										
$\mu(\mathcal{C}_D)$	0.319	**0.268**	**0.277**	**0.119**	0.374	**0.168**	**0.196**	**0.273**	*0.540*	0.394
$\sigma(\mathcal{C}_D)$	0.226	**0.077**	**0.098**	**0.066**	0.173	**0.090**	**0.063**	0.139	0.107	*0.274*
Signal Shape	Bias High	Bias High	Bias High	Stable	Bias High	Bias High	Bias High	Stable	Bias High	*Unstable*
Low Quality SROC on the LIVE database										
fastfading	0.662	0.893	**0.937**	0.614	0.788	0.877	0.611	0.863	*0.268*	0.731
gblur	*0.487*	0.872	**0.920**	0.669	0.732	0.847	*0.408*	0.849	*0.338*	0.582
jp2k	0.635	0.617	0.646	0.657	0.647	*0.397*	0.510	*0.391*	*0.070*	*0.427*
jpeg	*0.339*	0.699	0.829	*0.402*	0.595	0.735	*0.046*	0.702	0.720	0.668
wn	0.802	0.804	**0.911**	0.897	0.817	0.570	0.897	0.659	*0.290*	*0.231*
Low Quality SROC on the IVC-SelectEncrypt database										
iwind ec	*0.191*	*0.141*	0.518	*0.400*	0.523	*0.422*	*0.251*	*0.188*	*0.386*	*0.100*
iwind nec	0.694	0.823	0.732	0.628	*0.240*	*0.411*	*0.474*	0.562	*0.004*	0.655
resolution	0.644	*0.490*	0.823	0.897	0.875	0.551	0.688	0.507	*0.152*	*0.290*
trad	0.680	0.652	**0.913**	0.592	*0.386*	0.609	0.663	*0.416*	*0.166*	0.589
truncation	0.695	*0.181*	0.832	0.706	0.679	0.549	0.541	*0.022*	*0.272*	*0.291*
Comparison Score, -1 for insufficient performance, +1 for good performance, per above test, -1 for conflict in signal shape										
	SSIM	LEG	VIF	CPA1	LSS	ESS	PSNR	LFBVS	LE	NSD
Score	*-3*	1	**6**	0	*-3*	0	*-2*	*-5*	*-11*	*-12*

4. Conclusion

We have identified wavelet based codecs as a practical solution to the universal multimedia principle. We chose the MC-EZBC codec since it is mature and exhibits performance comparable to other state of the art codecs, most notably the standardized H.264/SVC. Two main challenges have been identified in order to utilize this codec in an application: Transporting a bitstream over hard- and software designed for the standardized H.264 codec; Encryption of the MC-EZBC with the goal of facilitating secure end-to-end connections as well as allowing for adaptation in the network.

We have solved the task of transporting the MC-EZBC based bitstream over hardware and software designed for the standardized H.264 codec. These methods use bitstream description protocols and an embedding of an MC-EZBC in a faux H.264 bitstream. Since the adaptation can be done on the fly on the server, the resulting computational load is low and allows the MC-EZBC to be utilized for UMA. This encapsulation into H.264 allows not only to use current hardware for transport but also allows the utilization of MANEs for JIT scaling and adaptation. This essentially allows an MC-EZBC bitstream to utilize the full range of options that a H.264 bitstream has on modern hardware. Nota bene: We specifically designed this encapsulation procedure for the MC-EZBC but the approach can easily be adapted to other wavelet based bitstreams.

In order to provide security to the streaming process we designed an encryption strategy for the MC-EZBC. This method allows for a secure end to end connection and is format compliant, i.e., scaling is possible on the encrypted bitstream without knowledge of the key or the need to decrypt the bitstream. Since the encrypted bitstream is format compliant we could eliminate possible points of attack from the network since the keys are only required at the endpoints of the secure channel. Furthermore, computational resources are reduced if in-network scaling is performed since scaling is done on the encrypted bitstream which allows to skip decryption and encryption on the MANE. This in turn leads to shorter frame delays and overall faster delivery of the bitstream.

To assess the security of the encryption method we did an in-depth analysis of the encryption method and identified that a confidential encryption is never possible with this method. Actually we showed that if format compliance with regard to scaling is required no encryption method can achieve confidential encryption, since information needed for scaling also allows identification of the content.

Since we utilized image metrics as security metrics, as is standard in literature, we found certain shortcomings of this evaluation method. Most notably the ability of image metrics to deal with low quality content is severely limited, but this is exactly the quality range which is of importance during security evaluations. Furthermore, we noted that there is no existing testing methodology for security metrics.

Consequently, we introduced a testing methodology for security metrics based on common application scenarios for said metrics. However, we could only concisely state a methodology for testing security metrics for transparent and sufficient encryption since we lack the ground truth to tackle content confidentiality. Furthermore, we tested image metrics, which are commonly used for security evaluation, as well as security metrics proposed in literature. The conclusion of this test unfortunately showed that almost none of the tested metrics are fit to perform security analysis. Only the LEG and VIF scored positive on the conducted tests, and

only the score of the VIF was not borderline.

During our work with image metrics and security evaluations we also found the lack of an image metric which is fast to compute and highly correlated with human observer scores. In order to rectify this we introduced the LEG image metric, based on local edge gradients. The LEG shows a high correlation with human judgement and is fast to compute, showing better performance in both fields than the SSIM which previously filled the role as a fast and robust image metric. Furthermore, during evaluation as security metric the LEG showed the second best performance, after the VIF.

Bibliography

[1] APOSTOLOPOULOS, J. Architectural principles for secure streaming & secure adaptation in the developing scalable video coding (SVC) standard. In *Proceedings of the IEEE International Conference on Image Processing, ICIP '06* (Oct. 2006), pp. 729–732.

[2] BAUGHER, M., MCGREW, D., NASLUND, M., CARRARA, E., AND NORRMAN, K. The Secure Real-time Transport Protocol (SRTP). RFC 3711 (Proposed Standard), Mar. 2004. `http://www.ietf.org/rfc/rfc3711.txt`.

[3] BELLARE, M., RISTENPART, T., ROGAWAY, P., AND STEGERS, T. Format-preserving encryption. In *Proceedings of Selected Areas in Cryptography, SAC '09* (Calgary, Canada, Aug. 2009), vol. 5867, Springer-Verlag, pp. 295–312.

[4] CAROSI, M., PANKAJAKSHAN, V., AND AUTRUSSEAU, F. Towards a simplified perceptual quality metric for watermarking applications. In *Proceedings of SPIE, Multimedia on Mobile Devices* (San Jose, CA, USA, Jan. 2010), vol. 7542, SPIE.

[5] CHEN, P., HANKE, K., RUSERT, T., AND WOODS, J. W. Improvements to the MC-EZBC scalable video coder. In *Proceedings of the IEEE Int. Conf. Image Processing ICIP* (Barcelona, Spain, 2003), vol. 2, pp. 81–84.

[6] CHEN, P., AND WOODS, J. W. Bidirectional MC-EZBC with lifting implementation. *IEEE Transactions on Circ. and Systems for Video Technology 14*, 10 (2004), 1183–1194.

[7] D. ROUSE, S. S. H. Natural image utility assessment using image contours. In *IEEE International Conference on Image Processing (ICIP'09)* (Cairo, Egypt, Nov. 2009), pp. 2217–2220.

[8] DAUGMAN, J. How iris recognition works. *IEEE Transactions on Circiuts and Systems for Video Technology 14*, 1 (2004), 21–30.

[9] DAUGMAN, J., AND DOWNING, C. Effect of severe image compression on iris recognition performance. *IEEE Transactions on Information Forensics and Security 3*, 1 (2008), 52–61.

[10] DUBOIS, L., PUECH, W., AND BLANC-TALON, J. Smart selective encryption of cavlc for h.264/avc video. In *Information Forensics and Security (WIFS), 2011 IEEE International Workshop on* (dec 2011), pp. 1 –6.

[11] HELLWAGNER, H., HOFBAUER, H., KUSCHNIG, R., STÜTZ, T., AND UHL, A. Secure transport and adaptation of MC-EZBC video utilizing H.264-based transport protocols. *Elsevier Journal on Signal Processing: Image Communication 27*, 2 (2011), 192–207.

[12] HOFBAUER, H., RATHGEB, C., UHL, A., AND WILD, P. Image metric-based biometric comparators: A supplement to feature vector-based hamming distance? In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'12)* (Darmstadt, Germany, Sept. 2012), pp. 1 – 5.

[13] HOFBAUER, H., RATHGEB, C., UHL, A., AND WILD, P. Iris recognition in image domain: Quality-metric based comparators. In *Proceedings of the 8th International Symposium on Visual Computing (ISVC'12)* (Crete, Greece, July 2012), pp. 1 – 10.

[14] HOFBAUER, H., STÜTZ, T., AND UHL, A. Secure Scalable Video Compression for GVid. In *Proceedings of the 3rd Austrian Grid Symposium* (Linz, Austria, 2009), J. Volkert, T. Fahringer, D. Kranzlmüller, R. Kobler, and W. Schreiner, Eds., vol. 269 of *books@ocg.at*, Austrian Computer Society, pp. 88–102.

[15] HOFBAUER, H., AND UHL, A. The cost of in-network adaption of the MC-EZBC for universal multimedia access. In *Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA '09)* (Salzburg, Austria, Sept. 2009).

[16] HOFBAUER, H., AND UHL, A. Selective encryption of the MC EZBC bitstream for DRM scenarios. In *Proceedings of the 11th ACM Workshop on Multimedia and Security* (Princeton, New Jersey, USA, Sept. 2009), ACM, pp. 161–170.

[17] HOFBAUER, H., AND UHL, A. Selective encryption of the MC-EZBC bitstream and residual information. In *18th European Signal Processing Conference, 2010 (EUSIPCO-2010)* (Aalborg, Denmark, Aug. 2010), pp. 2101–2105.

[18] HOFBAUER, H., AND UHL, A. Visual quality indices and low quality images. In *IEEE 2nd European Workshop on Visual Information Processing* (Paris, France, July 2010), pp. 171–176.

[19] HOFBAUER, H., AND UHL, A. An effective and efficient visual quality index based on local edge gradients. In *IEEE 3rd European Workshop on Visual Information Processing* (Paris, France, July 2011), p. 6pp.

[20] HOFBAUER, H., AND UHL, A. An evaluation of visual security metrics. *IEEE Transactions on Multimedia* (2013), 15 pages. submitted.

[21] HORITA, Y., SHIBATA, K., AND KAWAYOKE, Y. MICT image quality evaluation database. `http://mict.eng.u-toyama.ac.jp/mictdb.html`.

[22] HSIANG, S.-T., AND WOODS, J. W. Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank. *Signal Processing: Image Communication 16*, 8 (May 2001), 705–724.

[23] HUYNH-THU, Q., AND GHANBARI, M. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters 44*, 13 (June 2008), 800–801.

[24] ISO/IEC 14496-10. Information technology – coding of audio-visual objects – part 10: Advanced video coding, 2005.

[25] ISO/IEC 21000-7:2007. Information technology – Multimedia framework (MPEG-21) – Part 7: Digital Item Adaptation, Nov. 2007.

[26] ITU-R BT.500-11. Methodology for the subjective assesment of the quality of television pictures, 2002. `http://www.itu.int/rec/R-REC-BT.500/en`.

[27] ITU-T H.264. Advanced video coding for generic audiovisual services, Nov. 2007. `http://www.itu.int/rec/T-REC-H.264-200711-I/en`.

[28] ITU-T REC H.264. SERIES H: AUDIOVISUAL AND MULTIMEDIA SYSTEMS infrastructure of audiovisual services coding of moving video, Jan. 2012. `http://www.itu.int/rec/T-REC-H.264-201201-I`.

[29] ITU-T REC P.910. SERIES P: TELEPHONE TRANSMISSION QUALITY audiovisual quality in multimedia services, 1996. `http://www.itu.int/rec/T-REC-P.910/en`.

[30] KUSCHNIG, R., KOFLER, I., AND HELLWAGNER, H. An Evaluation of TCP-based Rate-Control Algorithms for Adaptive Internet Streaming of H.264/SVC. In *Proceedings of ACM Multimedia Systems (ACM MMSYS 2010)* (Feb. 2010).

[31] KUSCHNIG, R., KOFLER, I., RANSBURG, M., AND HELLWAGNER, H. Design options and comparison of in-network H.264/SVC adaptation. *Journal of Visual Communication and Image Representation 19*, 8 (Sept. 2008), 529–542.

[32] LE CALLET, P., AND AUTRUSSEAU, F. Subjective quality assessment IRCCyN/IVC database, 2005. `http://www.irccyn.ec-nantes.fr/ivcdb/`.

[33] LI, Q., AND COX, I. J. Using perceptual models to improve fidelity and provide resistance to valumetric scaling for quantization index modulation watermarking. *IEEE Transactions on Information Forensics and Security 2*, 2 (June 2007), 127–139.

[34] LIAN, S. Efficient image or video encryption based on spatiotemporal chaos system. *Chaos, Solitons & Fractals 40*, 5 (2009), 2509 – 2519.

[35] LIMA, L., MANERBA, F., ADAMI, N., SIGNORONI, A., AND LEONARDI, R. Wavelet-based encoding for HD applications. In *2007 IEEE International Conference on Multimedia and Expo* (July 2007), pp. 1351–1354.

[36] LOOKABAUGH, T. D., AND SICKER, D. C. Selective encryption for consumer applications. *IEEE Communications Magazine 42*, 5 (2004), 124–129.

[37] PANIS, G., HUTTER, A., HEUER, J., HELLWAGNER, H., KOSCH, H., TIMMERER, C., DEVILLERS, S., AND AMIELH, M. Bitstream syntax description: a tool for multimedia resource adaptation within MPEG-21. In *Special Issue on Multimedia Adaptation* (Sept. 2003), vol. 18 of *Signal Processing: Image Communication*, pp. 721–747.

[38] PONOMARENKO, N., BATTISTI, F., EGIZARIAN, K., ASTOLA, J., AND LUKIN, V. Color image database for evaluation of image quality metrics. In *Proceedings of International Workshop on Multimedia Signal Processing* (Australia, Oct. 2008), pp. 403–408.

[39] PONOMARENKO, N., BATTISTI, F., EGIZARIAN, K., ASTOLA, J., AND LUKIN, V. Metrics performance comparison for color image database. In *Fourth international workshop on video processing and quality metrics for consumer electronics* (Arizona, USA, Jan. 2009), p. 6 p.

[40] PONOMARENKO, N., LUKIN, V., ZELENSKY, A., EGIAZARIAN, K., CARLI, M., AND BATTISTI, F. TID2008 - a database for evaluation of full-reference visual quality assessment metrics. In *Advances of Modern Radioelectronics* (2009), vol. 10, pp. 30–45.

[41] RATHGEB, C., UHL, A., AND WILD, P. *Iris Recognition: From Segmentation to Template Security*, vol. 59 of *Advances in Information Security*. Springer Verlag, 2013.

[42] ROSS, A., NANDAKUMAR, K., AND JAIN, A. *Handbook of Multibiometrics*. Springer, 2006.

[43] SAID, A. Measuring the strength of partial encryption schemes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'05)* (Sept. 2005), vol. 2, pp. 1126–1129.

[44] SCHULZRINNE, H., CASNER, S., FREDERICK, R., AND JACOBSON, V. RTP: A Transport Protocol for Real-Time Applications. RFC 3550, July 2003. `http://www.ietf.org/rfc/rfc3550.txt`.

[45] SCHULZRINNE, H., RAO, A., AND LANPHIER, R. Real Time Streaming Protocol (RTSP). RFC 2326, Apr. 1998. http://www.ietf.org/rfc/rfc2326.txt.

[46] SCHWARZ, H., MARPE, D., AND WIEGAND, T. Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology 17*, 9 (2007), 1103–1120.

[47] SHANNON, C. E. Communication theory of secrecy systems. *Bell System Technical Journal 28* (Oct. 1949), 656–715.

[48] SHEIKH, H. R., AND BOVIK, A. C. Image information and visual quality. *IEEE Transactions on Image Processing 15*, 2 (May 2006), 430–444.

[49] SHEIKH, H. R., WANG, Z., CORMACK, L., AND BOVIK, A. C. LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality.

[50] STÜTZ, T., AND UHL, A. Efficient format-compliant encryption of regular languages: Block-based cycle-walking. In *Proceedings of the 11th Joint IFIP TC6 and TC11 Conference on Communications and Multimedia Security, CMS '10* (Linz, Austria, May 2010), B. D. Decker and I. Schaumüller-Bichl, Eds., vol. 6109 of *IFIP Advances in Information and Communication Technology*, Springer, pp. 81–92.

[51] STÜTZ, T., AND UHL, A. A survey of H.264 AVC/SVC encryption. *IEEE Transactions on Circuits and Systems for Video Technology 22*, 3 (2012), 325–339.

[52] SUN, J., XU, Z., LIU, J., AND YAO, Y. An objective visual security assessment for cipher-images based on local entropy. *Multimedia Tools and Applications* (Mar. 2010). online publication.

[53] TIZON, N., AND PESQUET-POPESCU, B. Scalable and media aware adaptive video streaming over wireless networks. *EURASIP Journal on Advances in Signal Processing Volume 2008*, Article ID 218046 (2008), 11 pages.

[54] TONG, L., DAI, F., ZHANG, Y., AND LI, J. Visual security evaluation for video encryption. In *Proceedings of the international conference on Multimedia* (New York, NY, USA, 2010), MM '10, ACM, pp. 835–838.

[55] VETRO, A., CHRISTOPOULOS, C., AND EBRAHIMI, T. From the guest editors - Universal multimedia access. *IEEE Signal Processing Magazine 20*, 2 (2003), 16 – 16.

[56] WANG, Y., HANNUKSELA, M., PATEUX, S., ELEFTHERIADIS, A., AND WENGER, S. System and transport interface of SVC. *IEEE Transactions on Circuits and Systems for Video Technology 17*, 9 (Sept. 2007), 1149–1163.

[57] WANG, Z., BOVIK, A., SHEIKH, H., AND SIMONCELLI, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (Apr. 2004), 600–612.

[58] WENGER, S., HANNUKSELA, M., STOCKHAMMER, T., WESTERLUND, M., AND SINGER, D. RTP Payload Format for H.264 Video. RFC 3984, Feb. 2005. http://www.ietf.org/rfc/rfc3984.txt.

[59] WHARTON, E. J., PANETTA, K., AND AGAIAN, S. Simultaneous encryption/compression of images using alpha rooting. In *2008 Data Compression Conference (DCC 2008), 25-27 March 2008, Snowbird, UT, USA* (2008), IEEE Computer Society, p. 551ff.

[60] WU, Y., GOLWELKAR, A., AND WOODS, J. W. MC-EZBC video proposal from Rensselaer Polytechnic Institute. *ISO/IEC JTC1/SC29/WG11, MPEG2004/M10569/S15* (Mar. 2004).

[61] ZHOU, Y., PANETTA, K., AND AAGAIAN, S. Partial multimedia encryption with different security levels. In *IEEE Conference on Technologies for Homeland Security 2008* (May 2008).

A. Appendix

A.1. Breakdown of Authors' Contribution

This section lists a breakdown of authors' contribution with respect to the papers included in this thesis. Author names on the papers are listed in alphabetical order.

Andreas Uhl is the thesis advisor of Heinz Hofbauer, likewise Hermann Hellwagner is the thesis advisor of Robert Kuschnig. Since the explicit contribution of an advisor cannot be stated for a single paper, it is omitted in the following breakdown.

Publication	Contribution (in %)						
	Hermann Hellwagner	Heinz Hofbauer	Robert Kuschnig	Christian Rathgeb	Thomas Stütz	Andreas Uhl	Peter Wild
HOFBAUER, H., AND UHL, A. Selective encryption of the MC EZBC bitstream for DRM scenarios. In *Proceedings of the 11th ACM Workshop on Multimedia and Security* (Princeton, New Jersey, USA, Sept. 2009), ACM, pp. 161–170		100					
HOFBAUER, H., AND UHL, A. The cost of in-network adaption of the MC-EZBC for universal multimedia access. In *Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA '09)* (Salzburg, Austria, Sept. 2009)		100					
HOFBAUER, H., STÜTZ, T., AND UHL, A. Secure Scalable Video Compression for GVid. In *Proceedings of the 3rd Austrian Grid Symposium* (Linz, Austria, 2009), J. Volkert, T. Fahringer, D. Kranzlmüller, R. Kobler, and W. Schreiner, Eds., vol. 269 of *books@ocg.at*, Austrian Computer Society, pp. 88–102		50			50		

Publication	Contribution (in %)						
	Hermann Hellwagner	Heinz Hofbauer	Robert Kuschnig	Christian Rathgeb	Thomas Stütz	Andreas Uhl	Peter Wild
HOFBAUER, H., AND UHL, A. Visual quality indices and low quality images. In *IEEE 2nd European Workshop on Visual Information Processing* (Paris, France, July 2010), pp. 171–176		100					
HOFBAUER, H., AND UHL, A. Selective encryption of the MC-EZBC bitstream and residual information. In *18th European Signal Processing Conference, 2010 (EUSIPCO-2010)* (Aalborg, Denmark, Aug. 2010), pp. 2101–2105		100					
HOFBAUER, H., AND UHL, A. An effective and efficient visual quality index based on local edge gradients. In *IEEE 3rd European Workshop on Visual Information Processing* (Paris, France, July 2011), p. 6pp		100					
HELLWAGNER, H., HOFBAUER, H., KUSCHNIG, R., STÜTZ, T., AND UHL, A. Secure transport and adaptation of MC-EZBC video utilizing H.264-based transport protocols. *Elsevier Journal on Signal Processing: Image Communication 27*, 2 (2011), 192–207		40	40		20		
HOFBAUER, H., RATHGEB, C., UHL, A., AND WILD, P. Iris recognition in image domain: Quality-metric based comparators. In *Proceedings of the 8th International Symposium on Visual Computing (ISVC'12)* (Crete, Greece, July 2012), pp. 1 – 10		33		33			33
HOFBAUER, H., RATHGEB, C., UHL, A., AND WILD, P. Image metric-based biometric comparators: A supplement to feature vector-based hamming distance? In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'12)* (Darmstadt, Germany, Sept. 2012), pp. 1 – 5		33		33			33

Publication	Contribution (in %)						
	Hermann Hellwagner	Heinz Hofbauer	Robert Kuschnig	Christian Rathgeb	Thomas Stütz	Andreas Uhl	Peter Wild
HOFBAUER, H., AND UHL, A. An evaluation of visual security metrics. *IEEE Transactions on Multimedia* (2013), 15 pages. submitted		100					