Evaluation of Domain Specific Data Augmentation Techniques for the Classification of Celiac Disease using Endoscopic Imagery

Georg Wimmer and Andreas Uhl
University of Salzburg,
Department of Computer Sciences,
Jakob Haringerstrasse 2, 5020 Salzburg, Austria
Email: {gwimmer, uhl}@cosy.sbg.ac.at

Andreas Vecsei
St. Anna Children's Hospital,
Department Pediatrics,
Vienna, Austria

***Abstract*—In this paper we evaluate the effects of various data augmentation techniques on the automated classification of celiac disease using endoscopic imagery in the circumstances of limited training data. The used data augmentation techniques range from standard augmentation techniques like cropping patches and flipping to augmentation techniques using the full spectrum of affine or even projective transformations. We also present a novel technique that adjusts the lighting conditions depending on the scale changes caused by the augmentation. These augmentation techniques aim to generate augmented images that model the mucosa shown in the original image when conditions like the rotation of the endoscope or its viewpoint and distance to the mucosal wall changes. Tests are carried out using 5 different image representations including two convolutional neural networks and three shallow image representations. Our experiments showed that CNN's clearly benefit from augmentation techniques using affine and projective transformation, especially when the lighting conditions are adjusted.**

I. Introduction

Data augmentation is used to artificially increase the number of training samples and has become common practice in the area of deep learning. It has been shown in previous research that data augmentation can increase the invariance of a classifier and can act as a regularizer in preventing overfitting in neural networks [1]. For example, the authors of the well known Alex-net [1] (a competition winning convolutional neural network) stated that without data augmentation their network would suffer from substantial overfitting, which would have forced them to use much smaller networks. Also the performance of shallow representations (e.g. improved fisher vectors) can be significantly improved by adopting data augmentation, typically used in deep learning as shown in [2]. The most common data augmentation steps are to crop image patches for training at different positions and to horizontally flip those patches [1]. Other common augmentation steps are to add randomly generated lighting which tries to capture invariance to changes in lighting and minor color variation [1], [3], [4], slight rotations of the images [5], [6] and scaling of the images [7], [3].

Which data augmentation is useful for a particular application is highly domain specific. For example in object recognition or for the recognition of handwritten digits, images should be either not rotated or only slightly rotated since digits and most objects usually have a common and only slightly differing orientation (e.g. rotate the digit '6'by 180° and it becomes the digit '9'). In other applications like endoscopic image classification [8], where there is no predefined orientation of the things shown in the image, it can be useful to arbitrary rotate the images for augmentation [8].

Also objects are usually gathered under different viewing conditions. For example scaling as part of the data augmentation for object recognition was applied in [7] by rescaling patches of different sizes to a fixed patch size, in [3] by stretching images, and in [6] by simply downscaling images. In [6] even homographic transformations (horizontal and vertical pannings) were used as data augmentation step to model viewpoint changes. In the recognition of handwritten digits, affine transformations like translation, small rotations,scaling and shearing in combination with elastic transformations turned out to improve the recognition rates [5], [9] when used as augmentation techniques. Other examples of augmentation techniques described in literature are to create synthetic images generated from non-photorealistic 3D CAD models [10] or to apply the augmentation in feature-space instead of data-space [11].

In this work we evaluate the effect of different kinds of data augmentation strategies (including novel augmentation techniques) on the classification of celiac disease using endoscopic imagery. As image representations we use CNN's as well as shallow image representations. To the best of our knowledge, there has been no evaluation on data augmentation techniques in the area of endoscopic image processing and only standard augmentation approaches like cropping image patches at different positions, horizontally flipping and rotations [8] have been used so far.

Which data augmentation is useful for a particular application is highly domain specific. In endoscopic imagery, mucosal texture is usually depicted under different orientations, spatial scales and angular views, depending on the camera perspective and distance to the mucosal wall (see Fig. 1). Therefore,

978-1-5090-3649-3/17/$31.00 @2017 IEEE

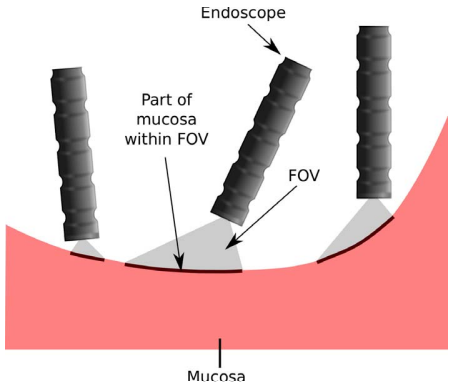


Figure 1. The field of view (FOV) depending on the endoscopic viewpoint and distance to the mucosal wall

the employment of affine and projective transformations as parts of the data augmentation is a highly intuitive idea to increase the number of training images and to improve the affine respectively projective invariance of a classifier while still modelling realistic viewing conditions. Using rotations as part of the augmentation models rotations of the endoscope around its own axis, scaling as part of the augmentation models varying distances of the endoscope to the mucosal wall, and using the full range of affine or projective transformations for augmentation models different viewpoints and distances of the endoscope to the mucosal wall.

Since the camera as well as the lights are both placed on the tip of the endoscope, the lighting conditions mainly depend on the distance from the endoscope to the mucosal wall. In this work we adjust the lighting conditions to the local scale changes caused by the affine or projective transformations in order to realistically model the changing scale conditions, a process which is novel to the best of our knowledge.

The amount of training data for the automated diagnosis of endoscopic images is usually limited to a few hundreds of images or even less. Consequently it is hard to achieve generalization with classifiers on such endoscopic data and to avoid overfitting to the training data corpus, especially for deep learning approaches like CNN's with millions of training parameters to be learnt. Data augmentation could help to reduce overfitting by increasing the number of training images and the variability of the training data.

The contributions of this manuscript are as follows:

- We evaluate which data augmentation strategies are most suited for the classification of celiac disease by performing experiments using 6 augmentation techniques reaching from standard augmentations (cropping and flipping) up to augmentation techniques using the full spectrum of affine or even projective transformations.
- Five state-of-the-art image representations for the diagnosis of celiac disease are applied in our experiments.
- Prior to the image transformations, the images are enlarged using a novel image preprocessing step that enlarges the images by means of reflections in order to enable all kinds of affine and projective augmentations.
- We apply a novel technique that adjusts the brightness of the images to the changing scale conditions caused by the affine and projective transformations.

II. Computer-Assisted Diagnosis of Celiac Disease

Celiac disease (CD) is a complex autoimmune disorder in genetically predisposed individuals of all age groups after introduction of gluten containing food. More than 2 million people in the United States have the disease, this is about one in 133. The gastrointestinal manifestations invariably comprise an inflammatory reaction within the mucosa of the small intestine caused by a dysregulated immune response triggered by ingested gluten proteins of certain cereals, especially against gliadine. During the course of the disease, hyperplasia of the enteric crypts occurs and the mucosa eventually looses its absorptive villi thus leading to a diminished ability to absorb nutrients. People with untreated celiac disease are at risk for developing various complications like osteoporosis, infertility and other autoimmune diseases including type 1 diabetes, autoimmune thyroid disease and autoimmune liver disease. This is why early diagnosis is of highest importance. Endoscopy with biopsy is currently considered the gold standard for the diagnosis of celiac disease.

Besides standard upper endoscopy, several new endoscopic approaches for diagnosing CD have been evaluated and found their way into clinical practice. The most notable techniques include the modified immersion technique (MIT) under traditional white-light illumination (denoted as $\mathrm{WL}_{\mathrm{MIT}}$), as well as MIT under narrow band imaging (denoted as $\mathrm{NBI}_{\mathrm{MIT}}$). These specialized endoscopic techniques were specifically designed for improving the visual confirmation of CD during endoscopy. Examples of healthy mucosa and mucosa affected by celiac disease for both endoscopy types are shown in Fig. 2.

A survey on computer aided decision support for the diagnosis of celiac disease can be found in [12].

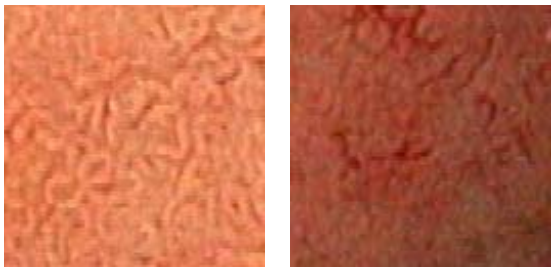
III. Image Augmentation

In this section we describe our data augmentation techniques as well as our technique to correct the lighting conditions. These techniques are only applied to the training portion of our database. We apply 6 augmentation strategies using the following image transformations additionally to the two most common augmentation steps (cropping and flipping):

1) no additional augmentation steps (basic augmentation)
2) rotations with multiples of 90°
3) arbitrary rotations
4) scaling combined with rotations
5) all affine transformations (scaling, rotations and shearing)
6) projective transformations (including all affine transformations)

Since CNN's and many other image representations require a constant image size, the size of the augmented image must not depend on the used affine or projective transformations. To solve this problem we add a preprocessing step that

(a) NBI_{MIT}, healthy (b) NBI_{MIT}, CD



(c) WL_{MIT}, healthy (d) WL_{MIT}, CD

Figure 2. Example images of the two classes healthy and CD using NBI_{MIT} and WL_{MIT}

doubles the size of the source image using reflections of itself as depicted in Fig. 3 (a). By applying the affine and projective transformations to the enlarged image, we are able to extract patches of the size of the source image from the transformed (and previously enlarged) image, even for arbitrary rotations, smaller zoom factors, shearing operations and projective transformations (see Fig. 3 (b)). The problem with this image enlargement approach is that reflection artifacts occur (consider the areas around the edges of the black square, which is marking the source image in Fig. 3 (a)). To crop as much parts as possible from the transformed source image and as less parts as possible from the synthetic enlarged part of the enlarged image, the position of the extracted patch is chosen so that its middle point corresponds to the transformed coordinate of the middle point of the untransformed (and enlarged) image. In that way we want to avoid reflection artifacts caused by the image enlargement as far as possible.

The affine and projective transforms are applied to the image using the following transformation matrix:

$$T = \begin{pmatrix} sc(x) * \cos(\alpha) * fl & -sh(y) * \sin(\alpha) & 0 \\ sh(x) * \sin(\alpha) * fl & sc(y) & 0 \\ pt(x) & pt(y) & 1 \end{pmatrix},$$

where $sc(x)$ and $sc(y)$ are the scale (zoom) factors in x-axis and y-axis, $sh(x)$ and $sh(y)$ are the shear (skew) parameters in x and y axis, α is the rotation angle and fl is the flipping parameter (either 1 or -1). $pt(x)$ and $pt(y)$ are the projective (homographic) parameters. They are zero in case of affine transformations and unequal zero in case of perspective warpings. Linear interpolation is used for the geometric transformation of the image (using Matlab's 'imwarp'function). Exemplar image transformations for the augmentation and the resulting images (marked as black squares) are shown in Fig. 3 (b). As we can see in this figure, shearing and projective transformations model the mucosa under different viewing angles from the endoscope towards the mucosa and the combinations of all affine or all projective transformations model the mucosa under different viewpoints and distances.

For our three augmentation strategies that are changing the scale of the images (scaling combined with rotations, all affine transformations, all projective transformations) we use a novel technique that adapts the lighting conditions of the transformed image I^T. First we have to compute the scale change caused by the transformation for each point of the image.

Let us define $\vec{x} = (x, y)$ as a arbitrary pixel coordinate of the Image I, $\vec{x_n} = (x+1, y+1)$ as the neighbored pixel of $\vec{x}$, $\vec{x}^T = (x^T, y^T)$ as the transformed image coordinate of $\vec{x}$ ($\vec{x}^T = \vec{x} \times T$) and $\vec{x_n}^T$ as the transformed image coordinate of $\vec{x_n}$. Let us furtherly define d as the euclidean distance between a pixel coordinate and its neighbor (e.g. $d(\vec{x^T}) = \sqrt{(x^T - {x_n}^T)^2 + (y^T - {y_n}^T)^2}$). The value of $d(\vec{x})$ is always $\sqrt{2}$ but the value of $d(\vec{x^T})$ is dependent on the transform T and can be different for different locations in the image. We define the local scale factor f of the transform T as $f(\vec{x}^T) = d(\vec{x}^T)/\sqrt{2}$. The brightness of the transformed image I^T is adapted as follows:

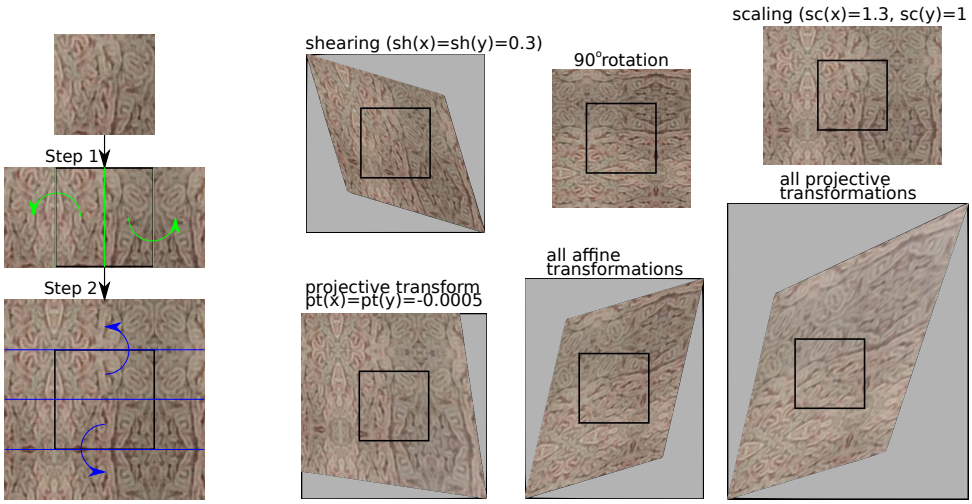
$$I_c^b(x,y) = \begin{cases} I_c^T(x,y) - I_c^T(x,y) * (1 - f(x,y)) * b & \text{for } f(x,y) < 1 \\ I_c^T(x,y) + (m - I_c^T(x,y)) * (f(x,y) - 1) * b & \text{else.} \end{cases}$$

I^b denotes the brightness adjusted image, c the color channel of the image (1,2 or 3 for our RGB images), m the highest possible brightness value of a pixel (255 in case of an unnormalized image) in the considered color channel, and b ($b = 0.4$) is the brightness adjusting factor (empirically determined). So increasing the scale (zoom) increases the brightness and decreasing the scale reduces the brightness. That means if the augmentation transform simulates a smaller (higher) distance between the endoscope and the mucosal wall then the brightness will be increased (decreased) in order to realistically model the brightness conditions to the changed distance of the lighting source (endoscope) to the mucosal wall.

IV. Image Representations

In this section we shortly describe the five used image representations.

VGG-CNN: The VGG net [2] consists of 5 convolutional layers and three fully connected layers with a final soft-max classifier. There are three versions of such nets presented in [2] and we only use the fastest of these nets (denoted as vgg-f in [2]) since this net turned out to be the most suited net for the classification of CD in [13]. As initialization for the convolutional layers we use the parameters that were learned on the ImageNet ILSVRC challenge data. Since the fully connected layers are more specific to the details of the classes contained in the ILSVRC challenge data, we randomly



(a) Enlargement (b) Augmentation

Figure 3. Fig. a): Synthetic image enlargement by reflecting the image outwards (first horizontally (Step 1) and then vertically (Step 2)). Fig. b): Illustration of four different image transformations applied to the enlarged image from (a) as well as the combination of all affine as well as all projective transformations.The black squares mark the resulting outputs.

initialize the coefficients of these layers. The training of the net on our CD database is performed as described in [13] (5000 iterations with decreasing learning rates).

CD-CNN: The celiac disease (CD) net [8] (denoted as Very-Deep net in [8]) consists of 4 convolutional layers and three fully connected layers with a final soft-max classifier and was developed for the classification of Celiac disease in [8] . The coefficients of the convolutional layers and the fully connected layers are randomly initialized. So contrary to the VGG-net, the CD-net is not initialized using pre-trained coefficients. The training of the net is performed as described in [8].

For both nets (CD and VGG), training is performed on batches of 128 images, which are for each iteration randomly chosen from the training data and subsequently augmented (using our 6 augmentation techniques).

Improved Fisher Vectors (IFV): Based on estimated Gaussian mixtures of locally pooled SIFT descriptors. The improved version based on a non-linear Hellinger's kernel and l^2 normalization with 16 clusters is used.

Dual-tree Complex Wavelet Transform (DT-CWT): DT-CWT [14]) is a multi-scale (4) and multi-orientation wavelet transform. The feature extraction is based on computing the means and standard deviations of the DT-CWT sub-bands. The concatenation of the extracted features from all subbands gives the final feature vector of an image.

Multiscale Block Binary Patterns: The Multiscale Block Binary Patterns (MB-LBP) [15] is based on the local binary patterns (LBP) operator and is applied using three different block sizes (3,9,15) and uniform LBP histograms.

V. Experimental Setup

The 1661 RGB image patches of size 128×128 pixels from our CD database are gathered from 353 patients by means of flexible endoscopes using NBI_{MIT} as well as WL_{MIT}. 1045 image patches are gathered by WL_{MIT} endoscopy (587 healthy images and 458 affected by celiac disease) and 616 image patches are gathered by NBI_{MIT} endoscopy (399 healthy images and 217 affected by celiac disease). So in total 986 image patches show healthy mucosa and the remaining 675 image patches show mucosa affected by celiac disease. The patches are extracted from regions exhibiting histological findings.

The CNN's are implemented using the *MatConvNet* framework [16], the IFV implementation is provided by *VLFeat* and we rely on in-house MATLAB implementations for MB-LBP and DT-CWT.

Due to the relatively small amount of data, we perform 3-fold cross-validation to achieve a stable estimation of the generalization error. For each of the three folds we took care that images of a single patient are either all in the training portion or all in the evaluation portion. The three folds are disjoint and each fold contains approximately one third of the images of the database. All image representations are trained using the training portion of our data corpus (two of the three folds). The final validation was performed on the left-out part. In our experiments, we compute the overall classification rate (OCR) for each fold and report the mean OCR over all three folds. We only used the OCR as performance measure because of the high number of results and its easy comparability over different augmentation strategies.

We use two classification strategies in this work:

1) *CNN soft-max classification:* In case of the CNN inbuilt soft-max classifier, our augmentation strategies are applied to the batches of images extracted for training. For each iteration and each image of a batch, the transform parameters are randomly chosen within the admissible values which are listed in Table I (CNN soft-max). The admissible values are chosen so that

the transformations clearly change the image without causing a blurry appearing output by increasing the size of the image too much and without decreasing the size of the image so far that too much reflection artifacts occur in the extracted patches. Examples for more extreme but still valid transform parameters are shown in Fig. 3 (b).

2) *SVM-classification:* SVM classification is applied to the shallow image representations as well as to the CNN's (additionally to the soft-max classification). The admissible transform parameter values are listed in Table I (SVM). The SVM classifier is provided by the *LIBLINEAR* library. [17]. The SVM cost factor (C) is found using cross validation on the training data. In case of the CNN's, the training and test samples are fed through the CNN's and the input of the last fully connected layer is extracted as feature for further SVM classification.

We use the same training and evaluation set portions for each image representation and classification strategy.

In the following, we describe the different augmentation strategies for soft-max as well as SVM classification:

1) *Basic augmentation:* The images are randomly horizontally flipped or not flipped. In case of the CNNs (SVM as well as soft-max), patches (112×112 pixels) are cropped at random positions. In case of the three shallow representations, the whole image is used (the idea behind cropping patches at different positions is to enhance the translation invariance of the classifier but the shallow representations are translation invariant and so cropping does not make sense.) The basic augmentation is a subset of each other augmentation strategy.
2) $90°$ *Rotations:* The images are randomly rotated using rotation angles that are multiples of $90°$ (4-fold increase of training samples).
3) *All Rotations:* The images are randomly rotated using arbitrary rotation angles in case of the soft-max classification and they are rotated using the 12 valid orientations shown in Table I in case of SVM classification (12-fold increase of training samples). Compared to the previous augmentation strategy, which is a subset of this strategy, we have the advantage of a higher range of rotation angles and the disadvantages that the augmented images contain reflection artifacts since parts of the augmented images are generated by the image enlargement.
4) *Scaling and* $90°$ *Rotations:* Expands on the augmentation strategy using $90°$ rotations and additionally randomly scales the images (in x and y direction). In case of the SVM classification, the original training data together with 19 augmented version of the training data is used for training. 19 randomly chosen and not repetitive combinations over all transform parameter options shown in Table I (SVM) are used to augment the images. This 19 combinations over all transform parameters are identical for the three augmentation strategies 'Scaling and $90°$ Rotations', 'Affine Transformations'and 'Projective Transformations'. The not required transform parameters for each of the three augmentation strategies are skipped.
5) *Affine Transformations:* Same as 'Scaling and $90°$ Rotations', but with shearing operations. So this augmentation strategy comprises all affine transformations.
6) *Projective Transformations:* Expands on the augmentation strategy using all affine transformations by adding projective transformations to model different viewpoints.

So in case of the CNN soft-max classification we have perfect conditions to apply and compare the different augmentation strategies since a maximum of different transform parameters are applied (randomly chosen for each image in each iteration) and the same numbers of training images are used for each augmentation strategy (128 augmented images per iteration). In case of the SVM classification, only a small selection of different transform parameters can be applied for augmentation (we cannot infinitely increase the number of training images by augmenting the training data using all combinations of augmentation transform parameter values since the system would run out of memory) and the number of augmented images varies depending on the augmentation strategy. So it is to be expected that the CNN's benefit more from the additional augmentation transforms since their augmentation was performed using a much higher variability of transform parameters as for the shallow image representations classified by SVM's. This also applies for the CNN's classified by SVM's, since features are extracted using the net which is trained with the respective augmentation strategy.

In case of the CNN's (SVM as well as soft-max classification), validation is performed using a majority voting over five crops from the validation image using the upper left, upper right, lower left, lower right and center part. In case of the shallow image representations, the whole images are used (like for training) and also no other augmentations are applied to the validation images.

VI. Results

In Table II we present the results for the different augmentation strategies. Results behind a dash are obtained using our technique to adjust the brightness. The best result for each image representation over all augmentation strategies is marked in boldface.

In case of the two CNN's, the best results were achieved using the three augmentation strategies 'Scaling and $90°$ Rotations', 'Affine Transformations'and 'Projective Transformations'. When we compare the results before and after a dash (without respectively with brightness adaption), we see that applying our proposed brightness adjusting technique furtherly improves the results for each of these augmentation techniques. The distinctly lowest results were achieved using only the basic augmentation steps cropping and flipping. The built-in soft-max classifier and the SVM classifier achieved nearly identical results. So the CNN's clearly profit from augmentation techniques including scaling, rotation transforms and brightness adjustments. Projective transformations were only favorable in case of the CD-net.

Method	90° rotations	all rotation	scaling	shearing	projective
CNN soft-max	$0°, 90°, ...270°$	$0° \leq \alpha \leq 360°$	$1/1.3 \leq sc \leq 1.3$	$-0.3 \leq sh \leq 0.3$	$-0.0005 \leq pt \leq 0.0005$
SVM	$0°, 90°, ...270°$	$0°, 22.5°, 45°, ...337.5°$	$1/1.25, 1, 1.25$	$-0.25, 0, 0.25$	$-0.0004, 0, 0.0004$

Table I
PERMITTED TRANSFORM PARAMETER RANGE AND TRANSFORM PARAMETER OPTIONS FOR THE SOFT-MAX (CNN) AND SVM CLASSIFICATION.

Method	Transformations					
	basic	90° rot	all rot.	scale & 90° rot	affine	projective
VGG-net soft-max	85.6	89.5	88.3	90.6 / **91.0**	90.5 / **91.0**	90.1 / 90.6
VGG-net SVM	85.7	89.6	88.6	90.8 / **91.1**	90.5 / 90.5	89.8 / 90.5
CD-net soft-max	85.6	87.6	87.2	87.7 / **88.9**	88.1 / 88.5	88.5 / **88.9**
CD-net SVM	85.2	87.8	87.2	88.3 / 88.5	88.1 / 88.4	88.5 / **88.8**
IFV	82.2	83.0	**84.9**	83.1 / 82.9	84.0 / 83.7	84.2 / **84.9**
DT-CWT	**74.2**	72.9	**74.2**	72.6 / 72.1	72.5 / 70.9	71.8 / 72.4
MB-LBP	**84.3**	82.4	81.4	79.7 / 79.7	80.3 / 79.4	80.2 / 81.0

Table II
MEAN ACCURACIES FOR THE DIFFERENT AUGMENTATION STRATEGIES ON THE CD DATABASE.

In case of the IFV, the highest result was achieved for the two augmentation approaches applying arbitrary rotations and projective transformations in combination with brightness correction, which clearly outperformed the basic augmentation approach.

DT-CWT and MB-LBP performed best using only the basic transformations and generally perform worse the more different image transformations are included in the augmentation strategy.

VII. CONCLUSION

Our experiments showed that CNN's clearly profit by using augmentation steps additional to the basic augmentation steps (cropping and flipping) on our celiac disease endoscopic image database. The three augmentation strategies combining scaling and rotations, all affine and all projective transformation achieved the highest results, especially in combination with our proposed brightness adjustment technique.

The shallow representations DTCWT and MB-LBP achieved the highest results for the basic augmentation approach (flipping) and did not profit from additional augmentation steps. IFV performed best using arbitrary rotations and projective transformations in combination with brightness correction and the worst result was achieved using the basic augmentation strategy.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5*, 2014.

[3] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," *CoRR*, vol. abs/1501.02876, 2015. [Online]. Available: http://arxiv.org/abs/1501.02876

[4] A. G. Howard, "Some improvements on deep convolutional neural network based image classification," *CoRR*, vol. abs/1312.5402, 2013. [Online]. Available: http://arxiv.org/abs/1312.5402

[5] J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR '12, 2012, pp. 3642–3649.

[6] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin, and C. Schmid, "Transformation pursuit for image classification," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14, 2014, pp. 3646–3653.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[8] G. Wimmer, A. Vecsei, and A. Uhl, "Convolutional neural network architectures for the automated diagnosis of celiac disease," in *Proceedings of the 3rd International Workshop on Computer-Assisted and Robotic Endoscopy (CARE)*, ser. Springer LNCS, Oct. 2016.

[9] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "High-performance neural networks for visual object classification," *CoRR*, vol. abs/1102.0183, 2011. [Online]. Available: http://arxiv.org/abs/1102.0183

[10] X. Peng, B. Sun, K. Ali, and K. Saenko, "Learning deep object detectors from 3d models," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. IEEE Computer Society, 2015, pp. 1278–1286.

[11] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" *ArXiv e-prints*, Sep. 2016.

[12] S. Hegenbart, A. Uhl, and A. Vécsei, "Survey on computer aided decision support for diagnosis of celiac disease," *Computers in Biology and Medicine*, no. 65, pp. 348–358, 2015.

[13] G. Wimmer, A. Vécsei, and A. Uhl, "Cnn transfer learning for the automated diagnosis of celiac disease," in *Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, ser. Springer LNCS, 2016, pp. 1–6.

[14] N. G. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters," in *Proceedings of the IEEE Digital Signal Processing Workshop, DSP '98*, Bryce Canyon, USA, Aug. 1998, pp. 9–12.

[15] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning multi-scale block local binary patterns for face recognition," in *Proceedings of the 2007 International Conference on Advances in Biometrics*, ser. ICB'07. Springer-Verlag, 2007, pp. 828–837.

[16] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," in *Proceeding of the ACM Int. Conf. on Multimedia*, 2015, pp. 689–692.

[17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.