To See or Not To See: Determining the Recognition Threshold of Encrypted Images

Heinz Hofbauer[1] • Florent Autrusseau[2] • Andreas Uhl[1]

[1] University of Salzburg, Austria

[2] University of Nantes, France

Abstract

Numerous standards and recommendations deal with the acquisition of visual quality assessment from human observers accounting for clearly visible images and trying to keep the just-noticeable-difference between quality steps as small as possible.

When it comes to the assessment of selective encryption schemes the question is the opposite. The quality is not really of interest, the question is rather if the content of the images is discernible at all.

There are no recommendations in literature for this kind of task. We outline different protocols and setups, test them and form a recommendation for the acquisition of the recognition threshold for encrypted images from human observers.

Main Results

- The **Match2** protocol is recommended since it gives a higher error rate allowing for a better differentiation between image recognition.
- Setup is not as important as for quality.
  - Little difference between the tested environments.
  - *UE* setup is fine, but unlimited viewing time might lead to viewer fatigue.
- A longer pre-test habituation period is better to familiarize obsevers with distortion types.
- For outlier detection hierarchical clustering methods are recommended.

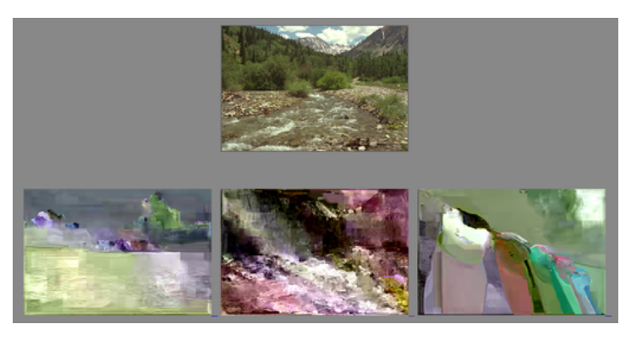
Protocols for Acquisition

Is the original still recognizable from the encrypted image?

- What we want to acquire is a score per image which reflects the recognizability.
- Traditional acquisition methods are quality estimation no finding the recognizability threshold.
- Direct comparison (traditional) suffers from apophenia, the tendency to perceive connections and meaning between unrelated things.
- A forced choice: the participant has to choose among a number of candidate images and identify the "correct" one (or guess).

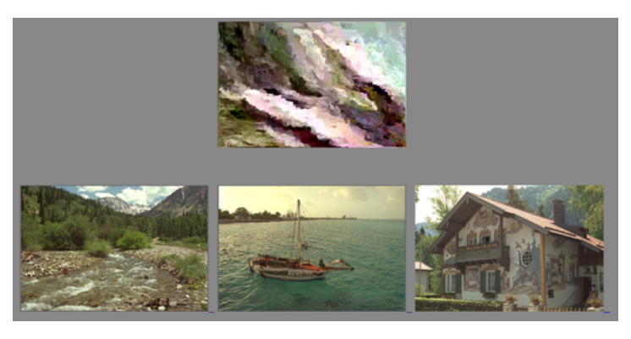
Three methods are conceivable

O3



Show a single original image and three encrypted images. The participant has to select the encrypted version of the original image.

3E



Three plain text images and one encrypted image is shown. The participant has to select the correct original image from which the encrypted image was derived.

Match2

Three originals and three encrypted images are shown. One pair of images is an original and derived encrypted image, the other four images have to be unrelated. The participant must select the matching pair.

For all three methods it is required to have images with similar encryption strengths to be shown simultaneously.

Environment for Acquisition

Three Environments to Test

In order to evaluate the environmental influence on the recording we will utilize three different setups:

Controlled (*CE*): The controlled environment uses a calibrated monitor in a closed room, i.e., no natural lighting is present, and a strictly controlled artificial lighting to conform to:

> ITU-R BT.500-13, *Methodology for the subjective assesment of the quality of television pictures*, http://www.itu.int/rec/R-REC-BT.500/en, Geneva, Switzerland, 2012

Semi-controlled (*SE*): A regular working space, some measures were taken to limit extraneous light, e.g., blinds were drawn.

Uncontrolled (*UE*): The uncontrolled environment is simply what was available at the users own PC. The experiment was set up to be used over the internet at the workstation of the users PC.

Conformance to Constraints by the Environments.

	Luminance	Viewing Distance	Scale	Vision Check	View Time	Observers
CE	✓	✓	✓	✓	8 sec	45
SE	✗	✓	✓	~	8 sec	30
UE	✗	✗	✗	✗	∞ sec	41

Viewing distance: The viewing distance was set to 6 times the images' height.

Scaling: Whether the displayed images need to be scaled to be displayed at once.

Illumination and Calibration: Monitor calibrated and environment conforming to ITU recommendations:

- illuminant white point CIE D65
- maximum screen luminance of 200 cd/m^2
- screen gamma function of 2.20
- contrast ratio/ black point of 2 cd/m^2
- background illumination of 10 lux.

Viewing Time: To combat viewer fatigue.

Vision Check: Testing for visual acuity and color vision.

Observers: Minimum number recommended by standards is 15.

Analysis of Data

Difference to Quality Tests

- Quality evaluation resulting in a score per image and observer.
- Recognition evaluation resuls in a binary decision per image and observer.
- Recognition: Final score is aggregate over all observers.
- Outlier detection in the classical sense will not work, no score to compare per observer.
- Error aggregate also won't work: two observers can have the same number of errors and not agreeing on a single image.

Outlier detection

- Assume a graph with:
  - Each observer is a node.
  - Distance is Hamming distance
- Perform a hierarchical clustering.
  - Start with smallest distance
  - Continue to cluster until all nodes in a single cluster
  - Remember the distance on join per node.

The outliers can then be detected based on statistics of similarity between observers O and the set of pairwise distances:

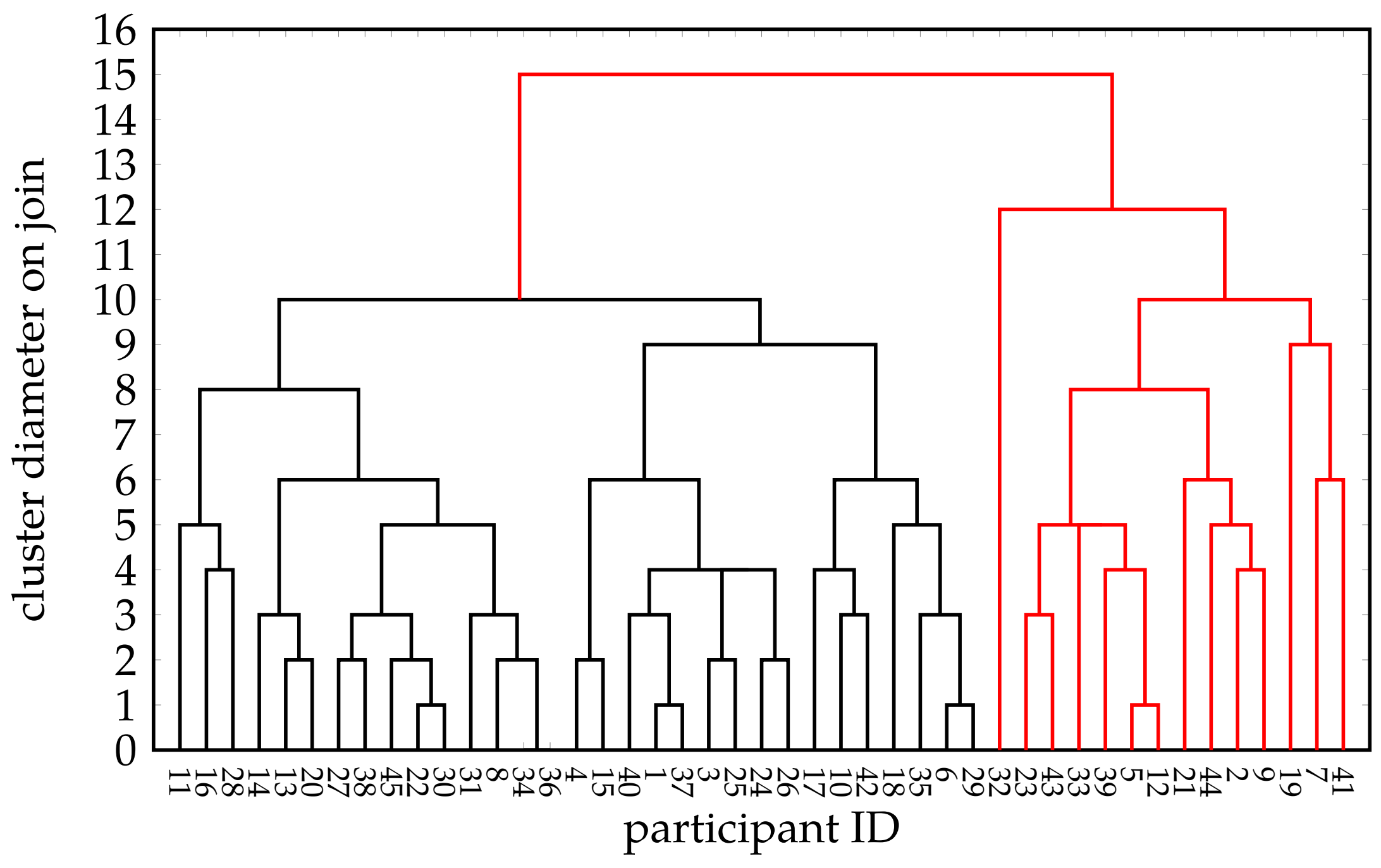
$$D = \{HD(O_i, O_j) \mid \forall O_i, O_j \in I, i \neq j\}$$

Use z-score on distances:

$$z_D = \mu(D) + 3\sigma(D)$$

Cut off cluster at z_D and discard smaller set as outliers.

Example based on *CE*



Dataset

Single image original and (jxr) encrypted variants.

- 14 images.
- 3 encryption types
- 6 different strengths per type

Recognition and Image Contents

Recognition, or rather encryption performance, is not independent of image content.,

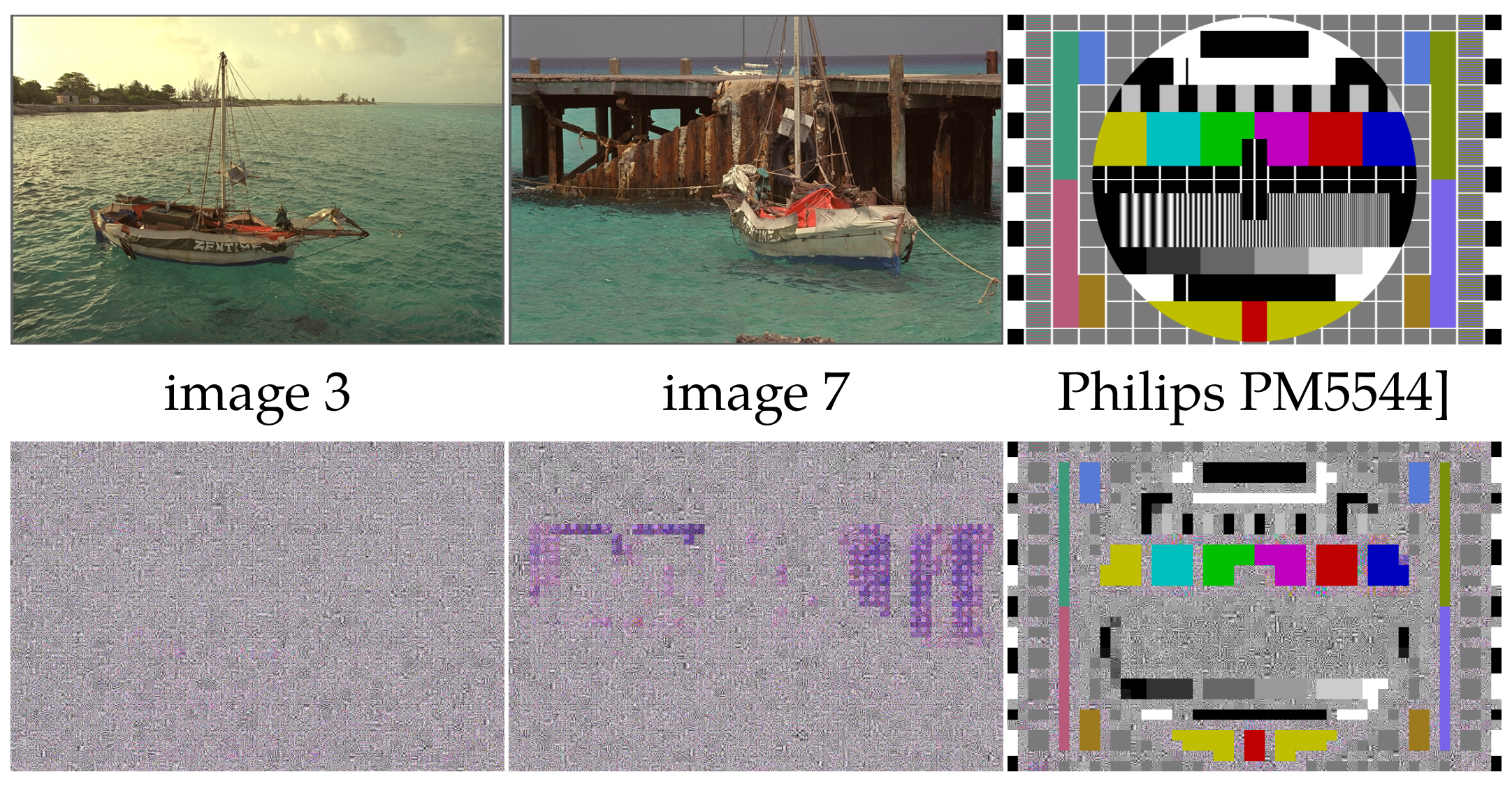


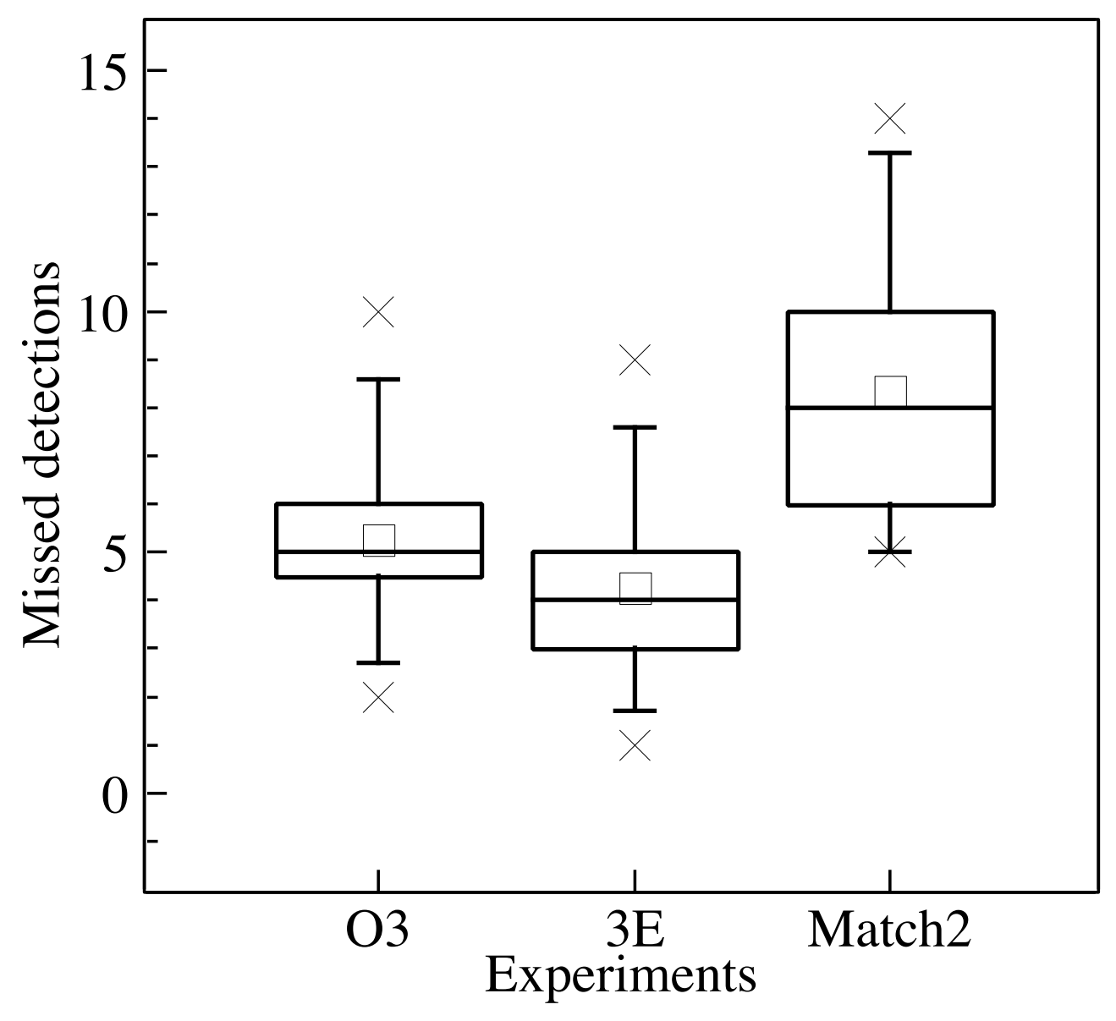
image 3 | image 7 | Philips PM5544]

RE = 0.874 | RE = 0.329 | RE = 0

Original images (top) and encrypted version (bottom) with the same encryption parameter. Recognition errors (RE) are given per image.

Evaluation of the Acquisition Protocol

- The collected outputs were the number of mis-detections.
- The decision appeared to be more difficult for the **Match2** protocol.



Repartition of the mis-detections across the three pre-tests.

- Goal:recognizability scores that will be continuously distributed.
- **Match2**: more missed detections, more errors and a higher number of scores which are between recognizable/nonrecognizable.

Analysis of Acquisition Environment

- linear correlation: same number of errors (as a representation of recognizability)
- rank order correlation: ordering images from least to most recognizable based on errors

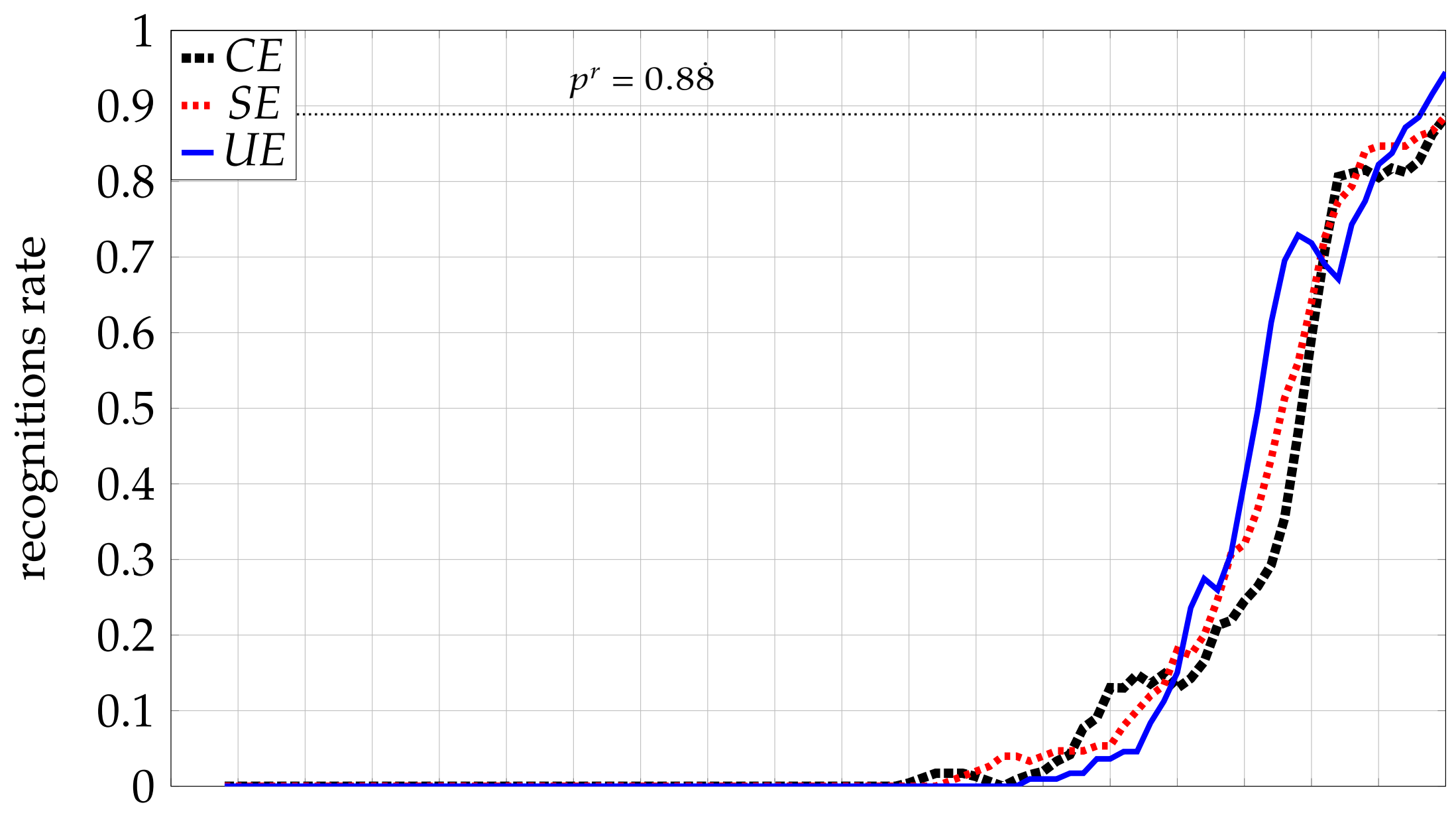
Agreement matrix between the acquisition environments based on linear and Spearman rank order correlation.

	linear correlation				*rank order correlation*		
	CE	***SE***	***UE***		***CE***	***SE***	***UE***
CE	1.000	0.984	0.978	***CE***	1.000	0.862	0.884
SE	0.984	1.000	0.978	***SE***	0.862	1.000	0.888
UE	0.978	0.978	1.000	***UE***	0.884	0.888	1.000

- Both methods agree on the outcome: the three environments are strongly related but there are differences.
- Overall all environments exhibit the same trend suggesting differences base on:
  - miss-clicks
  - the innate randomness in the recognizability study

Range of Responses

- Order images based on an error aggregate from all environments.
- Plot score per environment (smoothed).
- The plot was smoothed to suppress an extremely jaggedness due to miss clicks by observers.
- Recognition rate (RR) is the relative error over all observers: an RR of 0 means all observers recognized the image.



Images ordered by aggregate recognition rate

Plot of individual scores, per environment, compared to the overall ordering based on an aggregate over all environments.

All experimental setups

- show a very similar curve and the choice of environment does not seem to influence the results.
- show a gradient from recognizable to unrecognizable.
- are trending towards the probability of random choice (p^r).