You look blocky, is everything alright? Influence of video distortions on facial expression recognition & quality

Yohann Pitrey, Helmut Hlavacs

Entertainment Computing Research Group, University of Vienna, Austria
yohann.pitrey@univie.ac.at helmut.hlavacs@univie.ac.at

Abstract

The ability to read expressions on human faces is hardwired in the human brain, as a very powerful tool to sample non-verbal information about the emotional disposition of other people. Nowadays, more and more face-to-face interactions are happening over multimedia systems, including video conferencing, watching movies and interacting with virtual characters. Some of these systems usually rely on compression and other data reduction mechanisms to cope with limited transmission speeds. This paper evaluates the impact of typical video distortions, such as compression, spatial and temporal downscaling, on the ability of human subjects to recognize facial expressions. We use a realistic face simulator to generate four basic facial expressions, which we distort under typical scenarios, and measure both the ability of participants to recognize the facial expressions and the overall perceived quality. Our results show that the scene composition and the expressions have a significant impact on perception, and show some evidence on the relation between recognition rate and perceived quality.

Index Terms: perceived video quality, facial expression recognition, video compression, photo-realistic face simulation

1. Introduction

Psychology and cognitive sciences have long discovered that facial expressions are strongly linked with the underlying emotions and therefore play a vital role in interpreting one's mental disposition [1]. It is known that the ability to recognize a facial expression on a human face is both very fast and very efficient for healthy subjects [2].

Facial expression recognition is also needed in some multimedia applications. On one hand, more and more interactions between human subjects are dematerialized using multimedia systems. An example is videoconferencing, in which two or more participants communicate in real-time through a double-ended channel. Other broader contexts might involve one-way recognition, such as watching a movie picturing human characters. As a matter of fact, any type of video consumption scenario can potentially contain situations in which facial expression recognition is needed.

On the other hand, virtual characters have been used in human-computer interaction for many years [9, 7]. Virtual environments involving virtual characters can for instance be used in therapeutic contexts such as clinician-patient communication, or to ease social anxiety or fear of speaking in public [11, 12]. The behaviour of the virtual characters can be fully tailored to the needs and goals of the patient, providing a better control of exposure to fear and anxiety. A moajor challenge to be addressed by virtual characters is to mimic real human behavior in a credible way, to avoid introducing more stress on the users [8]. Particularly, the ability of avatars to reproduce realistic facial expressions has been identified as a major requirement for a natural communication [6]. Systems failing to meet this requirement might have to face what psychologists call the *uncanny valley*, which describes the uneasy feeling experienced by human subjects when facing virtual characters who look quite close to real humans, but who can still be identified as fake without ambiguity.

A classical problem encountered in video communications is the loss in quality induced by the use of compression and down-scaling techniques to cope with limited bandwidth, or induced by inconsistent transmission conditions. scenarios such as distant video communication, using virtual characters or not, are particularly likely to suffer from these distortions. In a broader sense, compressed movies and any kind of multimedia application using compression can be subject to such visual distortions.

In the particular case of facial expressions, the effect of these distortions can be two-fold. On one hand, they can induce a loss in recognition accuracy. The facial features used to process an expression might appear distorted and either yield recognition of an incorrect expression, or yield confusion if no natural expression can be recognized. The impact of compression might even bring uncanney valley's edge closer, as the distortions create eery faces with unnatural shape and motion. On the other hand, altering the process of expression recognition by introducing distortions might bias the overall perception of quality. The facial area is a natural visual attention at-

10.21437/PQS.2013-11

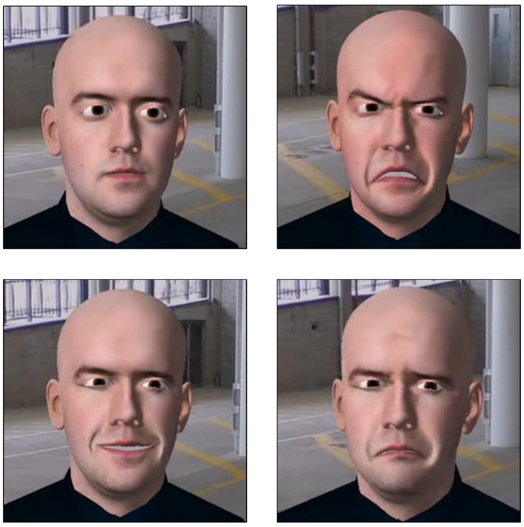

Figure 1: The four facial expressions used in our experiment. In reading order: *neutral*, *anger*, *joy*, *sadness*.

tractor, and its relative quality when compared to the rest of the scene might be of great importance in the feeling of overall quality experienced by the viewer [10].

The impact of compression on facial expression recognition has been studied in the past [3, 4]. It mainly focuses on the recognition performance of algorithms and does not involve subjective evaluation. To the best of our knowledge, the impact of distortions on the human ability to recognize facial expressions, and the relation between facial expression recognition and perceived quality have never been evaluated. In this paper, we present the results of a standardized subjective experiment in which participants were presented video sequences picturing facial expressions on a virtual animated character, using a realistic face model. The participants were asked to recognize the emotion associated with each expression, as well as to give their opinion of global quality. We used four basic expressions, varied the distance between the character and the camera, as well as the intensity of the expression. Each sequence was then encoded using six realistic scenarios introducing various levels of coding artifacts, blur due to spatial up-scaling and jerky motion due to lower temporal resolution.

The paper is organized as follows. In section 2, we present the realistic face model used in the experiment. The design of the experiment is described in section 3. In section 5, we analyze and discuss the experimental results, and section 6 concludes the paper.

2. Facial expression simulation

We built a face simulation framework based on the Ogre3D[1] rendering engine. Our simulator uses a three-dimensional mesh on which a skin texture is superimposed. One major particularity of our system is that it uses a sub-skin muscle structure in order to animate the facial features in a realistic way. The location and shape of the muscles are based on the work of the psychologist Paul Ekman, who devoted a great part of his life to studying human emotions and their link with facial expressions [1]. Our simulator can be manipulated using a graphical interface, allowing the experimenter to simulate the action of different muscles on the simulated face. Following the directives of Ekman, a given facial expression can easily be reproduced by moving a specific set of muscles.

As an example of the capabilities of our simulator, Figure 1 shows the four facial expressions we used in our experiment. The structure of the mesh on which the skin structure is superimposed was modeled from a real person's face. It is not perfectly symetrical and the skin colour is not perfectly uniform, which increases the realistic look of the character. Our system allows the experimenter to contract each muscle independently, as well as groups of muscles. The tension applied to a muscle can be varied progressively, which allows one to create expressions with arbitrary intensity. The simulator can also be used dynamically, in order to generate animated faces.

The four emotions were generated according to the directives of Paul Ekman's Facial Action Coding System [13], according to which a large set of emotions is described as a combination of facial muscles to be contracted. We obtained a first set of facial expressions, which we presented to a group of psychologists who were used to manipulate facial expressions as experimental material in their work. Some simple corrections were brought to the facial expressions after this meeting, in order to make them look as close as realistic and accurate as possible.

3. Experimental design

Experimental methodologies evaluating facial expression recognition under various scenarios have been designed in the past, such as the Emotion Recognition Task (ERT) [2]. This contribution was focused on the psychophysical and cognitive aspects involved in the process, and the authors explored the influence factors in terms of scene composition. This methodology provides a list of requirements in order to design a scientifically sound facial expression recognition experiment. The experiment we present in this paper is inspired by a subset of the ERT, which we combine with various video distortion scenarios and scene composition parameters.

We used our face simulator to generate three basic facial expressions illustrating basic emotions, namely *anger*, *joy* and *sadness*. The neutral face corresponding to no muscles activated was also used as a control condition in our experiment, increasing the number of distinct

[1] http://www.ogre3d.org

facial expressions to four. Figure 1 illustrates the four expressions generated using our simulator.

The face model and the attached virtual body can be animated and exported in a video sequence using the capacities of Ogre3D. Animated sequences of ten seconds each were created, with a panning camera motion and realistic lighting. For the needs of the experiment, the face model was mounted onto a virtual body. The body was modeled using traditional 3D modeling techniques, and we put some effort into making it appear as realistic as possible. We decided not to include hair on the head, as hair is relatively difficult to render in a realistic way. The virtual character was then positioned in front of a static background picture. The picture contains mostly shades of grey and light colours, and features relatively low semantic information, in order to avoid distracting the viewers. The camera motion was added to create some temporal activity, in order to impair the coding performances of the video encoder and reflect a more realistic scenario.

Within each ten-second sequence, the character's face is first shown neutral, then changes progressively to a full-blown expression, then returns progressively to a neutral expression (except of course for the neutral expression, in which the face remains neutral for ten seconds). This dynamic aspect of the expressions is mentioned in the ERT to be important in the recognition, as the movement of the facial features involved in the process provides vital clues for the human brain.

For each facial expression, we varied five parameters: distance to camera, intensity, spatial resolution, temporal frequency and encoding bitrate. Two scenarios were designed to evaluate the influence of the distance between the character and the camera. One scenario pictured the character in a head shot, the second scenario pictured a full-body shot. For the three non-neutral scenarios, we designed two scenarios for the intensity of the expression. One scenario pictured a high-intensity expression blown at 80% of the maximum intensity allowed by the face muscles involved; the second scenario pictured a low-intensity expression blown at 40% of the maximum intensity.

Each combination of expression, distance and intensity was then encoded according to six scenarios with various bit-rates, native spatial resolution and temporal frequency. The videos captured from the Ogre 3D rendering engine were in VGA format (640 × 480 pixels) at 24 Hz. They were then encoded using the x264 encoder[2], using standard parameter values. We used the available bit-rate control capability to encode the video sequences at constant bitrate. The six encoding scenarios are presented in Table 1. They involve different combinations of spatial and temporal down-scaling, as well as different bit-rates. All videos were presented to the viewers in VGA format at 24 Hz, which required spatial up-scaling and/or temporal up-scaling for some scenarios. The spatial up-scaling was performed using a standard Lanczos upscaler such as the one provided in the JSVM reference software suite for Scalable Video Coding [5]. Temporal upscaling was performed by repeating each frame in the scenarios at 12 Hz, in order to get video sequences at 24 Hz again.

#	Size	FPS	Bit-rate	Short notation
1	640 x 480	24	256 kbps	640x480@24:256
2	640 x 480	24	128 kbps	640x480@24:128
3	640 x 480	12	128 kbps	640x480@12:128
4	320 x 240	24	128 kbps	320x240@24:128
5	640 x 480	24	64 kbps	640x480@24:64
6	320 x 240	12	64 kbps	320x240@12:64

Table 1: Encoding scenarios used in our experiment for each expression and scene composition.

4. Experimental conditions

The video sequences were presented on a consumer-range 22 inches computer screen, in a quiet room with controlled light conditions. As stated earlier, the videos were in VGA format. The compressed video streams were first decoded and exported to YUV format, in order to avoid any decoding performance issues during playback. They were presented with no further scaling and surrounded with a neutral grey border. As this experiment was conducted in the context of a research project about perceived quality on mobile devices, we did not mean to reproduce strictly standardized viewing conditions. Moreover, the viewing distance was arranged so that the video sequences covered the same field of view as they would do on a standard mobile phone held at a distance of 30 centimeters from the viewer.

We used a slightly modified 5-level Absolute Category Rating methodology, in order to incorporate facial expression recognition. After watching a video sequence, the viewers first had to judge which emotion corresponded to the character's face. They were provided with a list of four choices, reflecting the four expressions of our test: anger, sadness, joy and neutral. The rest of the experiment followed the guidelines of the ITU-T P.910 recommendation. After recognizing the expression, the viewers had to give their overall opinion of quality on the sequence they just watched. They used a standard 5-level scale, with labels ranging from *excellent* to *bad*. The presentation order was randomized and different for each viewer, and we made sure that no videos picturing the same content (*i.e.* before compression) were displayed consecutively.

A total of 84 video sequences were presented to each viewer. A training phase was conducted prior to each test session, in which we presented five videos covering the whole quality scale and showing various levels

[2] http://videolan.org/x264

of difficulty for facial expression recognition. The sequences used for the training set were not presented in the remaining part of the experiment. During the training phase, viewers were instructed to try and recognize the displayed expression, and then to judge the overall quality of the video sequence. The first task was given a particular emphasis, as our focus was on the ability of viewers to recognize the expressions, and we wanted to investigate how the quality was affected when the first task was set as the main effort. It was made clear to the participants that the two tasks are not related to each other. We especially made it clear that they should vote on the visual quality and not on the ease of recognition of the facial expressions.

A total of 30 participants took part in the experiment. Most of them were students, and the gender distribution was slightly in favor of male subjects. All participants were checked to have corrected-to-normal vision and for colour blindness. None of the participants were rejected according to these criteria. Each test session lasted between 30 and 35 minutes, including approximately 10 minutes for viewer introduction and training.

5. Experimental results

Our demonstration is based on the analysis of the expression recognition rate and the Mean Opinion Scores (MOS) obtained on the stimuli. The emotion recognition rate corresponds to the percentage of correct answers given by the participants when asked to recognize the emotion displayed in a video sequence. The MOS is processed as usual, as the average of quality scores over a given set of videos by all participants. We also use the 95% confidence intervals to indicate the statistical significance of the comparisons we make between MOS values.

5.1. Recognition rate

The overall recognition rate in our experiment is 88.58%. The standard deviation among participants is equal to 5.42, showing a relatively consistent performance for the 30 participants involved in the test. The average recognition rate within each encoding scenario varies between 86.0% and 92.8%, however the standard deviation is quite high, varying between 12.7 and 26.4. With such an order of magnitude, the difference between encoding scenarios in terms of recognition rate does not appear to be significant.

We get more insight by analyzing the influence of the displayed emotion and the scene composition. Table 2 presents the confusion matrices obtained for the highest- and the lowest-quality scenarios. First we note that the neutral and angry expressions are almost always perfectly recognized. We observed that sadness is well recognized too, but suffers from moderate loss in some conditions. Joy however tends to be mistaken for anger and neutral, even in the high-quality scenario. When the quality is at its lowest, we observe that these two emotions tend to be mistaken for a neutral face. A possible explanation is that our voting interface only offers a choice among the four existing emotions, and no 'unknown' option is available. The neutral emotion might be considered as a fall-back in case the emotion is not recognized. Nevertheless, post-test interviews would be needed to confirm this hypothesis.

Ref→	640x480@24:256				320x240@12:64			
↓Rec	**Ang.**	**Joy**	**Neu.**	**Sad.**	**Ang.**	**Joy**	**Neu.**	**Sad.**
Ang.	**94.2**	5.8	0.0	2.5	**95.8**	11.7	1.7	2.5
Joy	0.0	**87.5**	0.0	0.8	0.0	**65.0**	0.0	0.0
Neu.	0.8	5.0	**98.3**	1.7	0.0	21.7	**98.3**	8.3
Sad.	5.0	1.7	1.7	**95.0**	4.2	1.7	0.0	**89.2**

Table 2: Confusion matrices for the highest- and the lowest-quality scenarios.

The substantially lower recognition rate for joy could be explained by the two following hypotheses. First, the facial expression of joy in our system might not be representative enough and the participants might have been misleaded. However, we followed the directives of Ekman's system in order to recreate the facial expressions and validated it during our meeting with psychologists. A more detailed analysis taking into account the distance between the character and the camera goes in the direction of invalidating this hypothesis, as we will demonstrate later in this section. Secondly, the low recognition rate for joy might be due to properties of the expression itself. Joy is the only of the three expressions (excluding neutral) for which the upper part of the face does not move. For both anger and sadness, the eyebrows move and some wrinkles appear on the forehead. For joy, the mouth widens and the eyes are half shut, but the eyebrows and forehead do not move. This hypothesis goes in the direction of the results presented in [6]. The upper part of the face seems to have an important role in the perception of facial expressions. Again, post-test interviews would be needed to understand what made participants fail at recognizing the right emotion in the case of joy.

Looking at the individual stimuli sheds light on the influence of the scene composition and encoding parameters on the recognition rate. Figure 2 presents the recognition rate for all four expressions in the scenarios with the highest and the lowest quality. At high quality, the intensity of the emotion does not appear to have a great impact on the recognition. A drop of about 10% in recognition for anger and joy in the head shot is the only noticeable difference. At low quality, a slightly higher impact can be noticed for joy in the full-body shot. The distance between the character and the camera seems to have a

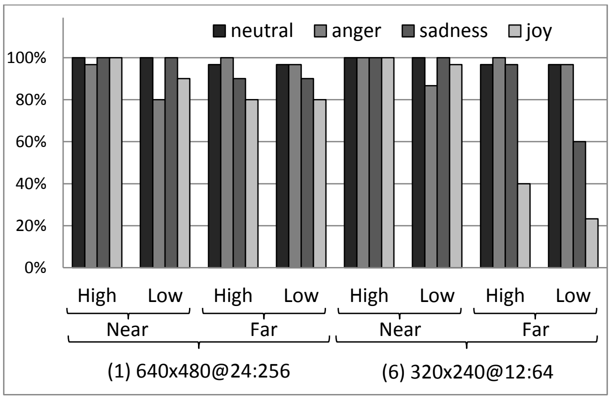

Figure 2: Recognition rate for the highest- and the lowest-quality scenarios. The stimuli are further grouped according to distance, intensity and associated emotion. "Near" and "Far" refer to the distance between the character and the camera. "High" and "Low" refer to expression intensity.

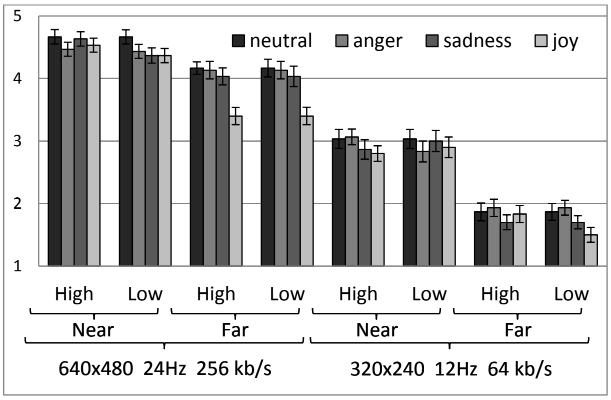

Figure 3: Mean Opinion Scores and their confidence intervals for the highest- and lowest-quality scenarios.

higher impact on recognition. The head-shot stimuli get a higher recognition than the full-body shot stimuli. Finally, it appears that distance is the main factor of loss in recognition for the stimuli picturing joy. This tends to invalidate the hypothesis of a non representative facial expression in favor of a dependency to distance, as the recognition rate is very high for the head-shot conditions.

Overall, the visual distortions affect the recognition rate, as expected. We noticed that the stimuli picturing sadness suffered a particularly high loss in recognition when the video sequence was spatially upscaled. We assume that the blur created by the upscaling step makes it difficult to identify the expression. All things being equal, temporal upscaling has the opposite effect. We observed an increase in recognition for joy and sadness in such conditions. As the bit budget per second is distributed among less frames in the configurations that use temporal upscale to get a video sequence at 24 Hz, the individual quality of each frame is higher, which appears to make recognition easier. Finally, encoding distortions affect the recognition rate, but not as heavily as blur due to upscaling. We observed that scenario 5 (640x480@24:64) allowed significantly better recognition than scenario 4 (320x240@24:128), even though the latter uses twice as much bitrate as the former. After a visual analysis of stimuli with a low bit-rate, we identified that the encoder was able to preserve the facial features, which are located in high spatial frequencies, in a fairly good way. However, upscaling blur affects every part of the image in the same way, which makes recognition of facial features more difficult.

5.2. Perceived quality

The relative quality of the encoding conditions confirm the results obtained in our previous work [14]. The encoding scenarios get decreasing MOS values as their resolution and bitrate decrease, with a wide spread along the entire quality scale and small variation among participants.

Looking at the MOS values of the individual stimuli brings more information regarding the influence of the facial expression and scene composition. Figure 3 presents the Mean Opinion Score obtained on the highest- and the lowest-quality scenarios. We note that within a given encoding scenario, the distance between the character and the camera has a significant impact on the perceived quality. When the character is far away from the camera, the facial features used to identify his expression are more likely to suffer from compression and blur, as they appear on a smaller area of the image.

The intensity of the expression does not have a systematic impact on the perceived quality. The expression itself does not have an impact in the case of netural, anger and sadness, as these three expressions reach equivalent levels of quality. However, the quality of joy is significantly lower than the other expresssions when combined with the full-body shot at high quality. This observation is less visible for the low quality scenario, which is imputable to the saturation effect at the lower end of the quality scale. However, we observed similar differences for the intermediate quality scenarios. The full-body shot for joy systematically gets a lower quality rating than other expressions. Some hypotheses to explain this phenomenon are given in the next section, where we analyse the link between recognition rate and perceived quality.

5.3. Quality versus recognition

We observed a moderate impact of the distance between the character and the camera on recognition. Distance also has an influence on the perceived quality. Furthermore, we showed that the stimuli picturing joy tend to have both low recognition rate and lower perceived quality. We would like to point out that we specifically instructed the participants that they should rate the quality of the stimuli according to the overall sequence, and not judge only on the face of the character. All things being equal, this might show a link between the ability to recognize an expression and the overall perception of quality. However, more evidence would be needed in order to give reliable conclusions.

6. Discussion & conclusion

In this paper we used a realistic face simulator to evaluate the influence of scene composition and encoding conditions on the ability of participants to recognize facial expressions. We also evaluated the impact of these parameters on the perceived quality. We identified a significant impact of the scene composition on both recognition and quality, through the distance between the character and the camera. The encoding scenario has an impact on the perceived quality, and an impact on recognition in extreme cases. Particularly, we showed that blur due to upscaling affects the recognition ability, and should be avoided.

We identified that the facial expression picturing joy suffers from lower recognition than the three other expressions. This could be linked either to the nature of the expression itself, or to the lack of representativeness of our implementation. This might also be a proof of the link between the ability to recognize an expression and the perceived quality. Indeed, we observe that those configurations that lead to a low recognition rate also get a significantly lower quality level.

The ability of human subjects to recognize facial expressions proves to be quite resistant to visual distortions. We observed that the lower-quality conditions substantially impaired the visibility of the facial features. However, the participants still reached impressive performance in most cases. We also noticed more observer variability for quality than for expression recognition, showing that the latter task is more natural for human viewers than the former.

This study could be extended in several ways. First, we could reconduct the experiment with more difficult conditions and measure the perception of quality when the distortions make recognition impossible. We could also investigate on the influence of other parameters such as lighting conditions or camera shaking, which are two major impairments in terms of quality. An interesting direction would be to let human subjects choose which combination of facial muscles is the most representative of each facial expressions, as a prior round of experiments. We would then get a more reliable set of expressions and could repeat our test. This study could also be repeated using human actors instead of a face simulator. This would alleviate the issues due to the uncanniness usually felt with virtual characters. However it would require more budget and more efforts, as working with human actors is not as flexible as working with a virtual environment.

7. Acknowledgements

We would like to address our grateful thanks to Leon Beutl who was working on the face simulator at the University of Vienna. He implemented the face simulator and its graphical interface, and helped us creating the stimuli used in our experiment. His reliability and responsiveness were greatly appreciated.

8. References

[1] P. Ekman. Facial expression and emotion. American Psychologist, 1993.

[2] B. Montagne, R. P. C. Kessels, E. H. F. De Haan and D. I. Perrett. The Emotion Recognition Task: a paradigm to measure the perception of facial emotional expressions at different intensities. Perceptual and motor skills, 2007.

[3] B. Klare and M. Burge. Assessment of H.264 video compression on automated face recognition performance in surveillance and mobile video scenarios. SPIE Biometric Technology for Human Identification, 2010.

[4] A. OToole and H. Abdi. Fusing face-verification algorithms and humans. Systems, Man, and Cybernetics, 2007.

[5] Joint Video Team. JSVM Reference Software V9.18, 2009. http://ip.hhi.de/imagecom_G1/savce/downloads/

[6] A. Tinwell, M. Grimshaw, D. Nabi, and A. Williams. Facial expression of emotion and perception of the Uncanny Valley in virtual characters. Computers in Human Behaviour, 2011.

[7] E. Reategui, E. Boff, and J. Campbell. Personalization in an interactive learning environment through a virtual character. Computers and Education, 2008.

[8] M. Powers and N. Francesca. Do conversations with virtual avatars increase feelings of social anxiety?, Journal of anxiety disorders, 2013.

[9] S. Kang and J. Watt. The impact of avatar realism and anonymity on effective communication via mobile devices, Computers in Human Behavior, 2012.

[10] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya. Visual attention in quality assessment. IEEE Signal Processing Magazine, 2011.

[11] S. Persky. Employing immersive virtual environments for innovative experiments in health care communication. Patient education and counseling, 2011.

[12] O. Kothgassner, A. Felnhofer, L. Beutl, H. Hlavacs, M. Lehenbauer and B. Stetina. A Virtual Training Tool for Giving Talks. International Conference on Entertainment Computing, 2012.

[13] P. Ekman et al. Facial Action Coding System. www.face-and-emotion.com, 2002.

[14] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pepion, P. Le Callet, Subjective quality of SVC-coded videos with different error-patterns concealed using spatial scalability, IEEE EUVIP, 2011.