International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:5, 2012

Video Quality Assessment using Visual Attention Approach for Sign Language

Julia Kucerova, Jaroslav Polec and Darina Tarcsiova

Abstract—Visual information is very important in human perception of surrounding world. Video is one of the most common ways to capture visual information. The video capability has many benefits and can be used in various applications. For the most part, the video information is used to bring entertainment and help to relax, moreover, it can improve the quality of life of deaf people. Visual information is crucial for hearing impaired people, it allows them to communicate personally, using the sign language; some parts of the person being spoken to, are more important than others (e.g. hands, face). Therefore, the information about visually relevant parts of the image, allows us to design objective metric for this specific case. In this paper, we present an example of an objective metric based on human visual attention and detection of salient object in the observed scene.

Keywords—sign language, objective video quality, visual attention, saliency

I. Introduction

THE recent century became a golden age in the area of technical innovations. One of the most widespread innovation is video in all its variations like cinema, television, videoconference etc. For some groups of people the video become important part of their lives, which can improves theirs quality of life. One of these groups are deaf and/or hard of hearing people.

Generally video consist from image and sound part. However deaf people usually are not affected by sound, so the subjective video quality evaluation can differ from hearing people. Also purpose of the video in sign language differs as it is the equivalent to sound channel in normal audiovisual recordings. In comparison with hearing people, for hearing-impaired people video quality means something different. It is very important for them whether they are able to understand the meaning.

The main difference between the terms quality and intelligibility is that the term quality describes the appearance of decoded video signal and the intelligibility is just one aspect of quality saying if the received information gives any sense.

Our main aim is to find the objective metric at the pixel level to evaluate the quality video signal quality encoded in various bit-rates, to achieve full intelligibility of Slovak (or other) sign language and finger alphabet.

J. Kucerova with the Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava (e-mail: kucerova@sccg.sk).
J. Polec is with the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology Bratislava (e-mail: jaroslav.polec@stuba.sk).
D. Tarcsiova is with the Faculty of Education, Comenius University in Bratislava (e-mail: tarcsiova@fedu.uniba.sk).

We decide to use visual attention method for detecting salient parts of the processed video and subsequently use the obtained information for new objective metric at pixel level.

This paper is organized as follows: The Sign Language and Finger alphabet is discussed in section II. In section III., the Intelligibility is described. In section IV., we describe objective methods for video quality evaluation for sign language. Algorithms for recognition of different regions of interest are described in section V. In section VI., are described reference objective method. Section VII. focuses on Visual attention approach. In section VIII., the paper presents the results via evaluation of quality and section IX. concludes the paper.

II. Sign Language and Finger Alphabet

Sign language is the primary communication tool deaf and/or hard of hearing people. It has its own grammar and vocabulary. Sign language is visual and spatial language which is independent of spoken language. It uses three-dimensional space (the sign space) for communication, which is defined horizontally and vertically. However sign language similarly to spoken language is not international, Slovaks used Slovak sign language. In sign languages, there are two types of components (parameters), which we can be analyzed [2]:

- manual parameters = location, handshape and movement
- non-manual parameters = facial expression, position of eyes, head, upper body, mouth movement

The basic communication element is sign. It is given by configuration (shape and placement) of the hands in the sign space, by palm and finger orientation, and also by movement of hand. To learn sign language it is necessary to have a personal presentation or video presentation as even slight difference in movement and location of the hand can change the meaning. This is the reason why the quality of video in sign language is very important.

III. The Intelligibility (Recognizability)

The intelligibility of the language (Z) can be defines as the percentage of correctly received elements or parts of speech (a) divided by their total number (b) [5]:

$$Z = \frac{a}{b}.100\% \quad (1)$$

Using the formula above, we can explore the intelligibility of video recordings: sentence and word intelligibility using sign.

IV. Objective Methods for Evaluating Video Quality for Sign Language

In this paper, we propose content based objective metric for evaluating video quality for sign language. Commonly the best results in evaluating video quality are provided by subjective methods. However in this specific case, this is not always true. The volunteers are not able to rate the quality as they do not know if they understand the sentence correctly, so the results have to be reviewed by someone else. That is why the subjective evaluation of quality is time consuming and administrative difficult to ensure that all viewers have standard viewing conditions according to ITU-T recommendations [3]. Due to this fact we have decided to design an objective content based method for evaluating video quality.

In our method, the visual attention and semantic information are used as a main criterion for quality evaluation. For observers, different parts of the image are important with different weights. The important parts are called salient regions and they can bring important information to observer. Taking into consideration the fact that the average human perceives different parts of video in different way, we have preprocessed the standard metrics PSNR, VQM [13] and SSIM [7] using the content based algorithm. In this paper, we used the algorithm that can detect salient regions of the considered scene.

Our visual attention model detects regions of different color, intensity and texture. As an semantic information, we used skin detection. To demonstrate that our method provides reliable results, we chose another objective method as a reference for comparison. Namely, we used PSNR, VQM, SSIM, 3-component SSIM index presented in [8] and other content-weighted video quality metrics presented in [1].

V. Algorithms for Recognition Different Regions of Interest

In comparative methods (3NCC, 4NCC etc.) we used two algorithms for recognition three and four different regions of interest. The main aim was to adapt to human perceiving quality as close as it is possible. That is why the algorithm was used to preprocess the video sequence and prepared it for quality evaluation made by mutual information. The first of mentioned algorithm is able to divide frame into edges, smooth region and texture region. It consists from following steps [8]:

1) Compute the gradient magnitudes for reference and distorted frame by using a Sobel operator.
2) Determine thresholds T_1 and T_2,where $T_1 = 0,12.g_m$ and $T_2 = 0,06.g_m$. Variable g_m represents maximum gradient magnitude value calculated from reference frame.
3) It this step each particular pixel is assigned to the three different regions (edges, smooth region and texture) in the following way:
   - Pixel $x(i,j)$ belongs to edge region if $g(i,j) > T_1$ or $\hat{g}(i,j) > T_2$.

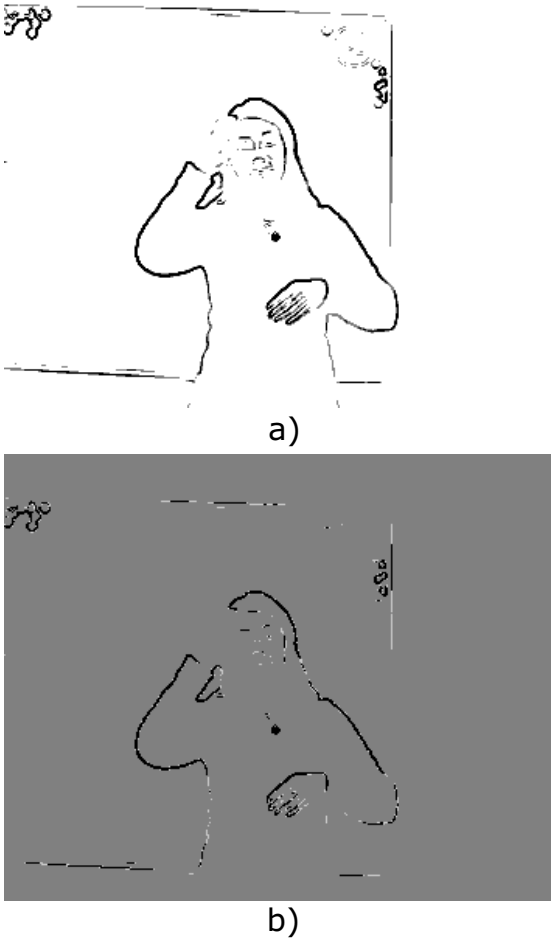


Fig. 1. Masks for 25th image taken from the original video and decoded video using h.264, QP30: a) three regions, b) four regions (image intensity: smooth 255, edge 0, texture 128, edge change 200)

   - If condition $g(i,j) < T_2$ or $\hat{g}(i,j) <= T_1$ is met consider pixel $x(i,j)$ as a part of the smooth region.
   - If the pixel $x(i,j)$ does not meet any of above criteria, it is part of the texture region.
4) After each pixel is assigned to the particular region, every pixel is weighted according to the region where it belongs.

The $g(i,j)$ is value of gradient magnitude of pixel at position (i,j) in original frame and $\hat{g}(i,j)$ stands for gradient magnitude of pixel at position (i,j) in distorted frame.

The second algorithm used for preprocess video sequence is able to divide video frame into four different region i.e. edge, smooth region, texture and changed edge. The steps are quite similar to the previous version of algorithm with three ROIs but there are some significant modifications [14]:

1) Compute the gradient magnitudes for reference and distorted frame by using Sobel operator. This step is

similar to previous algorithm.

2) Determine thresholds T_1 and T_2, where $T_1 = 0,12.g_m$ and $T_2 = 0,06.g_m$. Variable g_m represents maximum gradient magnitude value calculated from reference frame. This step is similar to previous algorithm.
3) Assign each particular pixel to four different regions (edges, smooth region, texture and changed edge). Conditions for assigning the pixel to regions differs from previous algorithm:
   - If $g(i,j) > T_1$ and $\hat{g}(i,j) > T_1$ consider pixel $x(i,j)$ as a part of edge region.
   - Pixel $x(i,j)$ belongs to changed edge region if the value of gradient magnitude meet conditions ($g(i,j) > T_1$ and $\hat{g}(i,j) <= T_1$) or ($\hat{g}(i,j) > T_1$ and $g(i,j) <= T_1$).
   - If $g(i,j) < T_2$ and $\hat{g}(i,j) > T_1$ consider pixel $x(i,j)$ as a part of smooth region.
   - Otherwise pixel $x(i,j)$ belongs to texture region.
4) After each pixel is assigned to the particular region, every pixel is weighted according to the region where it belongs.

In Figure 1 b) masks for 25th image from original and coded tested video sequence are shown. For coding h.264, QP30 coder was used. In the experiment follow settings were used: edge = 0.7, smooth = 0.6, texture = 1 and edge change = 0.2.

VI. Reference Objective Method

The reference objective method was proposed in [5]. It was used for examining the sentence intelligibility (as used in telephonometry for speech sentence articulation) with use of subjective ACR method (full categorical evaluation). It leans on the respondent's ability to rewrite the sentence which was presented in the Slovak sign language into the Slovak language. The main criterion of reference method was to consider if the sentence is correct.

The quality scale has four different numerical values and their representative descriptions. This quality scale is shown on following table:

TABLE I
Proposed Poting Options for Sentence Intelligimility Testing

1	Completely understandable
2	Partially understood, but understood the content
3	Partially understood, but misunderstood the content
4	Not understandable

The main disadvantage of the reference method is that the criteria for evaluating video quality allows to consider sentence as correct in many cases where the sentence should look very different as there is some natural variability in translation from and to the sign language. The way how the reference method evaluates quality is not very usable for the system in which the evaluation of the quality is needed in the real time.

VII. Visual Attention Approach

Attention is the process of concentrating on specific features of the environment, or on certain thoughts or activities. It has a large effect on what we are aware of, on perception, on memory, on language, and on solving problems [9]. Visual attention is the ability of a vision system to detect salient objects in an observed scene. Human visual system is sensitive to features like changes in color, shapes, intensity etc.

In past few years, a lot of different visual attention models were proposed. They use different features for detecting salient regions in scene, such as color, intensity, orientation, texture, etc.

For an observer, the important changes are in low level features, but from semantic point of view we are interested in detecting faces, humans, text etc. For the sake of our research we have decided to use model presented in [10], which is based on Hu's approach [11]. Hu's approach is based on detecting three features: color, intensity and texture. This model uses local context information to suppress spurious Attention Regions, while simultaneously enhancing the true Attention Regions. This is useful to capture visual attention in images containing small objects, but it fails in images containing faces, as it is not able to detect faces as salient objects.

This problem is solved in [10], where face detection is used as an additional attention cue. This particular model uses four features (color, intensity, texture, face detection) and their combination.

The model has three main parts:

- creation of contrast maps for color, intensity, texture and map for skin detection
- creation of map for suppression factor
- creation of saliency map

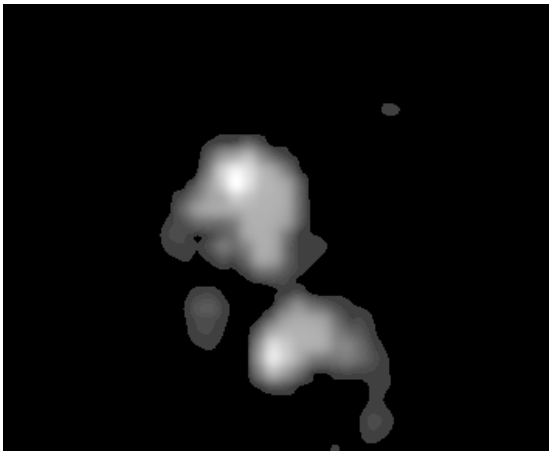
Creation of contrast maps

The detailed process of creating the texture map is described in [11].
Local context suppression strategy for adaptive combination of multiple attention cues like intensity, color and texture is describe here. Consider an image divided into blocks, called an *Attention Patches*, each containing $p \times q$ pixels. The contrast of particular feature at a patch centered at (i,j) is calculated as

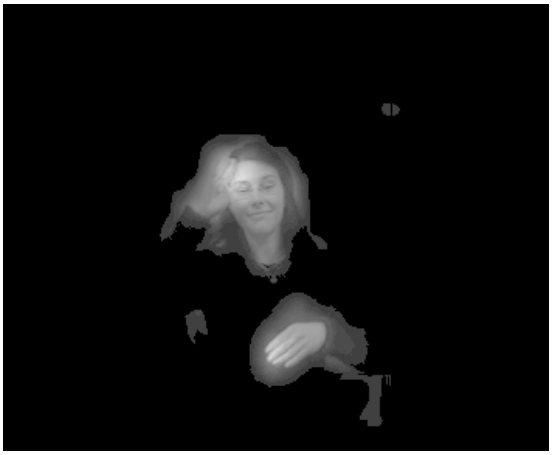
$$FV(i,j) = \frac{1}{N} \sum_{u,v} |MF(i,j) - MF(i+u, j+v)|, \qquad (2)$$

where $MF(i,j)$ is the mean of the feature in patch (i,j) and N is the number of patches in its neighbourhood. The contrasts at patch (i,j) for n features/attention cues are normalized to lie between $[0,1]$. Each patch is now represented by the n dimensional feature contrast vector which is compared with other feature contrast vectors in its neighbourhood and its contrast measure is suppressed if the patch and its neighbours

a)



b)



c)

Fig. 2. Image taken from the experiment: a) original b) saliency map c) product of the original image (only Y canal from YUV color space) and the saliency map.

are 'similar'. This similarity is estimated by the variance of data along eigen vectors of an $n \times n$ covariance matrix. This matrix is formed from the feature contrast vectors at a patch (i, j) and its neighbourhood. The eigenvalues $\bar{\lambda}$ of this matrix represent the extent of similarity or dissimilarity among the attention cues. For example a large eigenvalue indicates large variance along the direction of its corresponding eigen vector, which implies higher discriminating power [11].

On account of semantic amount of human faces and hands (in our testing data), we decided to replace face detection feature with skin detection feature. Skin color distribution used in this paper is modeled using a single 2D Gaussian distribution [12].

Suppression factor

The suppression factor SF for patch (i, j) is obtained as $\tau(i, j) = \prod_{u=1}^{p} \bar{\lambda}_u$, where the $\bar{\lambda}_u$'s are sorted in ascending order and the parameter p controls the degree of suppression. For obtaining the saliency $S(i, j)$ for patch (i, j) the multiple attention cues are linearly combined and the result is modulated by the SF as

$$S(i, j) = \pi(i, j) \times \sum_{u,v}^{k} FV_u(i, j), \tag{3}$$

The product of the combined map and the SF yields the final saliency map which contains the true Attention Regions. In the combined map there are spurious attention regions. Using Suppression Factor, these regions have been successfully removed [11].

Saliency map

To get the saliency map, we need to combine maps for features together. As a first step of feature combination, we sum together and normalize three contrast maps (color, intensity, texture) to get the Combined map. We derive the suppression factor by building up the suppression map from the three previously mentioned contrast maps. We combine Combined map with the skin detection map to get the suppression factor. Suppression factor is a map consisting of darker regions representing high suppression factor and brighter regions representing low suppression factor, i.e. brighter regions are more significant than darker regions. Consequently we multiply this suppression map with Combined map. Using this process we have constructed final Saliency map for input image [10].

In Figure 2 b), we can see the results of the saliency map applied to the input image Figure 2 a). The darker regions represent less important parts of the image, the brighter ones the important parts; values are normalized to the 0-1 interval. We combined the saliency map with the Y canal from the YUV color space to get the final image, see Figure 2 c). It is easy to see that our method suppresses less important features of the input image, and focuses on the significant parts, as desired.

Using this approach, we can divide the image information into 256 importance groups and use them for further utilization.

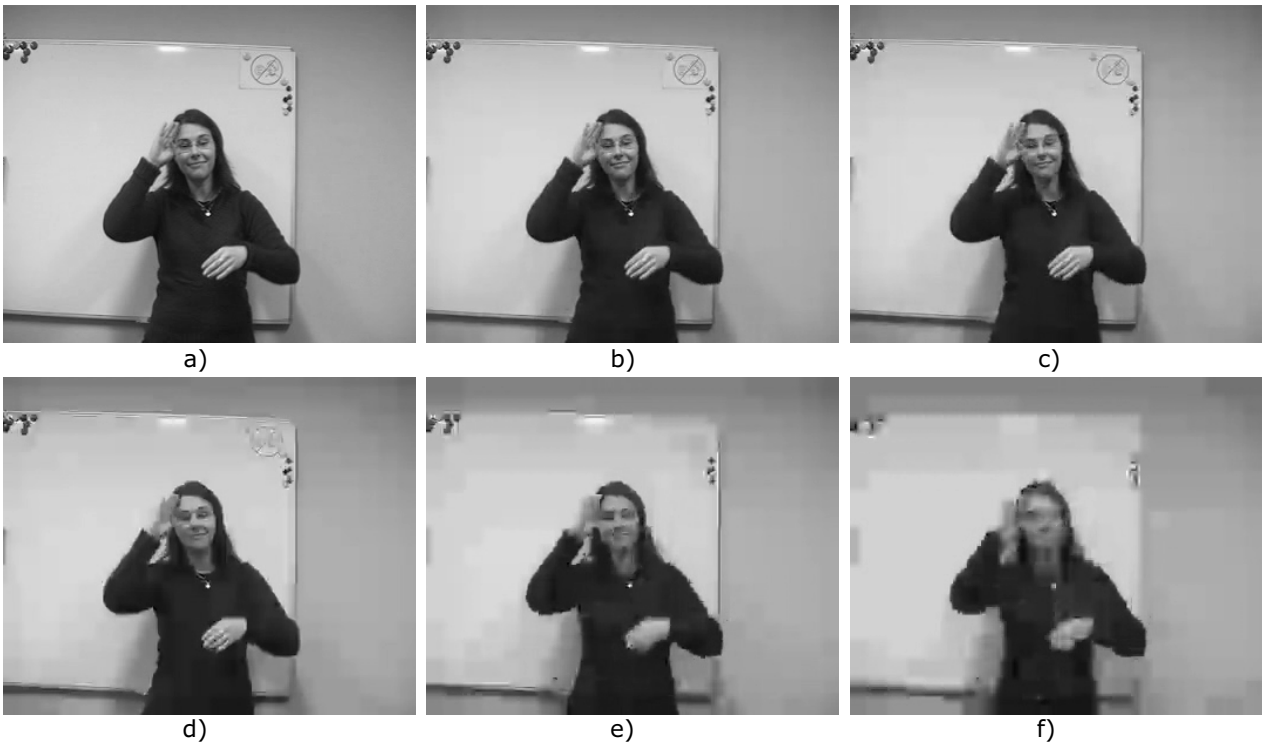


Fig. 3. 25th picture taken from the experiment: a) Original image; H.264 decoded frame with parameter: b) QP=30 c) QP=35 d) QP=40 e) QP=45 f) QP=50

VIII. Evaluation of Quality

Saliency regions detected by visual attention model give us information about important parts of the observed scene. The main goal of the proposed method is to design an objective method based on visual attention.

Results were obtained via processing of multiple video previews with different example phrases in Slovak sign language. The original input footage was examined in three speed variations - slow, normal and fast. For the experiment, the standard video format of 352x288 pixels per frame with 25 frames per second was used. Subsequently, these recordings were encoded by H.264 codec with various bit rates (QP = 30, 35, 40, 45, 50 that correspond to rates from 87 to 18 kbps respectively).

Each video sequence was shown to the group of 14 hearing impaired observers independently. Every observer stated, what was understandable for him. Afterwards, one independent observer evaluated the intelligibility of this video according to the Table I. For each QP was compute average intelligibility. Later it was used as an reference objective criterium for real-time objective methods, using Pearson's correlation.

Results from objective methods were processed according to [6]. The score of particular objective metric was calculated for each frame in every video sequence. Afterwards, the mean values were used for correlation coefficients calculation. The Pearson's linear correlation coefficient was calculated to describe the relations between the metrics and the model accuracy.

For comparison, our results were tested using the VQM, SSIM and 3SSIM and other metrics, such as: 3 and 4 - normalized cross-correlation measure (3NCC, 4NCC), 3 and 4 - angular distance (3ANG, 4ANG), 3 and 4 - moments of angle (3MOM, 4MOM). Our method is marked as SPSNR, SSSIM, SVQM respectively.

For SSIM1, we decided to use the *Attention Patches* containing 2×2 pixels. For SSIM2, we used the *Attention Patches* with 8×8 pixels.

Results for Pearson's correlation between intelligibility and other different metrics are shown in Table II. Results are listed in descending order by successfulness.

IX. Conclusion

In this paper, we have described the technique of real-time evaluation of the sentence intelligibility in gesture sign language, and we have also presented the obtained results. We focused mainly on the sentence intelligibility, where (under certain assumptions) it is possible to guess the missed words from the context.

In this paper is shown that the increasing number of weight per pixel gives better results than 3SSIM. We have formulated this argument for this specific video category. It would be very interresting to analyze the results of this method for other

TABLE II
PEARSON'S CORRELATION COEFFICIENT FOR EVERY OBJECTIVE METRICS

Objective metric	Pearson's correlation	Order
PSNR	0,91	13
SPSNR 1	0,91398	11
SPSNR 2	0,910857	12
SSIM	0,964678	7
3SSIM	0,93052	9
SSSIM 1	0,97869	2
SSSIM 2	0,976388	3
VQM	0,97624	4
SVQM 1	0,98447	1
SVQM 2	0,97434	5
3NCC	0,8403617	16
3ANG	0,907904	14
3MOM	0,9515367	8
4NCC	0,8141467	15
4ANG	0,924125	10
4MOM	0,9664193	6

different video input. We assume, that this method would be effective for any types of sign language communication (e.g. MAKATN), and in the video communication of the entire population, as well.

In our next work, we will further investigate the methodology of evaluating the quality of video signals based on logatom recognizability for one - handed double-handed finger alphabets.

ACKNOWLEDGMENT

Research described in the paper was financially supported by the Slovak Research Grant Agency VEGA under grant No. 1/0602/11.

REFERENCES

[1] M. Mardiak, *Video quality assessment*, (in Slovak), Slovak University of Technology. Bratislava, 2012.

[2] D. Tarcsiova, *Pedagogics of hearing-impaired*, (in Slovak), MABAG spol. s r. o., Bratislava, 2008.

[3] ITU-R, *Methodology for the subjective assessment of the quality of television pictures*, International Telecommunication Union Radiocommunication Sector, Tech. Rep. BT.500-11, 2002.

[4] F. Makan, *Elektroacoustics*, (in Slovak), Slovak University of Technology. Bratislava, 1995.

[5] P. Heribanova, J. Polec, S. Ondrusova, M. Hostovecky, *Intelligibility of cued speech on video*, World Academy of Science, Engineering and Technology, Iss. 79, pp. 492-496, 2011.

[6] ITU-T, *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*, Recommendation J.144, 2004.

[7] Z. Wang, L. Lu, and A. C. Bovik, *Video quality assessment based on structural distortion measurement*, Signal Process. Image Commun. 19, 2004, pp. 121132, 2004.

[8] Ch. Li, A. C. Bovik, *Content-weighted video quality assessment using a three-component image model*, Journal of Electronic Imaging, 19(1), 011003-1-9, , 2010.

[9] Goldstein E. B.: *Cognitive Psychology: Connecting Mind, Research and Everyday Experience*. ISBN-10: 0495095575 ISBN-13: 9780495095576, Thomson/Wadsworth, 2008.

[10] J. Kucerova, *Saliency Map Augmentation with Facial Detection*, CESCG 2011, Proceedings of the 15th Central European Seminar on Computer Grapgics. - Vienna : Institute of Computer Graphics and Algorithms, ISBN 978-3-9502533-7, Pages 61-66, 2011.

[11] Y. Hu et al., *Adaptive Local Context Suppression of Multiple Cues for Salient Visual Attention Detection*. In IEEE International conference on multimedia and expo, Pages 1-4, 2005.

[12] E. Sikudova, *Comparison of color spaces for face detection in digitized paintings*, In:Spring Conference on Computer Graphics : SCCG 2007 : Conference Proceedings. Bratislava : Comenius University, ISBN 978-80-223-2292-8, Pages 135-140, 2007.

[13] F. Xiao, *DCT-based Video Quality Evaluation*, MSU Graphics and Media Lab (Video Group), 2000.

[14] Ch. Li, A. C. Bovik, *Content-partitioned structural similarity index for image quality assessment*, Signal Processing: Image Communication, 25 (2010)517526.

Julia Kucerova was born in 1987 in Detva, Slovak Republic. She received M.Sc. degree in Geometry from the Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava in 2011. She is a PhD. student of Informatics at the same university. Her research interests include visual attention and image coding.

Jaroslav Polec was born in 1964 in Trstena, Slovak Republic. He received the Engineer and PhD. degrees in telecommunication engineering from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in 1987 and 1994, respectively. Since 1997 he has been associate professor and since 2007 professor at the Department of Telecommunications of the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology and since 1998 at the Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics of the Comenius University. His research interests include Automatic-Repeat-Request (ARQ), channel modeling, image coding, interpolation and filtering.

Darina Tarcsiova received the M.Sc. and PhD. degrees in Special Education from the Faculty of Education, Comenius University. She is professor at Institute of Special Education Studies of the Faculty of Education, Comenius University. Hers research interests include special education for deaf people (sign language, finger alphabets).