World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:7, No:6, 2013

Key Frames Extraction for Sign Language Video Analysis and Recognition

Jaroslav Polec, Petra Heribanová, and Tomáš Hirner

***Abstract*—**In this paper we proposed a method for finding video frames representing one sign in the finger alphabet. The method is based on determining hands location, segmentation and the use of standard video quality evaluation metrics. Metric calculation is performed only in regions of interest. Sliding mechanism for finding local extrema and adaptive threshold based on local averaging is used for key frames selection. The success rate is evaluated by recall, precision and F1 measure. The method effectiveness is compared with metrics applied to all frames. Proposed method is fast, effective and relatively easy to realize by simple input video preprocessing and subsequent use of tools designed for video quality measuring.

***Keywords*—**Key frame, video, quality, metric, MSE, MSAD, SSIM, VQM, sign language, finger alphabet.

I. Introduction

SIGN languages normally make use of both static and dynamic gestures to achieve communication goal. Our goal is to determine a set of keyframes that represent the sign of finger alphabet in video with Slovak sign language [19]. By [1], related key frame extraction techniques are based on approach: shot boundary [2], visual content [6], motion analysis [7], shot activity [8]. The first two approaches to key frame extraction are relatively fast. However, they do not effectively capture the visual content of the video shot, since the first frame is not necessarily a key frame. The last two approaches are more sophisticated due to their analysis of motion and activity. However, they are computationally expensive and their underlying assumption of local minima is not necessarily correct. Methods [9], [11], [13] are based on motion energy (ME) and on the simple idea that the more motion in the scene, the more interest of people should be attracted. Local maximal or minimal ME, related to the motion magnitude, is usually employed as the metric for key-frame extraction.

Song and Fan in [14] use a unified spatio-temporal feature space to characterize the video data. They jointly perform key frame extraction and object segmentation by maximizing the divergence between objects in the feature space.

An entropy-based method was introduced in [15] where the entropy of a grayscale frame is computed and compared with that of the previous frame. If the entropy difference is higher than a user defined threshold, then the new frame is assumed to be a key frame.

In [16], shot boundary detection and key frame extraction algorithm based on Singular Value Decomposition (SVD) is presented. The algorithm extracts low-cost, multivariate color features from the video and constructs a 2D feature matrix. The matrix is then factorized using SVD. The algorithm processes the video in an online fashion.

By [17] key frames selection approaches can be classified into cluster-based methods, energy minimization-based methods and sequential methods. The clustering techniques take all the frames of a shot together and classify them according to their content similarity. Then, key frames are determined as the representative frames of a cluster. The disadvantage of these approaches is that the temporal information of a video sequence is omitted. The energy minimization based methods extract the key frames by solving a rate-constrained problem. These methods are generally computational expensive, since they use iterative techniques to perform minimization. The sequential methods consider a new key frame when the content difference from the previous key frame exceeds a predefined threshold that is determined by the user. The proposed method belongs to sequential methods considering visual content.

II. Video Quality Metrics

The main goal in the objective quality assessment research is to design metric, which can provide sufficient quality evaluation in terms of correlation with the subjective results.

In this paper we proposed comparison of four objective metrics used for key frame extraction. Key frames represent the signs in sign language.

A. MSAD

MSAD is a widely used, extremely simple algorithm for measuring the similarity between image blocks. MSAD for images **X** and **Y** with dimension *M* x *N* is expressed as [12]:

$$MSAD = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\left|X(m,n)-Y(m,n)\right|}{M \cdot N} \qquad (1)$$

The value of this metric is the mean absolute difference of the color components in the correspondent points of image. This metric is used for testing codecs and filters.

J. Polec and T. Hirner are with the Institute of Telecommunications, University of Technology, Ilkovičova 3, 812 19, Bratislava, Slovak Republic (phone: +421268279409, e-mail: polec@ktl.elf.stuba.sk, tomas.hirner@gmail.com).

P. Heribanová is with the Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava (e-mail: petra.heribanova @gmail.com).

B. MSE

MSE is the simplest and the most widely used full-reference quality metric. The MSE can be calculated for two images as follows [12]:

$$MSE = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\left(X(m,n) - Y(m,n)\right)^2}{M \cdot N} \tag{2}$$

MSE is based on the assumption that human observer is sensitive to the summed squared deviations between reference and test sequences, and is insensitive to other aspects like spatial and temporal frequency or color of the deviations. This extremely simple assumption restricts its usage. It fails for some common circumstances.

C. SSIM

SSIM index (Structural similarity - based image quality assessment) is based on measuring of three components (luminance similarity, contrast similarity and structural similarity) and combining them into result value [4].

D. VQM

VQM uses DCT to correspond to human perception [5]. VQM is based on a simplified human spatial-temporal contrast sensitivity model.

III. The Proposed Method

Presented method is proposed with aim to find the particular signs in video sequence with sign language. It is based on comparison of two successive frames. The comparison is performed by metrics, which are commonly used for quality evaluation of coded video or video damaged by other way.

The block diagram of the proposed system is given in Fig. 1.

A. Region of Interest Selection

Spatial gesture segmentation is the problem of determining where the gesturing hand is located in each video frame. Various gray-level segmentation techniques, such as use of single threshold value, adaptive threshold, P-tile method, edge pixel method, iterative method and use of fuzzy set are available for object segmentation. In arbitrary environments however, neither skin color nor any other color can be guaranteed to appear only within the object of interest, which is the hand. The algorithms based on skin color distribution are used also for face detection [18], [20]. Because of that next processing is necessary to differentiate between selected regions.

The ROI determination for each frame is performed in two steps - hand tracking and segmentation:

1. Mean shift based tracking extracts the color distribution of target appearance, and is implemented using kernel histogram [3].
2. A single threshold value is assigned to every image [10]. Here it converts the RGB image to a binary image (Fig. 2).

Metric is then applied only to the region of ROI assigned to the first of compared frames.

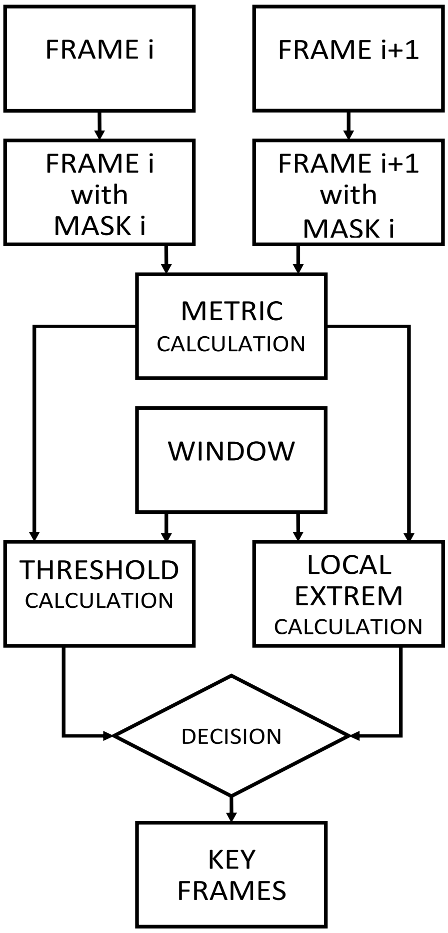
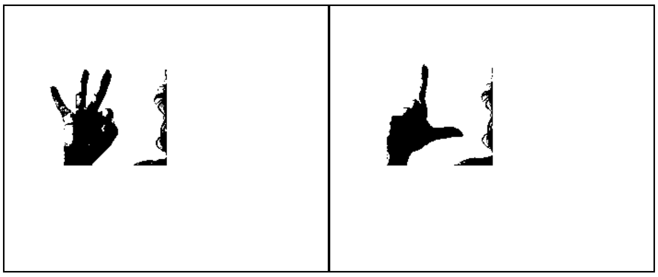


Fig. 1 Block diagram of the proposed system



56 565

Fig. 2 Region of interest masks

B. Decision Metric and Calculation of Threshold

We use the local threshold method whereby the frame difference of successive W frames is examined. The length of sliding window W has to be odd number, as the middle frame is examined and the threshold value is determined from left or right half of sliding window.

The middle sample represents a key frame if the conditions below are simultaneously satisfied:

1. The middle sample is the minimum in the window.
2. The middle sample is lower than threshold.

Mentioned method is valid for all metrics with minimal value for equality of compared sets (for example MSAD, MSE, RMSE, VQM).

For metrics, which reach maximum value for comparison of identical sets (for example SSIM, PSNR) is applied:

The middle sample represents a key frame if the conditions below are simultaneously satisfied:

1. The middle sample is the maximum in the window
2. The middle sample is greater than threshold

The local adaptive threshold is computed as

$$Th_L(i) = \frac{1}{\frac{W-1}{2}} \sum_{j=1}^{\frac{W-1}{2}} VQ(i-j) \tag{3}$$

$$Th_R(i) = \frac{1}{\frac{W-1}{2}} \sum_{j=1}^{\frac{W-1}{2}} VQ(i+j) \tag{4}$$

$$Th(i) = k \cdot \frac{Th_L(i) + Th_R(i)}{2} \tag{5}$$

where VQ is chosen metric for determination of compared frames dissimilarity.

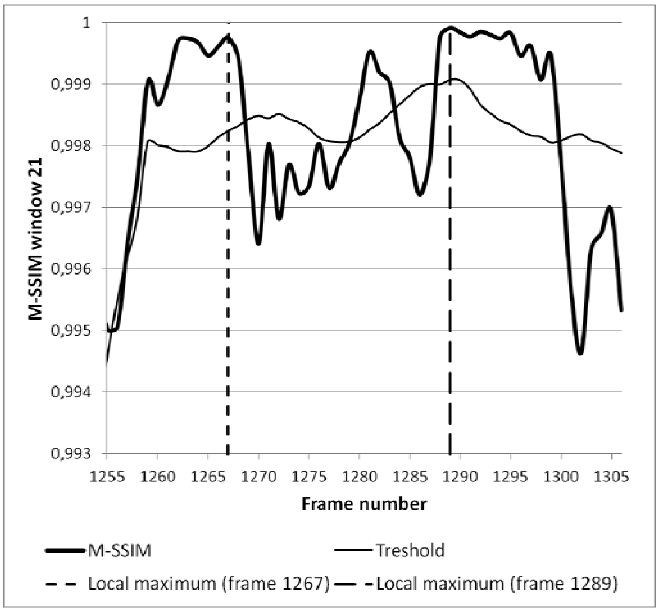


Fig. 3 Response of the decision metric in case of a key frame extraction

Fig. 3 shows the response of one of the decision metrics for a key frame extraction of the one-hand finger alphabet. The decision metric gives high values for similar scenes (one sign) and has a minimum peak in case of a sign change. Hence, a key frame decision can be given if the amplitude of the local maximum between peaks is higher than a certain threshold.

IV. Experiment

Three metrics used for the assessment of key frames extraction algorithms are the recall, precision and $F1$ rates. For the success rate evaluation of signs extractions we modify recall R, precision P and $F1$ rate as follows:

$$R = \frac{CS + DS + CP + DP}{CS + DS + CP + DP + MS + MP} \tag{6}$$

where CS denotes the number of correctly extracted signs, DS denotes the number of correctly extracted signs, where two frames are extracted for one sign, CP and DP analogically for word spaces, MS denotes the number of missed signs, and MP denotes the number of missed word spaces.

The recall measure, also known as the positive true function or sensitivity, corresponds to the ratio of correct experimental extractions over the number of all true extractions. The value of recall descends with ascending number of missed key frames.

$$P = \frac{CS + DS + CP + DP}{CS + DS + CP + DP + 0{,}5 \cdot DS + FP} \tag{7}$$

where FP denotes the number of false detected word spaces.

The precision measure is defined as the ratio of correct experimental extractions over the number of all experimental extractions. In other words, the value of precision is lower with higher amount of false extractions. We weight the false repetition of sign by half, because Slovak language has minimum of words, where the same letter is used two times consecutively.

$$F1 = \frac{2 \cdot R \cdot P}{P + R} \tag{8}$$

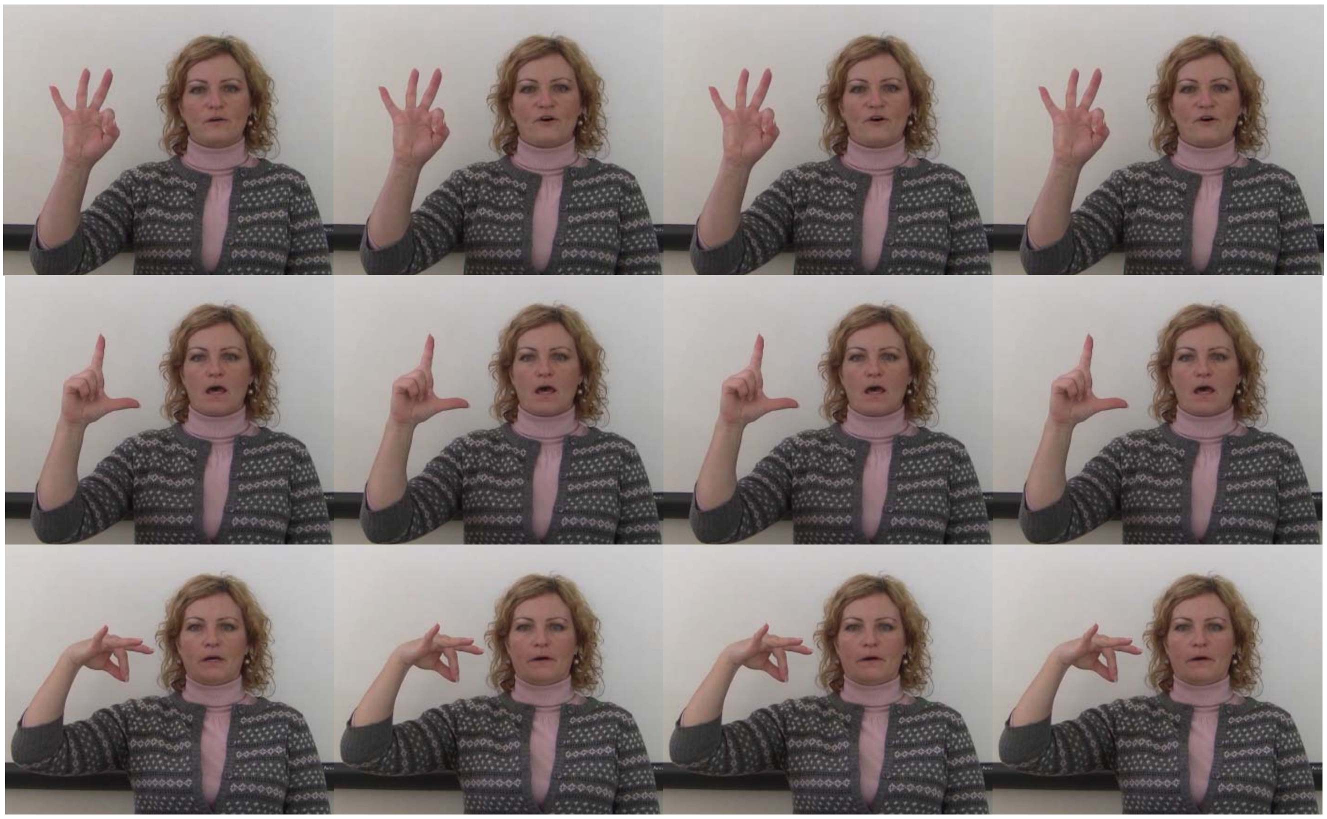
F1 score measure is a combined measure that results in high value if, and only if, both precision and recall result in high values. *F1* score gives more global look at the accuracy of examined key frame extraction algorithm, because it takes into account both missed key frames and false extractions.

Experiment was performed on video with one-hand finger alphabet. It contains 41 signs in seven logatoms (words without meaning) and 8 spaces.

For finding local extrema and threshold value determination we use windows W of size 19 and 21, the constant k from (5) was set to value 1.

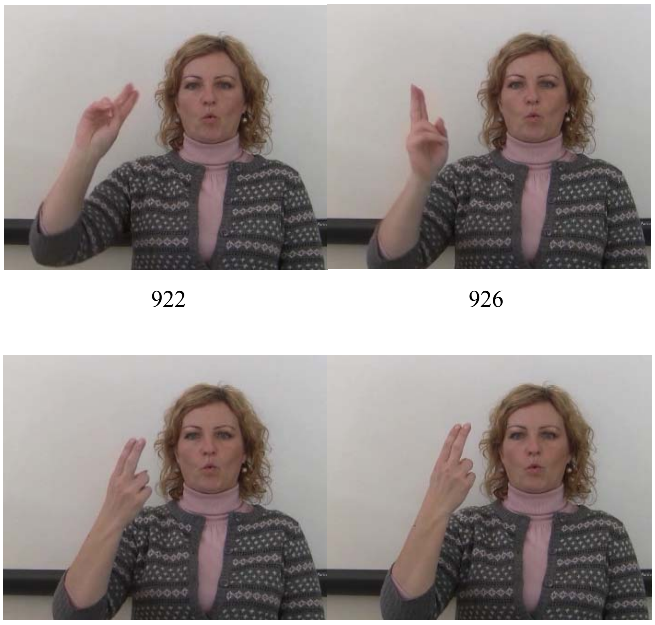
Fig. 4 shows some obtained key frames. The first M in metric name identifies, that the metric was applied only in ROI (mask). The following number represents the size of used window for threshold determination. Selected problematic signs, where key frame extraction can fail, are shown in Figs.

5 and 6. Numeric characteristics achieved in experiment are listed in Table I.



M-SSIM 21 M-VQM 21 M-MSE 19 M-MSAD 21

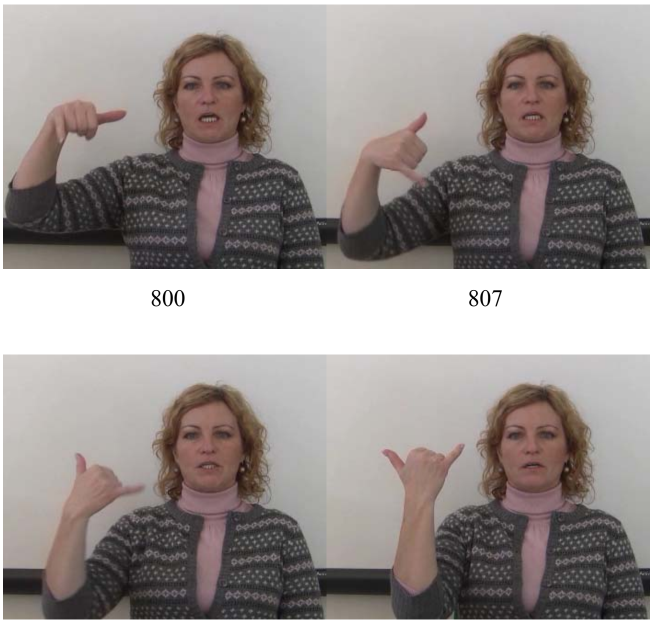
Fig. 4 Key frames obtained by different metrics



922 926

931 938

Fig. 5 An example of sign, which was often missed



800 807

812 822

Fig. 6 An example of sign, which was identified as two signs

TABLE I
KEY FRAME EXTRACTION PERFORMANCE COMPARISON

	MSAD	MSE	VQM	SSIM	M-MSAD	M-MSE	M-VQM	M-SSIM	M-MSAD	M-MSE	M-VQM	M-SSIM
	21	21	21	21	21	21	21	21	19	19	19	19
CS	35	29	34	35	36	35	37	37	34	35	32	33
DS	6	12	7	6	3	5	3	4	6	6	9	8
MS	0	0	0	0	2	1	1	0	1	0	0	0
CP	7	2	7	7	6	7	6	5	4	7	6	7
DP	1	6	1	1	2	1	2	3	4	1	2	1
FP	3	19	0	0	0	0	0	0	0	0	0	0
MP	0	0	0	0	0	0	0	0	0	0	0	0
R	1	1	1	1	0,959	0,979	0,979	1	0,979	1	1	1
P	0,890	0,662	0,933	0,942	0,969	0,950	0,969	0,960	0,941	0,942	0,915	0,924
F1	0,942	0,796	0,965	0,970	0,964	0,964	0,974	0,980	0,960	0,970	0,956	0,960

V. CONCLUSION

In this paper we presented a method for finding video frames representing one sign in the finger alphabet. The method is based on determining hands location, segmentation and the use of standard video quality evaluation metrics. Metric calculation is performed only in regions of interest. Sliding mechanism for finding local extrema and adaptive threshold based on local averaging is used for key frames selection. The success rate is evaluated by recall, precision and F1 measure. The method effectiveness is compared with metrics applied to all frames.

The achieved results indicate the proposed method is sufficiently successful and can be used for example during the process of quality assessment of video with sign language to evaluate only quality of individual video frames representing particular signs (key frames) instead of evaluating whole video sequence or eventually as preprocessing for automatic signs recognition in videos with sign language. For these purposes metric SSIM applied only to space of ROI - hand with adaptive threshold and windows of size 21 seems to be the most suitable.

ACKNOWLEDGMENT

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

REFERENCES

[1] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra, "Adaptive Key Frame Extraction Using Unsupervised Clustering", *Roc. of Int. Conf. on Image Proc.*, Chicago, Oct. 1998.

[2] A. Nagasaka, and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," in *Second Working Conference on Visual Database Systems*, 1992.

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 25, pp . 564-577, May 2003.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and Simoncelli, E.P., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 1-14, 2004.

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

[6] H. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, no. 4, pp. 643{658, 1997.

[7] W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE Int.* Conf. Acoust., Speech, and Signal Proc., 1996.

[8] P. O. Gresle, and T. S. Huang, "Gisting of video documents: A key frames selection algorithm using relative activity measure," in The 2nd Int. Conf. on Visual Information Systems, 1997.

[9] T. Y. Liu, X. D. Zhang, J. Feng, and K. T. Lo, "Shot reconstruction degree: a novel criterion for key frame selection," *Pattern Recogn. Lett.* 0167-8655 25, 1451–1457, 2004.

[10] Y. M. Abbass, W. Fakher, and M. Rashwan, "Arabic / English Identification in a hybrid complex documents images," *GVIP 05 Conference*, 19-21 December 2005, CICC, Cairo, Egypt.

[11] W. S. Chau, O. C. Au, and T. S. Chong, "Key frame selection by macroblock type and motion vector analysis," in 2004 *IEEE Int. Conf. on Multimedia and Expo*, Vol. 1, pp. 575–578.

[12] Cumar (22.10.2001), *An introduction to image compression* [Online]. Available: http://www.debugmode.com/imagecmp

[13] T. M. Liu, H. J. Zhang, and F. H. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Trans. Circuits Syst. Video Technol.* (10), 1006–1013 2003.

[14] X. Song, and G. Fan, "Key-frame extraction for objectbased video segmentation," in *IEEE Proc. Int. Conference on Acoustics, Speech and Signal Processing,* 2005.

[15] M. Mentzelopoulos, and A. Psarrou, "Key-frame extraction algorithm using entropy difference," in *Proceedings of the ACM SIGMM International workshop on Multimedia Information Retrieval*, 2004.

[16] W. Abd-Almageed, "Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing," In: *Proc. Image Processing, 2008*. ICIP 2008.

[17] A. Hanjalic, and H. Zhang, "An integrated scheme for automated video abstraction based onunsupervised cluster-validity analysis," *IEEE Trans. On Circuits And Systems For Video Tech*., vol. 9, no. 8, pp. 1280–1289, 1999.

[18] M. Beniak, J. Pavlovičová, and M. Oravec, "3D Chrominance Histogram Based Face Localization," In: *Int. Journal of Signal and Imaging Systems Engineering* (IJSISE). Vol. 4, No.1 pp. 3 - 12, 2011, www.inderscience.com/ijsise

[19] D. Tarcsiová, "*Communication System of Hearing Impaired Person*", Bratislava: Sapientia, p. 222, 2005.

[20] E. Šikudová, "Comparison of color spaces for face detection in digitized paintings", *SCCG - Spring Conference on Computer Graphic.* pp. 135 - 140, 2007.

J. Polec was born in 1964 in Trstená, Slovak Republic. He received the M.Sc. and PhD. degrees in telecommunication engineering from the Faculty of Electrical and Information Technology, Slovak University of Technology in 1987 and 1994, respectively. From 2007 he is professor at Institute of Telecommunications of Slovak University of Technology and at Department of Applied Informatic of Comenius University. His research interests include Automatic-Repeat-Request (ARQ), channel modeling and image coding.

P. Heribanová was born in 1986 in Kremnica, Slovak Republic. She received M.Sc. degree in Geometry from the Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava in 2010. She is a PhD. student of Geometry and Topology at the same university. Her research interests include image coding, reconstruction and quality evaluation.

T. Hirner was born in 1983 in Bratislava, Slovak Republic. He received his M.Sc. 2008 in Telecommunications from the Slovak University of Technology in Bratislava. From 2007 until the presence, he has been with the Slovak Telecom, a. s. He is also an external PhD. student at the Institute of Telecommunications, STU. His research interests are concentrated on Unequal error control coding for telemetric information and image transmission.