DE GRUYTER OPEN

Journal of ELECTRICAL ENGINEERING, VOL. 66, NO. 3, 2015, 142–148

DETECTION OF FRAMES WITH SINGLE COMPLETE SIGNS OF FINGER ALPHABET IN VIDEO SEQUENCE USING VISUAL ATTENTION APPROACH

Júlia Kučerová [*] — **Jaroslav Polec** [**]

Visual information is very important in human communication. It is used in any type of sign language communication, and in non-verbal communication of the entire population, as well. Therefore, visual information is crucial for communication of hearing impaired people. Video is the most common way to capture this type of information and it is very important to correctly process it. In this paper we propose a method for finding video frames representing single sign in the finger alphabet. The single sign is identified using standard video quality metrics. The calculations of the metrics are performed only within a region, which is determined by combination of object tracking and salient regions detection method based on human visual attention. For key frame selection, combination of sliding system for finding local extreme and adaptive threshold based on local averaging and variation is used. Proposed method is effective and achieves significantly better results in comparison with other commonly used methods.

K e y w o r d s: key frame, sign language, tracking, video quality metric, visual attention

1 INTRODUCTION

Sign language is the primary communication tool for hearing impaired people. In sign language, the static and dynamic gestures are used to achieve communication objectives. The goal of the proposed method is to determine a set of key-frames representing the single sign of finger alphabet in video with Slovak sign language and American sign language (ASL). According to [1], the techniques for key-frame extraction can be divided into four different groups: shot boundary [2], visual content [3], motion analysis [4] and shot activity [5]. The method based on shot boundary and visual content are relatively fast, however, the visual content of the video shot is not effectively captured. The motion analysis and shot activity are more sophisticated methods due to their analysis of motion and activity, yet these methods involve complex computations. The motion energy (ME) was used in the methods [6, 7] to determine the significance of the motion in the scene. Local maximal or minimal ME is related to the motion magnitude and it is usually used as the metric for key-frame extraction.

The unified spatio-temporal feature space was used in [8] to characterize the video data. In this method, the key-frame extraction and object segmentation was used by maximizing the divergence between objects in the feature space. In [9] the entropy-based method was used, where the entropy of the grayscale frame was computed and compared with previous frame. The processed frame was marked as a key-frame if the entropy difference was higher than a threshold. In [10], the real time key-frame extraction based on singular value decomposition (SVD) was presented. In this approach, the low-cost, multivariate color features are extracted from the video and the 2D feature matrix is created. Consequently, this matrix is factorized using SVD.

According to [11], the key-frame selection can be classified into three groups: cluster-based methods, energy minimization-based methods and sequential methods. The cluster-based methods classify all frames according to the content similarity. Consequently, the representative frames of each cluster are labelled as the key-frames. The main disadvantage of these methods is the absence of the temporal information of a video sequence. The methods based on the energy minimization extract key-frames by solving a rate-constrained problem. However, these methods are computationally expensive. In the sequential methods, the new key-frame is determined in the case when the content difference of the current frame with the previous key-frame exceeds a predefined threshold.

The method for key frame extraction using regions of interest and SSIM metric is presented in [12]. However, this method does not utilize tracking of the region of interest.

Nowadays, most of the video quality metrics process the whole image. However, this approach has major deficiencies in the fact, it does not take into account the visual content of the processed image. Therefore, in our

[*] Department of Applied Informatics, Comenius University in Bratislava, Mlynská Dolina, 842 48 Bratislava, Slovakia, kucerova@sccg.sk; [**] Institute of Telecommunications, Slovak University of Technology in Bratislava, Ilkovičova 3, 812 19 Bratislava, Slovakia, jaroslav.polec@stuba.sk

DOI: 10.2478/jee-2015-0022, Print ISSN 1335-3632, On-line ISSN 1339-309X

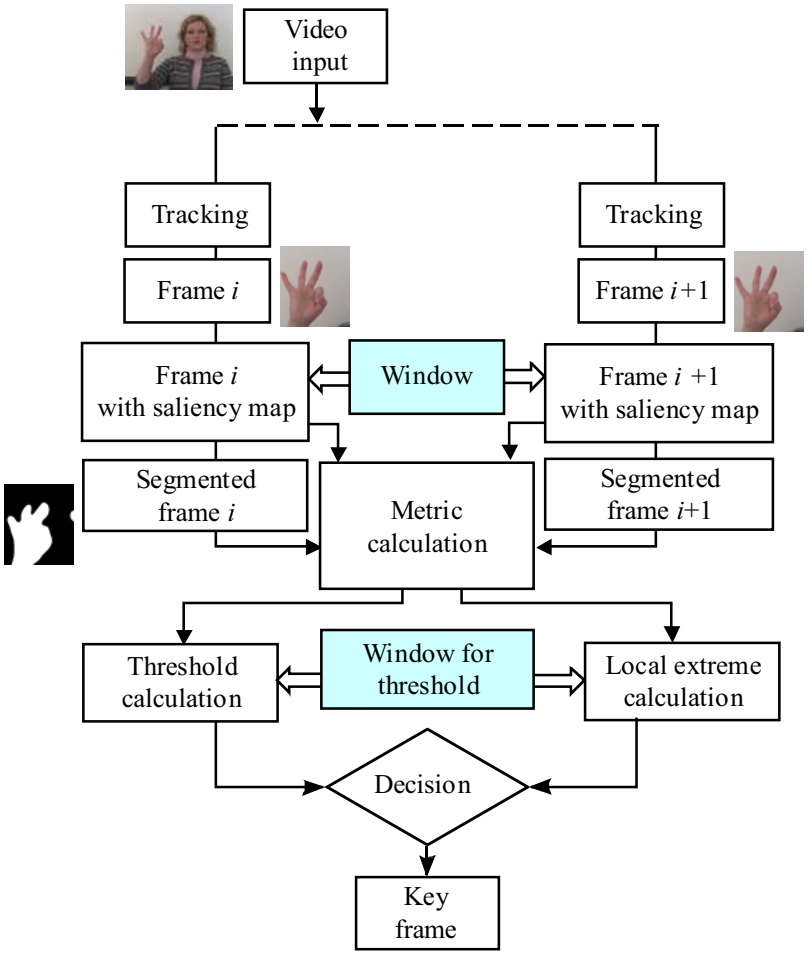

Fig. 1. Block diagram of the proposed system

research we decided to focus on salient regions using combination of object tracking, visual attention approach and video quality metrics.

The proposed method belongs to sequential methods considering visual content. We used visual attention approach for detecting salient regions in the frame. These regions were used to create a mask for further usage. Firstly, the object (hand) is tracked. Secondly, the video is centred to compensate movement of the region of interest. Therefore, comparison of successive frames is limited only on pixels from this region.

2 VIDEO QUALITY METRICS

The main goal in the objective quality assessment research is to design metric providing sufficient quality evaluation in terms of correlation with the subjective results. We performed comparison of two objective metrics used for key-frame extraction. Key-frames represent the signs in a sign language.

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 resultant value [13].

VQM (Video quality metrics) uses discrete cosine transform (DCT) to match the human perception [14]. VQM is based on a simplified human spatial-temporal contrast sensitivity model.

3 THE PROPOSED METHOD

Presented method is proposed with the goal of finding the particular signs in video sequence with sign language. It is based on comparison of two successive frames, which is performed by metrics commonly used for quality evaluation of coded video or video damaged in other way. The block diagram of the proposed system is given in Fig. 1.

3.1 Region of interest selection

Spatial gesture segmentation is the problem of determining where in each video frame the gesturing hand is located. 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

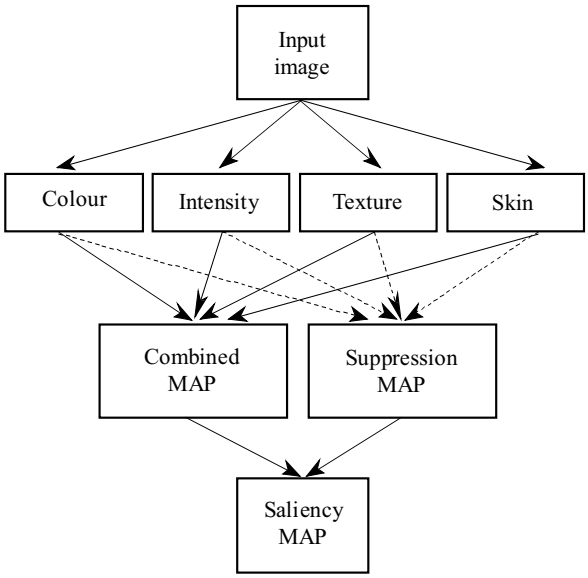

Fig. 2. Scheme of salient regions detection

fuzzy set are available for object segmentation. In arbitrary environment however, neither skin color nor any other color can be guaranteed to appear only within the object of interest – the hand. Algorithms based on skin color distribution are also used for face detection [15]. Because of that additional processing is necessary to differentiate between selected regions.

ROI determination for each frame is performed in three steps hand tracking, salient regions determination and segmentation:

- Mean shift based tracking extracts the color distribution of target appearance, and is implemented using kernel histogram [2].
- The salient regions determination for each frame.
- A single threshold value is assigned to every saliency map.

3.2 Visual Attention Approach

Visual attention is the ability of a vision system to detect salient regions in an observed scene. Human visual system is sensitive to features such as changes in color, shape, and illumination. Therefore for an observer, the important changes are in low-level features (*eg* color, intensity, orientation). However, from semantic point of view, the observer is also interested in faces, humans, text [16].

In this paper, we are using visual attention approach to improve key frame extraction in processed video. For the sake of our research we have decided to use model presented in [17], where local context information is used to suppress spurious attention regions, while simultaneously enhancing the true attention regions. In this model, three features are used: color, intensity and texture. This approach is very useful to capture visual attention in images containing small objects, but is insufficient in images containing more semantic features like face and skin. Therefore, on account of semantic value of human face and hands in our testing data, we decided to use skin detection as an additional feature.

In Fig.2, the scheme of used model is shown. As a first step, the contrast maps for features (attention cues FV) need to be created. The feature map for skin color distribution used in this paper is modeled using a single 2D Gaussian distribution [18]. The detailed process of obtaining the feature map for texture is described in [17]. The feature map for color is created using combination of H and S channels in HSV color space, intensity map is obtained using V channel.

Subsequently, local context suppression strategy for adaptive combination of multiple attention cues (intensity, color and texture) is performed. Consider an image divided into blocks, called Attention Patches, each containing $p \times q$ pixels. The contrast of particular feature at a patch centred at (i, j) is calculated as

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

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 contrast at patch (i, j) for n features attention cues is normalized to lie between (0,1). Each patch is 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 are similar using Suppression factor SF. The SF is derived by building up the Suppression map from all the previously mentioned attention cues. The SF for patch (i, j) is obtained as

$$\tau(i,j) = \prod_{u=1}^{p} \overline{\lambda}_u, \quad (2)$$

where the $\overline{\lambda}$'s are sorted in descending order and the parameter p controls the degree of suppression. The result is a map with values in range (0,1) for each pixel, where lower values represent higher suppression factor and higher values represent lower suppression factor.

To obtain the saliency $S(i, j)$ for patch (i, j) the multiple attention cues are linearly combined into a *Combined map* and the result is modulated by the SF as

$$S(i,j) = \tau(i,j) \times \sum_{u=1}^{k} FV_u(i,j). \quad (3)$$

The product of the *Combined map* and the SF yields the final *Saliency map* S which contains the true *Attention Regions*. Using *Suppression Map*, spurious attention regions in *Combined map* could be efficiently removed [17].

Saliency maps for selected frames of tested video are shown in Fig.3, along with different masks created by thresholding of the original saliency map and combination of thresholded map with the object (hand) tracking.

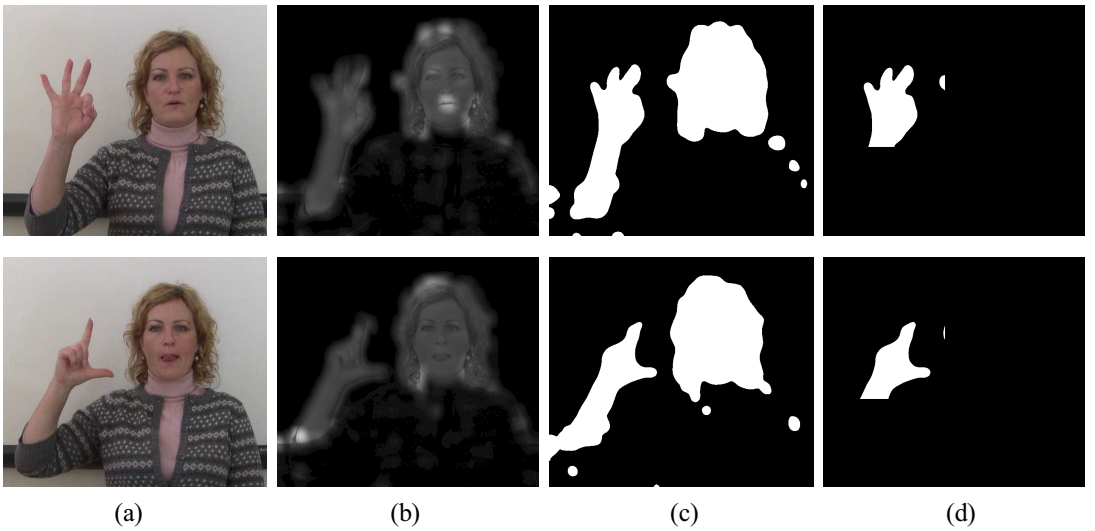

Fig. 3. (a) — original frame, (b) — weighted frame using saliency map, (c) — thresholded saliency map, (d) — segmented frame — combination of the thresholded map and mask of tracked object

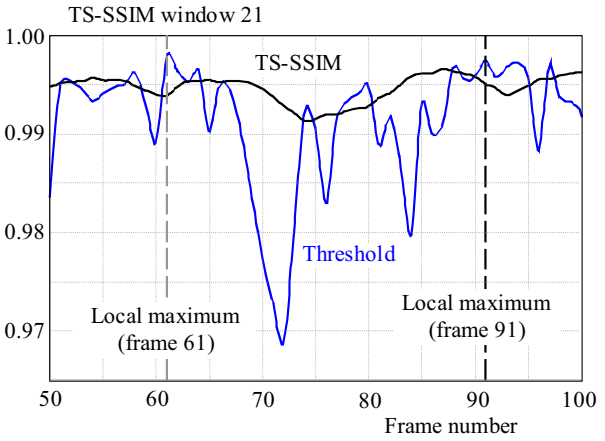

Fig. 4. Decision metric response in case of a key-frame extraction

3. 3 Metrics and Threshold Calculation

We use the local threshold method and examine the frame difference of successive frames. The sliding window length has to be an odd number, as the middle frame is examined and the threshold value is determined from left or right half of the sliding window [19].

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

- The metric value of middle frame is the minimum in the window W.
- The metric value of middle frame is lower than threshold. The value of this sample is reduced by the constant value. This will determine our new local minimum. Otherwise, the middle sample is not the key sample.

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

To metrics reaching maximum value for comparison of identical sets (for example SSIM, PSNR) following is applied:

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

- The metric value of middle frame is the maximum in the window W.
- The metric value of middle frame is greater than threshold. The value of this sample is enlarged by a constant value. This will determine our new local maximum. Otherwise, the middle sample is not the key sample.

For calculation of selected metrics we use the program MSU [20]. Proposed method is based on comparison of two successive frames and metric is then applied only to the region of ROI assigned to the first of compared frames. The local adaptive threshold is computed [21] for minimum as

$$m_T = \min\left\{{}^{VQ}\mu_{\text{left}} + k^{VQ}\sigma_{\text{left}}, {}^{VQ}\mu_{\text{right}} + k^{VQ}\sigma_{\text{right}}\right\}, \tag{4}$$

for maximum as

$$m_T = \max\left\{{}^{VQ}\mu_{\text{left}} + k^{VQ}\sigma_{\text{left}}, {}^{VQ}\mu_{\text{right}} + k^{VQ}\sigma_{\text{right}}\right\}, \tag{5}$$

(a) (b) (c)

Fig. 5. Testing videos, from left to right: video 1 [22], video 2 [23] and video 3 [24].

where μ is the mean value VQ for window W, α is the standard deviation and k is determined experimentally for each metric. VQ is our chosen metric for determination of the dissimilarity between compared frames. In the proposed method, the window W is asymmetric, because sought key sign is located further from the centre of the window.

The response of one of the decision metrics for a key frame extraction of one-hand finger alphabet is shown as a graph in Fig. 4. 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.

For the success rate of evaluation of signs extraction modifications of precision P, recall R and $F1$ rates were used.

The precision measure is defined as the ratio of correct extractions over the number of all extractions. The value of precision is lower with higher amount of false extractions. The Slovak language have minimal number of the words with the same letter being used two or more times consecutively. We can also find the same property in other languages. Therefore, for such languages, the calculation of the modified precision P is given as

$$P = \frac{CS + DS + CP + DP}{CS + DS + CP + 0.5 \cdot DP + FP}, \qquad (6)$$

where CS indicates the number of correctly extracted signs, DS indicates the number of correctly extracted signs, where two frames are extracted for one sign. CP and DP analogically represent word spaces and FP denotes the number of false detected word spaces.

The positive true function or sensitivity recall measure, corresponds to the ratio of correct experimental extractions over the number of all true extractions. The value of recall decreases with increasing number of missed key frames. The used modification of the recall R for the success rate evaluation of signs extractions is listed as follows

$$R = \frac{CS + DS + CP + DP}{CS + DS + CP + DP + MS + MP}, \qquad (7)$$

where MS denotes the number of missed signs and MP denotes the number of missed word spaces.

The relationship between precision and recall is given by $F1$ score measure. It is a combined measure that results in high value if and only if both precision and recall reach high values. Using $F1$ score, the more general view of the examined key frame extraction algorithm accuracy could be shown. The modified $F1$ score takes into account both missed key frames and false extractions and is calculated as follow

$$F1 = \frac{2 \cdot RP}{R + P}. \qquad (8)$$

4 EXPERIMENT

Experiment was performed on three video recordings. Video 1 contains signs of Slovak one-hand finger alphabet [22]. The testing video is in format CIF and its length is approximately one minute. It contains 41 signs in seven logatoms, *ie* words without meaning, and 8 word spaces. The two additional video recordings, obtained from youtube.com, contain American finger alphabet [23, 24]. In Fig. 5, selected frames for each used video are shown.

In this experiment, two different types of testing videos were created. Both types of testing video groups were created by combination of the original video and the mask. The mask for the first type of testing video was created by using visual attention approach, where salient regions of each frame in video were determined. Using this mask, the image information can be divided into 256 importance groups. The mask for the second type of the testing video was created by applying threshold on the original mask with salient regions and window.

We experimentally determined, that the window W size in the range from 21 to 49 was most suitable for the experiment. These windows W were used for finding local extreme and the threshold value. For the creation of the thresholded mask, the salient map was normalized into range [0,1]. The used method for salient regions detection is in this type of image input very precise, therefore the threshold value is set to 0.03, which guarantee that the needed salient region can be selected. The obtained results are present in Tables 1, 2 and 3, where TM represents approach with simple threshold mask and window, WS is based on [12] approach and TS represents

Table 1. Key frame extraction performance comparison for video 1 [22]

Metric	TM-VQM	TM-SSIM	WS-SSIM	TS-VQM	TS-SSIM
window	10+10	10+10	12+12	14+9	15+14
CS	37	37	25	37	41
DS	3	4	12	4	0
MS	1	0	4	0	0
CP	6	5	4	5	8
DP	2	3	4	3	0
FP	0	0	0	0	0
MP	0	0	3	0	0
R	0.980	1.000	0.918	1.000	1.000
P	0.970	0.961	0.833	0.961	1.000
F1	0.975	0.980	0.874	0.980	1.000

Table 2. Key frame extraction performance comparison for video 2 [23]

Metric	TM-VQM	TM-SSIM	WS-SSIM	TS-VQM	TS-SSIM
window	21+21	24+24	23+23	22+20	22+20
CS	26	25	25	26	26
DS	0	1	0	0	0
MS	0	0	1	0	0
CP	0	0	0	0	0
DP	0	0	0	0	0
FP	0	0	0	0	0
MP	0	0	0	0	0
R	1.000	1.000	0.962	1.000	1.000
P	1.000	0.981	1.000	1.000	1.000
F1	1.000	0.990	0.980	1.000	1.000

Table 3. Key frame extraction performance comparison for video 3 [24]

Metric	TM-VQM	TM-SSIM	WS-SSIM	TS-VQM	TS-SSIM
window	12+12	12+12	13+13	14+9	12+10
CS	24	24	19	25	25
DS	1	1	1	1	1
MS	1	1	6	0	0
CP	0	0	0	0	0
DP	0	0	0	0	0
FP	0	0	0	0	0
MP	0	0	0	0	0
R	0.962	0.962	0.769	1.000	1.000
P	0.980	0.980	0.976	0.981	0.981
F1	0.971	0.971	0.860	0.990	0.990

the proposed method. The best results for each video are underlined.

5 CONCLUSION

In this paper, we presented a method for finding video frames representing single signs in a one-hand finger alphabet. The proposed method is based on combination of object tracking, visual attention approach and standard video quality evaluation metrics. In this method, evaluation of the video quality metric is performed only in salient regions determined by the visual attention approach. The success rate is evaluated by recall, precision and F1 measure.

The main improvement presented in this paper is the combination of the standard video quality metrics with visual attention approach and object (hand) tracking. In

the experimental part of our research, the effectiveness of the proposed method was compared with metrics applied to all frames. The achieved results indicate that using combination of visual attention and standard video quality evaluation metrics is more successful in comparison to other methods.

Acknowledgement

Research described in the paper was financially supported by the Slovak Research Grant Agency VEGA under grant No. 1/0625/14 and UK/555/2014.

References

[1] ZHUANG, Y.—RUI, Y.—HUANG, T. S.—MEHROTRA, S.: Adaptive Key Frame Extraction using Unsupervised Clustering, In ICIP, vol. 1, 1998, pp. 866–870.

[2] NAGASAKA, A.—TANAKA, Y.: Automatic Video Indexing and Full-Video Search for Object Appearances, In VDB, vol. A-7 of IFIP Transactions, North-Holland, 1991, pp. 113–127.

[3] ZHANG, H. J.—WU, J.—ZHONG, D.—SMOLIAR, S. W.: An Integrated System for Content-Based Video Retrieval and Browsing, Pattern Recognition **30** (1997), 643–658.

[4] WOLF, W.: Key Frame Selection by Motion Analysis, In Acoustics, Speech, and Signal Processing 1996. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference on, vol. 2, May 1996, pp. 1228–1231.

[5] GRESLE, P. O.—HUANG, T.: Gisting of Video Documents: A Key Frames Selection Algorithm using Relative Activity Measure, In The 2nd Int. Conf. on Visual Information Systems, San Diego (USA), 1997, pp. 279–286.

[6] LIU, T.—ZHANG, H.—QI, F.: A Novel Video Key-Frame-Extraction Algorithm based on Perceived Motion Energy Model, IEEE Trans. Circuits Syst. Video Techn. **13** No. 10 (2003), 1006–1013.

[7] LIU, T.—ZHANG, X.—FENG, J.—LO, K.: Shot Reconstruction Degree: a Novel Criterion for Key Frame Selection, Pattern Recognition Letters **25** No. 12 (2004), 1451–1457.

[8] SONG, X.—FAN, G.: Joint Key-Frame Extraction for Object Based Video Segmentation, In IEEE Proc. Int. Conference on Acoustics, Speech and Signal Processing, vol. 2, Breckenridge (CO), 2005, pp. 126–131.

[9] MENTZELOPOULOS, M.—PSARROU, A.: Key-Frame Extraction Algorithm using Entropy Difference, In Multimedia Information Retrieval (Michael S. Lew, Nicu Sebe, and Chabane Djeraba, eds.), ACM, 2004, pp. 39–45.

[10] ABD-ALMAGEED, W.: Online, Simultaneous Shot Boundary Detection and Key Frame Extraction for Sports Videos using Rank Tracing, In ICIP, IEEE, 2008, pp. 3200–3203.

[11] HANJALIC, A.—ZHANG, H.: An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis, IEEE Trans. Circuits Syst. Video Techn. **9** No. 8 (1999), 1280–1289.

[12] MENDI, E.—BAYRAK, C.: Shot Boundary Detection and Key Frame Extraction using Salient Region Detection and Structural Similarity, In ACM Southeast Regional Conference (H. Conrad Cunningham, Paul Ruth, and Nicholas A. Kraft, eds.), ACM, 2010, p. 66.

[13] WANG, Z.—BOVIK, A. C.—SHEIKH, H. R.—SIMONCELLI, E. P.: Image Quality Assessment: from Error Visibility to Structural Similarity, IEEE Transactions on Image Processing **13** No. 4 (2004), 600–612.

[14] XIAO, F.: DCT-Based Video Quality Evaluation — final project for EE392J 2000.

[15] COMANICIU, D.—RAMESH, V.—MEER, P.: Kernel-Based Object Tracking, IEEE Transactions On Pattern Analysis and Machine Intelligence (PAMI) **25** No. 5 (2003), 564–575.

[16] GOLDSTEIN, B. E.: Cognitive Psychology: Connecting Mind, Research and Everyday Experience, Thomson Wadsworth, Belmont, 2008.

[17] HU, Y.—RAJAN, D.—CHIA, L.-T.: Adaptive Local Context Suppression of Multiple Cues for Salient Visual Attention Detection, In ICME, IEEE, 2005, pp. 346–349.

[18] ŠIKUDOVÁ, E.: Comparison of Color Spaces for Face Detection in Digitized Paintings, In Proceedings of the 23rd Spring Conference on Computer Graphics, SCCG'07, ACM, New York, NY, USA, 2007, pp. 219–224.

[19] POLEC, J.—HERIBANOVÁ, P.—HIRNER, T.: Key Frames Extraction for Sign Language Video Analysis and Recognition. Volume 7, World Academy of Science, Engineering and Technology, 2013.

[20] MSU Graphics and Media Lab (Video Group) Moscow. Msu Video Quality Measurement Tool (software). url: www.compression.ru/video/quality_measure/-video_measurement_tool_en.html, June 2014 [Online; accessed June-2014].

[21] DUGAD, R.—RATAKONDA, R.—AHUJA, N.: Robust Video Shot Change Detection, In IEEE Workshop on Multimedia Signal Processing, 1998.

[22] HERIBANOVÁ, P.: Video Quality Assessment Methods, PhD thesis, Comenius University, Bratislava, Faculty of Mathematics, Physics and Informatics, 2014.

[23] Sign Language Alphabet (2012). url: www.youtube.com/watch?v=xCm3wqF_doc, 2014 [Online; accessed July-2014].

[24] CASCIO, M.: Abc sign with me – baby sign language., url: www.youtube.com/watch?v=VCVbAct19A0, 2014 [Online; accessed July 2014].

Received 17 December 2014

Júlia Kučerová was born in 1987 in Detva, Slovakia. She received MSc 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, quality assessment and image coding.

Jaroslav Polec was born in 1964 in Trstená, Slovakia. He received the Eng 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 AutomaticRepeatRequest (ARQ), channel modelling, image coding, interpolation and filtering.