Selected Methods of Semantics Extraction

Vybrané metódy extrakcie sémantiky

Zuzana Černeková, Zuzana Haladová, Júlia Kučerová, Elena Šikudová
FMFI UK Bratislava

Abstrakt:

V tomto článku predstavujeme extrakciu sémantickej informácie z rôznych domén. Prinášame prehľad metód pre detekciu a popis zaujímavých bodov v obraze, modelovanie vizuálnej pozornosti a významných oblastí a podávame prehľad problémov riešených na TRECVID 2011, ktoré používajú sémantickú informáciu. Detektor a deskriptor pre kľúčové body sme otestovali v aplikáciách, ktoré v galériách rozpoznávajú maliarske diela a detekcia významných oblastí bola použitá v oblasti kompresie videa.

Kľúčové slová:

sémantika, detekcia objektov, významné oblasti

ACM classification: CCS → Information systems → Information retrieval → Retrieval tasks and goals → Information extraction, CCS → Information systems → Information retrieval → Specialized information retrieval → Multimedia and multimodal retrieval

Abstract:

The paper introduces extraction of semantic information in various domains. We bring an overview of the methods for detecting and describing interesting point in an image, modelling visual attention and saliency and review the tasks of the TRECVID 2011 that deal with using semantic information. We have tested the keypoint detectors and descriptor in an application that recognizes paintings in galleries. Detection of salient regions was used in compression of video sequences of sign language.

Keywords:

semantics, object detection, saliency

1. INTRODUCTION

An image is worth a thousand words. This is very true in the digital era, when people are browsing vast image databases on the internet. However, the problem of the image description remains unsolved. The semantic gap between the information that can be computationally extracted from the visual data and the interpretation that the user derives from the same data is crucial. When browsing, the users most of the time seek semantic similarity, but the databases provide similar images based only on low-level features as colour, texture, shape, etc.

In our paper, we will describe methods that identify important regions in images, extract features from these regions and assign basic semantic meaning to them. Important regions can

be found using low-level or higher-level properties. Low-level importance lies in local contrast, colour or texture difference, shape or orientation change, etc. The higher level takes into account the properties of human visual system and the concept of visual attention.

The first main part of our paper describes the method for detecting and describing interesting points in an image. The second part covers the area of visual attention and saliency. In the third part a brief overview of the TRECVID 2011 tasks using semantic information together with the best approaches is given.

2. Object detection

Object detection (finding if the object is presented on the image) and recognition (determining the object's category) is the key aspect of the semantics extraction process. The recognition can be seen from two different points of view:

1. Recognition of a concrete object instance (for example the mountain shelter Zamkovského chata).
2. Recognition of a class of objects (for example bugs).

Different input images for these two tasks can be seen in Figure 1.

Figure 1: *Examples of different object recognition tasks. Top line: Recognition of a generic class: bugs. Bottom line: Recognition of concrete object instance: Zamkovskeho chata.*

In general, we can say that in both tasks the ultimate goal is to recognize the object in all possible circumstances: different scale, rotation, background, composition with other objects, partial occlusion, varying illumination etc. This goal is very challenging, so nowadays some partial problems with imposed constraints (lighting, selected object category, etc.) are under investigation.

2.1 Feature extraction

The first step in the recognition process is usually the extraction of features, which describe the object to be recognized by the classifier. The features should be invariant to affine transformations, illumination and occlusions in order to recognize all instances of the object. Different types of features have emerged since the start of computer vision research, which can be generally divided in to three groups: colour, texture, and shape. We can also divide the features based on the area they describe into local and global ones.

Global features extract the information from the whole image. If we want to extract a feature, e. g the energy of the co-occurrence matrix, we first create the co-occurrence matrix for all pixels in the image and then compute the energy of this matrix.

Local features on the other hand extract information only from the parts of the image, which are „interesting". Interesting part is the part of the image with strong variation of intensity in the local neighbourhood. Most local feature detection methods use only intensity of the images. If we examine an image of a flat white wall, we will not detect any local features. Local features are extracted in two steps. Firstly, the interesting points are detected, then the features are computed for all detected points and finally feature vectors (descriptors) are created. The classical and most cited method for detection and description of the local features is the SIFT [1]. Nowadays the methods generating local features are very popular and many new ones are published every year.

There are many different methods in the area of the interesting points detection (called interest points detectors), however three of them are used the most:

The oldest method is the Harris' corner detector [2], which computes the eigenvalues of the second moment matrix of an image at some point. Harris' method was boosted in [3] where the authors propose taking the minimum of the eigenvalues and compare it to a given threshold. If it is bigger, the point is considered a corner. SUSAN (Smallest Univalue Segment Assimilating Nucleus) [4] is another description method, which utilizes second moment matrix. SUSAN utilizes a circular mask and compares the intensity of the pixels in the map with that of its nucleus.

The second method uses the approximation of Laplacian of Gaussian with the difference of Gaussians (DoG) and looks for the local extremes in the scale-space pyramid. Scale space pyramid consists of different image scales, so called octaves (scales of the image are 1, 1/4, 1/16 etc.), with each octave containing progressively smoothed image with a Gaussian kernel. This methods is used in the well-known SIFT's and SURF's detectors [1, 5].

The third method is based on the accelerated segment test (AST). This approach examines the neighbourhood of every point of the size of the Bressenham's circle with diameter d=7. The points are concerned as interesting if there are more than n (usually 8) points in the neighbourhood that fulfil the following criteria. The intensity difference between the examined pixel and the neighbourhood pixel must be larger than a given threshold. We can find this method in the FAST detector [6].

2.2 Feature description

Feature descriptors can be divided into two groups: integer and binary. The main advantage of binary descriptors is that two binary strings can be compared using the Hamming distance instead of the Euclidean distance. Hamming distance can be computed very fast and it saves the matching time. Integer description methods typically use the computation of the histogram of gradients (HoG) in the patch placed around the interesting point (for example the SIFT, SURF or DAISY descriptors [7]). On the other hand, binary methods use the binary intensity tests

which compare the line endings in the mikado like patch (for example BRIEF [8] or ORB [9] descriptors).

2.3 Feature description matching

Another important aspect of the object recognition using local features is the matching of the feature vectors. In the matching phase, the feature vectors extracted from the unknown object are matched with the database of the feature vectors extracted from the labelled objects. An example of matched correspondences between the labelled and the unlabelled image can be seen in Figure 2. The unknown object is labelled with the same label as the object with the most matches. This phase can be time consuming when performing all-to-all brute-force matching. Different methods for organizing of the database of features for faster search and match have been published. They are based on, to name few, kd-trees (in the later implementation of SIFT), random trees [10] or spectral hashing [11] .

Figure 2: *Correspondence of interest points between unlabeled image (on the left side) and labeled image (on the right side) matched with SIFT.*

2.4 Detector/descriptor evaluation

Our tests of selected local feature approaches proved that the detectors based on AST in combination with descriptors based on binary intensity tests are much faster than DoG based detectors and HoG based descriptors. We have evaluated three methods: ORB SIFT and SURF. The results were partially published in [12]. Our database consists of 100 tourists photographs of paintings acquired in galleries and 15 training paintings. We have classified the images into 16 classes (15 paintings and 1 for paintings not presented in the database). We have achieved 90% accuracy with the SIFT and SURF methods and 80% accuracy with the ORB method. On the other hand, ORB proved to be 80 times faster than SIFT and 30 times faster than SURF. SURF is known as a faster modification of the SIFT method.

2.5 Test case

In our work [13] we used a combination of local and global features to speed up the process of descriptor matching. We have tested the organization of the database by sorting of the labelled images of objects. We have decided to choose global features, which are fast and efficient to compute. In the pre-processing phase, we extract one chosen global feature for all images in the database. In the run time, prior to the matching phase we extract the same global feature from the image to be labelled. Then we sort our database based on the similarity according to the

global feature. Then we match the unlabelled image with the sorted database. The first labelled images with more matches than a threshold ***t*** is considered a correct match.

In order to extract the correct value of the global feature, we need to segment the object from the unlabelled image (to avoid the background of the image affecting the feature). In our study fine art paintings are used. The segmentation consist of finding the frame of the painting.

We have tested 9 global features: average intensity, percentage of light pixels, normalized intensity histogram, entropy, normalized hue histogram, number of pixels that belong to the most frequent hue, most populated hue, hue contrast, and hue count. The colour features were computed from the image transformed to CIE Lab colour space and hue was calculated as the four-quadrant arctangent of b/a. We evaluated individual features as well as their combinations to see which feature (combination) is the best to sort the database.

The tests on the database showed that after sorting according to the best global feature the number of needed local feature descriptors comparisons dropped to the half of the number needed in the matching without sorting. During the tests for our database, the height of the highest peak in the normalized histogram of grey values proved to be the fastest (in computation time), and the second most precise in sorting of the database. It also preserves the accuracy of the recognition at 80 and 90% in ORB and SIFT/SURF respectively.

2.6 Special object category – human face

One important type of objects for detection and recognition is the human face. Face detection and recognition is important in many human-computer interaction systems. Face detection is a difficult problem because of the wide variety of faces to match, variations in colour and shadows, presence of facial hair, partial occlusion by glasses, scaling and rotation, etc.

There are many different approaches for detecting faces in the images: knowledge-based methods, feature invariant approaches, template matching, appearance-based methods. A well-known method is the Viola/Jones' face detector. This system is used for real-time face detection. Training in this face detection system is slow, but the detection is very fast. The key ideas of this face detector are integral images for fast feature evaluation, boosting for feature selection and attentional cascade for fast rejection of non-face windows. The features used by this method represent difference of sums of image intensities of specific rectangular areas. The sums are easily computed using the integral image. An integral image is a grid data structure of the size of the original image and in each point (x,y) it contains the sum of intensities in the upper-left corner starting at (x,y) of the original image. During training weak classifiers are combined into stronger ones. This is done by using the AdaBoost algorithm [14].

Face detection in coloured images involves the knowledge of skin colour distribution. The simplest method to mark the skin locus in the chosen colour space is to design a boundary using simple thresholds or more complex curves. Skin colour can be also easily modelled by a histogram generated from pixels with known labels (skin pixels). But the most popular method for skin detection is the Gaussian density function; either unimodal or so called mixture of Gaussians. Other non-parametric skin modelling methods involve neural networks, support vector machines or Bayesian decision rules [15].

Face recognition can be used as an identification or verification tool. In face identification, the query face image is compared against all the images in the database to determine the identity of the query face. During face verification, the query face image is compared solely against the face image whose identity is being proved. There are several types of face recognition algorithm including PCA, ICA, LDA, graph matching, kernel methods, active appearance model, and many more.

Principal Component Analysis (PCA) finds a subspace whose basis vectors correspond to the maximum variance direction in the original image space. In the training phase the mean face is found and subtracted from the training data. Then the k biggest eigenvectors (principal components) of the covariance matrix are computed and used to project each training image onto the subspace stated by the principal components. In the case of face recognition the eigenvectors are called eigenfaces. In the recognition phase, also the novel image is projected onto the subspace and the closest training face (within a threshold) is identified as a match.

3. VISUAL ATTENTION AND SALIENCY

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 [16].

Humans cannot attend to all things at once. Their visual system has the ability to pay attention to some parts of the observed scene – salient objects. Visual attention models detect these salient objects in scene. There are two general visual processes for detecting salient objects:

- bottom-up
- top-down

The bottom-up process is task-independent. This process computes the saliency map by predicting which parts of the observed scene could attract more attention. It could be used in machine vision, automatic detection of objects in nature scenes, intelligent image compression, etc. Salient objects in scene are for example a burning candle in a dark room or the lips and eyes of a human face (because they are the most significant elements of the face). If there are many salient objects in the scene, they become obscure because of their big amount.

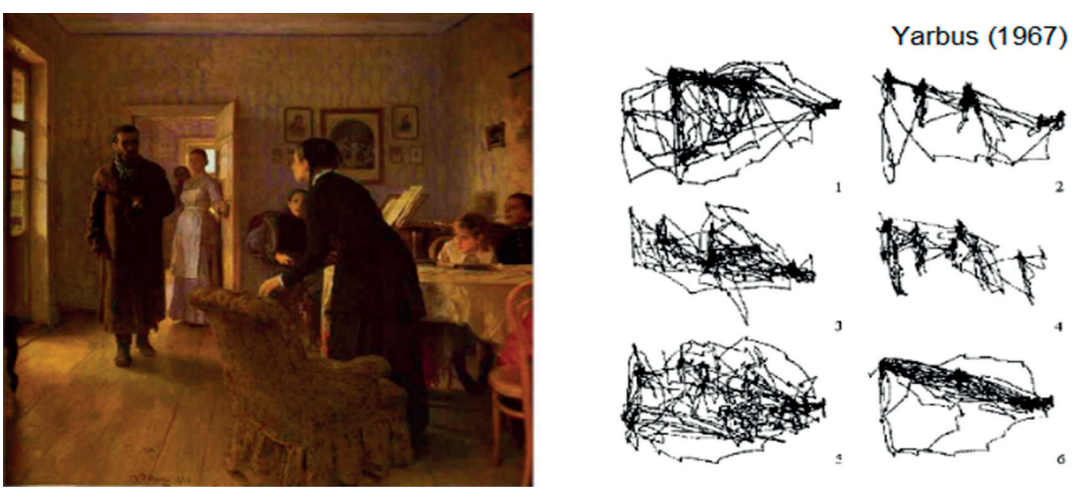

Figure 3: *Repin's picture was examined by subjects with different instructions;*
1. Free viewing, 2. Judge their ages, 3. Guess what they had been doing before the unexpected visitor's arrival, 4. Remember the clothes worn by the people,
5. Remember the position of the people and objects in the room,
6. Estimate how long the visitor had been away [22].

The top-down process is volition-controlled and task-dependent. The task and the volition drive the observer‘s attention to one or more objects that are relevant to the observers goal when studying the scene. For example, the task could be to find a red car on a car park, or to count

particular objects in a scene. When the observer is concentrated on finding some objects in the scene, he will fob off some salient objects. For that reason some objects that are salient in the bottom-up process could not be found with the top-down process. In 1967, the psychologist Yarbus recorded the eye movements of participants watching an image [16]. The subjects' task was to observe Repin's picture "An Unexpected Visitor" and to answer a number of different questions. Figure 3 show the painting and the observed eye movements for different questions.

3.1 Visual attention models

Visual attention has been studied for over a century. Early studies of visual attention were simple ocular observations. Since then the field has grown and nowadays it is involved in many scientific disciplines.

Detecting of the salient regions (which attract human vision) in the image using an eye tracking system is efficient but could be time and money consuming. Therefore, in past few years, many different visual attention models were proposed. These models are based on bottom-up, top-down visual processes and their combination.

Computational models based on the bottom-up visual process usually extract and combine low-level visual features such as colour, intensity, orientation, etc. One of the first models based on the bottom-up process was developed by Itti et al. [17]. In this model, the visual attention is based on the behaviour and the neural architecture of the early primate visual system.

Although models based on bottom-up approach [18, 19, 20, 21] are able to detect salient regions, they are just a basic description of the human vision. They are based on biological presuming of human visual attention, but in most of them, the importance of cognitive processing is missing. In visual observation of a scene there is a very important prior knowledge coming from our perceptual learning, our memory and our previous experience. The combination of low-level features and prior knowledge is a promising approach in visual attention detection.

One of the ideas of using more than low-level features is proposed in [22]. This research is based on the analysis of eye tracking data. The authors created a unique database of eye tracking data. By analysing these data they find out that observers focus their attention on faces, humans (as well as drawings and sculptures of humans) and text. They also used the data for creating a new visual attention model. In their study, they used low-level, mid-level and high-level features. This combination of features gave very good results compared to other visual attention models.

At the moment we are at the beginning of designing of a computational visual attention model that will use the prior knowledge. It is very difficult to detect all salient regions in observed scene and using prior knowledge will help to solve this challenge. Nowadays researchers focus on solving partial problems in this field.

3.2 Visual attention in image and video compression

Recently, image and video compression techniques have drawn much attention. A very popular approach for reducing the size of compressed image or video is selection of a small number of interesting regions (Regions of interest ROI) and to encode them in priority.

Regions of interest, such as humans, faces, text, etc., are very important in humans perceiving of a scene. Up to this day, many different approaches for ROI detection were proposed. Some of ROI detection approaches are very simple, other requires very difficult computations. In many approaches for image and video compression, the saliency map detection is used for ROIs determination.

The information about ROI is usually in binary form [23, 24, 25]. The compression based on this information gives very good results, but in some cases, binary information about ROI is deficient and more information is required. Therefore, in [26] different compression rate at different hierarchical salient locations is used. In this approach, original resolution is retained in the first salient region; the lowest resolution is applied in the unapparent salient regions and the middle resolution is decided by the saliency order from high to low. This method achieved variable resolution image compression by the model of visual attention.

3.3 Video quality assessment using visual attention approach

Visual information is very important for hearing-impaired people, because it allows them to communicate personally using the sign language. In our research [27] we focused on the fact that some parts of the person using the sign language are more important than others (e.g. hands, face). We presented a visual attention model based on detection of low-level features such colour, intensity and texture and combination with the prior knowledge – in our case information about skin in image (Figure 4). Information about the visually relevant parts allows us to design an objective metric for this specific case. We presented an example of an objective metric based on human visual attention and detection of salient object in the observed scene. The proposed metrics were compared to existing metrics and the results were very promising for the following research.

a) *b)*

Figure 4: *Image taken from the experiment [27]: a) original and b) product of the original image (only Y canal from YUV colour space) and the saliency map.*

4. SEMANTIC EXTRACTION FROM VIDEOSEQUENCE

There is a huge use of semantic information in video search. The ability to detect features is an interesting challenge by itself, but it takes on added importance to the extent it can serve as a reusable, extensible basis for query formation and search. Nowadays, the researchers focus mainly on solving the problems of finding the semantic information in video sequences. To promote progress in content-based retrieval from digital video via open, metrics-based evaluation is a goal of the TRECVID conference. The organizers of TRECVID want not only to provide a common corpus of video data as a testbed for different algorithms, but also to standardize and oversee their evaluation and to provide a forum for the comparison of the results [28]. In the last years, most of the TRECVID tasks focused on extracting semantic information from the video sequences. The next sections briefly describe the TRECVID 2011 tasks [29] together with the best approaches.

4.1 Semantic indexing

A potentially important asset to help video search and navigation is the ability to automatically identify the occurrence of various semantic features such as “Indoor/Outdoor”, “People”, “Speech” etc., which occur frequently in video information.

Systems developed for this task focused on robustness, merging many different representations, use of spatial pyramids, improved bag of word approaches, improved kernel methods, sophisticated fusion strategies, and combination of low and intermediate/high features. The best performance in semantic indexing was obtained using Gaussian mixture model (GMM) supervectors and tree-structured GMMs [30, 31]. GMM supervectors corresponding to six types of audio and visual features are extracted from video shots by using tree-structured GMMs. The extracted features are SIFT features with Harris-Affine detector, SIFT features with Hessian-Affine detector, SIFT and hue histogram with dense sampling, histogram of oriented gradients (HOG) with dense sampling, HOG from temporal subtraction images and Mel-frequency cepstral coefficients (MFCCs). The computational cost of maximum a posteriori (MAP) adaptation for estimating GMM parameters are reduced by tree-structured GMMs by keeping accuracy at high levels [32].

4.2 Known-item search

Imagine a situation in which someone has seen a video before, and they want to find it in a provided collection, but does not know where to look. To begin the search process, the searcher formulates a text-only description, which captures what the searcher remembers about the target video. This task is very different from the TRECVID ad hoc search task in which the systems began with a textual description of the need together with several image and video examples of what was being looked for.

The best result among all automatic search runs was achieved using the automatic text-based search system consisting of several main components, including text pre-processing, keywords extracting and processing, text-based retrieval, results fusion and re-ranking [33]. Authors of this approach proposed also a bio-inspired method. In this approach, a query topic is first parsed by a text analyser to produce several search cues, and then the cue-based bottom-up saliency map and the top-down cue-guided concept/object detection are fused and refined by the aid of context cues. This approach did not obtain as good results as the text-based search but can be promising if the attention model and knowledge base are further improved.

4.3 Instance search

In many situations involving video we need to find more video segments of a certain person, object, or place, given one or more visual examples of the specific item. Given a collection of test videos, a master shot reference, and a collection of queries that delimit a person, object, or place entity in some example video, the task is to locate for each query 1000 shots most likely to contain a recognizable instance of the entity.

The best results in this task were obtained using large vocabulary quantization by hierarchical k-means and weighted histogram intersection based ranking metric [34]. In the offline indexing phase the algorithm searches for matching in a computationally cheaper high dimensional bag-of-word feature space. Three frames per second are chosen from every video clips, and then SIFT descriptors are sparsely extracted. Then all SIFT descriptors are projected into the vocabulary tree and they get only one bag-of-words histogram as its representation. In the online searching phase, the SIFT features are extracted from the probe image and they are projected to the vocabulary tree. Thus, one histogram is obtained as the representation for

current probe topic. Histogram intersection metric is then taken to rank the similarity between each probe topic with every candidate video clip.

4.4 Multimedia event detection

A user searching for events in multimedia material may be interested in a wide variety of potential events. Since it is an intractable task to build special purpose detectors for each event a priori, a technology is needed that can take as input a human-centric definition of an event that developers (and eventually systems) can use to build a search query. The events for multimedia event detection were defined via an event kit, which consisted of:

- An event name which is a mnemonic title for the event.
- An event definition which is a textual definition of the event.
- An event explication which is an expression of some event domain-specific knowledge needed by humans to understand the event definition.
- An evidential description which is a textual listing of the attributes that is indicative of an event instance. The evidential description provides a notion of some potential types of visual and acoustic evidence indicating the event's existence but it is not an exhaustive list nor is it to be interpreted as required evidence.

The Raytheon BBN's VISER system [35] showed the best performance among all the submitted systems. The VISER system incorporates a large set of low-level features that capture appearance (SIFT, SURF, D-SIFT, CHoG), colour (RGB-SIFT, OpponentSIFT, and C-SIFT), motion (Space-Time Interest Points – STIP, D-STIP), audio, and audio-visual co-occurrence patterns in videos. The system also uses high-level (i.e. semantic) visual information obtained from detecting scene, object, and action concepts. Furthermore, the VISER system exploits multimodal information by analysing available spoken and videotext content. These streams are combined into a single, fixed-dimensional vector for each video. Two combination strategies are explored: early fusion and late fusion. Early fusion is implemented through a fast kernel-based fusion framework and late fusion is performed using both Bayesian model combination as well as a weighted-average framework.

5. CONCLUSION

Emerging new technologies demand tools for efficient indexing, browsing and retrieval of image and video data, which causes rapid expansion of areas of research where the semantic information is used. New methods that work with semantic information in image, video, and audio are developed frequently these days, which means that our list of methods is not final. Nevertheless, we picked the most used ones and tested them. We brought an overview of the methods for detecting and describing interesting points in an image, modelling visual attention and saliency and reviewed the tasks of the TRECVID 2011 that deal with semantic information. We have tested the keypoint detectors and descriptors in an application that recognizes paintings in galleries. We have evaluated the use of visual saliency for compression of video sequences containing sign language.

Acknowledgements

This work was funded from projects KEGA 068UK-4/2011 and VEGA 1/0602/11.

References:

[1] Lowe, D. G. 2002. ***Distinctive image features from scale-invariant keypoints.*** International Journal of Computer Vision, 60(2):91§110, 2002.

[2] Harris, C. – Stephens, M. 1988. *A combined corner and edge detector*. In Proc. of Fourth Alvey Vision Conference, pages 147§151, 1988.

[3] Shi, J. – Tomasi, C. 1994 *Good features to track, Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94, 1994 IEEE Computer Society Conference on* ,vol., no., pp.593-600, 1994.

[4] Smith, S. M. – Brady, J. M. 1995. ***Susan – a new approach to low level image processing***. International Journal of Computer Vision, 23:45§78, 1995.

[5] Bay, H. – Ess, A. – Tuytelaars, T. – Gool, L. V. 2008. ***Speeded-up robust features (surf)***.Computer Vision and Image Understanding, 110(3):346 § 359, 2008.

[6] Rosten, E. – Drummind, T. ***Machine learning for high-speed corner detection***. In A. Leonardis, H. Bischof, and A. Pinz, editors, Computer Vision – ECCV 2006, volume 3951 of Lecture Notes in Computer Science, pages 430§443. Springer Berlin / Heidelberg, 2006.

[7] Tola, E. – Lepetit, V. – Fua, P. 2008. *A fast local descriptor for dense matching*. In Conference on Computer Vision and Pattern Recognition, 2008.

[8] Calonder, M. – Lepetit, V. – Strecha, C. – Fua, P. 2010. ***Brief: Binary robust independent elementary features.*** In Computer Vision – ECCV 2010, volume 6314of Lecture Notes in Computer Science, pages 778-792. Springer Berlin / Heidelberg, 2010.

[9] Rublee, E. – Rabaud, V. – Konolige, K. – Bradski, G. 2011. *ORB: An efficient alternative to SIFT or SURF*, Computer Vision (ICCV), 2011 IEEE International Conference on, vol., no., pp.2564-2571, 6-13 Nov. 2011

[10] Lepetit, V. – Fua, P. 2006. ***Keypoint recognition using randomized trees***, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol.28, no.9, pp.1465-1479, Sept. 2006 doi: 10.1109/TPAMI.2006.188

[11] Ventura, J. – Hollerer, T. 2011. *Fast and scalable keypoint recognition and image retrieval using binary codes*, *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, vol., no., pp.697-702, 5-7 Jan. 2011 doi: 10.1109/WACV.2011.5711573

[12] Haladova, Z. – Sikudova, E. 2011. ***Limitations of the SIFT/SURF based Methods in the Classifications of Fine Art Paintings***. In Computer Graphics and Geometry. Vol.12 No. 1, 2010 the summer issue, s. 40-50. ISSN 1811-8992.

[13] Haladová, Z. – Šikudová, E. 2013. Combination of global and local features for efficient classification of paitings, in Proc. Spring Conference on Computer Graphics SCCG 2013, Bratislava, 2013, pp. 21-27

[14] Freund, Y. – Schapire, R. E. 1999. A Short Introduction to Boosting. In *Journal of Japanese Society for Artificial Intelligence*, 14(5):771-780, September, 1999.

[15] Šikudová, E. 2007. Comparison of color spaces for face detection in digitized paintings, In. Proc. Spring Conference on Computer Graphics SCCG 2007, pp. 135-140

[16] Yarbus, A.L. 1967. Eye movements during perception of complex objects. In Eye Movements and Vision, Plenum Press, New York, Chapter VII, pp. 171–196.

[17] Itti, L. et al., 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 11, 1254–1259.

[18] Bur, A. – Hugli, H. 2007. Optimal cue combination for saliency computation: A comparison with human vision. In Proc. . 2nd int. work-conference on Nature Inspired Problem-Solving Methods in Knowledge Engineering, *IWINAC '07*, pages 109-118.

[19] Le Meur, O. – Le Caller, P. – Barda, D. 2007. Predicting visual fixations on video based on low-level visual features. Vision Res.

[20] Oliva, A. – Torralba, A. – Castelhano, M. S. – Henderson, J. M. 2003. Top-down control of visual attention in object detection. In Proc. of the IEEE Int'l Conference on Image Processing (ICIP '03).

[21] Zhang, L. – Tong, M. – Marks, T. – Shan, H. – Cottrell G. 2008. Sun: A Bayesian framework for saliency using natural statistics. J Vis, 8(7):32.1§20.

[22] Judd, T. – Ehinger, K. – Durand, F. – Torralba, A. 2009. Learning to predict where humans look. In IEEE International Conference on Computer Vision (ICCV).

[23] Fukuma, S. –Tanaka, T. – Nawate, M. Switching wavelet transform for ROI image coding. IEICE Trans. Fundam. Electron. Comm. Comput. Sci., E88-A(7):1995-2006, July 2005.

[24] Ouerhani, N. – Bracamonte, J. – Hugli, H. – Ansorge, M. – Pellandini, F. Adaptive color image compression based on visual attention. In Proc. Image Analysis and Processing, pages 416-421, Sep 2001.

[25] Harding, P. – Roberston, N. Task-based visual saliency for intelligent compression. In Signal and Image Processing Applications (ICSIPA), 2009 IEEE International Conference on, pages 480 – 485, Nov. 2009.

[26] Wei, L. – Sang, N. – Wang, Y. – Wang, D. –Wang, F. Variable resolution image compression based on a model of visual attention. Pages 74950P, 2009.

[27] Kučerová, J. – Polec, J. – Tarcsiová, D. 2012. Video quality assessment using visual attention approach for sign language. volume 65, pages 194§-199. World Academy of Science, Engineering and Technology.

[28] Smeaton, A. F. – Over, P. – Kraaij, W. 2006. Evaluation campaigns and TRECVid, *MIR'06*: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, 2006, Santa Barbara, California, USA, ACM Press, pp 321—330,

[29] Over, P. et al. 2011. TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics, *TRECVID 2011*

[30] Inoue, N. – Shinoda, K. 2011. A Fast MAP Adaptation Technique for GMM-supervector-based Video Semantic Indexing Systems. In Proc. of *ACM Multimedia* (short paper), 2011

[31] Inoue, N. et al. 2010. High-Level Feature Extraction using SIFT GMMs and Audio Models. In Proc. of *ICPR*, 2010.

[32] Inoue, N. et al. 2011. TokyoTech+Canon, *TRECVID 2011*

[33] Zhao, Z. et al. 2011. BUPT-MCPRL, TRECVID 2011

[34] Le, D. – et al. 2011. National Institute of Informatics, Japan, *TRECVID 2011*

[35] Natarajan, P., 2011. BBN VISER TRECVID 2011 Multimedia Event Detection System, *TRECVID 2011*

Zuzana Černeková, Zuzana Haladová,
Júlia Kučerová, Elena Šikudová
FMFI UK Bratislava