KEY FRAME PARTITION MATCHING FOR VIDEO SUMMARIZATION

Olena Mikhnova, Nataliia Vlasenko

Abstract: *Summarization of video content is a complex task that requires feature selection and frame matching. To extract meaningful frames, named key frames, we have proposed partitioning of frames with Voronoi diagrams for further region matching throughout the video sequence. A unique partition metric has been used that takes into account color and textural, structural and geometric properties of Voronoi regions. Feature set designed for CBIR and CBVR has been analysed. The reasonable feature selection and incorporation into frame matching process has permitted to obtain competitive results. All these actions have allowed revealing significant changes in content while omitting slight deflections and repeats. Key frame extraction procedure has been described in detail. The proposed method has been checked on different test samples and compared with existing methods for precision and recall.*

Keywords: *key frame extraction, generator point detection, Voronoi diagram, partition metric, feature set.*

ACM Classification Keywords: *I.2.10 Vision and Scene Understanding (Video analysis), I.4.6 Segmentation (Edge and feature detection)*

Introduction

Multimedia utilization has greatly increased during the last decade, the amount of digital libraries had grown to enormous sizes and users started requiring instruments to deal with these data. Video is the most informative type of multimedia, as it consists of audio and graphical information simultaneously. Moreover, this information dynamically changes in time, which is one of the main difficulties for processing and analysis. Video summarization is among video recognition tasks that still lacks in performance and accuracy of computational procedures [Sonka, 2007].

Video summarization techniques are engaged in archiving, browsing and searching, cataloging and indexing, as well as improvement of information overload. It aids in maintenance of usability and accessibility to stored videos. Summarization may be of two types: static and dynamic. The last one is usually called skimming. The product of video summarization is a set of meaningful static images that depict video content, while the product of skimming is a shortened video [Laganière, 2008]. The subject of our research is the first type of summarization, static key frames.

The origin of video summarization and skimming comes from movie editing, when a film director decides which frames should be cut off. The relation of initial material to the resultant is usually not bigger than 20:1 [Rubin, 2005]. There are many movie editing techniques, but not a single rule of doing so. Acceptance of some frames and omission of others is a point of individual, cultural, or even political taste of an editor [Goldman, 2007]. American film director and scenario writer, John Huston stated that video editing should omit identical frames, as human brain recognizes graphical information by ignoring objects that have already been seen [Ward, 2008].

To address the issues mentioned above, video summarization has been attracting more research and development efforts in recent years. Despite of variety of already released methods for video summarization, the

main problem they face is the gap between information retrieved from video and semantic description required for efficient summarization [Laganière, 2008]. Another great challenge is that frames are obtained under different exposure, lighting conditions, aperture, focus and focal length of camera. That is why several kinds of features should be considered at the same time to obtain complete description and find similar frames. To create short and simultaneously comprehensive overview of a video, meaningful visual features are described in section one, and similarity measures for frame partitions are provided in section two. Experimental results are given in the third section, and the last section presents our conclusions.

Feature Set for Frame Partitioning

Spatio-temporal features are usually calculated for salient points [Laganière, 2008], local areas (detected objects or regions of interest) [German, 2005], or the whole frame [Lin, 2013]. In order to find significant features, we should look for parameters picked out by humans for visual information interpretation. Color, texture and context features are three main components used by people for video understanding. These parameters are not considered separately from each other, as they are closely related [Haralick, 1973].

Color features are traditionally analyzed with histograms that depict frequency of occurrence of one or another color tone in the region of interest. Sometimes only one color channel with the most significant changes is taken for analysis, sometimes an average, maximum or minimum value from a local range of a histogram is used. Color features are often considered in a form of intensity which is the simplest way of image color presentation. It is computed as an average from red r, green g and blue b component of an image with RGB color scheme [Bezdek, 2005].

$$\text{intensity} = \frac{r + g + b}{3}. \tag{1}$$

Comparison of histograms H_1 and H_2 can be obtained from their intersection or by calculus of correlation $C(H_1, H_2)$ between them, where μ_1 and μ_2 are average values for H_1 and H_2 respectively [Lin, 2013]. The higher correlation is, the more alike two histograms are.

$$C(H_1, H_2) = \frac{\sum (H_1 - \mu_1)(H_2 - \mu_2)}{\sqrt{\sum (H_1 - \mu_1)^2} \sqrt{\sum (H_2 - \mu_2)^2}}. \tag{2}$$

Textural features contain information about spatial distribution of changes in color tone for the whole image or its local part. In order to characterize texture, one may use any of 28 textural features proposed in [Haralick, 1973]. Though, it is important to note that they highly correlate with each other. Despite these features were proposed in 1973, many contemporary scientists turn towards them [Schonfeld, 2010]. Very often statistic degree of randomness, called entropy, is chosen for texture analysis.

From the point of video analysis, entropy describes spatial relations between brightness of frame pixel pairs, where $p(i, j)$ is an element of normalized matrix that describes spatial distribution of color tones in a frame (or local region) [Haralick, 1973].

$$E = -\sum_i \sum_j p(i, j) \log_2 (p(i, j)). \tag{3}$$

High entropy indicates large scatter of pixel values, while low entropy says about pixel homogeneity (and details consequently). Thus, entropy shows how much details consist in a local region, for which entropy value has been calculated [German, 2005]. This hypothesis can be easily proved by taking the same image with different resolution. The higher resolution is, the more details are visualized and the higher entropy value is (though such entropy changes are not that significant for an image with the same content). Along with the mentioned above

methods for texture analysis, there are methods based on auto-regression, Markov chains, mathematical morphology, fractals, wavelets, etc. [Sonka, 2007].

Such features as brightness, estimation of object borders, area, shape (using geometric matching), absolute and relative location, density and speed of motion, trajectory and many others are often used. Motion density and speed are usually estimated by optical flow [Schonfeld, 2010]. Though, by application of optical flow we bulk significantly the computational procedure compared with any other feature set. Trajectories of object motion are calculated with the help of differential images which may not account direction of motion. To save information about motion direction, cumulative differential image should be used. Such kind of images also enables to save some other temporal properties of motion, motion of huge objects and slow motion [Sonka, 2007]. Another interesting approach consists in analysis of structural features. Object shape is flooded with water-filling algorithm, filling time and shape length are considered [Zhou, 2001].

Great success has been achieved during the past years at the level of image understanding. Despite this, many questions remain undecided in context analysis, and researchers continue working on this field [Sonka, 2007]. Contextual features are referred to high level features, they include information from graphical data blocks around the area of interest. Such image description assumes model development for each recognized object, identification of regions with potential object samples.

High level description can be performed with low level features, assuming their absolute or relative spatial location on images, or by application of artificial intelligence methods for their processing (such as fuzzy production rules, heuristics, cluster analysis, neural nets, different filters, etc.) [Depalov, 2006; Zhang, 200;, Schonfeld, 2010]. Examples of 44 systems, where such an approach is realized, are given at [Veltkamp, 2001]. Dominant colors, histogram analysis of separate regions and the whole frames, histogram correlation, coherent color vectors, mean colors are used as color features here. Border pixel statistics, local binary patterns, random field, elementary textural features, wavelet analysis and Fourier transformation are used as textural features in CBIR systems. Ellipses and bounding boxes, Fourier descriptors, elastic models, different curves and patterns are often used to define shape features.

Another approach to high level description lies in assignment of textual labels for different image classes by construction of semantic nets based on thesaurus. Textual label correspondence to the particular image class is defined by users who train a system. Such recognition algorithms search for similarity inside semantic networks and consider integrated visual features [Carneiro, 2007; Divakaran, 2009].

Although, due to inability of full-scale recognition implementation (similar to human mental activity), it is used to speak about mid level features that link semantic description with low level features [Boureau, 2010, Schonfeld, 2010]. For the purpose of frame partitioning we consider traditional color, textural and spatial features, taking into account relative location of regions and their regional properties, which give us a chance to obtain meaningful segments and extract key frames in future.

Matching of Frame Partitions

Frame partitioning is proposed to be performed with Voronoi diagrams. This method has been chosen because of several reasons. First of all, frame partitioning into real objects is not reasonable because of their tremendous changes in time. For now it is impossible to process all of them efficiently at the same time. And their changes may cause fault detection and object mess. Secondly, this partition technique requires less computational resources than real object segmentation and much less than motion analysis. And thirdly, Voronoi diagrams have not been used yet for the purpose of video summarization, thus, we want to develop a novel method and check its efficiency by comparing with existing ones.

Voronoi diagrams were first mentioned by R. Descartes in 1644. Later, in 1850, they were declared by P.G.L. Dirichlet, and further named after Russian mathematician G.F. Voronoi [Okabe, 2000]. To give formal definition of a Voronoi tessellation, let us denote $D=[a,b]\times[c,d]$, $a,b,c,d=const$ as a field of view. Let $\{p_1, p_2, ..., p_n\}$ be a finite set of generator points selected by Harris method that takes into account pixel intensity and relative location of regions. (Harris method [Sonka, 2007] has been chosen as one of the most frequently used with good performance and relative simplicity.) Voronoi diagram is a field of view partition $V=\{v(p_1)\cap D, v(p_2)\cap D, \ldots, v(p_n)\cap D\}$ into convex polygons, s.t.

$$v(p_i) = \{z \in R^2 : d(z,p_i) \le d(z,p_j)\ \forall i \ne j\} \tag{4}$$

where $d(\circ,\circ)$ is a planar Euclidean metric [Okabe, 2000].

To define a key frame, let $B_k(z), z=(x,y)\in D$ be the k-th frame from video sequence Φ (here and subsequently $k=1,2,\ldots,K$ is a discrete time). If $1\le i<j\le K$ and $B_i(z), B_j(z)\in\Phi$ then we shall use notation $S_l(i,j)=[B_i(z),B_j(z)]$, $l=1,2,\ldots$, $i,j\in L_l$, $\sum L_l = K$ for a scene that is a set of sequential frames obtained after temporal segmentation into meaningful segments, s.t. $\forall l\ S_l(i,j)\ne\varnothing$, $\Phi=\bigcup_{l\in L} S_l(i,j)$, $\forall l', l''\ S_{l'}(i,j)\cap S_{l''}(i,j)=\varnothing$. For a fixed l, define a key frame as an image $B_r^*(z)\in S_l(i,j)$ with property

$$r=\arg\min_{r\in L_l}\left(\sum_{t\in L_l, r\ne t}\rho(B_r(z),B_t(z))\right) \tag{5}$$

where $\rho(\circ,\circ)$ is a metric. After all the key frames are extracted, we obtain the set $\{B_l^*(z)\}$ of key frames for video stream Φ. In other words, we extract a frame (or several) per scene, and each key frame extracted is the most representative one for its scene (or subscene).

Incorporation of matching procedure has not been done yet for Voronoi tessellations, except by Yukio Sadahiro [Sadahiro, 2011]. He introduced different methods of visual and quantitative analysis, including χ^2, Kappa index and their extensions, area and perimeter of tessellations, their variance and standard deviation, spatial mean of their gravity centers, etc. His idea was to implement granularity density measure and hierarchy relationships (overlay, partial overlay and inclusion) to compare different Japanese administrative region division systems, though the areal methods are quite ambiguous for video processing application, as objects may be shot at different zoom. Different objects in images may possess the same area. Thus, video objects cannot be traced with properties primarily based on area. In our case different attributes are needed to be considered. For our purposes we used spatial, textural and color features which are among the main attributes used for CBIR and CBVR.

To match frame partitions, consider two frames $B'(z), B''(z)$ with generator points $\{p'_1, p'_2, ..., p'_n\}$ and $\{p''_1, p''_2, ..., p''_m\}$ respectively, then spatial dissimilarity of frames can be approximately represented by partition metric $\rho_1(V',V'')$ [Mashtalir, 2006]

$$\rho_1(V',V'')=\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{card}(v(p'_i)\Delta v(p''_j))\,\mathrm{card}(v(p'_i)\cap v(p''_j)) \tag{6}$$

where $v(p'_i)\Delta v(p''_j)=(v(p'_i)\setminus v(p''_j))\cup(v(p''_j)\setminus v(p'_i))$ is symmetric difference that counts the number of elements on which $v(p'_i)$ and $v(p''_j)$ differ [Yianilos, 1991].

The above distance measure shows how two diagrams match each other in terms of regions. To take into account color and textural features, let us define two more metrics, $\rho_2(B'(z),B''(z))$ and $\rho_3(B'(z),B''(z))$

respectively which are defined in common regions of partitions. By measuring similarity in color and texture we observe changes between Voronoi regions of two frames being analyzed. Squared Euclidean distance has been used to incorporate more weight for distant color objects. Manhattan distance has been chosen for textural similarity measurement, as entropy values are calculated for the whole regions and they are presented by a single float value per region, while similarity in color is taken from each pixel present in both frames.

$$\rho_2(B'(z),B''(z))=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{x_q}\sum_{y_u}{}_{(x_q,y_u)\in(v(p'_i)\cap v(p''_j))}(B'(x_q,y_u)-B''(x_q,y_u))^2\ ,$$
$$\rho_3(B'(z),B''(z))=\sum_{i=1}^{n}\sum_{j=1}^{m}{}_{(v(p'_i),v(p''_j))\supseteq(v(p'_i)\cap v(p''_j))}\left|E(v(p'_i))-E(v(p''_j)\right|\ . \tag{7}$$

where $B'(x_q,y_u)$ is intensity value for pixels in a region $(v(p'_i)\cap v(p''_j)$, and $E(v(p'_i))$ is entropy value in a region $v(p'_i)$.

Thus, we have got non-normalized estimates. For this reason we offer to normalize formulas (6) and (7) to obtain values ranging from 0 to 1. Conversion of the above metrics to bounded forms assumes application of a function, named range compander [Yianilos, 1991], s.t. its combination with a metric still gives a metric which satisfies non-negativity, reflexivity, symmetry and triangle inequality rules.

$$\rho'(B'(z),B''(z))=\frac{1}{1+\rho(B'(z),B''(z))} \tag{8}$$

As non-negative linear combination of metrics is still a metric, we may propose the following resulting metric:

$$\hat{\rho}(B'(z),B''(z))=\alpha_1\rho'_1+\alpha_2\rho'_2+\alpha_3\rho'_3\ ,\quad \sum_{\gamma=1}^{\gamma=3}\alpha_\gamma=1\ ,\ \alpha_\gamma\geq 0 \tag{9}$$

where $\hat{\rho}(B'(z),B''(z))$ shows similarity between frames, and α_γ shows the impact of each feature in use.

In order to extract frames with lowest level of proximity, we should compare consecutive frames pair-wise. Tessellation matching algorithm for key frame selection is described below.

1. **Determine homogeneity of video content.** Calculate texture variance (dispersion of entropy) throughout the video sequence, and set a threshold value according to the following rule:

$$Threshold=\begin{cases}\frac{1}{4},\ \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z))-\frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2\to\infty\\ \frac{1}{2},\ \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z))-\frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2\to\frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\\ \frac{3}{4},\ \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z))-\frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2\to 0\end{cases} \tag{10}$$

where $E(B_k(z))$ is entropy for k-th video frame, K is a total number of frames in the video sequence.

A threshold value should be set according to video content. For videos with heterogeneous content and variety of scenes (see fig. 2) this value should be less than $\frac{1}{4}$, not to extract too much frames. On the contrast, for videos with homogeneous content (see fig. 3) and small number of scenes (or even a single scene) this value should be increased up to $\frac{3}{4}$, to extract a bit more frames.

2. Take the first ($B_k(z)$) and the second ($B_{k+1}(z)$) frames for comparison. Set $k = 1$.

3. **Frame matching.** According to formula (9), compute $\hat{\rho}(B_k(z), B_{k+1}(z))$ for two frames. If $\hat{\rho}(B_k(z), B_{k+1}(z))$ is less than the predefined threshold value, then extract both frames $B_k(z)$ and $B_{k+1}(z)$ as key frames $B_r^*(z)$ and $B_{r+1}^*(z)$ and go to step 4, otherwise extract only $B_k(z)$ as a key frame $B_r^*(z)$ and go to step 5.

4. Set $B_k(z) = B_{k+1}(z)$, $B_{k+1}(z) = B_{k+2}(z)$ and go to step 6.

5. Leave $B_k(z) = B_k(z)$ and set $B_{k+1}(z) = B_{k+2}(z)$.

6. Repeat step 3, until $B_{k+1}(z) \le K$.

7. **Inter-scene key frame comparison.** Thus, we have obtained a key frame per scene: $\{B_1^*(z) \in S_1(i, j)\} \ldots \{B_l^*(z) \in S_l(i, j)\}$. Compare key frames pair-wise between scenes using frame matching procedure, defined in step 3. Second identical key frame is to be deleted. Thus, the resulting key frame sequence will be $\{\hat{B}_1^*(z), \ldots, \hat{B}_l^*(z)\}$.

Experimental Results

The proposed method has been tested on low resolution Trecvid video samples, several commercials of medium resolution and self-made high definition videos. Key frames extracted from Chinese commercial about Mercedes Benz C-Class automobile are shown in fig. 2. Examples of partitioning of frames (with homogeneous content and high definition) using Voronoi diagrams are provided in fig. 3.

Test results have been compared with existing summarization techniques based on clustering, curve simplification, and motion analysis. This comparison has shown good balance between high precision and recall for the proposed procedure. The estimation has been performed by 10 respondents who knew nothing about the name of frame extraction method they tested.

Figure 2. Key frames extracted from Chinese commercial

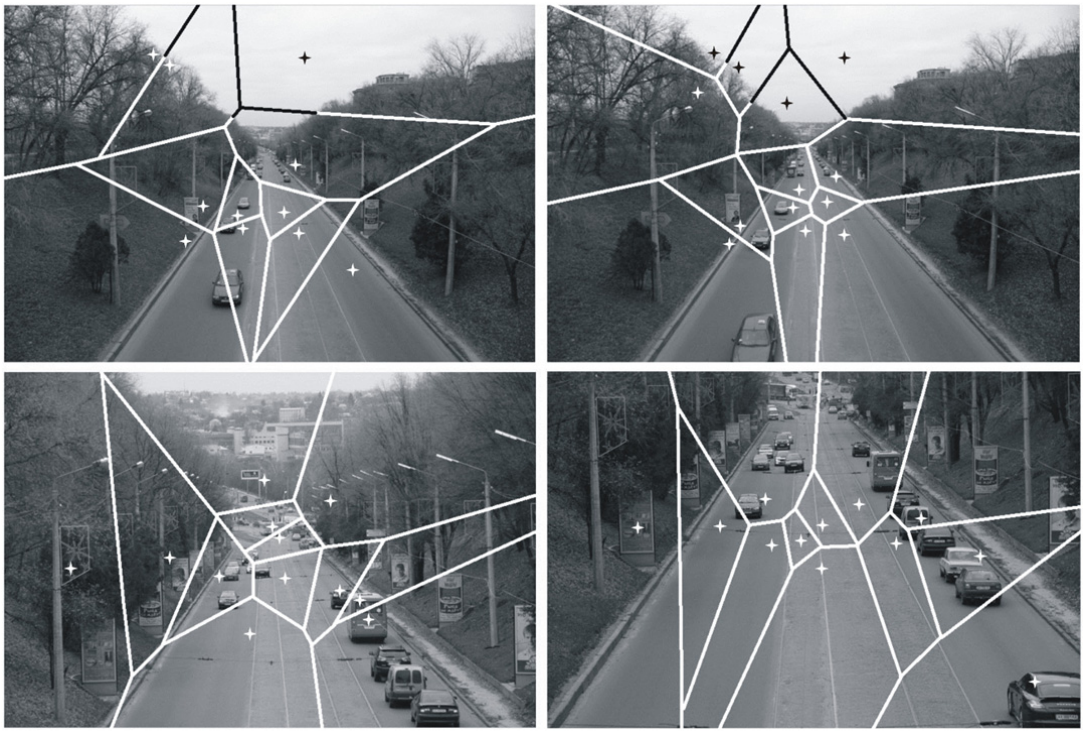


Figure 3. Examples of frame partitioning using Voronoi diagrams

Conclusion

By analyzing the difficulties faced by researchers during video summarization, we have come to the conclusion that the main problem lies in the gap between semantic content and low level frame presentation. To make an attempt of overcoming this gap, we have proposed a new method of key frame extraction based on Voronoi partitioning of frames, which assumes spatial features, color, texture, and relative location of regions.

The proposed method differs from existing ones in accuracy of results and computational uniqueness of matching procedure. The accuracy of results is reached due to generalized procedure of region processing. Existing algorithms reveal changes almost at each frame, though these changes may not be that important, while the proposed method returns only key frames with significant changes in content. Shape changes are dramatic at each frame, but Voronoi region is quite stable. It has been shown, that frames with identical content are partitioned in a similar manner with Voronoi diagrams.

The proposed method takes into account video content homogeneity by setting the appropriate threshold value before matching the frames. Key frames are compared with each other between video scenes, detected using the technique proposed in [Bodyanskiy, 2012]. Duplicate key frames are removed with the second pass of the algorithm.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[Bezdek, 2005] J.C. Bezdek et al. Fuzzy models and algorithms for pattern recognition and image processing. New York: Springer, 2005, 776 p.

[Bodyanskiy, 2012] Y. Bodyanskiy et al. On-line video segmentation using methods of fault detection in multidimensional time sequences. In: International Journal of Electronic Commerce Studies, 2012, Vol. 3, No. 1, pp. 1-20.

[Boureau, 2010] Y.-L. Boureau et al. Learning Mid-Level Features For Recognition. In: Computer Vision and Pattern Recognition, 2010, pp. 2559-2566.

[Carneiro, 2007] G. Carneiro et al. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 3, 2007, pp. 394-410.

[Depalov, 2006] D. Depalov et al. Perceptual feature selection for semantic image classification. In: IEEE International Conference on Image Processing, Vol. 2, 2006, pp. 2921-2924.

[Divakaran, 2009] A. Divakaran. Multimedia Content Analysis: Theory and Applications (Signals and Communication Technology). New York: Springer, 2009, 390 p.

[German, 2005] A. German, M.R. Jenkin, Y. Lesp´erance. Entropy-based image merging. In: 2-nd Canadian Conference on Computer and Robot Vision, 2005, pp.81-86.

[Goldman, 2007] D.R. Goldman. Framework for video annotation, visualization, interaction: Doctoral Thesis. Washington, 2007. – 140 p.

[Haralick, 1973] R.M. Haralick, K. Shanmugam, I. Dinstein. Textural features for Image Classification. In: IEEE transactions on systems, man and cybernetics, Vol. 3, No. 6, 1973, pp. 610-621.

[Laganière, 2008] R. Laganière et al. Video Summarization from Spatio-Temporal Features. In: 2-nd ACM TRECVid Video Summarization Workshop, 2008, pp. 144-148.

[Lin, 2013] G.-S. Lin, J.-F. Chang. Detection of frame duplication forgery in videos based on spatial and temporal analysis. In: International Journal of Pattern Recognition and Artificial Intelligence, Vol. 26, No. 7, 2013.

[Mashtalir, 2006] V. Mashtalir et al. A novel metric on partitions for image segmentation. In: IEEE International Conference on Video and Signal Based Surveillance, 2006.

[Okabe, 2000] A. Okabe et al. Spatial tessellations: Concepts and applications of Voronoi diagrams. – 2-nd ed. – Chichester: Wiley, 2000, 671 p.

[Rubin, 2005] M. Rubin. Droidmaker: George Lucas and the digital revolution. – Gainesville: Triad Publishing, 2005, 518 p.

[Sadahiro, 2011] Y. Sadahiro. Analysis of the relationship among spatial tessellations. In: Journal of Geographical Systems, Vol. 13, No. 4, 2011, pp. 373-391.

[Schonfeld, 2010] D. Schonfeld et al. Video Search and Mining. In: Studies in Computational Intelligence, Vol. 287, Berlin: Springer, 2010, 388 p.

[Sonka, 2007] M. Sonka, V. Hlavac, R. Boyle. Image Processing, Analysis, and Machine Vision, International Student Edition. – 3 ed. – Toronto: Thomson, 2007, 850 p.

[Veltkamp, 2001] R.C. Veltkamp, H. Burkhardt, H.-P. Kriegel. State-of-the-Art in Content-Based Image and Video Retrieval (Computational Imaging and Vision). Netherlands: Kluwer Academic Publishers, 2001, 343 p.

[Ward, 2008] K. Ward. Augenblick: The Concept of the 'Decisive Moment' in 19th and 20th Century Western Philosophy. Aldershot: Ashgate, 2008, 192 p.

[Yianilos, 1991] P.N. Yianilos. Normalized forms for two common metrics. In: NEC Research Institute, Report 91-082-9027-1, 1991, Revision 7/7/2002. Cambridge: Cambridge University Press, 1991, 7 p.

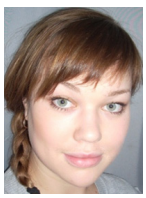
[Zhang, 2008] D. Zhang, Y. Liu, J. Hou. Digital Image Retrieval Using Intermediate Semantic Features and Multistep Search. In: Digital Image Computing: Techniques and Applications, 2008, pp. 513-518.

[Zhou, 2001] X.S. Zhou, T.S. Huang. Edge-Based Structural Features for Content-Based Image Retrieval. In: Pattern Recognition Letters, Vol. 22, No. 5, 2001, pp. 457-468.

Authors' Information

Olena Mikhnova *– PhD student of Informatics department in Kharkiv National University of Radio Electronics, P.O. Box: 61166, Ukraine, Kharkiv, Lenina av., 14; e-mail: elena_mikhnova@ukr.net*

Major Fields of Scientific Research: Image and video recognition, Data mining



Nataliia Vlasenko *– PhD student of Informatics department in Kharkiv National University of Radio Electronics, P.O. Box: 61166, Ukraine, Kharkiv, Lenina av., 14; e-mail: gorohovatskaja@gmail.com*

Major Fields of Scientific Research: Image recognition and analysis