TEMPORAL VIDEO SEGMENTATION VIA SPATIAL IMAGE SEGMENTATION

Dmitry Kinoshenko, Sergii Mashtalir, Vladislav Shlyakhov

Abstract: *Temporal video segmentation techniques provide unwavering support for video parsing and content analysis. The foremost challenge in storage and retrieval system creation is bridging the semantic gap between low-level features and high-level interpretations. The middle-level image partitions are proposed to be used for video segmentation. Such spatio-temporal segmentation is discussed and experimentally investigated.*

Keywords: *video, segmentation, frame.*

ACM Classification Keywords: I.2.10 Vision and Scene Understanding (Video analysis*)*

Introduction

Development and application explosion of multimedia technologies bring forward active research on the subject of Content Based Video Retrieval (CBVR) with digital video libraries and archives of immense size. There arise four main interdependent problems involved in CBVR:

i. content analysis, i.e. converting a query into frame or video segment features, e.g. color, texture, thresholds, shape, structure, layout, etc., and vice versa, as well as translating at least some results into semantic concepts, ipso facto bridging the semantic gap;
ii. video structure parsing, i.e. detection of shots that contain frames with similar visual contents, pattern composition of scenes, stories and, finally, video stream;
iii. summarization or abstraction, i.e. creating a rich presentation of visual information about the video sequence structure;
iv. indexing, i.e. support of similarity search in unstructured video collections.

It is evident that shots are the basic units for video content analysis but lap dissolve, fade, side curtain wipe, sluggish panning etc. put significant hindrances in valid shot boundary detection. In addition, a query ‘ad exemplum’ requires representations of story images and/or temporal aspects of the video. Semantic gap between low-level features and high-level concepts is challenged to generate more suitable feature spaces. This work was intended as an attempt to motivate reasonably double-ply segmentation: spatial segmentation of each frame and temporal segmentation of frame partition sequences to produce shot sets as the basis of high-level processing and interpretation.

A video can be analyzed at different levels of granularity [Tekalp, 2005]. For object tracking, the low-level is an individual frame that is usually used for extracting static visual features like color, texture, shape, or interest points. Videos, as a rule, can be decomposed into shots [Gauch, 1999] which show a sequence of frames captured by a single camera in a single continuous action in time and space. High-level techniques require determining more descriptive segments, e.g. well explored tools of time series analysis may be used for time varying nature of video analysis [Bodyanskiy, 2012].

The segmentation of each frame of a video into homogeneous regions is an important issue for many applications such that region-based motion detection [Varcheie, 2010], image enhancement (since different processing may be applied to different regions) [Jie, 2010], 2D to 3D conversion, human identification [Poignant, 2012], data context search [Chaisorn, 2003], image extraction [Caplier, 2001]. These applications make two main demands to segmentation: accuracy of region boundaries in spatial segmentation and temporal stability of segmentation from frame to frame.

Generally spatial segmentation can be divided into two main categories, namely, contour-based [Iannizzotto, 2000] and region-based methods [Nock, 2004]. In the first category, edges are computed and connected components are extracted. This approach cannot take benefit of statistical properties of the considered image regions. At the same time, region-based segmentation methods avoid these drawbacks by considering regions as basic elements [Snoek, 2005], hence any regional based matching of images is more reliable.

Various algorithms are known for video segmentation. The first class of techniques proposes to perform a 3D segmentation by considering the spatio-temporal data as a volume. We can cite the work of [Wang, 2003] that takes benefit of the 3D structure tensor for segmentation. Some other recent works propose 3D approaches using a mean-shift-based analysis 2 VLSI Design [DeMenthon, 2002]. Let us note that if each shot is segmented as a 3D volume, the number of frames to store may be unbounded for each segmentation.

The second class of methods concerns frame-by-frame algorithms. In these approaches, the spatial segmentation of the second frame is deduced from the spatial segmentation of the first one using motion estimation [Wang, 1998]. Matching is performed between regions of different frames in such approaches. All the regions are then linked and video object tracking algorithms [Patras, 2002] may take benefits of correspondence between regions.

We take the view that partition matching provides middle-level of feature analysis in the form of so-called image 'spatial content', that is a small but important step toward video understanding. The remaining sections are organized as follows. The next section deals with models of spatio-temporal video segmentation. Further, experimental results and discussion are presented.

Spatio-temporal segmentation

Considering existing approaches to video parsing, it is important to pay attention to possibility of video segmentation by means of analysis of separate frames, in particular, through their segmentation with further selection of different region features, and primarily, field of view partitions as a whole. So, we may talk about double-ply segmentation i.e. video temporal segmentation on the base of spatial frame segmentation matching. Thus, it is necessary to consider video analysis to take into account separate shot properties induced by traditional spatial segmentation. It is one of the basic objects of studying, at least, from the point of view of visual information 'spatial content' research.

Assume that field of view D at fixed time can be represented by partition $R=\{r_\beta\}$ composed of L regions ($\beta=\overline{1,L}$). Taking into account that $\mathrm{D}=\cup_\beta R_\beta$ and $\forall\beta',\beta''\in\{1,2,\ldots,L\}:\beta'\neq\beta''\Rightarrow r_{\beta'}\cap r_{\beta''}=\varnothing$, using e.g. ANOVA principles, we can indicate a criterion

$$C=\frac{1}{\mathrm{n}}\sum_{\beta=1}^{L}C_\beta\ \mathrm{card}\, r_\beta\ , \tag{1}$$

where C_β denotes nonnegative interregional and intraregional variance differences normalized by intraregional variance. It is necessary to minimize C on all possible segmentations $\{R\}$ for the fixed number of regions. Indeed, if segmentation is successful, i.e. detected regions are homogeneous, values C_β for every region are small, and consequently C tends to zero. The minimization of (1) can be fulfilled by various procedures, e.g. by well known color-texture algorithms, but with consideration of possible transitions 'partitions – coverings' under segmentation post-processing, passing to 'spatial contents' we may additionally use procedures intended for multiple meaning to obtain totally correct and complete segmentation of complex scenes. [Chupikov, 2007].

It should be emphasized that each detected region (equivalence class) can be interpreted as binary image, hence mathematical morphology tools are suitable to refine segmentations. Morphological filters are nonlinear filters suitable for filtering task. Firstly, a morphological filter may be used for restoring images corrupted by some type of noise. Secondly, a morphological filter may be used to selectively remove image structure or objects. Moreover, there exist many specific filters: rank opening [Ronse, 1986], segment based filters [Meyer, 1989]; orientation dependent filtering [Kurdy, 1989]); filters for graphs [Vincent, 1989] etc. Note, there are many approaches of morphological filtering like parallel combination, sequential combination, iterative combination, self-dual approach or toggle mapping (edge sharpening) [Soille, 2004]. But with regard to segmentation improvement we should talk only about second group of filters, allowing removing unwanted objects or their parts. Sequential filters are fully applicable in this case. For this purpose it is possible to use the derivative operations of mathematical morphology, such as opening γ and closing φ:

$$\gamma_{\mathrm{H}}(r_{\beta}) = \delta_{\mathrm{H}}(\varepsilon_{\mathrm{H}}(r_{\beta})),$$

$$\varphi_{\mathrm{H}}(r_{\beta}) = \varepsilon_{\mathrm{H}}(\delta_{\mathrm{H}}(r_{\beta})),$$

where $\delta_{\mathrm{H}}(r_{\beta})$=$\{z \in r_{\beta} : [(\mathrm{H}+z)\bigcap r_{\beta}] \subset r_{\beta}\}$ is the morphological dilation operation, $z = (x,y) \in \mathrm{D}$, $r_{\beta} \subseteq \mathrm{D}$ $\varepsilon_{\mathrm{H}}(r_{\beta}) = \{z \in r_{\beta} : (\mathrm{H}^{r} + z) \subset r_{\beta}\}$ is the morphological operation of erosion; H is a structural element and $\mathrm{H}^{r} = \{-z \in r_{\beta} : z \in \mathrm{H}\}$.

Opening can remove 'capes', 'isthmus' and 'islands' smaller than the structuring element and closing can fill 'gulfs', 'channels' and 'lakes' smaller than the structuring element. Then the simplest sequential filter can be built as follows: open-close or close-open. Difference between methods we'll see at fig.1. We will get different results depending on the sequence of operations. If in the first case we can unite part of objects lying close to each other (in the distance, not exceeding the sizes of structural element), then in the second case such objects are combined only in case when these objects are more then sizes of structural element (otherwise they will disappear after implementation of the first part of filtration). Thus, there are common situations: any small sized object (smaller than the structuring element), lying separately at the distance which exceeds the size of structural element, will disappear in any case. Another approach that can be applied is the usage of reconstruction

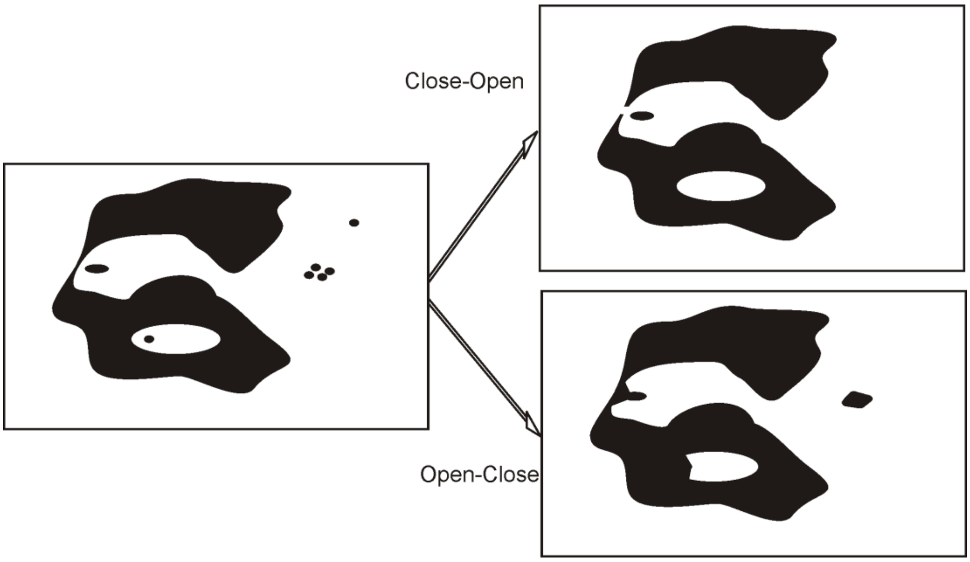

Figure 1. Difference between two methods

operators. In this case, we have iterative morphological filters with the following advantages: removing features smaller than the structuring element without altering the shape, reconstructing connected components from the preserved features. Reconstruction operators have greater exactness of element selection and removing of unnecessary objects, however, the process is iterative and its application is not the best one for video segmentation, as it requires considerable time and resource expenses that narrow the class of problems, where it can be used directly. This method used for segmentation improvement of some real image segment is illustrated on fig. 2.

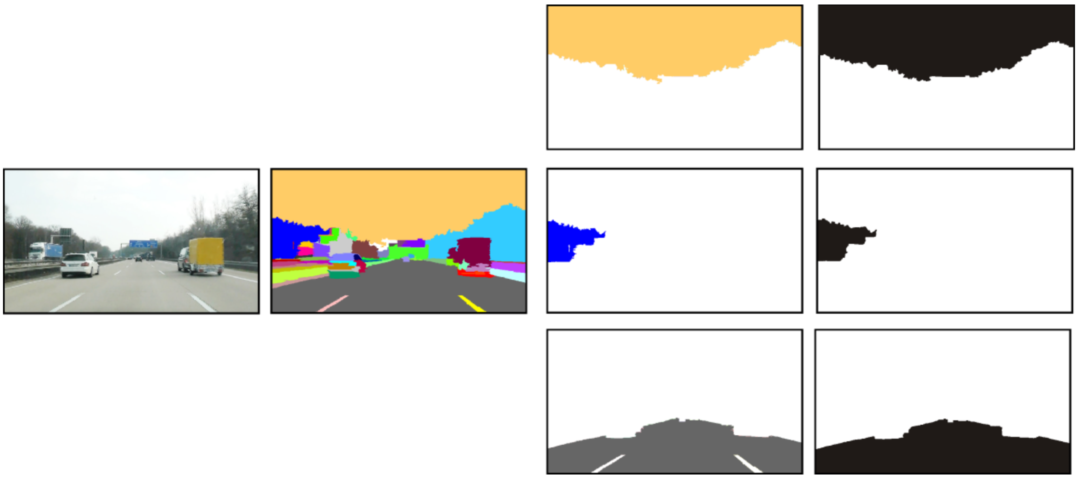

Figure 2 Morphological filtration for segmentation improvement

Before proceeding with experimental analysis of spatio-temporal video segmentation (or “double” segmentation), i.e. their granulation (in an ideal) on semantically equivalent sequences of shots, we will consider this process in a general view.

Consider video $\Phi = \{B_1(x,y), B_2(x,y),\ldots,B_K(x,y)\}$. Further we will disregard spatial coordinates $x,y \in D$, focusing attention on discrete time $k = \overline{1,K}$. If $i < j$ and $B_i, B_j \in \Phi$, then video segment $S[B_i,B_j]$ in video Φ is the set of ordered in times images: $\{B_k \in \Phi : k = \overline{i,j}\}$ where B_i is left bound of the segment $S[B_i,B_j]$, B_j is right bound of the same segment and $S(\Phi)$ is a power set of segments.

Let $P_i \subset \mathrm{R}^p$ be a feature vector (it is arbitrary but in fact we mean the partition P_i) which describes an image B_i, then each segment $S[B_i,B_j]$ is a set of values $\Pi_{ij} = \{(B_i,P_i),(B_{i+1},P_{i+1}),\ldots,(B_j,P_j)\}$. Using instantaneous (between two adjacent frames) dissimilarity of images, i.e. $\pi_k = \rho(P_{k-1},P_k)$, $k = \overline{2,K}$ where ρ is some metric on the feature set. Then we can describe feature changes in a segment $S[B_i,B_j]$ as

$$\Delta(\Pi_{ij}) = f(\rho(P_i,P_k),\pi_k),\ k = \overline{i+1,j}$$

where a function $f(\circ)$ takes into account similarity (dissimilarity) of segment based on current shot and similarity (difference) of two sequential in time images.

If value $\Delta(\Pi_{ij})$ is less than given or found from a priori defined threshold η, then $S[B_i,B_j]$ is the segment of video. There are sequences of frames, corresponding to the semantic concepts of ‘events’, ‘b-roll’ etc. based on segments. The number of stratification levels and their structure substantially depend on the current tasks of video data analysis.

We will distinguish only the concept of 'video group' $F_m \neq \varnothing$, inducing video segmentation

$$F_m = \bigcup_l S[B_{il}, B_{jl}]\,,\ \Phi = \bigcup_m F_m\,,\ F'_m \cap F''_m = \varnothing\,.$$

Generally, segments have to be sequential in time from the semantic point of view. For this reason, a video group is the basic object of video data analysis and a segment comes forward as a basic processing unit. Without loss of generality in consideration, we will suppose that a segment is the same as video group, uniting segments (incoherent in time, but close by metric) in a single one.

Quite often some key-frame $B_k \in S[B_i, B_j]$, which turns out to be a representative or synthesized from some averages in metric space, is a basic segment $S[B_i, B_j]$ description. On fig. 3 we can see video Φ, segments $S[B_i, B_j]$, key-frame B_k and spatial segmentation R_k by criteria (1).

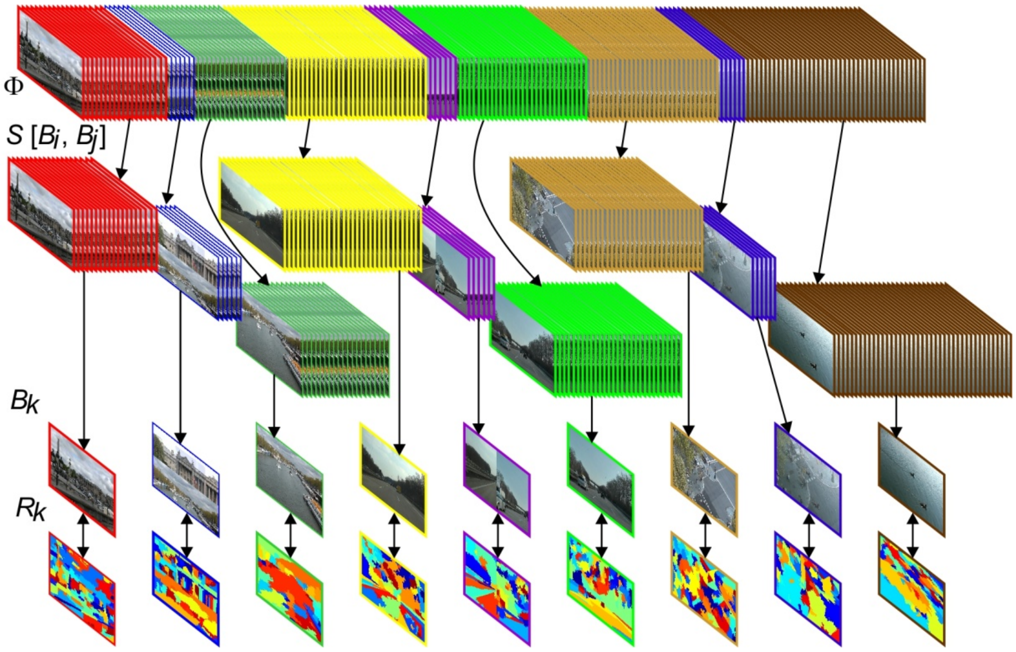

Figure 3 Model of video structure

Overall: first of all we need to get video granulation $\Phi = \{B_1, B_2, ..., B_K\}$ on segments $S_l[B_i, B_j]$, $l = 1, 2, ..., L$ such that $\Phi = \bigcup_m S_m$, $S_{m'} \cap S_{m''} = \varnothing$. As a criterion we will use dissimilarity between spatial frame segmentation R_k, $m = \overline{1, K}$.

Supposing the set cardinality as measure [Kinoshenko, 2007], we have

$$\rho(R', R'') = \sum_{i=1}^{n} \sum_{j=1}^{m} card(r'_i \Delta r''_j)\ card(r'_i \cap r''_j)\ , \tag{2}$$

where $R' = \{r'_i\}_{i=1}^{m}$, $R'' = \{r''_j\}_{j=1}^{n}$ is image $B', B'' \in \Phi$ partitions. Decision-making that an image belongs to the current segment is based on (2) by the one-dimensional time series analysis.

Thus, as a rule, methods on tracking signal are used, being an indicator of changes in controlled signals.

In practice the widest distribution was got by heuristic procedures, such as methods of Chow, Brown, Trigg-Leach etc. Authentication of gradual changes is performed by the basic Brown methods. Sharp fluctuations are better determined by means of Trigg-Leach tracking signal.

Experimental results

The aim of experimental researches is studying of video segmentation, specific for determination of segments with group correlation between objects.

All the experiments were held on video with 1500 frames. The duration of frame equals to a minute. Initial resolution is 1920×1080 pixels. Typical example of video is shown on fig. 3. It contains 9 segments, 3 of which actually are smooth interfering transitions from one to the next.

Obviously, that initial image quality is particularly significant for spatial segmentation of each frame. Fig. 4 illustrates changes in spatial frame segmentation (in pairs, consistently in time) for the same segmentation algorithm. Evidently, it is possible to find a reasonable compromise between complicated calculation and adequacy of results, in particular, the segments are better recognized from video data at resolution 480×272 and 720×412 pixels. Segments are practically not extracted at resolution 246×136 pixels. Also segments are difficult to extract at resolution 1280×720 and 1920×1080.

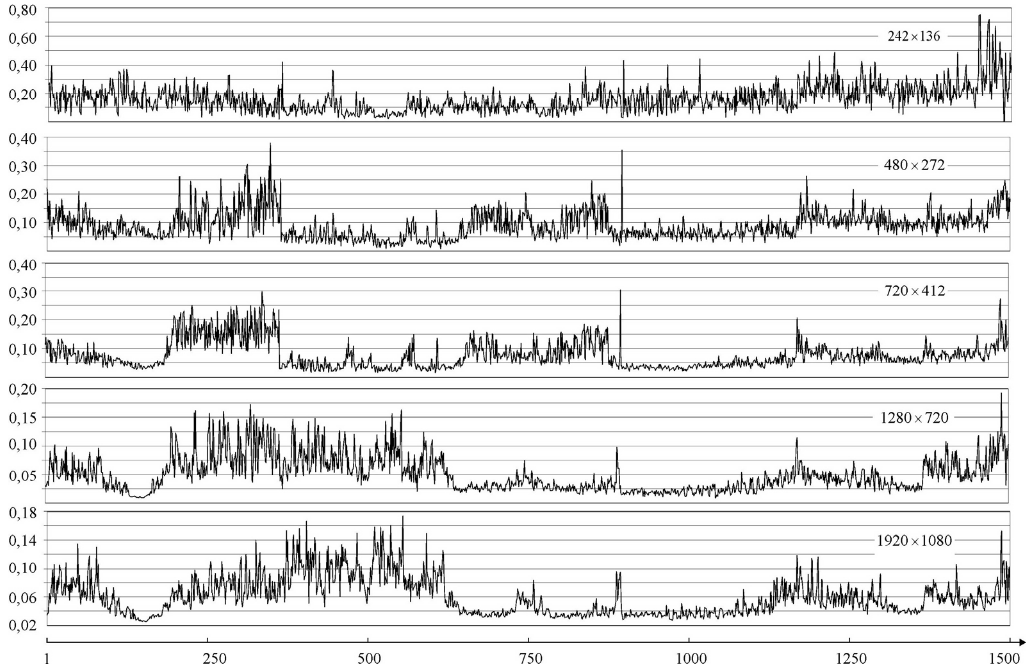

Figure 4. Spatial segmentation dependence on video resolution

It is necessary to underline that a choice of spatial frame segmentation algorithm influences on video temporal segmentation. Moreover, the same spatial frame segmentation algorithms with different parameters influence different object detailing. The influence of segmentation algorithm parameters is shown on fig. 5. As shown, a transition moment from one segment to another is clearly visible at the second and third charts, but for the first chart everything is not so obvious, though very distinct.

Conclusion

Video segmentation on the basis of spatial segmentation analysis of separate shots or so-called “double” segmentation (spatio-temporal) is considered. Initial video data is analyzed with its granulation of separate segments, characterizing fragments with single meaning. Initial video data influence on resolution choice for resultant spatio-temporal segmentation is considered, and also the algorithm is analyzed for parameter choice in terms of segmentation of separate shots. It should be noted that experimental results have shown that there is no need to take initial data with large resolution and conduct too detailed frame segmentation, as it results only in considerable increase of data processing without any visible improvement of video segmentation results (and maybe even with their degradation).

Consequently, this approach permits conducting video data segmentation in near to real-time mode that accordingly allows implementation for video data analysis. Thus, it can be used as an intermediate step on the way to key frame extraction from video data, as it allows allocating segments with similar content that facilitates key frame search.

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS (www.aduis.com.ua).

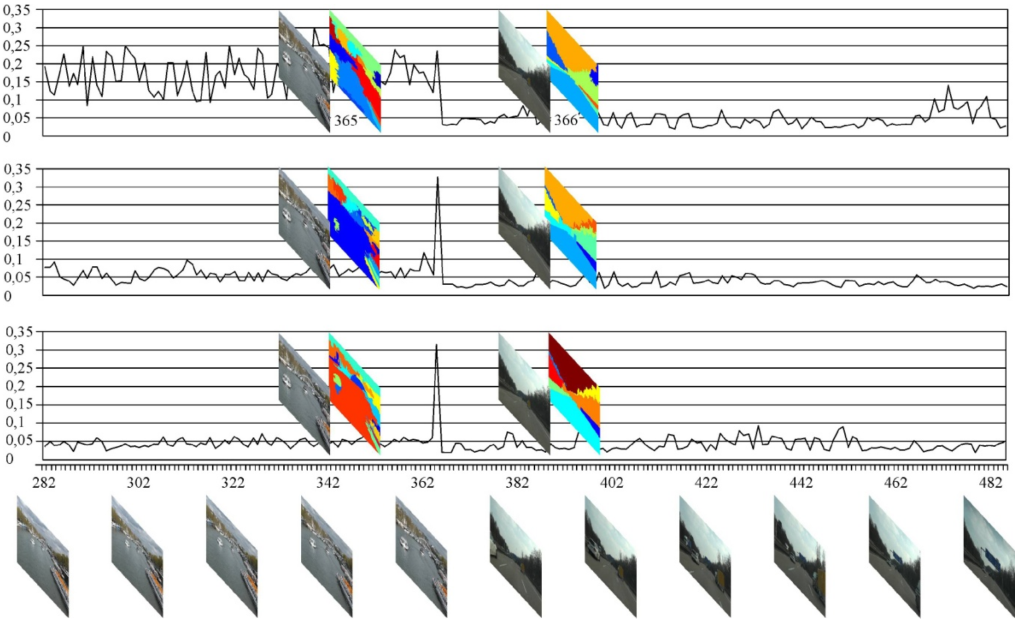

Figure 5. Example of segmentation parameter influence on distinctiveness of changes

Bibliography

[Bodyanskiy, 2012] Ye. Bodyanskiy, D. Kinoshenko, S. Mashtalir, O. Mikhnova On-line video segmentation using methods of fault detection in multidimensional time sequences / International Journal of Electronic Commerce Studies. – 2012. – Vol. 3, no 1, pp. 1-20.

[Caplier, 2001] A. Caplier, L. Bonnaud, and J. Chassery Robust fast extraction of video objects combining frame differences and adaptative reference image // Proceedings of the International Conference on Image Processing, Thessaloniki, Greece, – 2001. – Vol. 2, pp. 785–788.

[Chaisorn, 2003] L. Chaisorn, T. S. Chua, and C. H. Lee A multi-modal approach to story segmentation for news video / World Wide Web, – 2003. – Vol. 6, no. 2, pp. 187–208.

[Chupikov, 2007] A. Chupikov, D. Kinoshenko, V. Mashtalir, K. Shcherbinin Image retrieval with segmentation-based query // Adaptive Multimedia Retrieval: User, Context and Feedback / S. Marchand-Maillet et al. (eds.). – Lecture Notes in Computer Science. – Berlin-Heidelberg: Springer-Verlag. – 2007. – Vol. 4398. – pp. 208-222.

[DeMenthon, 2002] D. DeMenthon Spatio-temporal segmentation of video by hierarchical mean shift analysis // Proceedings of the Statistical Methods in Video Processing Workshop, Copenhagen, Denmark, 2002.

[Gauch, 1999] J.M. Gauch, S. Gauch, S. Bouix, and X. Zhu, Real time video scene detection and classification, / Information Processing and Management, – 1999. – Vol. 35, no. 3, pp. 381–400.

[Iannizzotto, 2000] G. Iannizzotto and L. Vita Fast and accurate edge-based segmentation with no contour smoothing in 2-D real images / IEEE Transactions on Image Processing, vol. 9, no. 7, pp. 1232–1237, 2000.

[Jie, 2010] W. Jie, W. Dada, Y. Wang, J. Li, W. Lei, H. Liang Industrial X-Ray Image Enhancement Algorithm based on AH and MSR / Engineering, – 2011. – Vol. 3, No. 10, pp. 1040-1044.

[Kinoshenko, 2007] D.Kinoshenko, V.Mashtalir, V. Shlyakhov A partition metric for clustering features analysis // International Journal "Information Theories and Applications" – Vol. 14. No 3. – 2007. – pp. 230-236.

[Kurdy, 1989] B. Kurdy and D. Jeulin Directional mathematical morphology operations. / Acta Stereologica. – 1989 – Vol. 8/2, pp. 473-480.

[Meyer, 1989] F. Meyer and J. Serra Contrasts and activity lattice / Signal Processing, – 1989. – Vol.16(4), pp. 303-317.

[Nock, 2004] R. Nock and F. Nielsen Statistical region merging // IEEE Transactions on Pattern Analysis and Machine Intelligence, – 2004. – Vol. 26, no. 11, pp. 1452–1458.

[Patras, 2002] I. Patras, E. A. Hendriks, and R. L. Lagendijk Video segmentation by MAP labeling of watershed segments / IEEE Transactions on Pattern Analysis and Machine Intelligence, – 2001. – Vol. 23, no. 3, pp. 326–332.

[Poignant, 2012] J. Poignant, L. Besacier, G. Qu´enot, and F. Thollard From text detection in videos to person identification // Proceedings of the IEEE International Conference on Multimedia and Expo, 2012.

[Ronse, 1986] C. Ronse Erosion of narrow image features by combining local rank and max filters // In Second International Conference on Image Processing and its Applications, London. –1986. – pp. 77-81.

[Snoek, 2005] C. Snoek, M.Worring, and A.W.M. Smeulders Early versus late fusion in semantic video analysis // Proceedings of the 13th Annual ACM International Conference onMultimedia, – 2005. – pp. 399–402.

[Soille, 2004] P. Soille Morphological image analysis: principles and applications. 2nd ed. Springer-Verlag, – 2004. – 391p.

[Tekalp, 2005] A.M. Tekalp Video segmentation / in Handbook of Image and Video Processing, Elsiever, Oxford, UK, – 2005.

[Varcheie, 2010] P.D.Z. Varcheie, M. Sills-Lavoie and G.-A. Bilodeau A Multiscale Region-Based Motion Detection and Background Subtraction Algorithm / Sensors, – 2010. – Vol. 10(2), pp.1041-1061.

[Vincent, 1989] L. Vincent Graphs and mathematical morphology / Signal Processing, – 1989. – Vol. 16, pp.365-388.

[Wang, 1998] D. Wang Unsupervised video segmentation based on watersheds and temporal tracking // IEEE Transactions on Circuits and Systems for Video Technology, – 1998. – Vol. 8, no. 5, pp. 539–546.

[Wang, 2003] H.-Y.Wang and K.-K.Ma Automatic video object segmentation via 3Dstructure tensor // Proceedings of the International Conference on Image Processing (ICIP '03) Barcelona, Spain, – 2003. – Vol. 1, pp. 153–156.

Authors' Information

Dmitry Kinoshenko – PhD, Informatics department, Kharkiv National University of Radio Electronics; Lenin av, 14, off. 288, Kharkiv-61166, Ukraine; e-mail: kinoshenko@kture.kharkov.ua
Major Fields of Scientific Research: Video processing

Sergii Mashtalir – PhD, Ass. prof. of Informatics department, Kharkiv National University of Radio Electronics; Lenin av, 14, off. 288, Kharkiv-61166, Ukraine; e-mail: mashtalir_s@kture.kharkov.ua
Major Fields of Scientific Research: Video processing, Multidimensional information research

Vladislav Shlyakhov – DSc, Full prof. of Higher Maths department, Kharkiv National University of Radio Electronics; Lenin av, 14. Kharkiv-61166, Ukraine; e-mail:
Major Fields of Scientific Research: Multi-algebraic systems