NEURAL NETWORK SEGMENTATION OF VIDEO VIA TIME SERIES ANALYSIS

Dmitry Kinoshenko, Sergey Mashtalir, Andreas Stephan, Vladimir Vinarski

Abstract: Semantic video retrieval which deals with unstructured information traditionally relies on shot boundary detection and key frames extraction. For content interpretation and for similarity matching between shots, video segmentation, i.e. detection of similarity-based events, are closely related with multidimensional time series representing video in a feature space. Since video has a high degree of frame-to-frame-correlation, semantic gap search is quite difficult as it requires high-level knowledge and often depends on a particular domain application. Based on principal components analysis a method of video disharmony authentication has been proposed. Regions features induced by traditional frame segmentations have been used to detect video shots. Results of experiments with endoscopic video are discussed.

Keywords: Video Data, Frames, Time series segmentation, Principal component

ACM Classification Keywords: I.2.10 Vision and Scene Understanding (Video analysis), G.3 Probability and Statistics (Time series analysis).

Introduction

Nowadays video retrieval methods are rapidly evolving as the modus operandi for information creation, exchange, storage and content search [Petkovic, 2004, Shanmugam, 2009]. There arises an access to a tremendous amount of video information so it is infeasible for a human to classify or cluster the video scenes, to find appropriate events. In contrast to searching data in a relational database, a content based video retrieval (CBVR) requires the search of similar objects as a basic functionality of the database system. There is a number of approaches on summarization, video data management, streaming media analysis, video coding, video indexing, video abstraction, video information retrieval, etc. [Hanjalic, 2004, Snoek, 2008].

One of the most important video analysis issues is an automat identification of semantic events without an operator having to view the video. Video content analysis consists of motion, style and object detection, events and objects recognition, etc. Multimedia content analysis of video data so far has relied mostly on the information contained in the raw visual, audio and text signals. Content-based (Concept-based) video retrieval techniques strive to accomplish this goal by using low level image features, such as colors, textures, shapes, motions, etc. [Snoek, 2008, Hanjalic, 2004, Geetha, 2008]. But for now more and more video content analysis is considered as a capability of video analysis with the view of detection and determination of temporal events not based on a single image (or single frame). Hereupon great attention is spared to the analysis of separate shots of video, rather to their semantic relations and changes in time. Such dependences are especially important at a content search in video data bases, discovering features unbalances which are produced by scene changes in input video data.

Acceptable mathematical model to search video is an expression in a form of multidimensional time series, describing scene changes in a feature space. Such approach allows to analyze video scene changes in time. Thus, we can find changes at some time intervals by an analysis of frames sequence. Determinations of features or descriptions for each frame and changes comparisons in these descriptions give possibilities to draw a conclusion about changes happening in the time series or differently speaking in video data. This approach has got a wide development in a number of video data processing problems, and especially in the areas related to data segmentation in time series [Liniker, 2000, Rao, 2000]. But there are a large number of issues that remain to be investigated, in particular problems of multidimensional time series processing are not fully described in respect to video understanding.

Mathematical models of multidimensional time series under video shot search

For cases when the number of observations is not fixed and grows in time, to find shot boundaries in video stream via analysis of arbitrary feature space, mathematical models of multidimensional non-stationary time series are discussed in the current section.

There exist a number of models introduced for description of multidimensional sequences or time series [Izermann, 1984, Nikifora, 1991, Bassevile, 1993, Kerestencioglu, 1993] that generally can be presented by two basic forms. The first one is structural

$$\sum_{l=0}^{p} B_l x(k-l) + Dz(k) = \eta(k). \quad (1)$$

Here B_l are matrix coefficients at intrasystem (endogenous) variables, B_0 is an nonsingular matrix at the endogenous variables of current time, D is a matrix of coefficients at exogenous variables, $z(k)$ is a vector of exogenous variables, including and their retarding values, $\eta(k)$ is a vector revolting signal with a zero maximal expectation and restricted second moments. The second one is normalized

$$x(k) = -B_0^{-1}(\sum_{l=1}^{p} B_l x(k-l) + Dz(k) - \eta(k)) \quad (2)$$

or

$$x(k) = CZ(k) + \xi(k) \quad (3)$$

where a vector $Z(k)$ plugs in itself both exogenous variables and retarding values of endogenous variables

$$\xi(k) = B_0^{-1}\eta(k).$$

The important problem is an authentication of parameters (1-3). Note, it often suffices to use indirect two-stage or three-stage least-squares methods, but they are intended for manipulations only with given dimensionality and predetermined data set [Izermann, 1984]. Algorithms intended for work in the sequential mode, such as [Nikifora, 1991] relaxation, recursive or method of fixed-point, are characterized by insufficient rate of convergence in order to provide signal processing in real-time.

In [Badavas, 1993] the model of multidimensional time series is examined

$$x(k) = \sum_{l=1}^{p} B_l x(k-l) + \sum_{p=1}^{q} D_p z(k-p) + F\psi(k-1) + \xi(k), \quad (4)$$

or

$$B(z^{-1})x(k) = D(z^{-1})z(k-1) + F\psi(k-1) + G(z^{-1})\eta(k) \quad (5)$$

where unknown coefficients, describing behavior of observed sequence, enter either into matrices B_l, D_p, F of corresponding dimensions or in matrix polynomials $B(z^{-1})$, $D(z^{-1})$, $G(z^{-1})$ from the backward shift operator z^{-1}, $\psi(k)$ that is a determinate function, describing a trend being in a signal $x(k)$. In addition it should be emphasized that the same time series $x(k)$ can be described by infinite great number of multidimensional equals (4) or (5) [Badavas, 1993].

There exist three basic parameters evaluation methods of these expressions: maximum likelihood method, Bayes approach and method of the restricted information. If first two from them are realized in a packet form, then third is a form of recurrent least-squares method which can process sequentially incoming data. Unfortunately, standard recurrent least-squares method, being in fact, an identifier with infinite memory, inherently is unsuitable for analysis of substantially non-stationary objects whose features can hardly change properties.

For finding out the changes of properties of multidimensional series, compact effective and easy-to-use so-called vector autoregressive models (VAR models) had been proposed in [Pouliezos, 1994, Juselis, 1994].

Generally VAR models links the past and current supervisions of vector signals $x(k)$ in the form

$$x(k) = B_0 + \sum_{l=1}^{p} B_l x(k-l) + \xi(k) \quad (6)$$

where $B_0 = \{b_{0i}\}$ is $(n \times 1)$ vector of mean values, $B_l = \{b_{lij}\}$ are $(n \times n)$ matrices of parameters, p is a model order. Except the formula (6) of VAR model can be compact described in a state space

$$\begin{cases} x(k) = \Pi x(k-1) + \Pi_0 + E(k), \\ y(k) = Cx(k) \end{cases} \quad (7)$$

where $x(k) = \begin{pmatrix} x(k) \\ x(k-1) \\ \vdots \\ x(k-p+1) \end{pmatrix}$, $\Pi_0 = \begin{pmatrix} B_0 \\ \vec{0} \\ \vdots \\ \vec{0} \end{pmatrix}$, $\Pi = \begin{pmatrix} B_1 & \cdots & B_{p-1} & B_p \\ I_n & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I_n & 0 \end{pmatrix}$, $E(k) = \begin{pmatrix} \xi(k) \\ \vec{0} \\ \vdots \\ \vec{0} \end{pmatrix}$,

$C = (I_n, 0, \ldots, 0)$, $x(k) - (np \times 1)$ is a vector of the states, Π is $(np \times np)$ translational matrix, $\vec{0}$ and 0 are $(n \times 1)$ and $(n \times n)$ zero vector and matrix accordingly.

Relations (7) allow to use the powerful mathematical tools of Kalman's filtration for the analysis of multidimensional signals.

Detection of properties changes of multidimensional time series $x(k)$ is related to analysis of each of its components $x_i(k), i = 1,2,\ldots,n$ and there arise three possible situations:

i) change of mean values of $l \le p$ components

$$b_{0i}(k) = \begin{cases} b_{0i}, \text{if } k < k_a, \\ b_{0i}^a, \text{if } k \ge k_a, \end{cases}$$

ii) change of descriptions (dispersions) $l \le p$ of perturbation $\xi_i(\sigma_i^2)$

$$x_i(k) = \begin{cases} b_{0i} + \sum_{l=1}^{p} \sum_{j=1}^{n} b_{lij} x_j(k-l) + \xi_i(k), \text{if } k < k_a, \\ b_{0i} + \sum_{l=1}^{p} \sum_{j=1}^{n} b_{lij} x_j(k-l) + \xi_i^a(k), \text{if } k \ge k_a, \end{cases}$$

iii) change of coefficients b_{lij}, causing the change of autocorrelation properties of non-stationary time series

$$x_i(k) = \begin{cases} b_{0i} + \sum_{l=1}^{p} \sum_{j=1}^{n} b_{lij} x_j(k-l) + \xi_i(k), \text{if } k < k_a, \\ b_{0i} + \sum_{l=1}^{p} \sum_{j=1}^{n} b_{lij}^0 x_j(k-l) + \xi_i(k), \text{if } k \ge k_a \end{cases}$$

where k_a is the instant time when measuring is executed.

Detection of multidimensional time series properties changes via principal components analysis

At the analysis of largescale (both on a volume and on a dimension) observations set in form of time series an important task lies in compression with the purpose of selection of latent factors qualificatory the underlying structure of the controlled signal, that pursues an aim to do initial time series more simply interpreted from the point of view of finding out of the properties changes.

One of the most effective going near the decision of this problem is a vehicle of factor analysis, within the framework of that the most wide distribution was got by the main components method especially in the problems of patterns recognition, image processing, spectrology etc. and known yet as Karhunen-Loeve's transform.

A $(k \times n)$ matrix of observations

$$X = \begin{pmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ x_1(2) & x_2(2) & & x_n(2) \\ \vdots & & & \\ x_1(u) & x_2(u) & \cdots & x_n(u) \\ \vdots & & \ddots & \\ x_1(k) & x_2(k) & \cdots & x_n(k) \end{pmatrix} \tag{8}$$

that is generated by an array from the k n-dimensional vectors containing observations $x(u) = (x_1(u), x_2(u), \dots, x_n(u))^T$ and also its $(n \times n)$ cross-correlation matrix

$$R(k) = \frac{1}{k}\sum_{u=1}^{k}(x(u) - \overline{x})(x(u) - \overline{x})^T = \frac{1}{k}\sum_{u=1}^{k}x^C(u)x^{CT}(u)$$

where $x^c(u) = x(u) - \overline{x}$ centered relatively basic data mean are input information for an analysis.

The kernel of PCA underlies in projection of observed data from input n-dimensional space into m-dimensional $(n > m \geq 1)$ output and is reduced to the system $w^1, w^2, \ldots, w^m$ of orthogonal eigenvectors of matrix $R(k)$ such that $w^1 = (w_1^1, w_2^1, \ldots, w_n^1)^T$ corresponds to the greater eigenvalue λ, matrices $R(k)$, $w^2 = (w_1^2, w_2^2, \ldots, w_n^2)^T$ corresponds to the second-largest eigenvalue λ etc. In any case, the problem is searching for matrix equation solution

$$(R(k) - \lambda_l I)w^l = 0$$

such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ and $\left\|w^l\right\|^2 = 1$.

Thus, in algebraic terms the problem solving is closely related to the goal seeking of eigenvalues and the rank of cross-correlation matrix; in geometrical sense it is a problem of passing to less dimension space with the minimal information loss; in statistical sense it is the search for orthonormal vectors set in input space assuming maximal data variations; and, finally, in algorithmic sense it is the problem of successive determination (excretions) of set of eigenvectors $w^1, w^2, \ldots, w^m$ by optimization of each local functionals providing global criteria

$$I_w(k) = \frac{1}{k}\sum_{l=1}^{m}\sum_{u=1}^{k}(x^{CT}(u)w^l)^2$$

with restrictions

$$\begin{cases} w^{lT}w^p = 0, \text{ with } l \neq p, \\ w^{lT}w^p = 1. \end{cases}$$

The first principal component bearing a maximum of information about the processed signal can be found by maximization of local criterion

$$I_w^1(k) = \frac{1}{k}\sum_{u=1}^{k}(x^C(u)w^1)^2$$

by standard undetermined Lagrange multipliers.

Further, the projection on the first principal component is subtracted from every vector $x^C(u)$ and the first principal component of remains, being the second principal component of basic data and orthonormal to first one, is calculated.

The third principal component is calculated by projection of each input vector on the first and the second principal components, projection subtraction from each $x^C(u)$ and search for the first principal component of remains. Eventually we arrive at the third principal component of basic data. Other principal components are calculated recursively in concordance with described procedure.

To the present time the developed mathematical and programming tools have the same disadvantage – the necessity to know matrix X with the fixed dimension to implement Karhunen-Loeve transform. If data acquisition is consequent, standard factor analysis procedures become out of operation in real time.

In this connection the use of recurrent procedures is reasonable to find eigenvectors of matrix $R(k)$ by the sequential processing of observations $x(1), x(2), \ldots, x(k), x(k+1)\ldots$ of multidimensional time series without a calculation cross-correlation matrix in order to achieve real time.

In [Cichocki, 1993] an artificial neuron is described on the basis of adaptive linear associators for the calculation of the first principal component in real time. On fig. 1 modified neuron for finding out the properties changes in a multidimensional signal on the basis of analysis of principal components is proposed.

For the preliminary centered data the learning algorithm looks like

$$\begin{cases} w^1(k+1) = w^1(k) + \eta(k+1)(x(k+1) - y(k)w^1(k))y(k+1), \\ y(k+1) = x^T(k+1)w^1(k), w^1(0) \neq 0, y'(1) = x^T(1)w^1(0) \end{cases} \quad (9)$$

where $\eta(k+1)$ is a step parameter of adaptation which is chosen enough small to provide the algorithm stability. Also algorithm (9) gives vector $w^1(k)$ normalization i.e.

$$\left\| w^1(k) \right\|^2 = 1$$

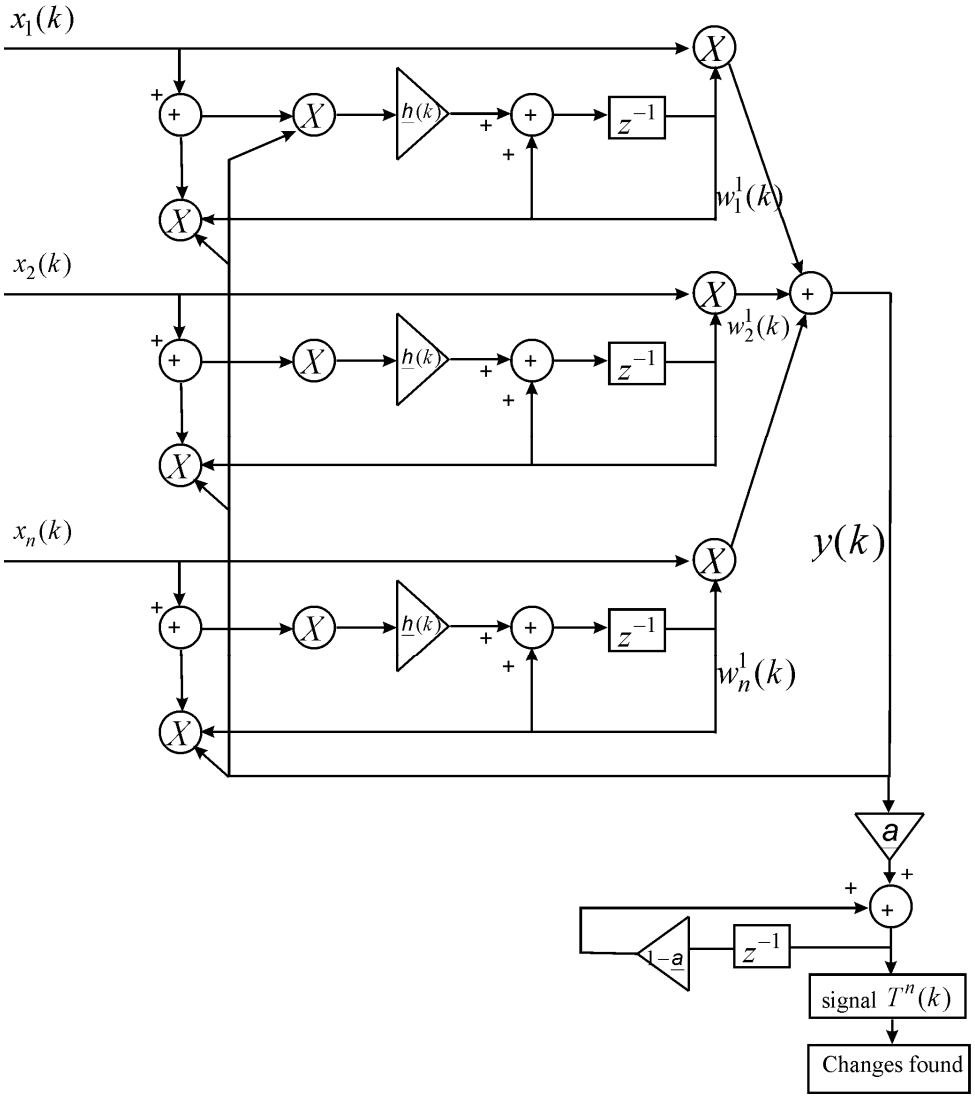

Fig.1 The modified neuron for finding out of principal component changes in multidimensional time series

Note the vector $w'(k)$ is the eigenvector of matrix $R(k)$, corresponding to the maximal eigenvalue and an output signal $y(k)$ is characterized by maximally possible dispersion, i.e. contains a maximum of information about a multidimensional input signal $x(k)$.

Further, an output signal $y(k)$ is processed with the exponential smoothing, filtering noise components $\xi(k)$, and finding out the properties change is produced by means of relation [Trigg, 1967]

$$T^{TL}(k) = \frac{T_i'(k)}{d_i(k)}$$

where $T_i'(k) = (1-\beta)|e_i(k)| + \beta T_i'(k-1)$, $d_i(k) = (1-\beta)|e_i(k)| + \beta d_i(k-1)$, $e_i(k)$ is a current estimation error, β is a smoothing parameter.

Discussion of experiments analysis

Experimental data set consists of different endoscopic videos each of which is composed by 550 frames. The goal of the experiment is to detect certain changes in the video streams. For this purpose at the first stage we built segmentation of initial video data frame by frame. Examples of input frame and its segmentation are presented on fig 2.

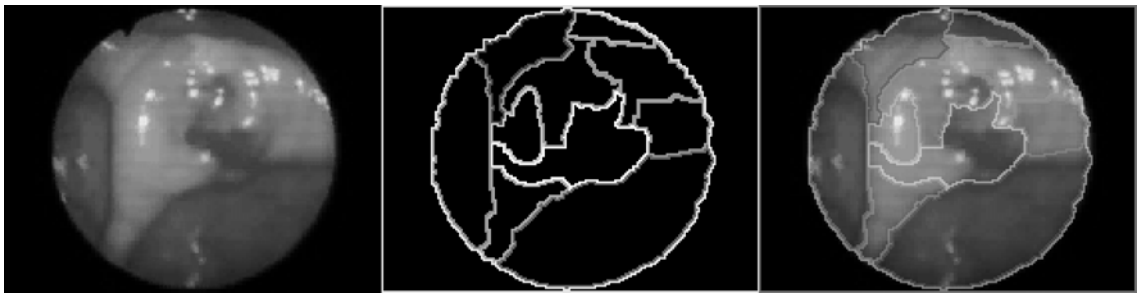

Fig. 2 Input frame and its segmentation

Extensively used Jseg algorithm [Comaniciu, 1997] was chosen for image spatial segmentation. The first part of experiments had the goal to find an influence of Jseg algorithm parameters (Scales, Quantresh, Merge) on the results of temporal segmentation of video in form of multidimensional time series. It is necessary to emphasized that the decline of threshold selection brings to appearance of more shallow regions and to an increase of its total number. Thus, it is possible to vary initial parameters in order to get more exact or more rough regions what is important for application areas, particularly for analyzed endoscopic video.

It should be noted that the change of some parameters insignificantly influences on the results of video data processing. In particular, if the parameter of Scales is varied and other parameters are invariable (see fig. 3), one can see the parameter values change, but the general trend of charts coincided.

Quantitative descriptions of the geometrical descriptions of regions that are produced by spatial segmentation were chosen to search video data changes. It should be noted that there exist a number of different descriptions of segmented images, in particular, shapes signature, polygonal approximation, moments, scale-space methods, shape transform domains etc [Mingqiang, 2008].

Numerous experiments allow us to assert that under considered frame region based video retrieval it suffices to find the integrated changes of such parameters for each region as area, perimeters, diameters of approximating circles, minimal and the maximal orthogonal projections, angle of slope to an axis and descriptions that are invariant to geometric transformations namely traditional moments features.

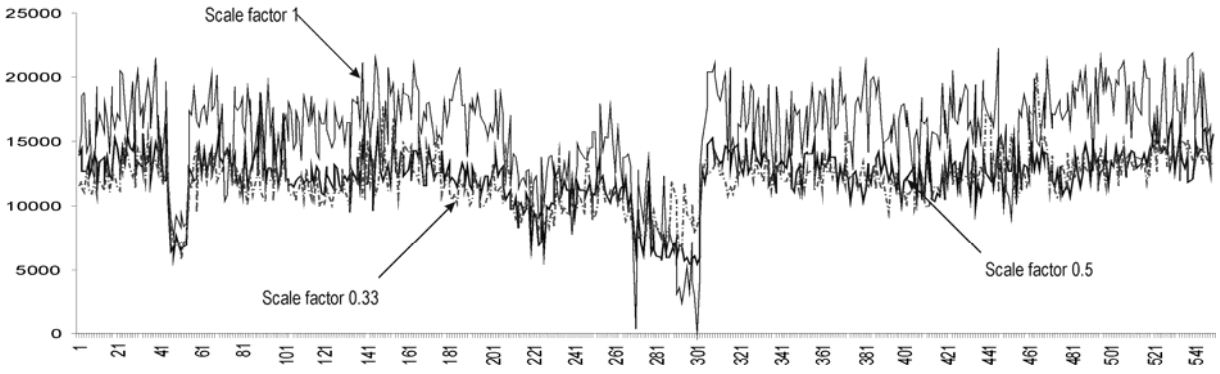

Fig. 3 Influence of segmentation parameters on video data processing

do not absolutely suit us they fully can appear unsensible to the substantial changes of initial video data. The incurrence of the chosen descriptions is equal to 11. After founding of these descriptions for all segments, the matrix of basic data (8) is formed, where as a line are values n of descriptions for one segment, and an amount of lines is a total regions number k at a frame. Thus, for each frame of video data we have a matrix of dimension $(k \times n)$ which we process in obedience to described methods. Further, using neural network (presented on fig. 1) we found one output parameter for each frame. An example of results is illustrated by fig. 4.

We can indicate that at the substantial changes of basic data, such as in a range a from 45 to 55 frame, we have shot boundaries produced by video artefacts or dramatic changes of basic data, or as in a district 300 frame we have the clear network output value overfalls, that also allows to see substantial changes of video. However in case of the small changes, for instance, related to shaking of camera, transition of eyeshot and other insignificant changes, search for shot boundaries remains enough difficult task.

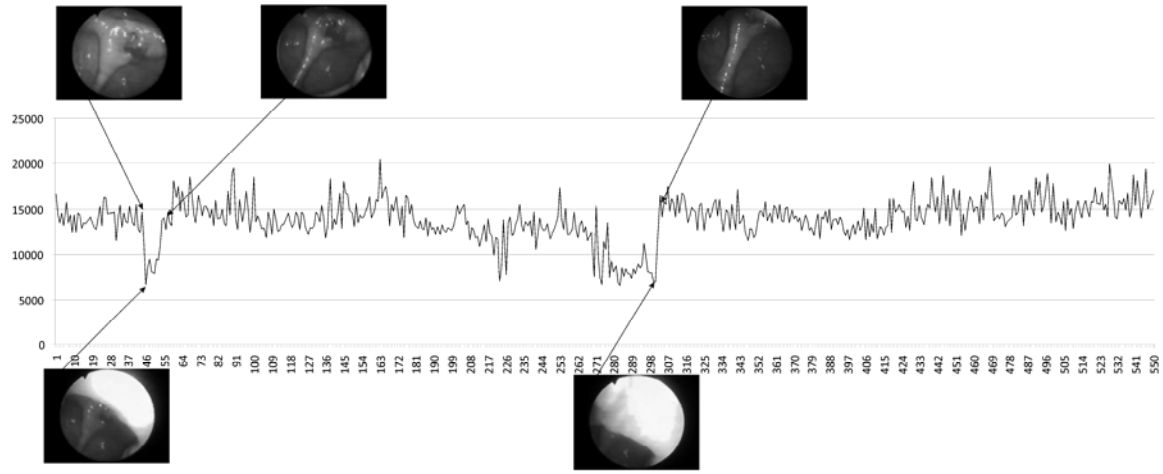

Fig. 4 Result of video data analysis

Conclusion

Content-based video retrieval systems whose effectiveness determines, in general, the success or failure in obtaining the required information are undergoing explosive growth due to the monotonic increase of accessible video data warehouses.

Based on low-level features the segmentation, breaking up the video sequences into temporally homogeneous segments (shots), is relatively simple and as a rule can be done automatically by shot-change detection algorithm. The development of shot-boundary detection algorithms has the longest and richest history in the video content analysis. But partitions a video stream into a set of meaningful and manageable segments video with a small intra-shot and large inter- shot variability generally were based on low-level spatial and rarely temporal features. Since humans perceive video as a complex interplay of cognitive concepts, a necessity of an access at the semantic level is obvious.

PCA based neural network temporal segmentation intended for real time has been proposed. Spatial segmentations of each frame or more exactly features extracted for each produced region of field of view in fact correspond to multidimensional time series describing video. All the more such partitions enough adequately reproduce 'spatial content' of an image and consequently frame sequences. Numerous experiments confirm a validity of theoretic results, however; offered approach solely gives possibility to find substantial changes in video data with filtration of marginal changes that can be important in some applications.

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).

Bibliography

[Badavas, 1993] P.C. Badavas. Real-Time Statistical Process Control. Eaglewood Cliffs, N.J.: Prentice-Hall, 1993.

[Bassevile, 1993] Bassevile M., Nikifora I. Detection of Abrupt Changes. Theory and Application. – Eaglewood Cliffs, N.J.: PTR Prentice-Hall, 1993.

[Cichocki, 1993] A. Cichocki, R. Unberhauen. Neural Networks for Optimization and Signal Processing. Stuttgart: Teubner, 1993.

[Comaniciu, 1997] D. Comaniciu, P. Meer. Robust Analysis of Feature Spaces: Color Image Segmentation In: IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[Geetha, 2008] P. Geetha, V. Narayanan. A Survey of Content-Based Video Retrieval. In: Journal of Computer Science, Vol.4, No 6, 2008.

[Hanjalic, 2004] A. Hanjalic Content-Based Analysis of Digital Video. Boston: Kluwer Academic Publishers, 2004.

[Izermann, 1984] Izermann R. Process Fault Detection Based Modeling and Estimating Methods – a Survey. In: Automatica, 20, No4, 1984.

[Juselis, 1994] Juselis K. The Cointegrated VAR-Model. N.Y.: Oxford University Press, 1994.

[Kerestencioglu, 1993] Kerestencioglu F. Change Detection and Input Design in Dynamical Systems. – Taunton, UK: Research Studies Press. – 1993.

[Liniker, 2000] F. Liniker, L. Niklasson Time Series Segmentation Using an Adaptive Resource Allocating Vector Quantization Network Based on Change Detection. In IEEE-INNS-ENNS International Joint Conference on Neural Networks, Vol. 6, 2000.

[Liu, 2007] Y.Liu, D. Zhanga, G. Lua, W.-Y. Ma. A Survey of Content-Based Image Retrieval with High-Level Semantics. In: Pattern Recognition. Vol. 40, No. 1, 2007.

[Mingqiang, 2008] Y. Mingqiang, K. Kidiyo, R. Joseph. A Survey of Shape Feature Extraction Techniques. In: Pattern Recognition Techniques, Technology and Applications, Vacuvar: Intech, 2008.

[Natsev, 2006] A. (P). Natsev. Multimodal Search for Effective Video Retrieval. In: Image and Video Retrieval. Berlin-Heidelberg: Springer-Verlag, Lecture Notes in Computer Science, Vol. 4071, 2006.

[Nikifora, 1991] Nikifora I.V. Sequential Detection of Changes in Stochastic Processes In: Prep. 9-th IFAC/IFORS Symp. 'Identification and System Parameter Estimation', Vol.1, 1991.

[Petkovic, 2004] M. Petkovic, W. Jonker. Content-Based Video Retrieval: A Database Perspective (Multimedia Systems and Applications). Boston-Dordrecht-London: Kluwer Academic Publishers, 2004.

[Pouliezos, 1994] A.D. Pouliezos, Y.S. Stavrakalis. Real Time Fault Monitoring of Clustering Processes. Dordrecht: Kluver Academic Publishers, 1994.

[Rao, 2000] Rao Y.N., J.C. Principe. A Fast On-line Generalized Eigendecomposition Algorithm for Time Series Segmentation. In: Adaptive Systems for Signal Processing, Communications, and Control Symposium, 2000.

[Shanmugam, 2009] T.N. Shanmugam, P. Rajendran An Enhanced Content-Based Video Retrieval System Based on Query-Clip. In: International Journal of Research and Reviews in Applied Sciences, Vol. 1, No 3, 2009.

[Snoek, 2008] C.G.M. Snoek, M. Worring. Concept-Based Video Retrieval. In: Foundations and Trends in Information Retrieval, Vol. 2, No. 4, 2008.

[Trigg, 1967] D.W. Trigg, A.G. Leach. Exponential Smoothing with an Adaptive Response Rate In: Operational Research, Vol. 18, No 1, 1967, P. 53-59.

Authors' Information

PhD Dmitry Kinoshenko – Informatics department, Kharkiv National University of Radio Electronics, Lenin Ave. 14, Kharkiv, Ukraine, kinoshenko@kture.kharkov.ua

Major Fields of Scientific Research: Video data analysis

PhD Sergey Mashtalir – Informatics department, Kharkiv National University of Radio Electronics, Lenin Ave. 14, Kharkiv, Ukraine, mashtalir_s@kture.kharkov.ua

Major Fields of Scientific Research: Video data analysis, Image processing

Dr. Andreas Stephan – CURATYS International, Nonnengasse, 5a, 99084, Erfurt, Germany, Stefan@curatys.de

Major Fields of Scientific Research: Time series analysis

Dr. Vladimir Vinarski – University of Applied Sciences and Arts, Fachhochschule Hannover Ricklinger Stadtweg 120 30459 Hannover, Germany, vladimir.vinarski@fh-hannover.de

Major Fields of Scientific Research: Video data analysis