NOVEL APPROACH TO CONTENT-BASED VIDEO INDEXING AND RETRIEVAL BY USING A MEASURE OF STRUCTURAL SIMILARITY OF FRAMES

David Asatryan, Manuk Zakaryan

Abstract: *Extracting a small number of key-frames that can abstract the content of video is very important for efficient browsing and retrieval in video databases. Most research on video content involves automatically detecting the boundaries between camera shots. After shot boundary detection there is a need for shots indexing. In this paper, we present the gradient field based algorithm of shot detection and new method of key-frames determination. We provide a novel algorithm aimed to find a compact set of key-frames that can represent a video segment for a given degree of fidelity. The advantages of proposed approach are high performance of detection of the key-frames and linear speed of retrieval from the databases. Also we provide a new concept of the shot segment analysis and interpretation, using graphical and numerical methods.*

Keywords: *shot detection, similarity measure, indexing, key-frame, content-based retrieval.*

ACM Classification Keywords: *Image Processing and Computer Vision*

Introduction

Multimedia information indexing and retrieval are required to describe, store and organize multimedia information and to assist people in finding multimedia resources conveniently and quickly [Weiming, 2011, Lew, 2006]. Content-based video indexing and retrieval have a wide range of applications such as quick browsing of video folders, analysis of visual electronic commerce, remote instruction, digital museums, news event analysis, intelligent management of web videos etc.

The framework includes the following: 1) structure analysis to detect shot boundaries, extract key-frames, and segment scenes; 2) feature extraction from segmented video units (shots or scenes): these features include static features in key-frames, object features, motion features; 3) query: the video database is searched for the desired videos using the index and the video similarity measures; 4) video browsing and feedback.

A shot is a consecutive sequence of frames captured by a camera action that takes place between start and stop operations, which mark the shot boundaries. Methods for shot boundary detection usually first extract visual features from each frame, then measure similarities between frames using the extracted features, and, finally, detect shot boundaries between frames that are dissimilar. In the following, we discuss the main three steps in shot boundary detection: feature extraction, similarity measurement, and detection. The features used for shot boundary detection include color histogram or block color histogram, edge change ratio, motion vectors, together with more novel features such as scale invariant feature transform, corner points etc.

After the shot boundaries are identified, most of the existing works for video abstraction generally go through the following two steps: first, select the key-frames in each shot, and then cluster the similar shots based on the key-frames to construct the hierarchical or transition representation of video. Key-frames are a small set of images that can represent the visual content of a video. They can be used to compute the similarity between two video sequences, as well as to browse the video based on its content.

Following to [Truong, 2007], current approaches to extract key-frames are classified into six categories: sequential comparison-based, global comparison-based, reference frame-based, clustering based, curve simplification-based, and object/event-based.

1) **Sequential Comparison of Frames**: In these algorithms, frames subsequent to a previously extracted key-frame are sequentially compared with the key-frame until a frame which is very different from the key-frame is obtained. This frame is selected as the next key-frame. For instance, Zhang et al. [Zhang, 1997] used the color histogram difference between the current frame and the previous key-frame to extract key-frames. The merits of the sequential comparison-based algorithms include their simplicity, intuitiveness, low computational complexity.

The limitations of these algorithms include the following: a) key-frames represent local properties of the shot rather than the global properties, b) the irregular distribution and uncontrolled number of key-frames make these algorithms unsuitable for applications that need an even distribution or a fixed number of key-frames, c) a redundancies can occur when there are contents appearing repeatedly in the same shot.

2) **Global Comparison of Frames**: The algorithms based on global differences between frames in a shot distribute key-frames by minimizing a predefined objective function that depends on the application. In general, the objective function has one of the following four forms [Truong, 2007].

a) Even temporal variance: These algorithms select key-frames in a shot such that the shot segments, each of which is represented by a key-frame, have equal temporal variance;
b) Maximum coverage: These algorithms extract key-frames by maximizing their representation coverage, which is the number of frames that the key-frames can represent;

c) Minimum correlation: These algorithms extract key-frames to minimize the sum of correlations between key-frames (especially successive key-frames), making key-frames as uncorrelated with each other as possible;

d) Minimum reconstruction error: These algorithms extract key-frames to minimize the sum of the differences between each frame and its corresponding predicted frame reconstructed from the set of key-frames using interpolation.

3) **Reference Frame**: These algorithms generate a reference frame and then extract key-frames by comparing the frames in the shot with the reference frame. For instance, Ferman and Tekalp [Ferman, 2003] construct an alpha-trimmed average histogram describing the color distribution of the frames in a shot. Then, the distance between the histogram of each frame in the shot and the alpha-trimmed average histogram is calculated. Key-frames are located using the distribution of the distance curve. The merit of the reference frame-based algorithms is that they are easy to understand and implement. The limitation of these algorithms is that they depend on the reference frame.

4) **Clustering**: These algorithms cluster frames and then choose frames closest to the cluster centers as the key-frames. Girgensohn and Boreczky [Girgensohn, 2000] select key-frames using the complete link method of hierarchical agglomerative clustering in the color feature space. The merits of the clustering-based algorithms are that they can use generic clustering algorithms, and the global characteristics of a video can be reflected in extracted key-frames. The limitations are: first, they are dependent on the clustering results, but successful acquisition of semantic meaningful clusters is very difficult, and second, the sequential nature of the video cannot be naturally utilized.

5) **Curve Simplification**: These algorithms represent each frame in a shot as a point in the feature space. The points are linked in the sequential order to form a trajectory curve and then searched to find a set of points which best represent the shape of the curve. Calic and Izquierdo [Calic, 2002] generate the frame difference metrics by analyzing statistics of the macro block features extracted from the MPEG compressed stream. The merit of the curve simplification-based algorithms is that the sequential information is kept during the key-frame extraction. Their limitation is that optimization of the best representation of the curve has a high computational complexity.

6) **Objects/Events**: These algorithms [Kang, 2005] jointly consider key-frame extraction and object/event detection in order to ensure that the extracted key-frames contain information about objects or events. Calic and Thomas [Calic, 2004] use the positions of regions obtained using frame segmentation to extract key-frames where objects merge. The merit of the object/event-based

algorithms is that the extracted key-frames are semantically important, reflecting objects or the motion patterns of objects. The limitation of these algorithms is that object/event detection strongly relies on heuristic rules specified according to the application.

Once video indices are obtained, content-based video retrieval can be performed. On receiving a query, a similarity measure method is used, based on the indices, to search for the candidate videos in accordance with the query. The retrieval results are optimized by relevance feedback, etc. In the following, we review query types, similarity matching, and relevance feedback.

Similarity Measure and Algorithm for Shot Detection

Shot detection algorithm is usually based on consecutive determination of similarity of neighboring frames and detecting abrupt dissimilarities between them. When the level of similarity measure exceeds some predefined threshold t_c, then corresponding frame is considered as a cut frame. The quality of a decision rule depends on used similarity measure.

The measure which is applied in this paper is based on the structural properties of an image [Asatryan, 2009]. The mentioned measure is described below. We consider a model of image structure based on the set of edges which are determined by the gradient field of the image.

Let $\|G_H(m,n)\|$ and $\|G_V(m,n)\|$ (m = 0, 1, ..., M-1, n = 0,1,...,N-1) at a point (m, n) of an image be the horizontal and vertical gradients, determined by one of known gradient methods, and the matrix of gradient magnitude $\|\Delta(m,n)\|$, where

$$\Delta(m,n) = \sqrt{G_H^2(m,n) + G_V^2(m,n)} \tag{1}$$

We suppose that the gradient magnitude (1) is a random variable of two-parameter Weibull distribution density with parameters c > 0 and b > 0. As a measure of structural similarity of two images with probability distribution functions of gradient magnitude $f_1(x;b_1,c_1)$ and $f_2(x;b_2,c_2)$ accordingly, we accept

$$W^2 = \frac{\min(b_1,b_2)\min(c_1,c_2)}{\max(b_1,b_2)\max(c_1,c_2)},\ 0 < W^2 \le 1, \tag{2}$$

where the parameters $b_j, c_j, j = 1,2$ are statistically estimated by corresponding samples of gradient magnitudes of comparing images. This approach of images similarity assessment was successfully applied to different problems of image processing, see for example [Asatryan, 2009, Asatryan, 2010].

The algorithm for shot detection was described in our previous papers [Asatryan, 2014]. It was shown the advantages of algorithm against other algorithms based on mean-square deviation between images.

Key-frame Determination Algorithm

According to analysis of key-frame extraction algorithms, which have been described above, we can consider that the main difficulty of existing algorithms is the huge amount of comparisons of the content of frames inside each shot and between adjacent shots as well.

Our proposed algorithm of key-frame extraction have a big advantage in comparison with others, considering the fact that the content of each frame is characterized by only two parameters, which have already been determined during the shot detection procedure. The count of calculation needed to determine the similarity measure between any two shots is done using formula (2).

Key-frame detection is based on determining the parameters of the hypothetical frame, which are calculated as arithmetic mean of corresponding parameters b and c in the current shot. As there may not exists a frame with such parameters in the considered shot, we consider as a key-frame the frame which parameters b and c are the most near to parameter of hypothetical frame.

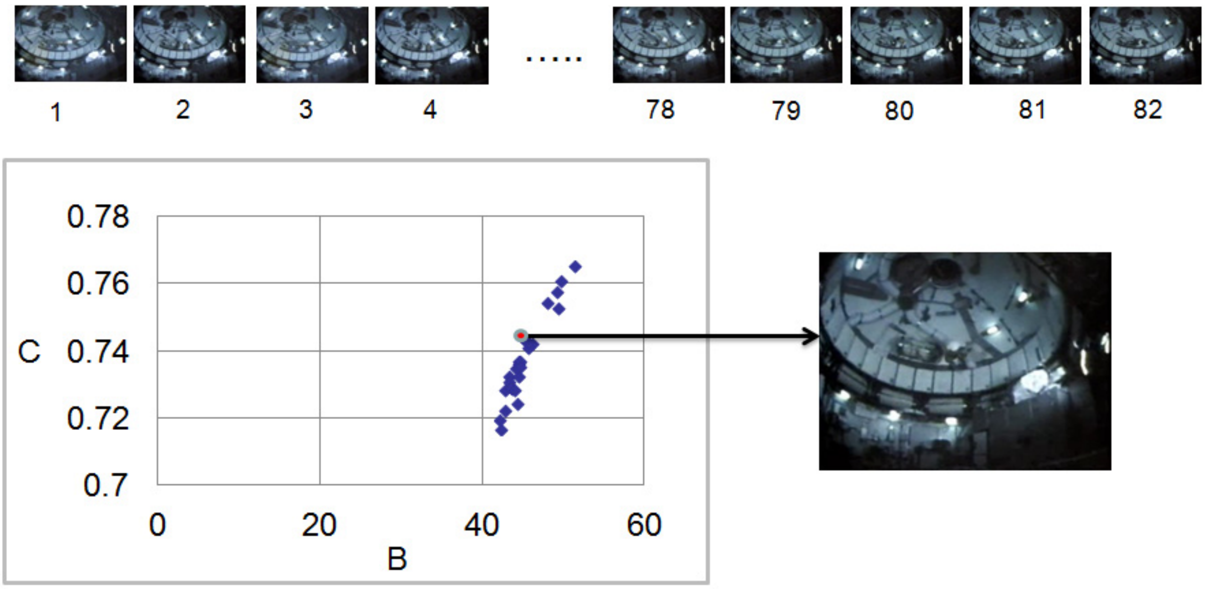

Figure 1. Illustration of key-frame

Figure 1 shows the process and the result of key-frame extraction from the given shot. The points in the graph corresponds to parameters b and c for each frame of the given shot above (the vertical axis shows the parameter c, the horizontal one - the parameter b).

It is also important to mention one more advantage of current approach. The thing is that the allocation of the points placed in the graph contains additional information about the content modifications of the frames inside the given shot. Therefore arises the problem of formal analysis of the parameters distribution in the (b, c) plot and fully interpretation of transitions inside the shot. Experimental results of key-frame extraction and corresponding formal analysis are given in the next section.

Results of Experiments

Described method of shot boundary detection was tested for various video sequences and some results have been given in our previous articles [Asatryan, 2014]. Here we graphically illustrate the results, which we got for key-frame extraction for exact video sequence, and also the analysis of the shot behavior using statistical regression technique for dependency between specified parameters.

The considered test video has 5 shots, which we determined by method described in section 2. Here in Figure 2 and Figure 3 the dependency graphs between b and c parameters as example for first and second shots are illustrated.

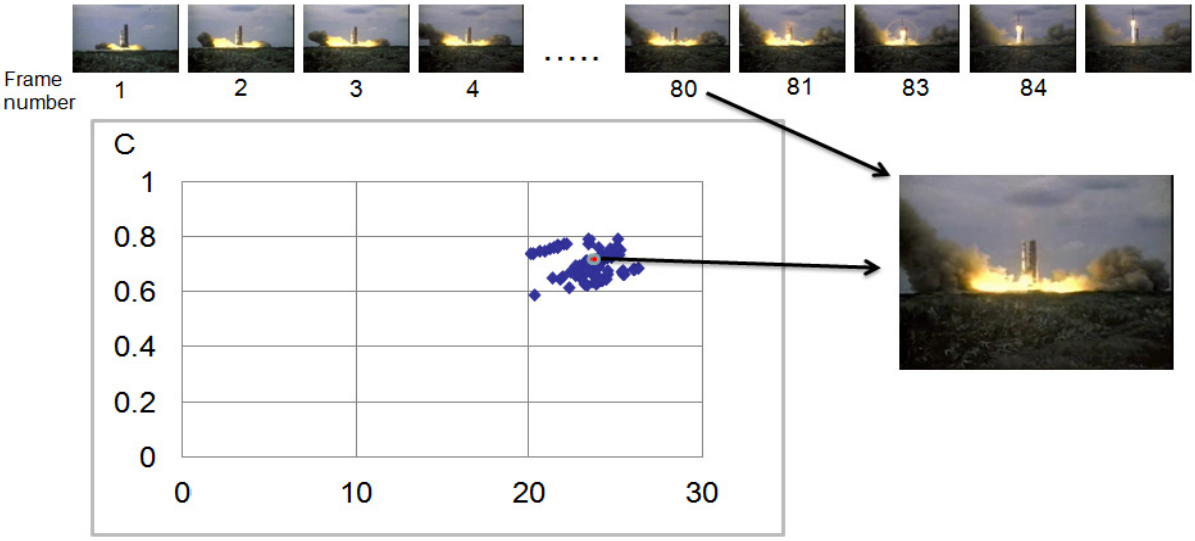

Figure 2. Distribution of points (b, c) for 1st shot

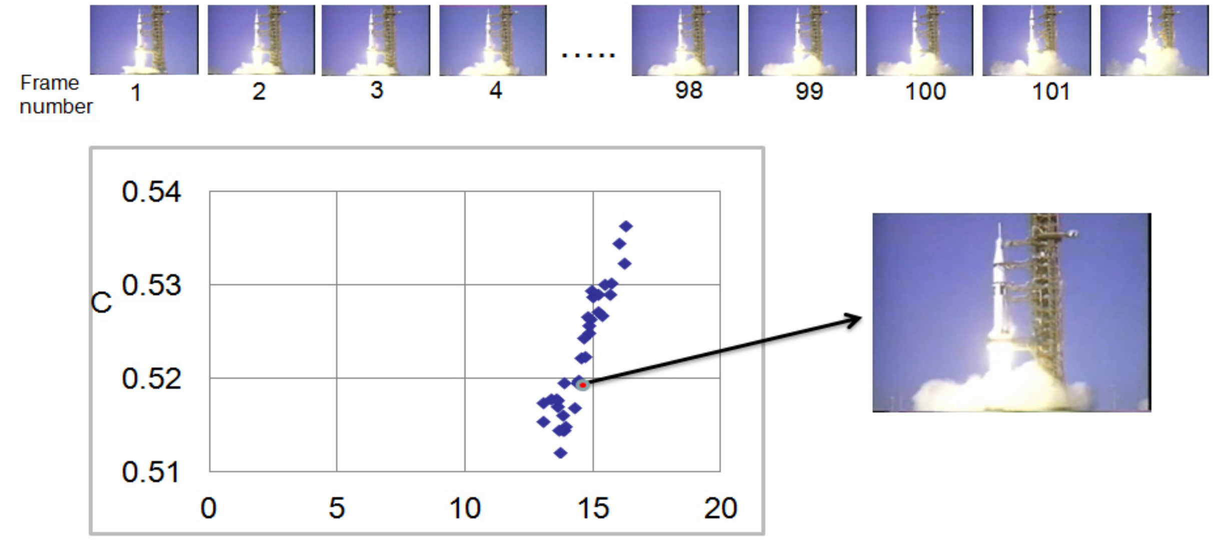

Figure 3. Distribution of points (b, c) for 2nd shot

In Table 1 are given the averaged values of parameters b and c for all frames in each shot, and also the maximum value of W^2 which are calculated during comparison of averaged values of parameters with parameters of all frames of corresponding shot. In the last column of Table 1 there are mentioned the frame numbers which are chose as key-frames.

Table 1. Key-frames which has been selected as maximum similar to frames of each shot

Shot number	$\bar{b}$	$\bar{c}$	$W^2_{\max}$	Number of chosen frame
1	23.44	0.7044	0.997	42
2	14.53	0.5223	0.999	101
3	28.25	0.5405	0.938	163
4	45.20	0.7435	0.991	219
5	118.77	0.9011	0.997	234

In the Figures 2 and 3 it is shown the distribution of points (b,c) for 1st and 2nd shots correspondingly. The visual analysis of the point's allocation in the (b, c) plot shows that the most convenient mathematical model of investigation for this kind of problem is the regression analysis. For simplicity we consider a linear regression analysis for the stochastic dependency between parameters b and c, wherein the absence of significant value of regression indicates the slight changeability of the frames content inside the shot, and on the contrary, the existence of significant regression may evidence about quit rapid changeability of frames.

In Figures 4 - 7 the experimental results of regression analysis for corresponding four shots is illustrated. For more visibility we also bring the determined linear regression graphs, and also the correlation coefficient R and F-ratio are given.

We would like to mention that the given results of formal analysis are considered as auxiliary characteristics, and therefore they should be accompanied with informative analysis of the video sequence itself.

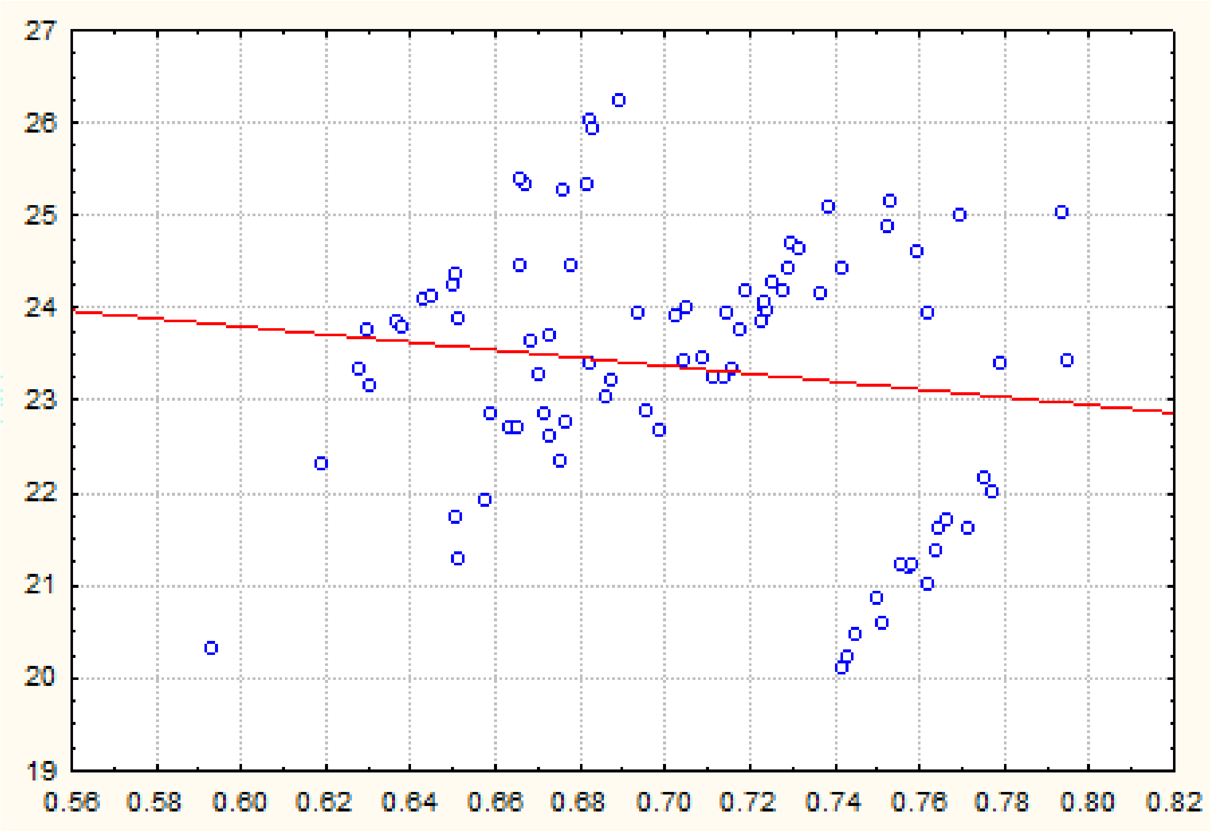

Figure 4. Regression analysis for 1st shot (R=0.138, F=1.61)

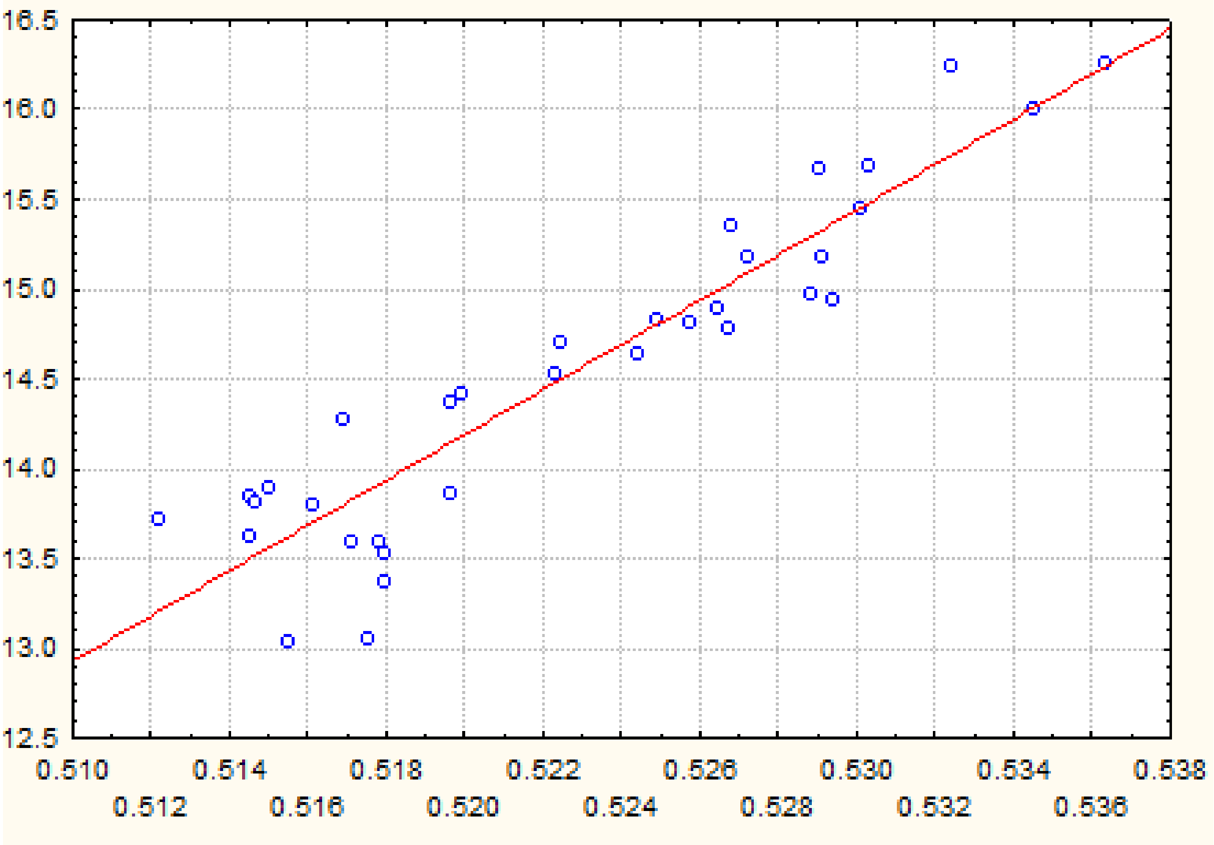

Figure 5. Regression analysis for 2nd shot (R=0.929, F=200.5)

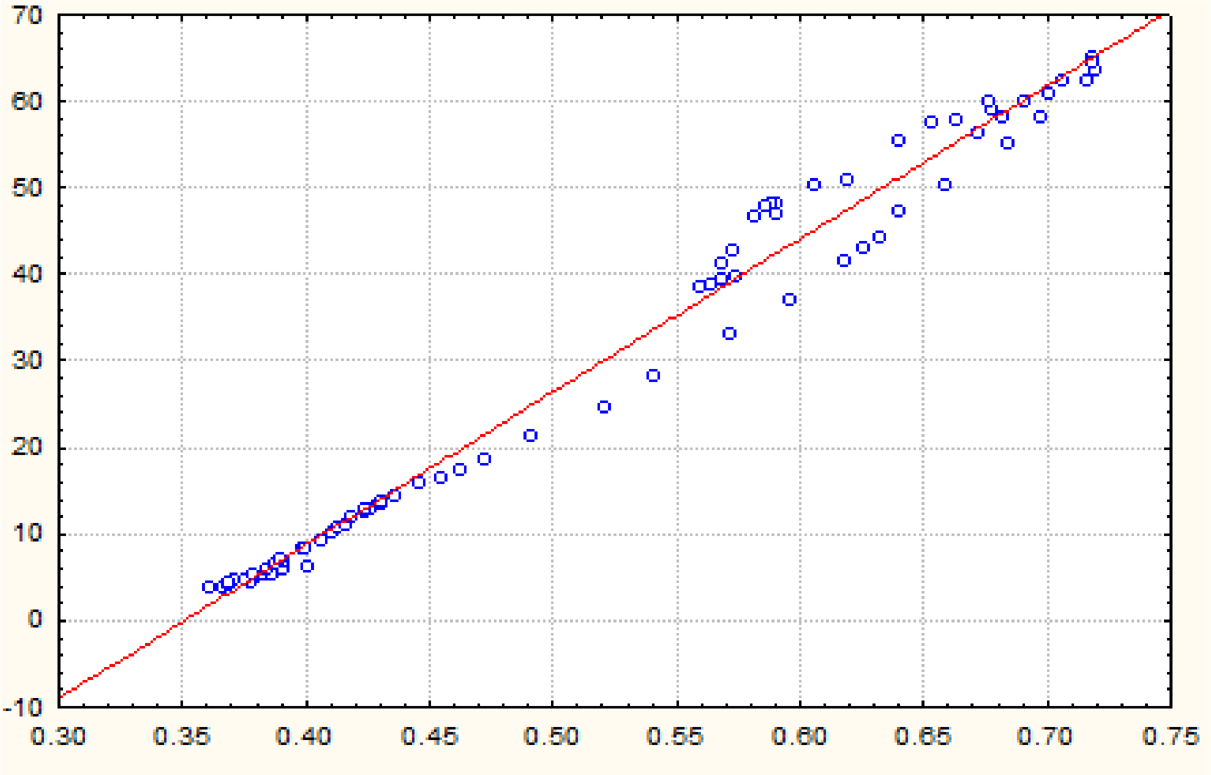

Figure 6. Regression analysis for 3rd shot (F=4559, R=0.991)

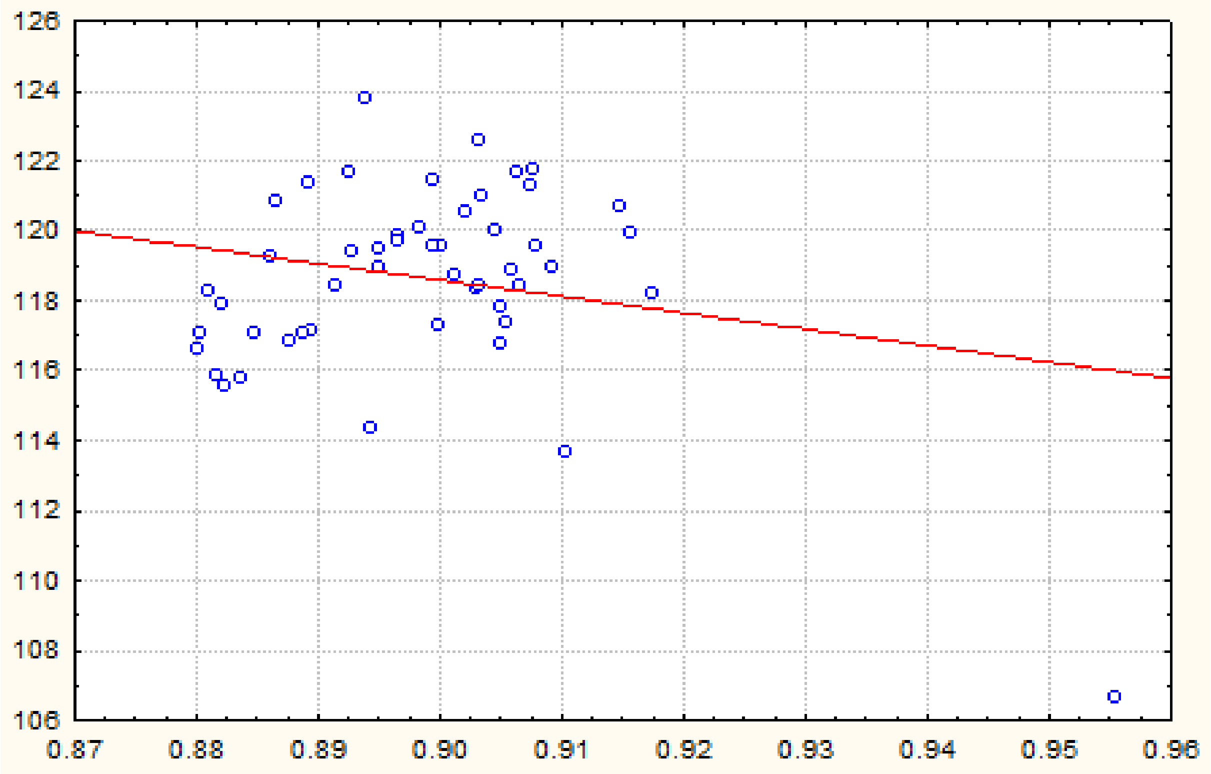

Figure 7. Regression analysis for 4rd shot (R=0.224, F=2.53)

Conclusion

In this paper, we have proposed a novel approach to content-based video indexing and retrieval by using a measure of structural similarity of frames. The first step of the proposed procedure is a shot detection; the second one is the key-frames determination for each shot. The key-frames are determined by using only the information obtained in the first step and its corresponding statistical analysis. The main advantages of proposed approach are the computational low time and keeping only two parameters for each shot which are very important for video retrieval tasks.

As a side effect during this investigation, we provide a new concept of shot segment behavior by analyzing the dependence between specified parameters.

Bibliography

[Asatryan, 2009] D. Asatryan, K. Egiazarian. “Quality Assessment Measure Based on ImageStructural Properties”. Proc. Of International Workshop on Local and Non-Local Approximation in Image Processing. Finland, Helsinki, pp. 70-73, 2009.

[Asatryan, 2010] David Asatryan, Karen Egiazarian, Vardan Kurkchiyan. Orientation Estimation with Applications to Image Analysis and Registration. International Journal “Information Theories and Applications”, Vol. 17, Number 4, pp. 303-311, 2010.

[Asatryan, 2014] D.G. Asatryan, M.K. Zakaryan. Improved Algorithm for Video Shot Detection. International Journal "Information Content and Processing", vol. 1, pp. 66-72, Number 1, 2014.

[Asatryan, 2014] D.G. Asatryan, M.K. Zakaryan. Method for Video Shot Detection and Separation. International Journal "Information Models and Analyses" Volume 3, pp. 247-251, Number 3, 2014.

[Calic, 2002] J. Calic and E. Izquierdo, “Efficient key-frame extraction and video analysis,” in Proc. Int. Conf. Inf. Technol.: Coding Comput., Apr. 2002, pp. 28–33.

[Calic, 2004] J. Calic and B. Thomas, “Spatial analysis in key-frame extraction using video segmentation,” in Proc. Workshop Image Anal. Multimedia Interactive Services, Lisbon, Portugal, Apr. 2004.

[Ferman, 2003] A.M. Ferman and A.M. Tekalp. “Two-stage hierarchical video summary extraction to match low-level user browsing preferences,” IEEE Trans. Multimedia, vol. 5, no. 2, pp. 244–256, Jun. 2003.

[Girgensohn, 2000] A. Girgensohn and J. Boreczky, “Time-constrained key-frame selection technique,” Multimedia Tools Appl., vol. 11, no. 3, pp. 347–358, 2000.

[Kang, 2005] H.W. Kang and X.S. Hua. “To learn representativeness of video frames,” in Proc. ACM Int. Conf.Multimedia, Singapore, 2005, pp. 423-426.

[Lew, 2006] M.S. Lew, N. Sebe, C. Djeraba, and R. Jain, “Content-based multimedia information retrieval: State of the art and challenges,” ACM Trans. Multimedia Comput., Commun. Appl., vol. 2, no. 1, pp. 1–19, Feb. 2006.

[Truong, 2007] B.T. Truong and S. Venkatesh, “Video abstraction: A systematic reviewand classification,” ACM Trans. Multimedia Comput., Commun. Appl.,vol. 3, no. 1, art. 3, pp. 1–37, Feb. 2007.

[Weiming, 2011] Weiming Hu, Senior Member, IEEE, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank, “A Survey on Visual Content-Based Video. Indexing and Retrieval”. IEEE transactions on systems, man, and cybernetics-Part c: Applications and reviews, vol. 41, no. 6, November 2011, pp 797-813.

[Zhang, 1997] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, “An integrated system for content-based video retrieval and browsing,” Pattern Recognit., vol. 30, no. 4, pp. 643–658, 1997

Authors' Information

***David Asatryan** – Professor, Head of group of the Institute for Informatics and Automation Problems of NAS Armenia, 1, P.Sevaki Str., 0014, Yerevan, Armenia; e-mail: dasat@ipia.sci.am*

Major Fields of Scientific Research: Digital signal and image processing.

***Manuk Zakaryan** – Ph.D. student at Russian – Armenian (Slavonic) University, Software Developer at EGS Armenia; e-mail: zakaryanmanuk@yahoo.com*

Major Fields of Scientific Research: Digital signal and image processing, Software developing.