IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Sequence Matching for Content-Based Video Retrieval

Sergii Mashtalir
Informatics Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
sergii.mashtalir@nure.ua

Olena Mikhnova
Cibernetics Department
Kharkiv Petro Vasylenko National Technical University of Agriculture,
Kharkiv, Ukraine
elena_mikhnova@ukr.net

Mykhailo Stolbovyi
Informatics Deptment
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
d_inf@nure.ua

***Abstract*— In this paper the authors propose a novel technique for comparing video frame sequence presented in an arbitrary metric space. By reviewing existing best practices in spatio-temporal video segmentation and frame matching, the authors suggest mathematical grounding for efficient video content analysis. Variants of relationships are observed between the frame sequences under comparison (perfect match, inclusion, equality of cardinality of sets). Examples of application as well as estimation metrics are also provided.**

***Keywords*— Video Content Matching, Spatio-Temporal Segmentation, Set Theory, Metric Space**

I. Introduction

Great diversity of artificial intelligence problems emerged during the last two decades. Researchers around the globe are trying to streamline people's activities by introducing contemporary methods and techniques that aid machines in decreasing human mental workload. Multimedia processing is among leading areas of research and development in numerous companies. Motorola Multimedia Research Lab, IBM Research, FX Palo Alto Laboratory, Google, just to name a few.

In this paper we consider frame sequence matching which may turn out quite complicated for a machine because of blends and dissolves that make color and texture changes almost impossible to track under some visual conditions [1-3]. The concept of video segmentation and frame matching is also typical for any frame extraction procedure. In video processing, the closer the frames are to each other in terms of some metric, the harder is to pick up a boundary between them. The nature of segments and segmentation process itself is crucial for successful matching. The simplest way of segmenting video content is dividing it into fragments of equal length. Despite it may reduce time needed for processing for more than a half part, of course it is not a perfect idea of segmentation as a scene may appear in consecutive segments or several scenes may be contained in one segment. A strong post-processing is needed after such a temporal segmentation, which is very hard to ensure [4, 5].

Inter-frame difference (for cuts) and skipping frame difference (for dissolves) are considered to be the main sources for spatio-temporal segmentation. It can be measured by dissimilarity of pixels, frame blocks, or the whole frames. Among color feature comparison techniques, histogram difference gains most popularity and simplicity. Most of the current methods can cope with both kinds of inter-frame transitions (cuts and dissolves). When cuts occur, a method should detect changes in two consecutive frames, while for dissolves a method should analyze a number of consecutive frames to detect a new scene. Dissolves are harder to detect with traditionally used color-based methods only (or skipping frame difference should be used), but a good method should distinguish both kinds of transitions at the same video, as no one knows editor's plans of video organization [4, 6].

More than fifty spatio-temporal segmentation and frame matching techniques are briefly observed in [7-15]. Such a large number approaches to the video segmentation and the shot detection indicate, on the one hand, the interest of scientists in these methods development, and on the other, the need to develop new ones, because there are no universal approaches suitable for analyzing arbitrary video data. The main problem they face lies in a variety of video content genres, without mentioning object and camera motion, flashes and other changes in lighting conditions. Most of the available methods take into account frame difference, without paying great attention to content which changes in time. Fig. 1 below details how frame matching techniques are distributed according to their popularity. The most widely used techniques that remain fundamental parts of the most successful approaches turn out to be color histograms and machine learning. Other techniques such as detecting camera flashes or working only in the compressed domain are not yet of widespread applicability [4, 7].

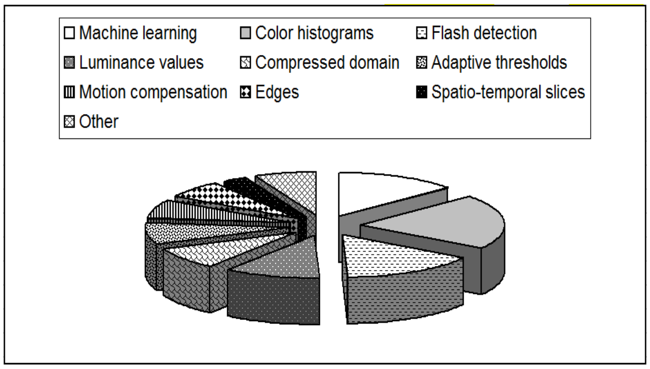

Fig. 1. Pie chart of the most popular frame matching techniques

It is important to figure out how video frames should be compared to each other to determine exact and fuzzy matches. The next section outlines the novel unique model proposed for such purposes. It has been already mentioned that a good spatio-temporal segmentation is delimited by fades and wipes along with lighting conditions, camera angle change, etc. With this in mind, such a model should be

978-1-5386-2874-4/18/$31.00 ©2018 IEEE

constructed that solves the aforementioned problems from mathematical and applied points. The model should also comply with existing standards along with contemporary level of video feature presentation and processing. After introducing the model, the latter section of the paper testifies its application significance and provides measures for estimation.

II. Video Segment Matching Model

Consider an arbitrary metric space Ω with a specified metric $\rho(x,y)$ where $x, y \in \Omega$. In addition, assume a set F which elements are finite sequences of elements from Ω. In fact, the set F represents temporal segment in video Ω, which is a tuple of frames $x_1, x_2, ..., x_n$. To be more precisely, this means $\overline{x} = (x_1, x_2, ..., x_n) \in F$ if $x_1, x_2, ..., x_n \in \Omega$, n is an arbitrary non-negative integer that is greater than zero, and the order of elements $x_1, x_2, ..., x_n$ is essential. In other words, they cannot be rearranged without changing the element from the set F. In this case, such sequences will be called tuples. Thus, the set F is constructed from element tuples of the metric space Ω. Then, assume the following pattern on the set F. Introduce the notion of distance matrix for the tuple pair.

Definition 1. Suppose matrix $A(\overline{x}, \overline{y})$ is the distance matrix for the pair of elements $\overline{x}, \overline{y} \in F$, which is constructed in the following way:

Compare $card(\overline{x}) = n$ and $card(\overline{y}) = m$. Suppose the first tuple is less or equal to the second one $n \le m$, then the fist row of distance matrix $A(\overline{x}, \overline{y})$ looks as follows: $\rho(x_1, y_1) \quad \rho(x_2, y_2) \dots \rho(x_n, y_n)$. The second row looks like: $\rho(x_1, y_2) \quad \rho(x_2, y_3) \dots \rho(x_n, y_{n+1})$, etc. Then, s-th row is $\rho(x_1, y_s) \quad \rho(x_2, y_{s+1}) \dots \rho(x_n, y_{s+n-1})$. Consequently, the number of rows in $A(\overline{x}, \overline{y})$ is equal to $s = m - n + 1$, and the matrix itself looks as follows:

$$A(\overline{x}, \overline{y}) = \begin{pmatrix} \rho(x_1, y_1) & \dots & \rho(x_n, y_n) \\ \rho(x_1, y_2) & \dots & \rho(x_n, y_{n+1}) \\ \dots & \dots & \dots \\ \rho(x_1, y_s) & \dots & \rho(x_n, y_{s+n-1}) \end{pmatrix}. \quad (1)$$

The size of the above matrix is $s \times n$, taking into consideration that $n \le m$, $s = m - n + 1$, $n = card(\overline{x})$, $m = card(\overline{y})$. Assume the following properties of the marix $A(\overline{x}, \overline{y})$.

Property 1. If the distance matrix contains a zero row, then it means that the smaller tuple (in terms of cardinality) is fully included somewhere in the bigger one. In this case $\overline{x} \subset \overline{y}$.

Property 2. If the total number of elements in the first and the second tuple are equal to each other $card(\overline{x}) = card(\overline{y})$, then the distance matrix $A(\overline{x}, \overline{y})$ is constructed from a single row:

$$A(\overline{x}, \overline{y}) = (\rho(x_1, y_1), \dots, \rho(x_n, y_n)). \quad (2)$$

Property 3. When the two tuples are fully equal to each other $\overline{x} = \overline{y}$, then the distance matrix $A(\overline{x}, \overline{y})$ is constructed from a single zero row, and it looks like this:

$$A(\overline{x}, \overline{y}) = (0, \dots, 0). \quad (3)$$

Property 4. $A(\overline{x}, \overline{y}) = A(\overline{y}, \overline{x})$ for $\forall\, \overline{x}, \overline{y} \in F$.

All of the above properties follow from the definition of the distance matrix and from the fact that $\rho(x, y)$ is initially a metric. Now, consider a series of functionals for the set $F \times F$, i.e. for its Cartesian square. Suppose $\overline{x}, \overline{y} \in F$ and $card(\overline{x}) = n \le m = card(\overline{y})$, then the following functionals are correspondent with them:

$$\begin{gathered} g_1(\overline{x}, \overline{y}) = \sum_{i=1}^{n} \rho(x_i, y_i), \\ g_2(\overline{x}, \overline{y}) = \sum_{i=1}^{n} \rho(x_i, y_{i+1}), \\ \dots \\ g_s(\overline{x}, \overline{y}) = \sum_{i=1}^{n} \rho(x_i, y_{i+s-1}) \end{gathered} \quad (4)$$

where $s = m - n + 1$. The following theorem is held.

Theorem 1. Each functional specified by the equations (4) is a metric on the set F.

The above theorem is easily proved by ensuring reflexivity, symmetry and triangle inequality.

Reflexivity. If the two tuples are fully equal to each other $\overline{x} = \overline{y}$, then from the Property 3 it follows that $A(\overline{x}, \overline{y}) = (0, \dots, 0)$, i.e. there exists $g_1(\overline{x}, \overline{x}) = 0$.

Symmetry. Symmetry apparently follows from the Property 4.

Triangle inequality. Triangle inequality can be explained using the example $g_1(\overline{x}, \overline{y})$. Suppose there are three tuples $\overline{x}, \overline{y}, \overline{z}$. The following Fig. 2 illustrates these in a schematic manner.

Then, it is clear that:

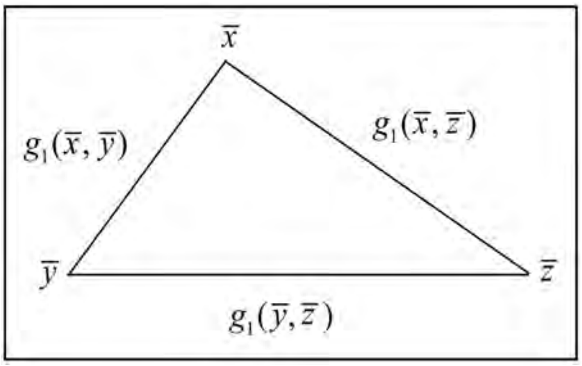
$\overline{x}$
$g_1(\overline{x}, \overline{y})$
$g_1(\overline{x}, \overline{z})$
$\overline{y}$
$\overline{z}$
$g_1(\overline{y}, \overline{z})$

Fig. 2. Triangle inequality for the three tuples

$$\begin{gathered} g_1(\bar{x},\bar{y}) = \sum_{i=1}^{n} \rho(x_i, y_i), \\ g_1(\bar{x},\bar{z}) = \sum_{i=1}^{n} \rho(x_i, z_i), \\ g_1(\bar{y},\bar{z}) = \sum_{i=1}^{n} \rho(y_i, z_i) \end{gathered} \tag{5}$$

where $card(\bar{x}) = n,\ card(\bar{y}) = m,\ card(\bar{z}) = k$ and $n \le m \le k$. For the sake of certainty, consider $g_1(\bar{x},\bar{y}) + g_1(\bar{x},\bar{z})$, then the following equation is held from (5):

$$g_1(\bar{x},\bar{y}) + g_1(\bar{x},\bar{z}) = \sum_{i=1}^{n} [\rho(x_i, y_i) + \rho(x_i, z_i)]. \tag{6}$$

As $\rho(x, y)$ is a metric, then the following is held for any i:

$$g_1(\bar{x},\bar{y}) + g_1(\bar{x},\bar{z}) = \sum_{i=1}^{n} [\rho(x_i, y_i) + \rho(x_i, z_i)]. \tag{7}$$

By taking equation (6) into account, the following can be obtained:

$$g_1(\bar{x},\bar{y}) + g_1(\bar{x},\bar{z}) \ge g_1(\bar{y},\bar{z}). \tag{8}$$

The other two couples of summands needed in the triangle inequality may be proved the same way, i.e. $g_1(\bar{x},\bar{y})$ is a metric. Now, we shall explain why this theorem is held for all the other functionals in (4). With this, we understand it in case of their existence for particular tuple cardinalities. The following Fig. 3 shows this relation in a form of a schema.

The number of functionals in (4) corresponds to the number of times the smallest tuple can be included into the medium-sized tuple. For those number of functionals it is essential to consider the triangle inequality. At the end, we may conclude that $g_1(\bar{x},\bar{y})$ always exists. The theorem is proved.

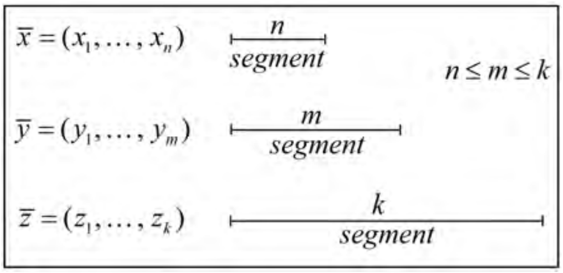
$\bar{x} = (x_1, \dots, x_n)$
n
segment
$n \le m \le k$
$\bar{y} = (y_1, \dots, y_m)$
m
segment
$\bar{z} = (z_1, \dots, z_k)$
k
segment

Fig. 3. Schema of possible relations between the three tuples

Consider the functional $g(\bar{x},\bar{y})$ as the sum of all the elements in the distance matrix. It is easily seen that this functional is symmetric and reflexive as well as $g_1(\bar{x},\bar{y})$. The triangle inequality should be observed separately. Suppose $card(\bar{x}) = card(\bar{y}) = card(\bar{z}) = n$, then the distance matrices $A(\bar{x},\bar{y}), A(\bar{x},\bar{z}), A(\bar{y},\bar{z})$ will look as (2), and the functionals $g(\bar{x},\bar{y}), g(\bar{x},\bar{z}), g(\bar{y},\bar{z})$ will look as (4). By analogy to the theorem 1, the triangle inequality is grounded for $g(\bar{x},\bar{y})$. By transferring to a more general case, consider $card(\bar{x}) = n,\ card(\bar{y}) = m,\ card(\bar{z}) = k$. Assume the following relations are held:

$$\begin{cases} n \le m \le k, \\ m - n = s_1, \\ k - m = s_2. \end{cases} \tag{9}$$

It is clear that $k - n = s_1 + s_2$, and the distance matrices will look as follows:

$$\begin{gathered} A(\bar{x},\bar{y}) = \begin{pmatrix} \rho(x_1, y_1) & \dots & \rho(x_n, y_n) \\ \vdots & \ddots & \vdots \\ \rho(x_1, y_s) & \cdots & \rho(x_n, y_{s_1+n-1}) \end{pmatrix}, \\ A(\bar{x},\bar{z}) = \begin{pmatrix} \rho(x_1, z_1) & \dots & \rho(x_n, z_n) \\ \vdots & \ddots & \vdots \\ \rho(x_1, z_{s_1+s_2}) & \cdots & \rho(x_n, z_{s_1+s_2+n-1}) \end{pmatrix}, \\ A(\bar{y},\bar{z}) = \begin{pmatrix} \rho(y_1, z_1) & \dots & \rho(y_m, z_m) \\ \vdots & \ddots & \vdots \\ \rho(y_1, z_{s_2}) & \cdots & \rho(y_m, z_{s_2+m-1}) \end{pmatrix}. \end{gathered} \tag{10}$$

The following equations can be obtained from the above:

$$\begin{aligned} & g(\bar{x},\bar{y}) = [\rho(x_1, y_1) + \dots + \rho(x_n, y_n)] + \dots \\ & +\ [\rho(x_1, y_{s_1}) + \dots + \rho(x_n, y_{s_1+n-1})], \\ & g(\bar{x},\bar{z}) = [\rho(x_1, z_1) + \dots + \rho(x_n, z_n)] + \dots \\ & +\ [\rho(x_1, z_{s_1+s_2}) + \dots + \rho(x_n, z_{s_1+s_2+n-1})], \\ & g(\bar{y},\bar{z}) = [\rho(y_1, z_1) + \dots + \rho(y_m, z_m)] + \dots \\ & +\ [\rho(y_1, z_{s_2}) + \dots + \rho(y_m, z_{s_2+m-1})]. \end{aligned} \tag{11}$$

The equations in (11) enable checking triangle inequality directly. By rearranging the summands and assuming that $\rho(x, y)$ is a metric, this check testifies fulfillment of the triangle inequality in a general case. The following section provides information on the model implementation on video sequences and estimation of the results. To enhance this paradigm in future, it may be interesting to divide the first (smaller) set into subsets and perform search of these smaller subsets in the second (bigger) set. The practical application of it seems quite trivial as not always the whole video scene is repeated, but a small fragment of it.

III. Experimental Result

Assume Ω is a video sequence of frames. Let $\rho(x, y)$ be a metric or a distance between the two elements from this set Ω . Here, x and y are minimum possible video elements, i.e. frames in case of video. Suppose $\overline{x} = (x_1, x_2, ..., x_n) \in F$, $\overline{y} = (y_1, y_2, ..., y_m) \in F$ where F is a set of all the scenes in a video. Thus, $\overline{x}$ and $\overline{y}$ are the two video segments (fragments or scenes) for comparison, and $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_m$ are the frames in these segments, which number equals to n and m respectively. The above matrix (1) indicates how two scenes match each other.

For example, as a first sequence, a fragment of a documentary was taken, illustrating the rescue helicopter crash. In Fig. 4 a) the complete video sequence is shown in the form of 501 frames, which can be divided into a set of segments by the video spatio-temporal segmentation approaches proposed in [4,16]. As a result, you can get the following partition: the first segment, 1..66 frames are illustrating the lone flight of the helicopter; 67...151 frames are helicopter flight against the shore; 152...260 frames are the landing process and rescuers work; 261...332 frames are departure of the rescue helicopter; 333...451 frames are the process of crash; 451...501 frames are segment showing the result of the rescue helicopter crash into the water.

Next, for the experiment, the first and last segments from the original data were taken, with a length of 66 and 49 frames, respectively. Examples of frames from this sequence are shown in Fig. 4 b) and c).

To these segments, the same approach to segmentation was applied similarly and the results corresponding to these segments were obtained. The graphs in Fig. 5 illustrate result of video segmentation for all 3 video. Despite the fact that the segmentation approach was applied to different sequences, the values obtained for segments 4b) and c) correspond to values in the intervals from 1 to 66 frames and from 452 to 501 frames of the original video data. It is quite logical that when we using the proposed metric for comparison with the initial sequence (Fig. 4a)) one of the analyzed segments (Fig. 4b) or Fig. 4c)), we obtained at the appropriate places of the distance matrix (1) a zero sequences values 66 and 49 frames, respectively.

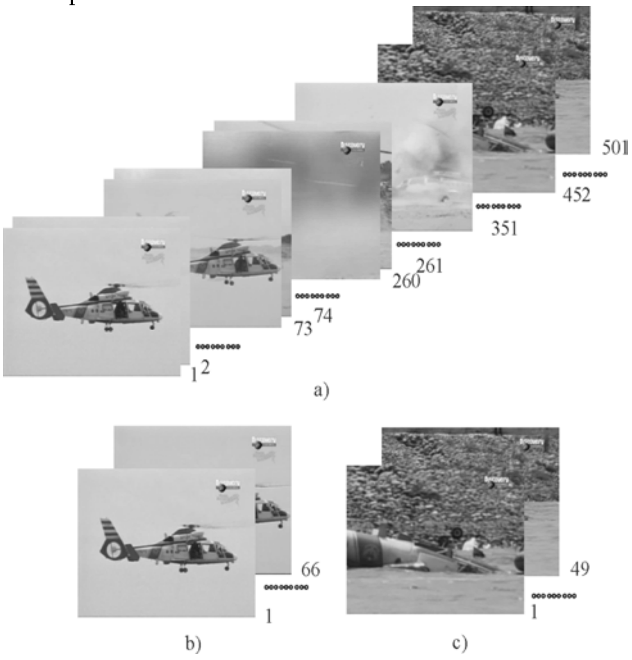

Fig. 4. Example of initial video data and segments to compare

Thus, we can compare video sequences and find the same or similar if we establish a certain threshold value for possible differences in the distance matrix (1). In other words, if we compare the results of different video sequences segmentation with the proposed approach, we obtain a certain sequence in the distance matrix whose values do not exceed the established threshold, then we can say about the similarity of the compared data, and in the case of obtaining a zero sequences about the conjunction of the compared data corresponding parts.

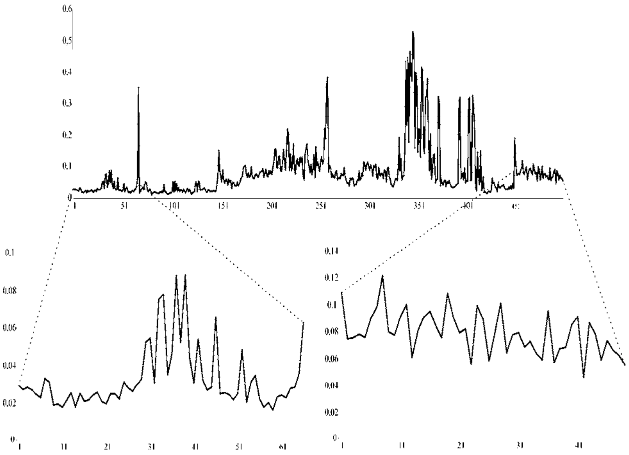

Fig. 5. Result of spatio-temporal video segmentation

IV. Conclusion

The novel proposed model may be effectively implemented for video scene comparison. The effectiveness of it can be testified by traditional precision-recall metrics in terms of finding perfect matches for the subsets under analysis. For video content presentation and scene comparison, temporal segmentation plays the key role because extraction of correct scene fragment duplicate is basically what we are trying to reach using the model. Precision-recall metrics reveals correspondence of multidimensional data processing results to human expectations from mental analysis of such data. Although this estimation is performed by human experts, which may be subjective in a sense, the combination of metric parameters more precisely indicates opportunities of the model. The true positives show how many relevant frame sequences are extracted. The false positives are considered being mismatches of extracted segments. The false negatives are the omitted segments that should actually be extracted. The true negatives are not being used as they are the inverse from the above [18, 19]. The only drawback of the estimation is that it does not take into account fuzzy matches and half-satisfaction of the experts, which may be the topic of further research.

References

[1] D. Schonfeld, et. al., Video search and mining. Studies in Computational Intelligence. Springer, Berlin, 2010.

[2] R. Szeliski, Computer vision. Algorithms and applications. Springer, London, 2011.

[3] L. Chen, and F. W. M. Stentiford "Video sequence matching based on temporal ordinal measurement," Pattern Recognition Letters., vol. 29, pp. 1824-1831, 2008.

[4] S. Mashtalir, and O. Mikhnova, "Key frame extraction from video: framework and advances," J. Computer Vision and Image Processing. vol. 4(2), pp. 67-78, 2014. (https://www.igi-global.com/article/key-frame-extraction-from-video/115840)

[5] H. Lu, and Y.-P. Tan, "An effective post-refinement method for shot boundary detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 15(11), pp. 1407–1421, November, 2005.

[6] W. Heng, and K. Ngan, "Shot boundary refinement for long transition in digital video sequence", IEEE Transactions on Multimedia, vol. 4(4), pp. 434-445, December, 2002.

[7] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," J. Computer Vision and Image Understanding. vol. 114(4), pp. 411-418, 2010.

[8] Zhang Y.-J. (ed.), Advances in image and video segmentation. Hershey- London-Melbourne-Singapore: IRM Press, 2006.

[9] S. Porter, M. Mirmehdi, and B. Thomas, "Temporal video segmentation and classification of edit effects", Image and Vision Computing., vol. 21, pp. 1097-1106, December 2003.

[10] S. Piramanayagam, E. Saber, N. D. Cahill, and D. Messinger, "Shot boundary detection and label propagation for spatio-temporal video segmentation" Proc. SPIE 9405, Image Processing: Machine Vision Applications VIII, 94050D 7 p., February 2015.

[11] S. Thakare, "Intelligent processing and analysis of image for shot boundary detection," International Journal of Emerging Technology and Advanced Engineering., vol. 2, no. 2, pp. 208-212, Mar.-Apr. 2012.

[12] R. Vázquez-Martín, and A. Bandera, "Spatio-temporal feature-based keyframe detection from video shots using spectral clustering," Pattern Recognition Letters, vol. 34, no. 7, pp. 770-779, 2013.

[13] G. I. Rathod, and D.A. Nikam, "An algorithm for shot boundary detection and key frame extraction using histogram difference," Int. J. Emerging Technology and Advanced Engineering, vol. 3(8), pp. 155-163, August, 2013.

[14] J. Nesvadba, F. Ernst, J. Perhavc, J. Benois-Pineau, and L. Primaux, "Comparison of shot boundary detectors", Int. Conf. on Multimedia and Expo, IEEE Press, Amsterdam, pp. 6-8, 2005.

[15] H. Jiang, G. Zhang, H. Wang and H. Bao, "Spatio-temporal video segmentation of static scenes and its applications" IEEE Transactions on Multimedia., vol. 17, no. 1, pp. 3-15, January, 2015.

[16] Y. Bodyanskiy, D. Kinoshenko, S. Mashtalir, and O. Mikhnova, "On-line video segmentation using methods of fault detection in multidimensional time sequences", Int. J. of Electronic Commerce Studies, vol. 3(1), pp. 1-20, 2012.

[17] O. Mikhnova, and N. Vlasenko, "Key frame partition matching for video summarization," Int. J. of Information Models and Analyses, vol. 2(2), pp. 145-152, 2013.

[18] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge University Press, Cambridge, 2008.

[19] S. V. Mashtalir, and O. D. Mikhnova, "Stabilization of key frame descriptions with higher order Voronoi diagram", J. Bionics of intelligence. vol. 1, pp. 68-72, 2013.