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

Representative Based Clustering of Long Multivariate Sequences with Different Lengths

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

Volodymyr Mashtalir
Informatics dept.
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
volodymyr.mashtalir@nure.ua

Mykhailo Stolbovyi
Informatics dept.
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine

***Abstract*—Video streams as unstructured or poorly structured data issue a challenge to create a unified framework capable to depict and convey high-level stories. Up-to-date indexing and search techniques to manage video data are able to operate the voluminous amounts of contained in video information in order to detect spatial and temporal events. Nevertheless, bridging semantic gap between the low-level frame or video features and high-level semantic concepts necessitates extremely high-speed procedures of temporal unlabeled data. Automatic video annotation in visual forms appears one of the promising approaches representing most pertinent and crucially important information. This goal is achieved by (among others) clustering large collections of video data.**

Keywords—video stream, clustering, metric

I. INTRODUCTION

To detect spatial and temporal events in video streams as a whole, to be able to have easily understandable visual abstracts of video collections, it is essential to have the means for creation of summaries in forms of video skims (frame sequences composed with excerpts of the original video) and key frames (the most semantically typical frame or several ones from the consecutive image sequence) sets. Summarizing video content is substantial for at least such applications [1,2] as video browsing archiving content based retrieval, access to teleconferences, video mail and video news, etc. It should also be noted that advanced tools for film trailers making, generating of TV programs previews and other similar treatments to quick overview of the content can not be simple reductions of the original. Video processing requires to keep, at least, the restricted inquiries content aspects and to reduce the semantic gap between the features extracted from images (or/and their time sequences) [3,4].

There arise remarkable achievabilities in video processing approaches based on specific feature integration via CNN (Convolutional Neural Networks) to extend, in some measure, capabilities of task-dependent and scene-dependent explanations of visual spatial and temporal events to index, store, linear and nonlinear edit, retrieve and present video records and excerpts in convenient mode [5].

Nevertheless, promising way of solving video streams analysis problems remains in temporal (pre-, on line, post) processing of multivariate time series. Such series are induced by frame sequences in various ways. First of all, it is necessary to allocate time series in terms of local and global features concerning color, texture, shape extracted from each frame. In the second, it is it should be emphasized time series discovered in matrix form which are agreed with one or more subframes covering field of view or surrounding some content associated localities in the image. Local features may be related with edge, corner, ridge, blob detection, the histogram of oriented gradients (HOG), the gradient location and orientation histogram (GLOH), the scale-invariant feature transform (SIFT), the speeded up robust features (SURF) etc. [1,6].

It is absolutely recognizable that perceptual process is closely related with salient regions of the visual scene. To consider all the visual impressions contained in videos it is required to take into most complete account useful for human attention concept with the object of perceptual video complexity reduction and support event decision on the base of the relevant parts of sensory input selections. The ability to extract the relevant slices of an image is of large interest, especially if each frame from video stream is unavoidable processed in real time. So, prevalent belief, that visual attention models are useful starting-point for arbitrary type of content based video parsing, predefines objects of an automatic frame by frame analyze. In field of view such objects can be represented as some significant feature points and their neighborhoods [7].

The points that satisfy the properties of repeatability distinctiveness, stability, uniqueness, interpretability are usually called saliency points [6,8]. The saliency map can be elucidated as spatially arranged point collection that provides semantic stability of a visual scene. Moreover, temporal proximity assumption ensures that the salient points of the current frame are localized in windows surrounding the locations of each previously found salient point. Thereby, any saliency model based on spatiotemporal salient points should predict what attracts the (human) attention and sequence of saliency maps or window sets, covering a subset of salient points, produce a multivariate time series what gives serious reason to get brief valid video summary via temporal clustering.

Clustering of time series data induced by video streams, first and foremost, is aimed to an exploration of video tagging. Shots and key frames, as a rule, are adapted as units for excerpts labeling when segmenting video and searching for sets of similar temporal fragments, desirably, with nearly equivalent content [4,9]. If any known clustering algorithms can be applied when using key frames (at least in feature spaces) [10, 11], then the clustering of time series is associated with a number of difficulties , the main one of which lies in different lengths of the sequences to be processed [12]. Thus, techniques advancement to get a quick

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

video overview is remained profound challenge.

The remainder of the paper is structured as follows. The second section is devoted to available capabilities destined for different lengths streams matching and additional refinement of the series under consideration. The third section proposes a clustering technique providing similar temporal segment of video grouping. Further, the results of video data clustering are discussed and the tasks to be solved in the future are defined.

II. Preliminary Remarks and Statement of the Problem

To estimate the similarity of two multidimensional time series induced by video streams, one can use a modification of the popular DTW (Dynamic Time Warping) method [11, 13], which was, by and large, widely used to estimate the distance between arbitrary sequences of different lengths. The essence of it is as follows. Introduce two multidimensional sequences $X=\{x(1),...,x(k),...,x(M)\}$ and $Y=\{y(1),...,y(l),...,y(N)\}$, $N\neq M$ where $x(k)$ and $y(k)$ can be represented either vectors $x(k)=(x_1(k),...,x_n(k))$, $y(l)=(y_1(l),...,y_n(l))$ in any feature space $\Omega\subset\mathbb{R}^n$ or matrices (subframe set for each frame from video stream) $x(k)=\{x_{i_1i_2}(k)\}$, $y(l)=\{y_{i_1i_2}(l)\}$, $x(k),\ y(l)\in\mathbb{R}^{h\times v}$ mostly in the image space. Next, $(N\times M)$ matrix of distances in the chosen local (element-wise) metric with elements $d(x(k),y(l))$, $k=\overline{1,N}$, $l=\overline{1,M}$ between all elements of the sequences to be matched is introduced into consideration. On the basis of this matrix a warping path is constructed as distance sequence $W=\{w_1,w_2,...,w_q,...,w_L\}$, $w_q=d(x(k),y(l))_q$, $max\{N,M\}\le L\le M+N-1$, $q=\overline{1,L}$ which, in fact, determines the similarity between X and Y on the basis of the accumulative distance

$$D(k,l)=d(x(k),y(l))+min\{D(k,l-1),D(k-1,l),D(k-1,l-1)\}. \quad (1)$$

It should be emphasized that the resulting warping path, generally speaking, is a proximity measure, but not a metric.

As the distance between local elements of X and Y, the Euclidean metric is usually used

$$d(x(k),y(l))=\|x(k)-y(l)\|_2 \quad (2)$$

which, when processing subframes of images from video, takes the form of the Frobenius norm metric

$$d(x(k),y(l))=(Sp(x(k)-y(l))(x(k)-y(l))^{\mathrm{T}})^{\frac{1}{2}}. \quad (3)$$

In situations where the processed data are noisy and have outliers, it is appropriate to use the Manhattan metric, which has some robust properties. In this case, (2) corresponds to expression

$$d(x(k),y(l))=\sum_{i=1}^{n}|x_i(k)-y_i(l)|, \quad (4)$$

and for (3) we have

$$d(x(k),y(l))=\sum_{i_1=1}^{n}\sum_{i_2=1}^{v}|x_{i_1}(k)-y_{i_2}(l)|. \quad (5)$$

Now turn to the clustering problem. Introduce a set of multidimensional time series $X_1,X_2,...,X_q,...,X_Q$ that must be grouped into clusters. It is assumed that each of the clusters contains a different number of observations, $N_1,...,N_q,...,N_Q$ respectively. Direct clustering of the original sequences seems to be ineffective, since calculation (1) is based on dynamic programming on long series produced by video streams and such clustering is associated with high computational complexity. In addition, the processed sequences are, as a rule, nonstationary, i.e. different segments can belong to different classes.

In such situations, it is possible to take advantage of the window-type approach, when each series X_q, $q=1,2,...,Q$ is divided into P window sections, resulting in a set of new series $WX_{q1},X_{q2},...,WX_{qp},...,WX_{qP}$, $p=1,2,...,P$, which are further considered as independent sequences to be clustered. it is not difficult to understand that if the windows WX_{qp} of one sequence X_q fall into different clusters, which indicates that the original sequence is nonstationary, then the resulting segments can be classified as signals independent of each other.

Thus, QP signals are introduced, each of which contains samples, and the ultimate goal is a partition of these series in the self-learning mode in homogeneous in the sense of (1) – (5) classes.

III. Clustering Based on Representatives

Among the known clustering algorithms, the most widely explored ones are based on prototype-centroids due to the simplicity of computational models and the interpretability of the results obtained [10,11]. However, these methods are of little use for clustering multidimensional time series produced by video streams, since the processed patterns in the known approaches have the same dimension, but subject to video processing the center of attention is different lengths of the series: WX_{qp} has N_{pq} elements.

In such situations, approaches connected with sample exemplary usage may be more preferable when instead of the computed prototype- centroid one of the vectors (matrices), available in the processed sample $WX_{11},...,X_{1p},...,WX_{21},...,\ WX_{qp},...,WX_{QP}$ with QP temporal segments, is selected. Consider such a clustering procedure.

The algorithm functioning begins with the selecting of the initial representatives $WX_{qp}^1(0),WX_{qp}^2(0),...,WX_{qp}^m(0)$ where m is the given number of clusters. Here, it is important to emphasize that the choice of m would not be productive without deep preliminary semantic analysis of the video parsing goals. In the capacity of $WX_{qp}^1(0)$, the sample component furthest from all others is chosen, i.e. $\forall q,p\in\{1,2,...,Q\}$, $\forall r,s\in\{1,2,...,P\}$

$$\mathrm{DTW}(\mathrm{W}\,\mathrm{X}^1_{qp}(0), \mathrm{W}\,\mathrm{X}_{qp}(0)) > \mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{rs}, \mathrm{W}\,\mathrm{X}_{qp}).$$

Further, $\mathrm{W}\,\mathrm{X}^1_{qp}(0)$ is temporarily excluded from the sample and the second representative is selected, as before, the most distant from the elements left in the sample, i.e.

$$\mathrm{DTW}(\mathrm{W}\,\mathrm{X}^1_{qp}(0), \mathrm{W}\,\mathrm{X}_{qp}(0)) > \mathrm{DTW}(\mathrm{W}\,\mathrm{X}^2_{qp}(0), \mathrm{W}\,\mathrm{X}_{qp}(0)) > \\ > \mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{rs}, \mathrm{W}\,\mathrm{X}_{qp}).$$

This selection procedure (with at each step the exception of the already selected representatives, being the starting centers of clusters) is repeated m times till to the formation of initial patterns such that

$$\mathrm{DTW}(\mathrm{W}\,\mathrm{X}^1_{qp}(0), \mathrm{W}\,\mathrm{X}_{qp}(0)) > ... > \\ > \mathrm{DTW}(\mathrm{W}\,\mathrm{X}^m_{qp}(0), \mathrm{W}\,\mathrm{X}_{qp}(0)) > \mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{rs}, \mathrm{W}\,\mathrm{X}_{qp}).$$

This procedure for selecting initial patterns is effective if the original data do not contain outliers. Otherwise, the initial representatives can be randomly chosen, just as it happens in greedy clustering algorithms, e.g. in the classical k-means. With respect to the video streams processing, such outliers can be, e.g. temporal segments of inserted advertising products, and clustering with such comparable attractors is semantically meaningless. In other words, the preprocessing of sequence $\mathrm{W}\,\mathrm{X}_{11}, ..., \mathrm{X}_{1p}, ..., \mathrm{W}\,\mathrm{X}_{21}, ..., \mathrm{W}\,\mathrm{X}_{qp}, ..., \mathrm{W}\,\mathrm{X}_{QP}$ passes into meaningful procedures.

In the second stage, the remaining $QP - m$ patterns are allocated over clusters Cl_j, $j = \overline{1, m}$ according to the relation: $\forall j \neq l \in \{1, 2, ..., m\}$ $\mathrm{W}\,\mathrm{X}_{qp} \in Cl_j$ if

$$\mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{qp}, \mathrm{W}\,\mathrm{X}^j_{qp}(0)) < \mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{qp}, \mathrm{W}\,\mathrm{X}^l_{qp}(0)).$$

Thus, all available patterns have been collected in the neighbourhoods of each of initial representatives $\mathrm{W}\,\mathrm{X}^j_{qp}(0)$.

At the third stage in each of the groups formed, a new representative is determined, which selects as the observation with the minimum total distance to all points of the initial cluster. In other words, for the improved representative for all admissible q, p, r, s, j relations

$$\sum \mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{qp} \in Cl_j, \mathrm{W}\,\mathrm{X}^j_{qp}(1)) < \\ < \sum \mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{qp} \in Cl_j, \mathrm{W}\,\mathrm{X}_{rs} \in Cl_j).$$

have to be hold.

After discovery m refined representatives $\mathrm{W}\,\mathrm{X}^j_{qp}(1)$, $j = \overline{1, m}$ the return to the second stage takes place, where the patterns are attributed to the newly formed patterns $\mathrm{W}\,\mathrm{X}^j_{qp}(1)$ according to the rule:

$$\mathrm{W}\,\mathrm{X}_{qp} \in Cl_j \text{ if } \mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{qp}, \mathrm{W}\,\mathrm{X}^j_{qp}(1)) < \\ < \mathrm{DTW}(\mathrm{W}\,\mathrm{X}_{qp}, \mathrm{W}\,\mathrm{X}^l_{qp}(1)) \; \forall j \neq l \in \{1, 2, ..., m\}.$$

This process is repeated until all representatives cease to change, i.e. for all j from 1 to m the stabilization condition $\mathrm{W}\,\mathrm{X}^j_{qp}(\alpha + 1) = \mathrm{W}\,\mathrm{X}^j_{qp}(\alpha)$ is fulfilled, where $\alpha = 0, 1, 2, ...$ is the iteration number.

The original series X_q are obtainable from $\mathrm{W}\,\mathrm{X}_{qp}$ at the end of the procedure. If for any q all elements belong to the same cluster, it means that the series X_q is stationary and does not change properties in the time interval from 1 to N_q. Otherwise, we have that in such segments there have been changes in the properties of the time series, and a more detailed analysis of the corresponding segments of the series is required.

Turning to the data under examination viz to video streams, one significant for applications detail must be noted. Plausible content-identification is required when shots are split up separated by different tricks e.g. gradual effects (fade, dissolve, wipe) and inheritances of lens distortion, magnify blur and sharpen as well as temporal smoothing, shadow/highlight effects. Due to the difficulties and challenges of such insignificant information for abstracting video (but not for segmentation of time series), the need for a more acceptable content based filtering has arisen. There exists sufficient ground to delete video transitions at clustering stage and serious attention should be paid to the detection shots equiprobably belonging to two consecutive segments, which are obtained by editing the video and in fact are obstacles to the content analysis.

IV. Results and Discussions

Associated with video clustering experiments were conducted based on 40 episodes “Destroyed in seconds” (Discovery Channel) with time length about 22 minutes each, with a resolution of 688×422 and a frame rate of 25 per second. The peculiarity of the video data used is the sufficiently rich availability of trailers that were manually excluded from further processing as well as the transitions between consecutive pairs of shots that seems quite reasonable for maintaining the adequacy of delivering the core information. Previously, corresponding to the episode time series, produced by simplest shape features of each frame segmentation, was segmented on the base of spatio-temporal approach [14]. In a number of cases, this temporal segmentation was repeated for already found shots (if their length exceeded 1500 frames).

Fig. 1 illustrates a typical shot, the aggregate of which generates multidimensional time series to be clustered. Fig. 2 shows the final representative-shot, which is used as an element in summarizing sequence.

Because of the great length and rich content of video data clustering can be repeated for the found representative-shots until the required duration of result (given or found empirically) will be obtained. At this point one important detail must be emphasized, substantial difficulties remain in the construction of semantic structure.

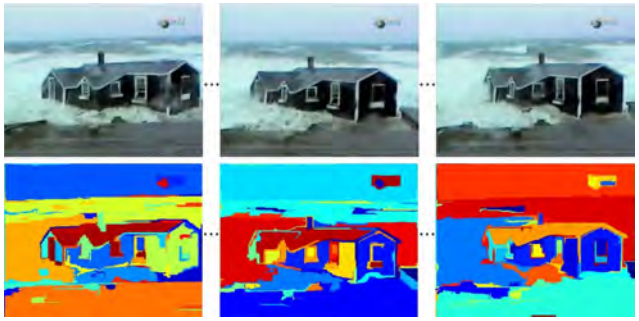

Fig. 1. Example of a shot (with usage shape features).

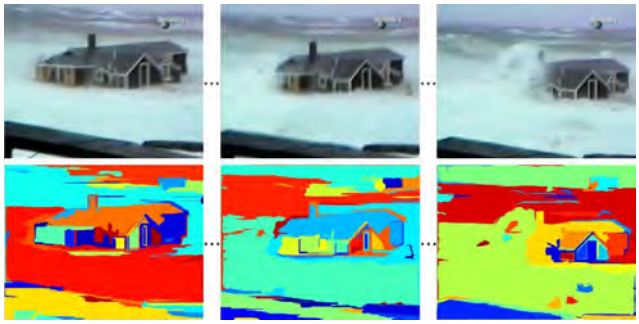

Fig. 2. Example of a key shot as clustering representative.

Similar results, but with some semantic deterioration and a noticeable reduction in computational complexity, were obtained by temporal segmentation of matrix pixel series and clustering of subframe set sequences, examples of which are shown in Fig. 3. To further elucidate clustering results, it is necessary to consider disjoint matrix sequences in time separately and together.

Fig. 3. Subframe examples for video segmentation and clustering.

Seen as a whole, the results obtained highlight the potential of the clustering approach to video summarizing. A few words of comment are necessary here about DTW. DTW opens up a whole range of new opportunities to match time series with different lengths and under certain circumstances fully valid comparisons take place. However, if there exists a considerable difference in the sequences lengths, there can be a peculiar averaging when a point (frame) from a sequence is associated with a significant number of points (frames) from another sequence. It should also be noted that the proposed approach operates with a known number of clusters, which can not be a priori determined under large video streams processing and there are no clear ways of reasonable choice.

Summing up, it can be argued that the main findings of the study reveal a clustering procedure intended to analysis of long multivariate sequences (including matrix type series) with different lengths. Further investigation will stimulate an understanding of rational detection and application of content based saliency maps with possible backward analysis of partly other regions of visual attention.

REFERENCES

[1] C. Liu, Recent Advances in Intelligent Image Search and Video Retrieval. Intelligent Systems Reference Library, vol. 121, Cham: Springer, 2017.

[2] G. Csurka, Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition, Cham: Springer, 2017.

[3] B. T. Truong, and S. Venkatesh, “Video abstraction: A systematic review and classification,” ACM Trans. on Multimedia Computing, Communications, and Applications, vol. 3, iss. 1, pp. 1–37, 2007.

[4] S. Mashtalir, and O. Mihnova, “Key frame extraction from video: framework and advances,” Int. J. of Computer Vision and Image Processing, vol. 4, iss. 2, pp. 68-79, 2014.

[5] T. Wiatowski, and H. Bölcskei, “A mathematical theory of Deep Convolutional Neural Networks for feature extraction.” IEEE Trans. on Information Theory, vol. 64, iss. 3, pp. 1845–1866, 2018.

[6] F. Shih, Image Processing and Pattern Recognition: Fundamentals and Techniques, Hoboken: John Wiley & Sons, Inc., 2010.

[7] L. Elazary, and L. Itti, “Interesting objects are visually salient,” J. of Vision, vol. 8, iss.3, pp. 1–15, 2008.

[8] D. Liu, G. Hua, and T. Chen “A hierarchical visual model for video object summarization,” IEEE Trans. on PAMI vol. 32, iss. 12, pp. 2178–2190, 2010.

[9] Ye. 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, iss. 1, pp. 1-20, 2012.

[10] C.C. Aggarwal, Data Mining. Cham: Springer, 2015.

[11] G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Probability. Philadelphia : SIAM, VA, 2007.

[12] Z. Hu, S. V. Mashtalir, O. K. Tyshchenko, and M. I. Stolbovyi “Video shots’ matching via various length of multidimensional time sequences,” Int. J. of Intelligent Systems and Applications (IJISA), vol. 9, iss. 11, pp.10–162, 017.

[13] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, “Experimental comparison of representation methods and distance measures for time series data,” Data Mining and Knowledge Discovery, vol. 26, iss. 2, pp. 275–309, 2012.

[14] S. Mashtalir, and V. Mashtalir, “Sequential temporal video segmentation via spatial image partitions,” IEEE First Int. Conf. on Data Stream Mining and Processing (DSMP’2016), Lviv, Ukrane, pp. 239-242, 2016.