A new method using moments correlation for action change detection in videos

Imen Lassoued, Ezzeddine Zagrouba, Youssef Chahir

A new method using moments correlation for action change detection in videos

Imen Lassoued, Ezzeddine Zagrouba
Team of research SIIVA, RIADI laboratory
High Institute of computer science (ISI),
University of Tunis El Manar
Ariana, Tunis
lassoued.imen@yahoo.fr,
ezzeddine.zagrouba@fsm.rnu.tn

Youssef Chahir
Team of research IMAGE, GREYC laboratory
University of Caen basse Normandie,
Caen, France
chahir@unicaen.fr

Abstract— **Automated characterization of human actions plays an important role in video indexing and retrieval for many applications. Action change detection is considered among the most necessary element to ensure a good video description. However, it is quite challenging to achieve detection without prior knowledge or training. Usually humans are practicing different actions in the same video and their silhouettes give significant information for characterizing human poses in each video frame. We have developed an approach based on pose descriptors of these silhouettes, cross correlations matrices and Kullback-Leibler distance to detect action changes. In this paper, we will focus firstly on the specific problem of change detection in videos. After that, the proposed approach for action change detection will be detailed and tested on Weizman dataset. Finally, experimental results has been analyzed and showed the good performance of our approach.**

Keywords- Action change, correlation matrix, Pose representation

I. INTRODUCTION

Action change detection is a key parameter in video identification and recognition domains. Currently, most works on the temporal processing of videos separate sequences into shots, i.e. groups of frames filmed from the same camera or viewpoint. This segmentation in shots is not efficient and not suitable for video action recognition and classification domain because action can be practiced in one or more shots. Furthermore, a shot is a technical unit that is often characterized by a short duration and does not really take into account the progress of the action. In this paper, we have developed a new method for action change detection based on a correlation matrix between descriptors of poses of silhouette and the global video. To describe silhouette poses we have used 2D Krawtchouk moments. We have used also a spatio-temporal Krawtchouk moments proposed in our previous work [9] to describe a global video.

The remainder of this paper is organized as follows. Section II will present an overview of existing methods for change detection in videos. In the third section, the different steps of the proposed method based on correlation matrix and poses representation will be described. Section IV will be dedicated to experimental results and analysis. In the last section, conclusion and perspectives will be given.

II. STATE OF THE ART

The technology for organizing and searching images and videos based on their content is still an open research domain. This is especially true in multimedia applications where the difficulty of searching and editing data is often the largest cost factor. Detecting scene, shot or action changes in videos is the first step of extracting content-based information from sequences. In the following, we present some works that deal with action, scene and shot change detection problems.

In dynamic scene analysis, motion detection is often tied to change. Literatures present several approaches for scene segmentation problem. Proposed approach in [1] consists to transform scene segmentation into a graph partitioning problem. A shot similarity graph is constructed, where each node represents a shot and edges between shots depict their similarity based on color and motion information. Authors in [2], propose a detecting boundary method based on logical story units by linking similar shots and connecting overlapping links. Key frames for different shots are merged into a larger image and the similarity between shots is computed by comparing these shot images. They use also, approach proposed in [3] and a transition graph which is divided into connected subgraphs representing the scenes. Authors in [4] propose a method that uses Markov chain Monte Carlo to determine scene boundaries.

Action change detection is the input for many works such as videos actions classification, indexation and retrievals. Literatures show that scene and shots are usually used for video segmentation. Actions in videos are becoming as a mean for video segmentation. Guo in [5] proposes a method based on silhouettes framework for action representation and comparison. He uses a non-parametric statistical framework to learn the distribution of the distance between covariance descriptors.

III.PROPOSED APPROACH

The developed approach is composed of several steps represented by Fig.1. First, human silhouettes are extracted in

the space-time volumes. Then, we compute poses descriptors and a global video descriptor. The next step consists to calculate a Kullback-Leibler distances between cross correlation matrices of previous descriptors. Finally, we applied a learning algorithm to detect different action change frames.

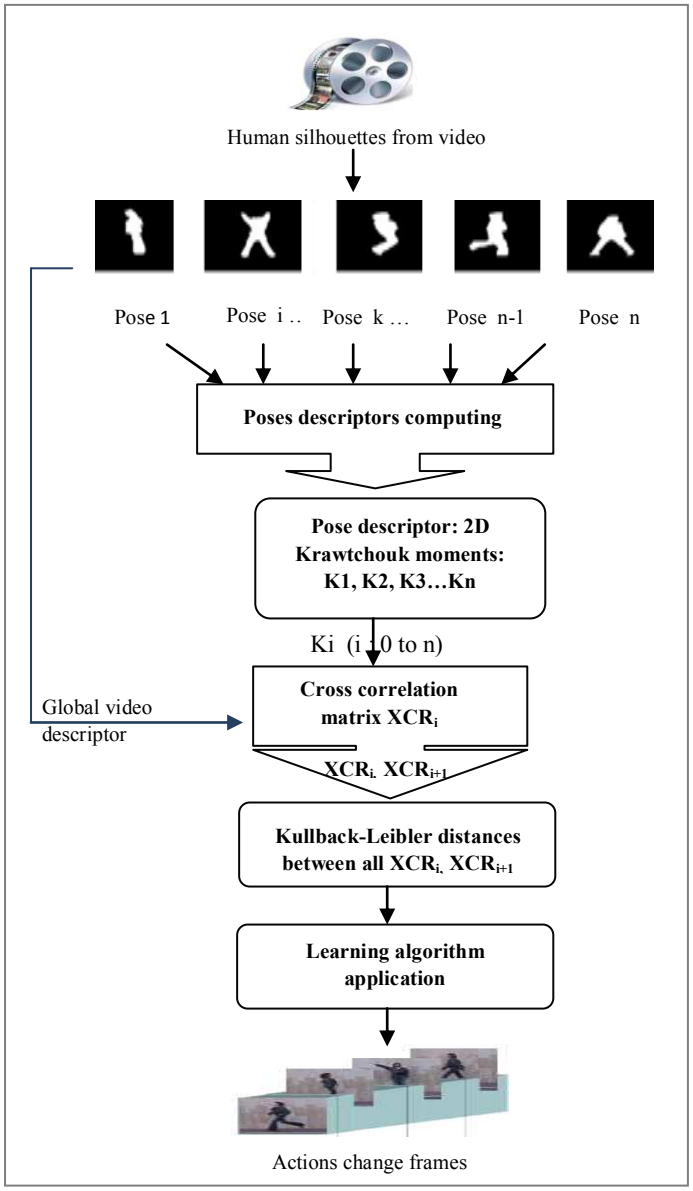

Figure 1. General architecture for the proposed approach

A. Pose representation and description in space time volume

Human action in a video is composed by a set of poses in different frames. Each action is made by the repetition of a series of poses represented by silhouettes. We choose 2D Krawtchouk moments to describe a silhouette pose in each frame. In fact, Krawtchouk moments are considered as a good form descriptor [6, 7] and give good results in object recognition domain. We define the n-th order of Krawtchouk polynomial in the variable x as described in [6].

$$K_n(x; p_x N_x) = \sum_{j=0}^{n} (-1)^j \, (p_x - 1)^{n-j} \binom{N_x - x}{n-j} \binom{x}{j} \tag{1}$$

Where $0 \le n \le N_x$ and $p_x \in (0, 1)$

$$\binom{x}{j} = \begin{cases} \frac{x(x-1)..(x-j+1)}{j!} & if\ j \ge 1 \\ 1 & if\ j = 0 \end{cases} \tag{2}$$

The set $S = \{K_n(x; p_x, N_x)\}$ has (N_x+1) Krawtchouk polynomials witch satisfy a discrete orthogonality relation of the form

$$\sum_{z=0}^{N} w(x; p_z, N_z) \, K_i(x; p_z, N_z) K_j(x; p_z, N_z) = \varphi(i, j) \tag{3}$$

Where

$$\varphi(i, j) = h(i, p_x, N_z)\delta_{ij} \tag{4}$$

Where i, j = 1 ...Nx and w (x; p_x, N_x) is a weight function in x and h is a function depending on i:

$$w(x; p_z, N_z) = \binom{N_x}{x} p_z (1 - p_z)^{N_p - z} \tag{5}$$

δij is the Kronecker delta function and

$$h(i; p_x, N_z) = \frac{1}{\binom{N_z}{i}} \left(\frac{1-p_z}{p_z}\right)^i \tag{6}$$

The conventional method of avoiding numerical fluctuations for moment computations is by means of normalization by the norm. The normalized Krawtchouk polynomials with respect to the norm $K_n(x; p, N)$ is defined as:

$$\widetilde{K_n}(x; p, N) = \frac{K_n(x; p, N)}{\sqrt{\rho(n; p, N)}} \tag{7}$$

The set of weighted Krawtchouk polynomials $K_n(x; p, N)$ is defined by:

$$\widetilde{K_n}(x; p, N) = \sqrt{\frac{w(x; p_z, N_z)}{h(n, p_z, N_z)}} K_n(x; p_z, N_z) \tag{8}$$

The orthogonality condition becomes

$$\delta_{nm} = \sum_{x=0}^{N} \widetilde{K_n}(x; p, N) \widetilde{K_m}(x; p, N) \tag{9}$$

Krawtchouk moments have the interesting property of being able to extract local features of an image. The Krawtchouk moments of order (m+n) in terms of weighted Krawtchouk polynomials, for an image with intensity function, f(x,y) is defined as:

$$Q_{nm} = \sum_{x=0}^{N}\sum_{y=0}^{M} \widetilde{K_n}(x;p_1,N)\ \widetilde{K_n}(y;p_2,M)\, f(x,y) \tag{10}$$

The parameters N and M are substituted with N-1 and M-1 respectively to match the NxM pixel points of an image.

B. Pose correlation with a global video

We use a cross correlation 'XCR' to study the correlation for each pose relative to global video sequences. To compute cross correlation, we apply the 2D Krawtchouk moment descriptor for each pose. Spatio-temporal Krawtchouk descriptor [8] are used to describe global silhouettes volume of video.

Cross correlation is a standard method of estimating the correlation degree between each silhouette pose and a global volume of video. Consider two series x(i) and y(i) where x(i) is the descriptor vector for each pose and y(i) is the descriptor vector of the global silhouettes volume. {i=0,1,2...N-1, N is the descriptor vector length. The cross correlation XCR at delay d is defined as [11]

$$XCR = \frac{\sum_i [(x(i)-mx)*(y(i-d)-my)]}{\sqrt{\sum_i (x(i)-mx)^2}\sqrt{\sum_i (y(i-d)-my)^2}} \tag{11}$$

Cross correlation is a measure of similarity between two different non-identical signals. They detect the presence of one signal in another signal. In our case, XCR gives an idea on the correlation of different poses with the global video. If the same signal is buried in both signals, it will be reinforced in the cross correlation function, whereas the noise which is uncorrelated will be reduced. We use different computing XCR to characterize each silhouette pose in video sequence.

C. Silhouette similarity metric

The metric Kullback-Leibler has been chosen to measure distance between pose descriptors in each frame. Kullback-Leibler distance is very good in quantifying the amount of information present in a sample correlation matrix with respect to an hypothetical reference model. Kullback-Leibler distance has been used previously in probability domain and Tumminello and al [9] compute its analytical form. They show that it gives very good results when applied for measurement between correlation matrices. We define the Kullback-Leibler metric as described in [9].

$$\begin{aligned} l(\mathrm{XCR}_1||\mathrm{XCR}_2) &= \sum_{i=1}^{n} \mathrm{XCR}_1(i)\log\left(\frac{\mathrm{XCR}_1(i)}{\mathrm{XCR}_2(i)}\right) \\ &= \sum_{i=1}^{n} [-\mathrm{XCR}_1(i) log\ (\mathrm{XCR}_2(i)) \\ &\quad - (-\mathrm{XCR}_1(\mathrm{i})\log(\mathrm{XCR}_1(i))] \end{aligned} \tag{11}$$

Where n is lines number in cross correlation matrix.

It is noted that the Kullback-Leibler distance takes into account the statistical nature of correlation matrices. Indeed $l(XCR_1||\ XCR_2)$ is well defined only when matrices XCR_1 and XCR_2 are positive definite. This property is not common to other measures of distance between matrices which are based generally on isomorphism between the matrix space and a vector space.

D. Action change detection algorithm

The developed approach for action change detection is based on the spatio-temporal volumes of human silhouettes. In this work, we have supposed that actions persist for some time and they are not changing rapidly and suddenly. The first step in the process of action change detection is to form a first learning set to detect the first action. This set contains the "M" first frames in video. In the second step, we compute pose descriptors presented previously for each silhouette in the video. A descriptor for the global sequence is also computed. After that, we calculate correlation matrices between the global video descriptor and each silhouette pose descriptor. The Kullback-Leibler distances 'D' between these correlation matrices are determinate. The Next step consists of computing the derivative in time of the matrix distance 'D' to appear peaks changes. The Last step is to apply training algorithm from the frame number (M+1). The next learning set begins just after action change detection frame and contains 'M' frames.

Below, we present the training algorithm for detecting frames of action change.

Change_gpe : change detection frames

Input: d(N) derivative distances between different correlation matrix

For each frame 'j' from M+1 to N

K= learning set: first M frames after detecting change

If(d j> max (K) + epsilon)
Change_gpe+ = dj
Endif

K= [dj+1….dj+M]

III. Experimental results

A. Weizmann Dataset

The action change process was evaluated on a publicly available benchmark dataset of human action: "Weizmann dataset" (Fig.2). This dataset composed by 90 low resolution videos (180 * 144, 50 fps) where 9 persons performs 10 actions (bend, jack, jump, jump in place, run, side, skip, walk, wave1hand et wave2hand) .

B. Experimental process

In order to test the performance of the developed approach, we concatenate a series of videos with a time continuous camera.

Illustrative example of concatenated videos is shown in Figure 2. We use the subtraction method described in [10] to obtain moving object silhouettes (silhouette tunnels). Then, we aligned the centroid of individual 2D silhouettes to compensate global object displacements. Precision of detected action boundaries depends on the length of action segments and pose descriptor precision. Clearly, a good descriptor of silhouette poses and a large segment length reduce the uncertainty of action boundary location. We applied our action change detection algorithm using 30 frames for a training set. A careful choice of training set frames number is essential for the performance of our algorithm.

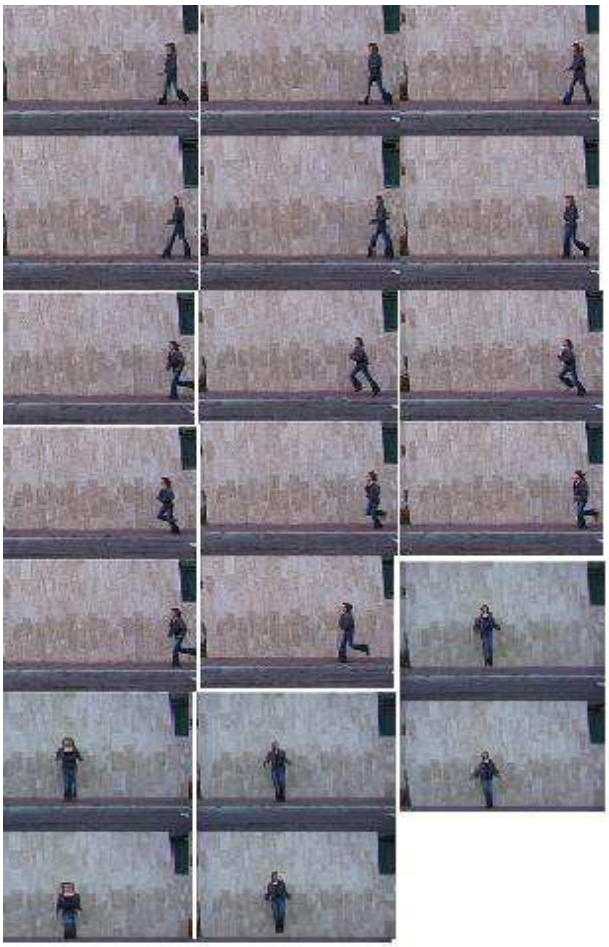

Figure 2. Example of sequences containing human practicing three successive actions: Walk, Run and jack

We created 9 video sequences test containing single-person practicing several actions. We concatenate all video sequences where the same person is practicing different actions. We have taken the same measures used in [5] in order to do comparison with our approach.

Two measures has been used for action detection errors:

- False negative: Frame does not contain action change will be considered as a frame contain action change
- False positive: Frame containing action change will be considered as a frame that does not contain it.

Figure 3 shows a curve describing the derivative of the distance matrix 'D' vs frames index for "Daria Weizmann video action". The derivative allows to appear peaks change in videos.

After applying proposed approach to the nine videos action tests, we observe that, our approach has detected all action changes frames in the video, Obtained false positive rate is null. But in some videos, frames do not contain action change they were considered as a frame containing action change. For example the "Daria video" test contain two false negative frames (see figure 3).

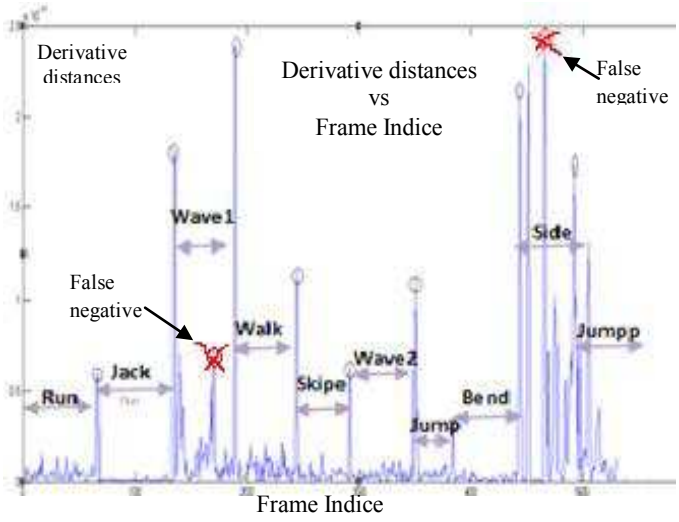

Figure 3. Derivatives distances between poses descriptors vs frames of videos

VI. Conclusion

In this paper, a new method for action change detection in video has been proposed. This method is based on cross correlation measures between silhouette poses and global video descriptor. We describe each pose using 2D Krawtchouk moments. We develop a training algorithm for detecting frames of action change. The approach was tested with a video sequences containing single-person practicing several actions created from the Weizmann dataset. Experimental results show that our approach detect all action changes in most tested videos.

References

[1] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," IEEE Trans. Multimedia, vol. 7, no. 6, pp. 1097–1105, Dec. 2005.

[2] A. Hanjalic, R. L. Lagendijk, and J. Biemond, “Automated high-level movie segmentation for advanced video-retrieval systems,” IEEE Trans. Circuits Syst. Video Technol., vol. 9, no. 4, pp. 580–588, June 1999.

[3] M. Yeung, B. Yeo, and B. Liu, “Segmentation of videos by clustering and graph analysis,” Comput. Vis. Image Understand., vol. 71, no. 1, pp. 94–109, July 1998

[4] Y. Zhai and M. Shah, “Video scene segmentation using Markov chain Monte Carlo,” IEEE Trans. Multimedia, vol. 8, no. 4, pp. 686–697, 2006.

[5] K. Guo , P.Ishwar, J. Konrad, « Action change detection in video by covariance matching of silhouette tunnels”, Conference on Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International , pp 1110 – 1113, 2010

[6] P. T. Yap, R. Paramesran and S. H. Ong, Image Analysis by Krawtchouk Moments,. IEEE Transactions on Image Processing, Vol. 12, No. 11, pp. 13671376, November 2003

[7] W. Gautschi, ”Orthogonal Polynomials: Computation and Approximation”, Journal of Computational and Applied Mathematics, vol. 178 pp.215234, 2005.

[8] I. Lassoued, E. Zagrouba and Y. Chahir, « Video Action Classification: A New Approach combining Spatio-temporal Krawtchouk Moments and Laplacian Eigenmaps », 7th International Conference on Signal Image Technology & Internet-Based Systems, Dijon, FRANCE, pp. 291-297, November, 2011.

[9] M. Tumminello, F. Lillo, R. Mantegna, « Kullback-Leibler distance as a measure of the information filtered from multivariate data », *Physical Review E* 2007, 76:256-67.

[10] K.Kim, T.Chalidabhongse, D.Harwood, and L.Davis, ”Real-time Foreground-Background Segmentation using CodebookModel”, Real-time Imaging, vol 11,pp 167-256, 2005

[11] P.Bourke, “Cross correlation,” http://astronomy. swin.edu.au /˜pbourke/ analysis/correlate, August 1996.