MDS-Based Multi-Axial Dimensionality Reduction Model For Human Action Recognition

Redha Touati and Max Mignotte
Département d'Informatique et de Recherche Opérationnelle (DIRO)
Université de Montréal, Faculté des Arts et des Sciences,
Montréal H3C 3J7 QC, Canada
Email: touatire@iro.umontreal.ca

***Abstract*—In this paper, we present an original and efficient method of human action recognition in a video sequence. The proposed model is based on the generation and fusion of a set of prototypes generated from different view-points of the data cube of the video sequence. More precisely, each prototype is generated by using a multidimensional scaling (MDS) based nonlinear dimensionality reduction technique both along the temporal axis but also along the spatial axis (row and column) of the binary video sequence of 2D silhouettes. This strategy aims at modeling each human action in a low dimensional space, as a trajectory of points or a specific curve, for each viewpoint of the video cube in a complementary way. A simple K-NN classifier is then used to classify the prototype, for a given viewpoint, associated with each action to be recognized and then the fusion of the classification results for each viewpoint allow us to significantly improve the recognition rate performance. The experiments of our approach have been conducted on the publicly available Weizmann data-set and show the sensitivity of the proposed recognition system to each individual viewpoint and the efficiency of our multi-viewpoint based fusion approach compared to the best existing state-of-the- art human action recognition methods recently proposed in the literature.**

***Keywords*-Human action recognition; Multi-axial reduction of dimensionality; Gesture recognition; Multidimensional scaling; FastMap; Weizmann data-set; K-Nearest Neighbor.**

I. Introduction

The proliferation of video content on the web or in everyday life makes human action recognition, one of the key prerequisites for video analysis and understanding with many important computer vision applications, such as video surveillance, indexing and browsing, human-computer interfacing, recognition of gesture and analysis of sport events, etc. [1], [2], [3], [4], [5], [6].

The goal of any unsupervised human recognition system is to be able to automatically recognize low-level actions such as running, walking, hand clapping, etc. from an input video sequence and the main difficulty of this human motion categorization [7] lies in representing the different types of human motion with effective models both taking into account the intra-class variations in appearance and size of different individuals, and between-class variations, *i.e.,* in different action types with similar body shapes.

Various approaches for human action recognition have been already proposed in the literature, and a way to classify them into several categories may be considered depending on the type of (*e.g.*, static, dynamic or spatial-temporal) features extracted from the spatial temporal information of the video sequence and intended to model the human action *via* the local or global description of the (spatial) human body information and its (temporal) motion information [1].

Some approaches rely on local features to represent the motion patterns and to capture local events in video. In this way, Schuldt *et al.* [3] have used the spatial Harris 3D detector. Building on the success of the histogram of gradient (HOG) based descriptor for static images, an extension of the SIFT descriptor to 3D was proposed in [9] as a new local spatial-temporal descriptor for video sequences, which was also further generalized in [7] for a quantization without singularities based on regular polyhedrons. Jhuang *et al.* [4] model each action class with a multilayer model based on the set of spatial-temporal features extracted by the Gabor filters. Niebles *et al.* [8] have proposed a hierarchical model that can be characterized as a constellation of bags-of-features and that is able to combine both spatial and spatial-temporal features.

Mid-level motion features constructed from low-level optical flow features can be also used as in [10].

Global temporal approaches rely on global features computed on the whole time span of the action. In this context, some authors have proposed to regard each human action as 3D shape induced by the set of spatial silhouettes and propose to extract a set of local and global spatial-temporal features from this space-time shape with the generalization of the Poisson equation [2], [11] or with a (3D) distance transform [12]. Tseng *et al.* [5] have suggested to construct, in a reduced dimensional space, a spatial and temporal action graph which connects the different (dynamic) shape variation of human silhouettes of a same human action. Saad *et al.* [1] have proposed to use a set of spatial-temporal kinematic features that intend to capture the representative dynamics of the optical flow of the video sequence in the form of its dominant kinematic modes. Each video is then embedded into a kinematic-mode-based feature space and

978-1-4799-4337-1/14 $31.00 © 2014 IEEE
DOI 10.1109/CRV.2014.42

the coordinates of the video in that space are then used for classification. Bobik and Davis [6] have exploited a temporal image template for stored instances of views of known actions where the value at each point is a function of the motion properties at the corresponding spatial location in an image sequence.

The proposed method first relies on the exploitation of a reliable, compact and discriminative representation of the data cube containing the sequence of binarized silhouettes. To this end, a set of (at most three) prototypes is thus generated by using an MDS-based dimensionality reduction technique with respect first to the time axis but also through the spatial (row and column) axis of the binary video sequence of 2D silhouettes. This strategy aims at modeling each human action in a low dimensional space, as an ordered set (or trajectory) of a few points (or a specific curve), for each viewpoint of the video cube in a very complementary way and thus with a minimal loss of reliable information. A K-nearest neighbor classifier will be used to classify an action on each view point and a simple and intuitive fusion technique which allows us to achieve a recognition rate performance close to the best state-of-the-art methods.

II. Method

A. Description

Our method is based on three stages:

- Preprocessing
- Prototype extraction
- Classification and Fusion

B. Preprocessing

The first step of this preprocessing consists in obtaining the binary video sequence of 2D silhouettes (indicating only the body position) for each human action. To this end, we have subtracted the median background from each image of the sequence and have then used a simple thresholding technique. Once the body silhouette extraction is achieved, an additional step of filtering by a classical 3×3 median filter is then used to remove some misclassified pixels inside and outside the binarized body silhouettes. The last step consists in centering the gravity center of each silhouette inside a rectangular fixed size bounding box ($N_l \times N_w$) with a translation vector (cf. Fig. 1). Finally we consider only the first $N_i = 13$ frames of each sequence (with a step size variable according the class), which corresponds approximately, for the Weizmann data-set, to the number of frames typically occurring during a periodic cycle of human action.

C. Prototype Extraction

This stage consists in building a set of (at most three) prototypes or, more precisely, a set of reliable, compact and discriminative representations of the data cube containing the sequence of binarized silhouettes. To attain this goal, this set of prototypes are herein generated by an MDS[1]-based non linear dimensionality reduction technique [14] from different view-points of the video cube containing the sequence of binarized silhouettes, namely

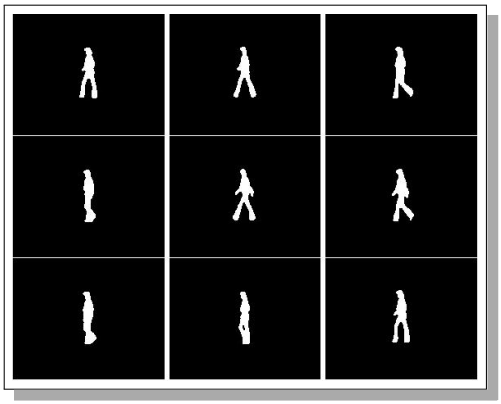

Figure 1: Example of video sequence from the Weizmann data-set after the preprocessing step to extract and center the body silhouettes.

- The first viewpoint aims at reducing the dimensionality of the image cube along the temporal axis of the video sequence. To this end, every N_i silhouette image frames in the $N_l \times N_w$ dimensions is converted into a N-dimensional ($N = N_l \times N_w$) vector in a raster scan manner and reduced to 3 dimensions[2] by the FastMap technique [15] (which is a fast alternative to the MDS algorithm with a linear complexity). This strategy aims at modeling each human action in a low 3D dimensional space, as an ordered set of N_i points or a specific curve (possibly periodic if N_i is greater, in terms of number of images that the considered human action cycle, cf. Fig. 3).
- The second viewpoint aims at reducing the dimensionality of the image cube through the spatial (line or column) axis of the set of 2D binarized silhouettes (cf. Fig. 2). For example, if we consider the axis of lines,

[1]MDS is a dimensionality reduction technique that maps objects lying in an original high N dimensional space to a lower dimensional space, but does so in an attempt that the between-object distances are preserved as well as possible. The original MDS algorithm is not appropriate for large scale applications because it requires an entire $N \times N$ distance matrix to be stored in memory (with a $O(N^3)$ complexity). The FastMap [15] is a fast alternative to the MDS algorithm with a linear complexity ($O(pN)$) with p, the dimensionality of the target space) and an algebraic procedure that determines one coordinate at a time by examining a constant number of rows of the distance matrix. In FastMap, the axis of target space are then constructed dimension by dimension. More precisely, it implicitly assumes that the objects are points in a p-dimensional Euclidean space and selects a sequence of $p \leq N$ orthogonal axes defined by distant pairs of points (called pivots) and computes the projection of the points onto the orthogonal axes.

this amounts to considering the set of N_l images which are laterally created in this data cube. These images are then converted into a $N_i \times N_w$-dimensional vector in a raster scan manner and reduced to 3 dimensions by the FastMap, thus generating, for this viewpoint, a trajectory of N_l ordered points in a 3D dimensional space[2].

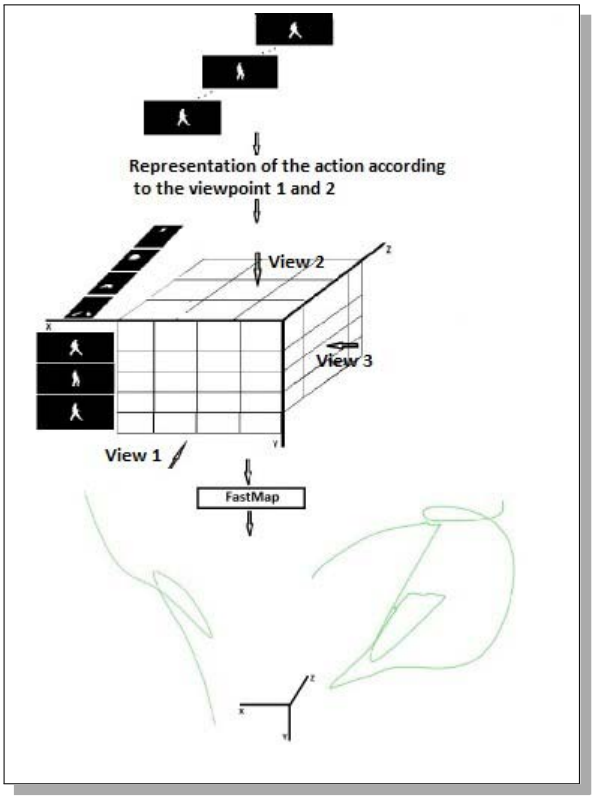

Figure 2: Set of two prototypes or 3D curves generated by a MDS-based dimensionality reduction according two viewpoints or through two axis (namely; the temporal axis and the line axis) on a sequence of binarized silhouettes.

This multi-axial non-linear dimensionality reduction strategy has several advantages. It allows us to obtain a set of compact representations, retaining the most significant information in each human action (by removing redundancy in the data) in two different and complementary ways (with a minimal loss of reliable information) while preventing, to some extend, the classification model from over-fitting in the training phase. In our application, this will allows us to generate a compact and discriminative (set of) prototype(s) which will be similar and consistent between the same action of two different persons and is also somewhat invariant with respect to variations in the speed of the actions.

[2]In our application, experiments have shown that the use of a higher dimensional space (greater than 3) does not allow us to provide a better representation and generalization (without over-fitting) of the human action.

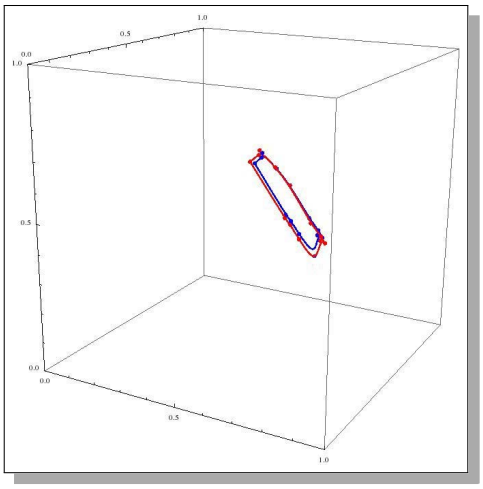

Figure 3: Periodic prototype or 3D curve Modeling two human action cycles of the SIDE class (normalized between $[0, 1]$ for the visualization).

We can evaluate the efficiency of the FastMap technique in its ability to reduce the dimensionality reduction of a sequence of binarized silhouettes in two different and complementary ways when this is achieved according to different axis. To this end, we can easily compute the correlation metric [13] which is simply the correlation of the Euclidean distance between each pairwise vectors in the high dimensional space (let X be this vector) and their corresponding (pairwise) Euclidean distances in the low (3D) dimensional space (let Y be this vector). The correlation ρ can be estimated by the following equation:

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \, \sigma_Y} = \frac{X^t Y / |X| - \bar{X}\bar{Y}}{\sigma_X \, \sigma_Y} \quad (1)$$

where X^t, $|X|$, $\bar{X}$ and σ_X respectively represent the transpose, cardinality, mean, and standard deviation of X. This correlation factor (Pearson) will specifically quantify the degree of linear dependence between the variables X and Y and quantify how the FastMap technique is able to give a low dimensional mapping in which each object is placed such that the between-object distances (in the original high dimensional space) are preserved as well as possible [14]. A perfect correlation $\rho = 1$ indicates a perfect relationship between original data and reduced data and a correlation of $\rho = 0$ indicates a total loss of information. The following table shows the mean correlation coefficient obtained on the Weizmann data-set for each viewpoint:

Table I shows us first that the MDS procedure is able to preserve a large quantity of structural information of the original data set and that the main efficient way to reduce the dimensionality of the information contained in the video cube consists in doing this for each image of the video

Class	Viewpoint 1 (%)	Viewpoint 2 (%)
walk	83	69
run	79	68
skip	83	66
side	79	68
jack	80	68
wave1	88	64
wave2	91	63
jump	81	67
pjump	89	70
bend	80	68

Table I: Mean correlation rate in percentage for the MDS-based dimensionality reduction with the FastMap technique according to two viewpoints; namely the temporal axis (viewpoint 1) and the line axis (viewpoint 2) for each human action class.

cube commonly generated along the temporal axis direction (viewpoint 1) compared to the line-axis direction (viewpoint 2). Nevertheless, we will see, in the following section, that the second viewpoint generates a second prototype which is very complementary to the first one, in terms of classification accuracy.

D. Classification and Fusion

Let us recall that, in our application, the first and second prototype is respectively a set of N_i and N_l ordered points in a (reduced) 3D dimensional space.

For the first prototype, each i-th point ($1 \leq i \leq N_i$) of a test prototype (to be classified) is used to feed a non-parametric K-Nearest Neighbor (KNN) classifier using a 3D Euclidean distance between each i-th point (and thus trained with the set of i-th points of each prototype belonging to the training set). For the first prototype, the result of these N_i KNN classification results allows us to generate a score vector both indicating the class and the sum (over the K-nearest neighbors) of the number of nearest points of the prototype (of the training set) with the most nearest points.

Our fusion procedure then consists simply in adding the two score vectors generated by the first and second viewpoint and to classify a test prototype by the majority class. If there is no majority class, our recognition system will produce a classification error.

III. Experimental Results

To evaluate the efficiency of our human action recognition system we validate our approach on the famous Weizmann data-set [11]. This data-set contains 10 action classes performed by 9 different human subjects. The actions include bending (bend), jumping jack (jack), jumping-forward-on-two-legs (jump), jumping-in-place-on-two-legs (pjump), running (run), galloping-sideways (side), skipping (skip), walking (walk), waving-one-hand (wave1) and waving-two-hands (wave2). There are totally 93 video sequence (180×144, 25 fps) since some types of actions are performed twice by some individuals. In order to validate our procedure, we replicate the scenario proposed in [8], [9], [7], [5], [10], [12], [11]. More precisely, for every video sequence, we perform a leave-one-out procedure, *i.e.,* we remove the entire sequence from the database while other actions of the same individual remain. Each video cube of the removed sequence is then compared to all the remaining video cube examples in the database and is classified, in our application, with our KNN-based fusion procedure.

The confusion matrix for each viewpoint is shown in Tables II and III and illustrate the different classification results obtained with $K = 1$ for each class. The confusion matrix given by our fusion procedure combining these two viewpoints is shown in Table IV (in our application, $K = 1$ allows us to obtain the best classification accuracy). Finally, the Table V shows us the recognition rate obtained respectively for each viewpoint and for the fusion procedure combining these two viewpoints. In our application, the third viewpoint according to the column axis does not allow us to improve the recognition rate.

	walk	run	skip	jack	jump	pjump	side	wave1	wave2	bend
walk	0.89		0.11							
run	0.11	0.67	0.22							
skip	0.11	0.22	0.45	0.11	0.11					
jack				0.67					0.33	
jmup					0.78			0.11	0.11	
pjump				0.11		0.78		0.11		
side							0.89	0.11		
wave1								0.89	0.11	
wave2								0.33	0.67	
bend									0.11	0.89

Table II: Confusion matrix associated to viewpoint 1.

	walk	run	skip	jack	jump	pjump	side	wave1	wave2	bend
walk	0.78	0.11					0.11			
run		0.89	0.11							
skip		0.22	0.67		0.11					
jack		0.11		0.45	0.11			0.11	0.22	
jump			0.33		0.34	0.22	0.11			
pjump			0.11		0.11	0.78				
side	0.11						0.89			
wave1						0.22		0.66	0.22	
wave2			0.11					0.11	0.78	
bend									0.11	0.89

Table III: Confusion matrix associated to viewpoint 2.

	walk	run	skip	jack	jump	pjump	side	wave1	wave2	bend
walk	1.0									
run		1.0								
skip		0.11	0.89							
jack				0.78					0.22	
jump					0.78			0.11	0.11	
pjump			0.11			0.89				
side							1.0			
wave1								1.0		
wave2								0.11	0.89	
bend										1.0

Table IV: Confusion matrix associated to the fusion of the two viewpoints.

	View point 1	View point 2	Fusion
Number of K-NN	K=1	K=1	K=1
Recognition rate	75.8	71.3	**92.3**

Table V: Table show the recognition rate of each view point and the fusion.

IV. Discussion

It can be observed (see V) that we can recognize all types of actions existing in the data-set with very good performance results on most of the actions except some of them (e.g., JACK and JUMP). We can also notice that some misclassified action classes, given by our system, are generally and very logically in the action classes which are physically the closest to the test class; this is the case for instance of the class WAVE1 or WAVE2 which are mechanically and visually similar to the action class JUMP or class SKIP and RUN. These action classes are very similar between them in the way the subjects move and bounce across the video sequence. We can also note that the classification results obtained by each individual viewpoint seem complementary. We have compared the accuracy of our approach with other state-of-the-art (recently published) methods using the *leave-one-out* procedure and the Weizmann data-set [8], [9], [7], [5], [10], [12], [11] in the Table VI.

Method	Accuracy
Our method	**92.3%**
Fathi *et al.*	99.9%
Gorelick *et al.*	97.8%
Grundmann *et al.*	94.6%
Jia *et al.*	90.9%
Klaser *et al.*	84.3%
Scovanner *et al.*	82.6%
Niebles *et al.*	72.8%

Table VI: Comparison with other state-of-the-art methods [8], [9], [7], [5], [10], [12], [11].

V. Conclusion

In this paper, we have presented an original and simple human action recognition system based on a set of (two) compact and discriminative prototype models which is similar and consistent between the same action of two different persons. In our application, these two prototype models are generated from an MDS-based multi-axial non-linear dimensionality reduction strategy which has several advantages. It allows us to retain the most significant information in each human action in two different and complementary ways and also give a better representation of the action in low dimension, while preventing, to some extend, the prototype model-based classification scheme from over-fitting in the training phase. In addition of its simplicity (and speed), our method method does not exploit a representation or spatio-temporal characteristics of action, and is also somewhat invariant with respect to variations in the speed of the actions. This set of two prototypes contains rich and descriptive information about the action performed and this is clearly demonstrated by the success of the relatively simple classification scheme used in our application (KNN classification and Euclidean distance). Experimental results demonstrate that our method can accurately recognize human actions and outperforms some, more complex, state-of-the-art recognition methods on a publicly available action data-set. Finally, it is also worth mentioning that our recognition performance can also be easily improved by using a more sophisticated and powerful classification strategy such as the SVM classifier or a deep neural network.

References

[1] S. Ali and M. Shah, *Human action recognition in videos using kinematic features and multiple instance learning*, IEEE Trans. Pattern Anal. Mach. Intell., 32(2):288-303, 2010

[2] M. Blank and L. Gorelick and E. Shechtman and M. Irani and R. Basri, *Actions as space-time shapes*, Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV'05, pp. 1395–1402, 2005.

[3] C. Schuldt and I. Laptev and B. Caputo, *Recognizing human actions: A local SVM approach*, Proceedings of the 17th International Conference on Pattern Recognition, IEEE Computer Society, ICPR'04 Vol.3, pp. 32–36, 2004

[4] H. Jhuang and T. Serre and L. Wolf and T. Poggio, *A biologically inspired system for action recognition*, IEEE Proceedings of the 11th International Conference on Computer Vision, ICCV'07 pp. 1–8

[5] K. Jia and D.Y. Yeung, *Human action recognition using local spatio-temporal discriminant embedding*, Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, CVPR'08

[6] A.F. Bobick and J.W. Davis, *The recognition of human movement using temporal templates*, IEEE Trans. Pattern Anal. Mach. Intell., 23(3):257–267, 2001

[7] A. Klaser and M. Marszalek and C. Schmid, *A spatio-temporal descriptor based on 3D-gradients*, Proceedings of the 19th British Machine Vision Conference, BMVC'08, pp. 1–10, 2008.

[8] J.C. Niebles and L. Fei-Fei, *A hierarchical model of shape and appearance for human action classification*, Proceedings of IEEE International Conference in Computer Vision and Pattern Recognition, CVPR'07

[9] P. Scovanner and S. Ali and M. Shah, *A 3-dimensional SIFT descriptor and its application to action recognition*, Proceedings of the 15th international conference on Multimedia, MM'07 pp. 357–360, 2007.

[10] A. Fathi and G. Mori, *Action recognition by learning mid-level motion features*, Proceedings of IEEE International Conference in Computer Vision and Pattern Recognition CVPR'08

[11] L. Gorelick and M. Blank and E. Shechtman and M. Irani and R. Basri, *Actions as space-time shapes*, IEEE Trans. on Pattern Anal. Mach. Intell., 2007 29(12):2247–2253, 2007

[12] M. Grundmann and F. Meier and I. Essa, *3D shape context and distance transform for action recognition*, Proceedings of IEEE International Conference in Pattern Recognition, ICPR'08 pp. 1–4, 2008

[13] P. Jacobson and M.R. Gupta, *Design goals and solutions for display of hyperspectral images*, IEEE Trans. Geosci. Remote Sens., 2005

[14] F.T. Cox and M.A.A. Cox, *Multidimensional Scaling*, Chapman and Hall/CRC, 2000

[15] C. Faloutsos and K. Lin, *FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets*, Proceedings of 1995 ACM SIGMOD, pp. 163-174, 1995