An Efficient Approach for Video Action Classification Based on 3D Zernike Moments

I. Lassoued[1], E. Zagrouba[1], and Y. Chahir[2]

[1] SIIVA Unit of Research, Institut Supérieur d'Informatique
Ariana, Tunis
[2] Laboratoire GREYC, Université de Caen
14050 Caen Cedex, France
lassoued.imen@yahoo.fr, ezzeddine.zagrouba@fsm.rnu.tn,
youssef.chahir@info.unicaen.fr

Abstract. Action recognition in video and still image is one of the most challenging research topics in pattern recognition and computer vision. This paper proposes a new method for video action classification based on 3D Zernike moments. These last ones aim to capturing both structural and temporal information of a time varying sequence. The originality of this approach consists to represent actions in video sequences by a three-dimension shape obtained from different silhouettes in the space-time volume. In fact, the given video is segmented in space-time volume. Then, silhouettes are extracted from obtained images of the video sequences volumes and 3D Zernike moments are computed for video, based on silhouettes volumes. Finally, least square version of SVM (LSSVM) classifier with extracted features is used to classify actions in videos. To evaluate the proposed approach, it was applied on a benchmark human action dataset. The experimentations and evaluations show efficient results in terms of action characterizations and classification. Further more, it presents several advantages such as simplicity and respect of silhouette movement progress in the video guaranteed by 3D Zernike moment.

Keywords: Actions classification, Zernike moments, LS-SVMs, 3D silhouette shape.

1 Introduction

Audiovisual contents volumes don't stop growing. In fact, the problem is how to navigate and to look for exactly these contents within large collections. So, video indexation and retrieval by automatic content analysis is now one of the major goals in information systems. In this context, it is very important to extract automatically high-level information which can describe the semantic content of the given video. Indeed, many works were interested by the events or actions in video for several applications such as: sports analysis, visual surveillance or human computer interaction. Action classification consists to classify videos based on action of the object detected in a given video. In this paper, we present a new approach of video action classification based on 3D Zernike moments. This last ones are computed using

J.J. Park, L.T. Yang, and C. Lee (Eds.): FutureTech 2011, Part II, CCIS 185, pp. 196–205, 2011.

silhouettes volumes. The remainder of this paper is organized as follows: the next section presents an overview of existing methods of action recognition and classification. In the third section, the proposed approach based on 3D Zernike moment and Least Square Support Vector Machine is detailed. The experimental results and evaluations are shown in section4. Finally, conclusion and perspectives are drawn in section5.

2 Related Works

The existing approaches in video action recognition can be classified in three main categories based on the used classification descriptor.

2.1 Optical Flow, Gradient or Intensity Based Methods

There are many existing works in action recognition which are based on global features such as optical flow, gradient histogram and intensity. Zelnik-Manor and Irani [1] used marginal histograms of spatiotemporal gradients at multiple temporal scales to cluster video events. Wu [2] develop an algorithm to automatically extract figure-centric stabilized images from the tracking system. They also propose to use Decomposed Image Gradients (DIG), which can be computed by decomposing the image gradients into four channels, to classify the person's actions. Efros et al. [3] proposed a descriptor based on blurred optical ?ow measurements, and applied it to recognize actions on ballet, tennis and football datasets. Dollar et al. [4] proposed to characterize behaviors through spatiotemporal feature points, in which a behavior was described in terms of the types and locations of feature points present. For this class of methods, the recognition results depend greatly on the recording conditions.

2.2 Feature Tracking Based Methods

Many activity recognition methods are based on feature tracking in either 2-D or 3-D space [5,6]. Rao and Shah [7] proposed an approach based on the trajectory of a tracked hand to differentiate between actions. Song et al. [21] used spatial arrangements of the tracked points to distinguish between walking and biking.Song and al [8] proposed to represent an action by 40 curves derived from the tracking results of five body parts using a cardboard people model. 3-D information is also used to establish motion descriptors based on positions, angles and velocities of body parts [9], [10]. Ali and Aggarwal [10] used the angles of inclination of the torso, the lower and upper parts of legs as features to recognize activity. Feature tracking is complex due to the large variability in the shape and articulation of the human body. In particular, perfect limb tracking is not yet well solved.

2.3 Silhouette-Based Methods

Many researches in action classification are now oriented to silhouette-based methods [11]-[12]. In fact, human actions can be characterized as motion of a sequence of human silhouettes over time. Gorelick et al[13] regard human actions as a three-dimension shape induced by the silhouettes in the space-time volume and utilizes

properties of the solution to the Poisson equation to extract space-time features such as local space-time saliency, action dynamics, shape structure and orientation. In [14] an action is viewed as a temporal sequence of local shape-deformations of centroid-centered object silhouettes, i.e., the shape of the centroid-centered object silhouette tunnel. Each action is represented by the empirical covariance matrix of a set of 13-dimensional normalized geometric feature vectors that capture the shape of the silhouette tunnel. Kellokumpu et al. [15] proposed a human activity recognition method from sequences of postures. A SVM was used for posture classification and then the discrete HMMs were used for activity recognition. Sminchisescu et al [16] recognized human motions based on discriminative conditional random field (CRF) and maximum entropy Markov models (MEMM), using image descriptors combining shape context and pairwise edge features extracted on the silhouette. For human silhouette extraction from videos, it is easy for current vision techniques, especially in the imaging setting with fixed cameras. So, the method that we present here directly relies on moving silhouettes.

3 Proposed Approach

The main goal of the proposed approach is to detect specific actions in videos. The contribution consists to use efficient 3D Zernike moments features to classify actions. In our approach, the human action in a video sequences is represented by a space-time shape in the space-time volume. These shapes are induced by a concatenation of 2D silhouettes in the space-time volume and contain both the spatial information about the pose of the human figure at any time as well as the dynamic information. The general architecture of the proposed approach is decomposed to several stages. In fact, given an input video, it is segmented to obtain images volumes. These last ones are used to construct image foreground and to obtain 3D silhouettes. Then, 3D Zernike moments are calculated for each obtained silhouette series. Finally, the LS-SVM classification method is applied to obtain the actions classes found in the input video. Figure 1 presents the different steps of the proposed approach.

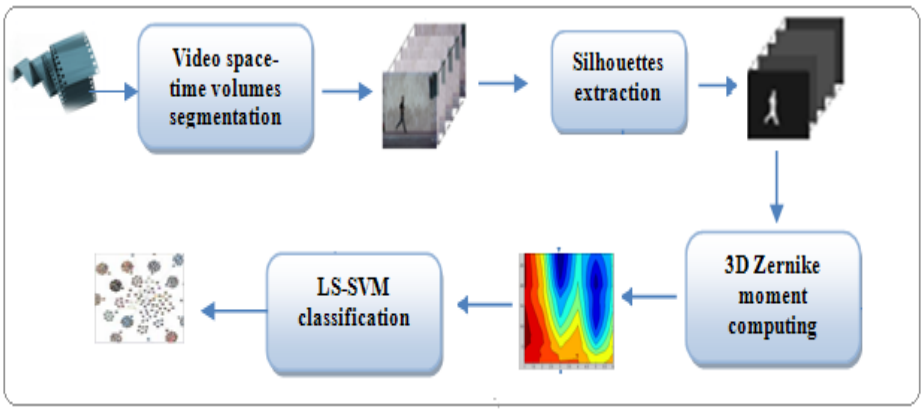

Fig. 1. General Architecture of the proposed approach

3.1 Segmentation and 3D Silhouettes Extraction

Given that video contains body separable actions and that the camera is fixed, the background can be rebuilt easily. In the proposed work, the median value of all pixels in the temporal direction is simply used to obtain the background. The shape of the body is computed by subtracting the background from each frame as shown in Figure 2 where the first row presents video into space-time volumes, the second row presents a static background and the last row explains a silhouette volumes obtained by static background subtracting from image. The last row in Figure 2 present volumes of extracted silhouettes, this obtained silhouettes are stored at the same resolution as original image.

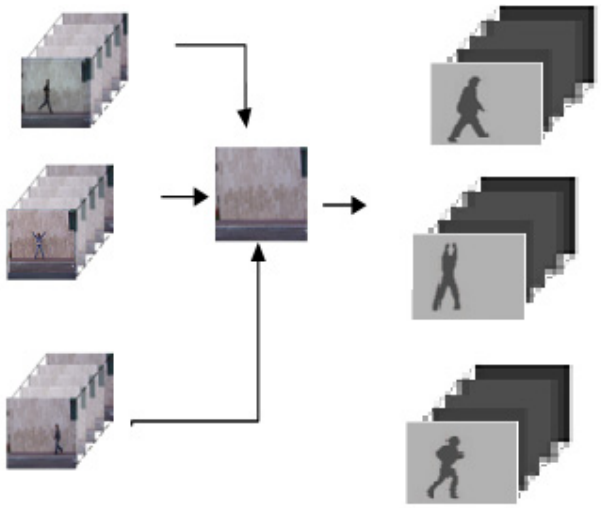

Fig. 2. Background subtraction

3.2 3D Zernike Moments Computing

The main contribution of the proposed approach is to use 3D Zernike moments also known as Zernike velocity moments for video action classification. Indeed, Zernike moments are the best among multiple invariant moments in terms of overall performance. Zernike moments are a class of orthogonal moments and have been considered as effective in terms of image and shape representation. In more, Zernike moments are rotation invariant and can be easily constructed to an arbitrary order. Although higher order moments carry more fine details of an image, they are also more susceptible to noise. Shutter and Nixon in [17] proved that Zernike moments perform well when applied to analyzing walk sequences resulting in a good recognition rate and a compact description. In the proposed approach, 3D Zernike moments are chosen to characterize the actions in video sequences. Therefore, Moments are calculated for each silhouettes series and experimented with different orders of 3D Zernike moments to determine the optimal order which can resolve the

proposed problem. To compute Zernike moments for the obtained silhouette volumes, the are two axis. The first one treat time as the zaxis when applied to images into three-dimensional XYT (x,y and time) block. However, this method confounds the separation of the time and space information, as they are embedded in the data and not specific to the descriptor. Time must be treated separately to space because they are fundamentally different. For our work, we choose to apply the second axis which resolves this problem. In fact, it reformulate the moment descriptor to incorporate time, enabling the separation and/or combination of the time and spatial descriptions. If Zernike velocity moments describe just spatial information (no motion) of a moving rigid shape, then the correlation between images is exploited, and is advantageous, refining the description of the rigid shape as the sequence increases in length. The final Zernike velocity moments of this sequence can be considered as refined (or averaged) Zernike moments of a single image, the descriptions of which are orthogonal. In our work, the shape is moving and deforming (non-rigid), such as a person walking. Then the spatial correlation between consecutive image descriptions is reduced. The Zernike velocity moments are a weighted sum of the Zernike moments over multiple consecutive images. The weighting (velocity) is real-valued and scalar, and the spatial description of each consecutive image in the sequence are orthogonal. The final descriptors of the moving and deforming shape are temporally correlated due to the use of the image sequence. The Zernike velocity moments are expressed as:

$$A_{mn} = \frac{m+1}{\pi} \sum_{i=2}^{\text{images}} \sum_{x} \sum_{y} U(i, \mu, \gamma) S(m,n) P_{ixy} \tag{1}$$

They are bounded by x2 + y2 ≤1, while the shape's structure contributes through the orthogonal polynomials:

$$S(m,n) = [V_{mn}(r,\theta)]^* \tag{2}$$

where * denotes the complex conjugate. The Zernike polynomials [3] Vmn(x; y), expressed in polar coordinates are:

$$V_{mn}(r,\theta) = R_{mn}(r) \exp(jn\theta) \tag{3}$$

where (r; θ) are defined over the unit disc and $R_{mn}(r)$ is the orthogonal radial polynomial, defined as:

$$Rmn(r) = \sum_{s=0}^{\frac{m-|n|}{2}} (-1)^s F(m,n,s,r) \tag{4}$$

Where $$F(m,n,s,r) = \frac{(m-s)!\, r^{m-2s}}{s!\left(\frac{m+|n|}{2}-s\right)!\left(\frac{m-|n|}{2}-s\right)!} \tag{5}$$

Figure 3 presents the series of silhouette in the first line and their associate 3D Zernike moments in the second line.

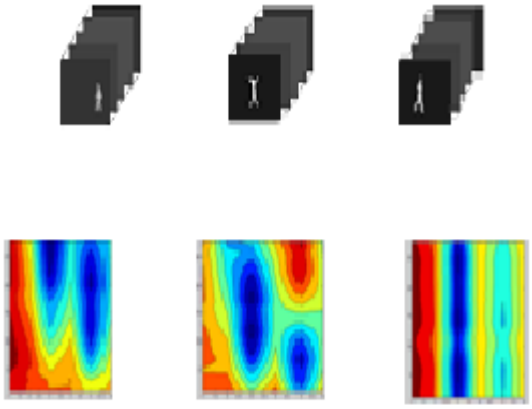

Fig. 3. Series of silhouette and her associate 3D Zernike moments

3.3 LS-SVM Classification

To classify actions in videos, many works are proposed using several attributes such as neural networks [18], GMM [19], etc. The most used classifier is SVM which allows obtaining efficient results to classify actions. It is a powerful classifier used successfully in many pattern recognition problems. For this reason, the Least squares support vector machines LSSVM method is used in the proposed work. It is a variant of standard SVM which simplify the SVM formulation without losing any of its advantages. This method is proposed by Suykens and Vandewalle [20] where the training algorithm solves a convex problem like SVM. It has been shown by a meticulous empirical study that the generalization performance of LS-SVM is comparable to that of SVM [21]. In addition, the training algorithm of LS-SVM is very simplified since a linear problem is resolved instead of a quadratic programming (QP) problem in the SVM case. Given a training set (x; y), i=1,2,,l with input data xi $\in$ R^n and output label data $y_i \in R^n$, and the classifier takes the following form:

$$y(x) = sign\ [\omega^T \varphi(x) + b] \tag{6}$$

Where $\varphi(.)$: Rn $\rightarrow$ Rn is the mapping to the high dimensional and potentially infinite dimensional feature space. In the primal weight space, the optimization problem becomes:

$$\min_{\omega,b,\xi} j(\omega,\xi) = \frac{1}{2}\omega^t\omega + \gamma\frac{1}{2}\sum_{i=1}^{l}\xi_i^2 \tag{7}$$

$$y_i[\omega^t\varphi(x_i)+b] = 1-\xi_i, i = 1,2,...,l \tag{8}$$

Where $\xi_i > 0$ denotes a real constant used to control the punishment degree for misclassification. Because ω becomes infinite dimensional, this primal problem cannot directly be solved. Therefore, let us proceed by constructing the following Lagrangian:

$$L(\omega, b, \xi, \alpha) = J(\omega, \xi) - \sum_{i=1}^{l} \alpha_i (y_i [\omega^t \kappa(x_i) + b] - 1 + \xi_i) \quad (9)$$

Where the values are Lagrange multipliers, which can be positive or negative due to the equality constraints [21].

4 Experimental Results

To evaluate the proposed approach, it was applied on publicly available benchmark dataset proposed by weizmann [22]. This last one is usually used to test action classification models, that originally contained 81 low resolution videos (180* 144). This dataset consists of 93 videos, where 9 people perform 9 different actions: running, bending, waving with one hand, jumping in place, jumping jack, jumping, walking, skip, and waving with two hands. Each video clip contains one subject performing a single action. Illustrative examples for each of these actions are shown in Figure 4. Each video in the dataset is segmented in the space-time volume. Then, the associated silhouettes are extracted. 3D Zernike moments are calculated for each silhouette volumes to describe video. These moments are calculated for square images. For this reason, we have resized the frames series a common size of 140*140 pixels. In the experiments, we select 1/3 of the silhouettes series from each action category to form the training set (40 silhouette sequences) which are used for training LS-SVM, and the rest of series are used for testing LS-SVM. The outputs of LS-SVM process present the corresponding classes of actions.

Fig. 4. Example of actions in the weizmann dataset

To choice the optimal and representative order of Zernike moments which allows to resolve the proposed problem, different orders are tested. The experimentations show that order seven is the best. In fact, it permits to obtain the best values on the diagonal of the confusion matrix. In more, the order higher than seven didn't permit more improvement for classification. The following array presents the confusion matrix for the third order (a1-"walk," a2-"run," a3-"skip," a4-"jack," a5-"jump," a6-"jump in place," a7- "wave with one hand," a8- "wave with two hands," and a9-"bend").

	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	85	10	2	0	0	0	3	0	0
a2	7	80	0	0	0	0	8	5	0
a3	10	0	83	7	0	0	0	0	0
a4	0	3	0	75	4	3	4	5	6
a5	8	11	3	8	70	0	0	0	0
a6	13	12	3	0	7	65	0	0	0
a7	7	12	0	0	7	0	68	6	0
a8	11	0	9	0	0	0	9	71	0
a9	0	0	0	0	3	0	9	7	81

Array (a)

	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	100	0	0	0	0	0	0	0	0
a2	2	98	0	0	0	0	0	0	0
a3	0	3	97	0	0	0	0	0	0
a4	0	0	0	100	0	0	0	0	0
a5	10	0	0	0	83	7	0	0	0
a6	0	0	2	0	0	98	0	0	0
a7	0	0	0	0	2	0	96	0	2
a8	0	0	0	0	0	0	4	96	0
a9	0	0	0	0	0	0	1	0	99

Array (b)

	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	88	11	0	0	0	1	0	0	0
a2	4	93	3	0	0	0	0	0	0
a3	4	8	66	0	11	11	0	0	0
a4	0	0	0	93	0	0	4	3	0
a5	0	0	7	0	91	2	0	0	0
a6	0	0	0	0	9	91	0	0	0
a7	0	0	0	0	0	0	90	10	0
a8	0	0	0	0	0	0	9	91	0
a9	0	0	3	0	3	0	0	0	94

Array (c)

	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	82	1	3	0	0	14	0	0	0
a2	2	35	51	0	10	2	0	0	0
a3	0	41	44	0	9	0	6	0	0
a4	0	0	0	96	0	1	0	1	2
a5	14	14	26	0	29	0	16	1	0
a6	0	0	0	13	0	85	0	1	1
a7	0	0	0	1	0	7	90	2	0
a8	0	1	0	1	0	0	44	52	2
a9	0	0	0	5	3	0	3	5	87

Array (d)

Matrix (a) show that, using order three for Zernike moments, some actions are misclassified. For example, jumping in place action presents a modest classification ratio (65). Matrix (b) show that the seventh order is the best order. In fact, it allows to obtain the better classification ratio for almost all actions. To improve these efficient results, they are compared with two other works [24][1]. The following arrays (c) and (d) present confusion matrix obtained with. The comparative study with other works prove that the proposed approach improve classification results for the most of actions.

5 Conclusion

This paper proposed a new approach for action classification in video based on 3D Zernike moments. In fact, after segmentation of the input video and a subtraction of the background, 3D Zernike moments are calculated on the 3D shape from silhouettes

found in the space-time volume. Finally LS-SVM is used to classify actions. The proposed method has been evaluated by carrying out experiments on the Weizmann databases. The results show a good classification rate for most of tested actions. In future work, we will extend the approach to the classification of videos containing body which makes different actions at same time and videos containing several bodies in action. In more, we can also treat the occlusion effect on the characterization of action.

References

1. Zelnik-Manor, L., Irani, M.: Event-Based Analysis of Video, Computer Vision and Pattern Recognition, Computer Vision and Pattern Recognition. In: Proceedings of the 2001 IEEE Computer Society Conference, vol. 2, pp. 123–130 (2001)
2. Wu, X., Ngo, C.W., Hauptmann, A.G., Tan, H.K.: "Real-Time Near-Duplicate Elimination for Web Video Search with Content and Context. IEEE Transaction on Multimedia 11, 196–207 (2009)
3. Efros, A., Breg, C., Mori, G., Malik, J.: Recognizing Action at a Distance. In: Computer Vision Proceedings. Ninth IEEE International Conference, vol. 2, pp. 726–733 (2003)
4. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005)
5. Gavrila, D.: The visual analysis of human movement: A survey. Computer Vision Image Understand. 73, 82–98 (1999)
6. Cedras, C., Shah, M.: Motion-based recognition: A survey. Image Vision Computer 13, 129–155 (1995)
7. Rao, C., Shah, M.: View-invariance in action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 316–321 (2001)
8. Song, Y., Goncalves, L., Perona, P.: Unsupervised learning of human motion. IEEE Transaction on Pattern Analyses and Machine Intelligence 25, 814–827 (2003)
9. Yacoob, Y., Black, M.: Parameterized modeling and recognition of activities. Computer Vision on Image Understand. 73, 232–247 (1999)
10. Ali, A., Aggarwal, J.: Segmentation and recognition of continuous human activity. In: Procedure of Intelligent Workshop on Detection and Recognition of Events in Video, pp. 28–35 (2001)
11. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Transaction on Pattern Analyses and Machine Intelligence 23, 257–267 (2001)
12. Weinland, D., Ronfard, R., Boyer, E.: Motion history volumes for free viewpoint action recognition. Presented at the IEEE Workshop Modeling People and Human Interaction, pp. 87–89 (2005)
13. Gorelick, L.: al Actions as Space-Time Shapes. IEEE Transactions Pattern Analysis and Machine Intelligence 29, 2247–2253 (2007)
14. Guo, K., Ishwar, P., Konrad, J.: Action Recognition in Video by Covariance Matching of Silhouette Tunnels. In: XXII Brazilian Symposium on Computer Graphics and Image Processing, pp. 299–306 (2009)
15. Kellokumpu, V., Pietikainen, M., Heikkila, J.: Human activity recognition using sequences of postures. In: IAPR Conference on Machine Vision Applications, p. 6570–6573 (2005)

16. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional models for contextual human motion recognition. In: Procedure International of Conference on Computer Vision, vol. 2, pp. 1808–1815 (2005)
17. Shutler, J.D., et al.: Statistical gait recognition via temporal moments. In: Image Analysis and Interpretation 4th IEEE Southwest Symposium, pp. 291–295 (2000)
18. Baccouche1, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Une approche neuronale pour la classification d'actions de sport par la prise en compte du contenu visuel et du mouvement dominant. In: Compression et Représentation des Signaux Audiovisuels (CORESA), pp. 25–30 (2010)
19. Tian, Y., Liu, Z., Yao, B., Zhang, Z., Huang, T.: Action Detection Using Multiple Spatial-Temporal Interest Point Features. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 340–345 (2010)
20. Suykens, J.A.K., Vandewalle, J.: Least Squares Support Vector Machines. Neural Processing Letters 9(3), 293–300 (2002)
21. Gestel, T.V., Suykens, A.K.: Benchmarking Least Squares Support Vector Machine Classifiers, Technical report of Internal Report 00-37 on ESAT-SISTA 54, 5–32 (2000)
22. Blank, M., Gorelick, L., Shechtman, M., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on pattern Analysis and Machine Intelligence 29, 2247–2253 (2005)
23. Green, R., Guan, L.: Quantifying and recognizing human movement patterns from monocular video images. IEEE Transaction Circuits on System Video Technologys 14, 179–190 (2004)
24. Dhillon, C.H., Nowozin, P.S., Lampert, S.: Combining appearance and motion for human action classification in videos. In: Computer Vision and Pattern Recognition Workshops, pp. 22 – 29 (2009)