Moment Shape Descriptors Applied for Action Recognition in Video Sequences

Katarzyna Gościewska and Dariusz Frejlichowski(✉)

Faculty of Computer Science and Information Technology,
West Pomeranian University of Technology, Szczecin,
Żołnierska 52, 71-210 Szczecin, Poland
{kgosciewska,dfrejlichowski}@wi.zut.edu.pl

Abstract. Algorithms for recognition of human activities have found application in many computer vision systems, for example in visual content analysis approaches and in video surveillance systems, where they can be employed for the recognition of single gestures, simple actions, interactions and even behaviour. In this paper an approach for human action recognition based on shape analysis is presented. Set of binary silhouettes extracted from video sequences representing a person performing an action are used as input data. The developed approach is composed of several algorithms including those for shape representation and matching. It can deal with sequences of different number of frames and none of them has to be removed. The paper provides some initial experimental results on classification using proposed approach and moment shape description algorithms, namely the Zernike Moments, Moment Invariants and Contour Sequence Moments.

Keywords: Action recognition · Shape descriptors · Video sequences · Binary silhouettes

1 Introduction

The category of human activities includes gestures (elementary human body movements executed for a short time), actions (composed of multiple temporarily organized gestures performed by a single person, such as walking, running, bending or waving), interactions (human-human or human-object activities such as carrying/abandoning/stealing an object) and group activities (i.e. activities performed by groups consisting of multiple objects) [1]. If silhouettes are used for action classification then the human movement can be represented as a continuous pose change and silhouettes extracted from consecutive video frames can be applied to obtain action descriptors used in traditional classification approaches [2]. The approach presented in this paper addresses the problem of recognising an action of a single person and uses information contained in a sequence of binary silhouettes. Each silhouette is firstly processed individually using particular shape description algorithm. Shape representations are

N.T. Nguyen et al. (Eds.): ACIIDS 2017, Part II, LNAI 10192, pp. 197–206, 2017.
DOI: 10.1007/978-3-319-54430-4_19

compared within a sequence to obtain sequence representation. Then sequence representations are processed and final representations are obtained. These are further subjected to the process of classification based on the template matching approach.

The variety of surveillance system applications affects the diversity of activity recognition approaches. In [1] methods for activity recognition were classified into hierarchical (statistical, syntactic and description-based) and non-hierarchical (spatio-temporal and sequential). Hierarchical methodologies enable the recognition of complicated and complex human activities, including interactions and group activities. In turn, non-hierarchical solutions aim to recognize short, primitive actions and repetitive activities—the recognition process is based on the analysis of unknown sequences using an algorithm that matches data to the predefined activity classes. One of the common spatio-temporal technique using accumulated silhouettes was proposed in [3] and is based on motion energy image (MEI, shows where the movement occurs) and motion history image (MHI, shows how the object is moving). MEI and MHI are used to create a static vector-image (temporal template), scale and shift invariant descriptor is used for representation and Mahalanobis distance is used for matching. Another space-time approach was introduced in [4]. It utilizes Poisson equation for feature extraction and human actions are represented as three-dimensional shapes (silhouettes accumulated in the space-time volume). In [5] action sequence is represented by a History Trace Template composed of the set of Trace Transforms extracted for each silhouette. In [6] an action is represented by a set of SAX (Symbolic Aggregate approXimation) vectors which are based on one-dimensional representations of each silhouette. Some other action recognition approaches utilizes only characteristic frames—action is recognized based on selected key poses, e.g. [7–9].

The rest of the paper is organized as follows: Sect. 2 describes the developed approach, Sect. 3 presents algorithms used for shape representation, Sect. 4 presents experimental results on action classification, and Sect. 5 concludes the paper.

2 Proposed Approach

The developed approach is composed of several steps, starting from representing shape information of each silhouette in the dataset and ending on preparing final representation of each sequence, which is then used for action classification. In our approach it is assumed that one video sequence represents one action and it corresponds to a set of binary images—one image contains one silhouette and can be understood as a foreground mask.

In step 1 each silhouette is represented by one shape descriptor using information about its contour or region. In many cases, one shape description algorithm can be employed to obtain representations of different size, e.g. by calculating various order of moments for Zernike Moments or taking various subparts of Fourier Transform-based descriptor. Thank to this we have an opportunity to investigate many shape representations and to select the smallest one which simultaneously carries the most information. If needed, the resultant shape representation is transformed into a vector.

Step 2, for a single sequence, includes calculation of dissimilarities between first frame and the rest of frames using Euclidean distance. The resulting vector containing distance values (normalized to interval $[0, 1]$) is a one-dimensional descriptor of a sequence (a distance vector). The number of its elements equals the number of silhouettes in the input sequence. This vector can be plotted and analysed visually in terms of similarities between actions and their characteristics.

Step 3 aims to convert distance vectors into the form and size that enables the calculation of similarity between them. Therefore, distance vectors were treated as signals and it turned out that the best way to transform such a signal was to use the magnitude of the fast Fourier Transform and a periodogram. Periodogram is a spectral density estimation of a signal and it can determine hidden periodicities in data [10]. Moreover, the periodogram helped to equalize the size of final representations.

Step 4 includes the process of action classification based on sequence descriptors using template matching approach and correlation coefficient. Here template matching is understood as a process that compares each test object with all templates and indicates the most similar one, which corresponds to the probable class of a particular test object. This is a traditional classification solution when only one template set is used. However, some initial tests showed that final results depend on templates. Therefore, we have decided to perform the experiment several times using k-fold cross-validation technique [11] and different set of templates in each iteration. The final recognition effectiveness is then the average of all iterations. For instance, in the first iteration objects with numbers from 1 to k are used as templates and objects with numbers from $k + 1$ to n as test objects, then in the second iteration objects with numbers from $k + 1$ to $2 * k$ are used as templates and remaining objects are used for testing, and so on. Then the results can be interpreted and analysed in three different ways (considering only correct classifications—'true positive'):

1. Recognition effectiveness for each shape descriptor, averaged for all classes and all iterations,
2. Recognition effectiveness for each iteration, each shape descriptor and averaged for all classes,
3. Classification accuracy for each class, each shape descriptor and averaged for all iterations (or for one selected iteration only).

3 Selected Shape Descriptors

3.1 Zernike Moments

The Zernike Moments are orthogonal moments which can be derived using Zernike orthogonal polynomials and the following formula [12]:

$$V_{nm}\left(x,y\right)=V_{nm}\left(r\cos\theta,\sin\theta\right)=R_{nm}(r)\exp\left(jm\theta\right), \tag{1}$$

where $R_{nm}(r)$ is the orthogonal radial polynomial [12]:

$$R_{nm}(r)=\sum_{s=0}^{(n-|m|)/2}\left(-1\right)^{s}\frac{(n-s)!}{s!\times\left(\frac{n-2s+|m|}{2}\right)!\left(\frac{n-2s-|m|}{2}\right)!}r^{n-2s}, \tag{2}$$

where $n=0,1,2,\ldots$; $0\leq|m|\leq n$; $n-|m|$ is even.

The Zernike polynomials are a complete set of functions orthogonal over the unit disk $x^2+y^2<1$. The Zernike Moments are rotation invariant and resistant to noise and minor variations in shape. The Zernike Moments of order n and repetition m of a region shape $f(x,y)$ are derived using the following formula [12]:

$$Z_{nm}=\frac{n+1}{\pi}\sum_{r}\sum_{\theta}f(r\cos\theta,r\sin\theta)\cdot R_{nm}(r)\cdot\exp(jm\theta),\ r\leq1. \tag{3}$$

3.2 Moment Invariants

Moment Invariants can be applied for grayscale images or objects (both region and contour) and are described below based on [13–15]. Firstly, general geometrical moments are calculated using a following formula (discrete version):

$$m_{pq}=\sum_{x}\sum_{y}x^{p}y^{q}f(x,y). \tag{4}$$

The value of function $f(x,y)$ equals 1 if a pixel belongs to the object (silhouette) and 0 for background. Then, to make representation invariant to translation, the centroid is calculated:

$$x_c=\frac{m_{10}}{m_{00}},\quad y_c=\frac{m_{01}}{m_{00}}. \tag{5}$$

In the next step, Central Moments are calculated using the centroid:

$$\mu_{pq}=\sum_{x}\sum_{y}(x-x_c)^{p}(y-y_c)^{q}f(x,y). \tag{6}$$

Then, the invariance to scaling is obtained by central normalised moments:

$$\eta_{pq}=\frac{\mu_{pq}}{\mu_{00}^{\frac{p+q+2}{2}}}. \tag{7}$$

Ultimately, Moment Invariants are derived (usually seven first values are used in pattern recognition applications):

$$\begin{gathered}
\phi_1 = \eta_{20} + \eta_{02} \\
\phi_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2] \\
+(3\eta_{21} - \eta_{03})(\eta_{03} + \eta_{21})[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2] \\
\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03}^2] \\
+4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21}) \\
\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2] \\
-(\eta_{30} - 3\eta_{12})(\eta_{03} + \eta_{21})[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2]
\end{gathered} \tag{8}$$

3.3 Contour Sequence Moments

Another moment shape descriptor is Contour Sequence Moments (based on shape contour only). The method is described below based on [16]. In the first step, the contour is represented as ordered sequence $z(i)$ which elements are the Euclidean distances from the centroid to particular N points of the shape contour. The one-dimensional normalised contour sequence moments are calculated using following formulas:

$$m_r = \frac{1}{N}\sum_{i=1}^{N}[z(i)]^r, \tag{9}$$

$$\mu_r = \frac{1}{N}\sum_{i=1}^{N}[z(i) - m_1]^r. \tag{10}$$

The r-th normalised contour sequence moment and normalised central sequence moment are:

$$\bar{m_r} = \frac{m_r}{(\mu_2)^{r/2}}, \quad \bar{\mu_r} = \frac{\mu_r}{(\mu_2)^{r/2}}. \tag{11}$$

The final shape description consists four values:

$$F_1 = \frac{(\mu_2)^{1/2}}{m_1}, \quad F_2 = \frac{\mu_3}{(\mu_2)^{3/2}}, \quad F_3 = \frac{\mu_4}{(\mu_2)^2}, \quad F_4 = \bar{\mu_5}. \tag{12}$$

4 Experimental Conditions and Results

4.1 Data and Conditions

Several experiments have been carried out in order to verify the effectiveness and accuracy of the proposed approach using moment shape descriptors. Each experiment consisted of four steps described in Sect. 2, except that for each experiment different shape description algorithm was used: Moment Invariants, Contour Sequence Moments and Zernike Moments (orders from 1st to 15th). The experiments were performed using a part of the Weizmann dataset [17]. The original Weizmann dataset contains 90 low-resolution (180 × 144, 50 fps) video sequences of 9 actors performing 10 actions. The corresponding binary masks extracted using background subtraction are available and were used as input data. We have selected five types of actions for the experiments: run, walk, bend, jump and one-hand wave (see Fig. 1 for exemplary frames and Fig. 2 for exemplary silhouettes). Therefore, our experimental database consisted of 45 silhouette sequences of 9 actors performing 5 actions. The number of frames (silhouettes) in a sequence varied from 28 to 125. During the experiment each subgroup of 5 action sequences was iteratively used as a template set. Experimental results are presented and described in the next subsection.

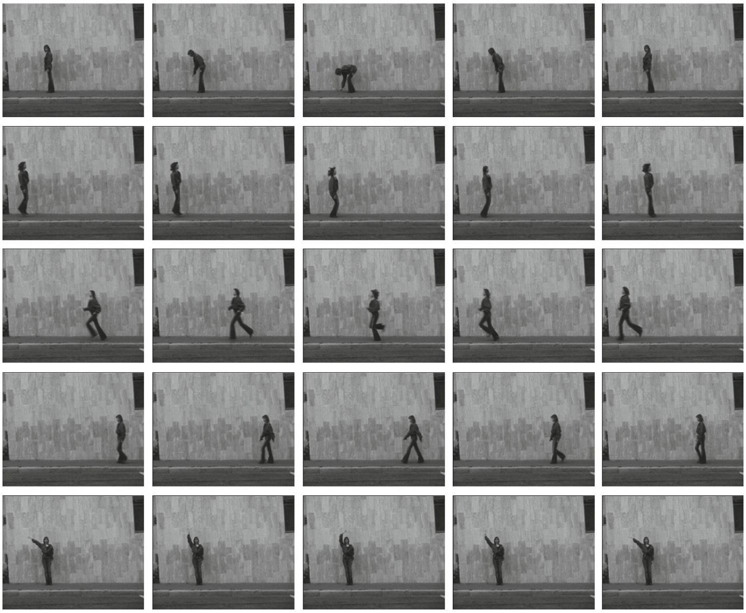

Fig. 1. Exemplary video frames from the Weizmann dataset [17]—images in rows correspond to (from the top) bending, jumping, running, walking and waving actions respectively.

4.2 Results

In this subsection some initial experimental results are provided. The goal of the first experiment was to identify the best approach for further work. Therefore, the results correspond to the average recognition effectiveness values for each shape descriptor, all classes and all iterations, and are as follows:

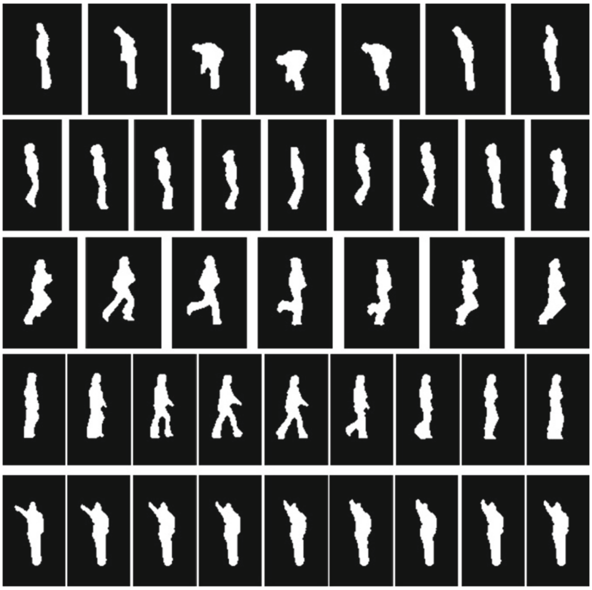

Fig. 2. Exemplary silhouettes extracted from video sequences presented in Fig. 1—silhouettes in rows correspond to (from the top) bending, jumping, running, walking and waving actions (images come from the Weizmann dataset [17]).

- 38.05% for Moment Invariants;
- 40.55% for Contour Sequence Moments;
- 32.5%, 37.22%, 29.44%, 33.61%, 30.55%, 34.16%, 32.22%, 40.55%, 43.05%, 45.83%, 47.5%, 46.66%, 50.27%, 47.77% and 48.88%, for Zernike Moments of orders from 1st to 15th respectively;

For Zernike Moments the use of 13th order gives the best results and only this order for feature representation will be further analysed. Table 1 contains the results of the second experiment—percentage recognition effectiveness values obtained in each iteration, for each shape descriptor and averaged for all classes. It can be seen that the classification accuracy values vary between iterations and that the best result is obtained in iteration no. 7. This can be interpreted in such a way that templates used in this iteration are represented by the most distinctive features enabling proper class indication.

In Table 2 the results of the third experiment are provided—the averaged classification values for all iterations. It can be clearly seen that 'bend' action is the most recognizable one, while the 'jump' action is the least distinctive. Based on Table 2, Zernike Moments can be selected as the best shape descriptor, except for the classification to 'wave' class which was more effective while Contour Sequence Moments descriptor was used. The same dependencies can be indicated for Table 3 where the classification accuracy values for iteration no. 7 are depicted.

Table 1. Recognition effectiveness for each iteration, each shape descriptor and averaged for all classes

Iteration no.	Moment Invariants	Contour Sequence Moments	Zernike Moments
1	27.5%	32.5%	27.5%
2	42.5%	35.0%	55.0%
3	25.0%	35.0%	47.5%
4	42.5%	50.0%	37.5%
5	35.0%	42.5%	37.5%
6	50.0%	40.0%	60.0%
7	50.0%	42.5%	67.5%
8	40.0%	50.0%	60.0%
9	30.0%	37.5%	60.0%

Table 2. Classification accuracy averaged for all iterations

Class	Moment Invariants	Contour Sequence Moments	Zernike Moments
‘bend’	56.9%	48.6%	90.3%
‘jump’	18.0%	22.2%	23.6%
‘run’	26.3%	36.1%	40.3%
‘walk’	43.0%	40.2%	48.6%
‘wave’	45.8%	55.5%	48.6%

Table 3. Classification accuracy for iteration no. 7

Class	Moment Invariants	Contour Sequence Moments	Zernike Moments
‘bend’	75.0%	37.5%	100%
‘jump’	0.0%	12.5%	37.5%
‘run’	37.5%	25.0%	37.5%
‘walk’	37.5%	37.5%	75%
‘wave’	100.0%	100.0%	87.5%

There is another interesting element of the approach—normalized distance vectors can be plotted and compared visually. Figure 3 contains five plots of distance vectors corresponding to five silhouette sequences used in Fig. 2 to illustrate exemplary frames. The differences between plots are clearly visible which relates to variability of silhouettes within a sequence and reveals periodicities in actions. Low peaks correspond to silhouettes that are most similar (due to the use of Euclidean distance) to the first silhouette in a sequence. Some distinctive features of the plots could be further employed to improve classification results. For instance, the faster the action, the more densely arranged low or high peaks are obtained.

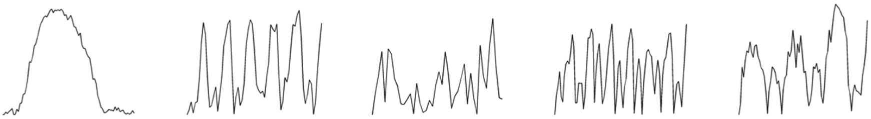

Fig. 3. Exemplary plots of distance vectors for five actions performed by the same actor: bending, jumping, running, walking and waving respectively. The exemplary distance vectors were obtained using Zernike Moments of order 13th.

5 Summary and Conclusions

In the paper, an approach for action recognition based on silhouette sequences has been presented. It uses various shape description algorithms to represent silhouettes and Euclidean distance to estimate dissimilarity between first and the rest of frames. Normalized distance vectors are further processed using fast Fourier transform and periodogram in order to obtain final sequence representations. These representations are compared using template matching approach and correlation coefficient. The best experimental results in terms of classification accuracy were obtained using Zernike Moments of order 13th.

Generally, the initial results are promising, although the proposed approach requires improvements and should be examined using more data. To make the approach more effective, future works include experimental verification of other shape representation algorithms and matching measures. Moreover, the problems which cause lower classification accuracy should be identified and solved—the use of additional processing step, another shape feature or different classification process should be investigated. Also, not only shape descriptors may be used, but other features, such as centroid of an object which may help to distinguish e.g. jumping and running actions. Any modifications should be verified for their influence on final effectiveness of the approach and time consumption.

References

1. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. Vis. Comput. **29**, 983–1009 (2012)
2. Borges, P.V.K., Conci, N., Cavallaro, A.: Video-based human behavior understanding: a survey. IEEE Trans. Circuits Syst. Video Technol. **23**, 1993–2008 (2013)
3. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 257–267 (2001)
4. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 2247–2253 (2007)

5. Goudelis, G., Karpouzis, K., Kollias, S.: Exploring trace transform for robust human action recognition. Pattern Recogn. **46**, 3238–3248 (2013)
6. Junejo, I.N., Junejo, K.N., Aghbari, Z.A.: Silhouette-based human action recognition using SAX-shapes. Vis. Comput. **30**, 259–269 (2014)
7. Baysal, S., Kurt, M.C., Duygulu, P.: Recognizing human actions using key poses. In: 20th International Conference on Pattern Recognition, pp. 1727–1730 (2010)
8. Liu, L., Shao, L., Zhen, X., Li, X.: Learning discriminative key poses for action recognition. IEEE Trans. Cybern. **43**, 1860–1870 (2013)
9. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: Silhouette-based human action recognition using sequences of key poses. Pattern Recogn. Lett. **34**, 1799–1807 (2013)
10. Chitode, J.: Digital Signal Processing. Technical Publications, Berlin (2009)
11. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of 14th International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137–1143 (1995)
12. Zhang, D., Lu, G.: Shape-based image retrieval using generic Fourier descriptor. Sig. Proc.: Image Commun. **17**, 825–848 (2002)
13. Rothe, I., Susse, H., Voss, K.: The method of normalization to determine invariants. IEEE T. Pattern Anal. **18**, 366–376 (1996)
14. Hupkens, T.M., de Clippeleir, J.: Noise and intensity invariant moments. Pattern Recogn. Lett. **16**, 371–376 (1995)
15. Liu, C.B., Ahuja, N.: Vision based fire detection. In: Proceedings of 17th International Conference on Pattern Recognition, vol. 4, pp. 134–137 (2004)
16. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. PWS—an Imprint of Brooks and Cole Publishing, Belmont (1998)
17. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: 10th IEEE International Conference on Computer Vision, pp. 1395–1402 (2005)