Action Recognition Using Silhouette Sequences and Shape Descriptors

Katarzyna Gościewska and Dariusz Frejlichowski(✉)

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

Abstract. The paper provides an approach for human action recognition based on shape analysis. The developed approach is intended for specific type of data, namely sequences of binary silhouettes representing a person performing an action, and consists of several processing steps including shape description as well as similarity or dissimilarity estimation. The approach can deal with sequences of different length without removing any frames. The paper also provides some experimental results showing the classification accuracy and overall recognition effectiveness of the proposed approach using several popular shape description algorithms, namely the Two-Dimensional Fourier Descriptor, Generic Fourier Descriptor, Point Distance Histogram and UNL-Fourier Descriptor.

1 Introduction

Human activity recognition has found applications in many areas, especially in video surveillance systems. Nowadays surveillance and security solutions are associated with advanced video content analysis algorithms which support human operators in observation of many scenes and detection of various events related e.g. to abnormal/unusual activities [17,20]. Human activity recognition can be performed on various levels of complexity, depending on the type of activity. According to [18] the simplest level includes recognition of gestures, that is elementary human body movements executed for a short time. In turn, an action is composed of multiple temporarily organized gestures performed by a single person, such as walking, running, bending or waving.

The literature provides a number of shape-based methods that are applied for the recognition of actions using silhouette sequences. For instance, in [11] Trace Transform for each silhouette is extracted, and the whole action sequence is then represented by a final History Trace Template composed of the set of transforms. In [12] an individual silhouette is converted into a one-dimensional representation and then transformed into symbolic vector called SAX (Symbolic Aggregate approXimation). An action is represented by a set of SAX vectors. Some other approaches limit the number of silhouettes and action recognition is based on selected key poses (characteristic frames), e.g. [1,5,15]. Silhouettes can also be accumulated in order to generate spatio-temporal features. A common technique in this category uses motion energy images and motion history images,

R.S. Choraś (ed.), *Image Processing and Communications Challenges 8*,
Advances in Intelligent Systems and Computing 525, DOI 10.1007/978-3-319-47274-4_21

and was proposed in [3]. The basis of the representation is a static vector-image (temporal template) and the vector value at each pixel corresponds to motion properties at that pixel location in the image. The authors of [10] introduced other space-time approach, which utilizes Poisson equation to extract several features, among others, space-time saliency and action dynamics. This approach regards human actions as three-dimensional shapes—accumulated silhouettes in the space-time volume.

This paper focuses solely on action recognition where the amount of processed information about an action is limited to a sequence of binary silhouettes which are then represented using selected shape description algorithms. These representations are subjected to further processing and ultimately are compared on the basis of template matching approach in order to identify overall recognition effectiveness and final classification accuracy for a particular shape descriptor. The rest of the paper is organized as follows: Sect. 2 presents the consecutive steps of the proposed approach and describes the algorithms use for shape representation, Sect. 3 presents some experimental results on action classification and recognition, and Sect. 4 concludes the paper.

2 Developed Approach—Data Processing Steps

The developed approach addresses the problem of recognising an action of a single person and uses information contained in a sequence of binary silhouettes. According to [4], the use of silhouettes for action classification assumes that human movement can be represented as a continuous pose change. Then action descriptors can be obtained based on silhouettes extracted from consecutive video frames and traditional classification approaches can be applied. The proposed approach has been already tested using a part of the Weizmann dataset [2]. 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 (see Fig. 1 for example), walk, bend, jump and one-hand wave. Some data and results will be used in this section to illustrate several processing steps of our approach. Therefore, the following description also includes explanation for the research experiment.

Fig. 1. Exemplary silhouettes from a running action sequence (images come from the Weizmann dataset [2])

2.1 Step 1—Calculation of Shape Descriptor for Each Silhouette

In this step, each silhouette is represented by one shape descriptor using information about its contour or region. We have already tested four shape description algorithms, namely the Two-Dimensional Fourier Descriptor, Generic Fourier Descriptor, Point Distance Histogram and UNL-Fourier Descriptor. All selected algorithms enable to calculate shape representations of different size and in a form of a vector. It is important due to the fact that we are trying to select the smallest descriptor which simultaneously carries the most information. The selected algorithms have been previously successfully employed for shape analysis and in template matching approaches, e.g. in [7,9], and are described below.

The Two-Dimensional Fourier Descriptor is applied to a region shape. It is calculated as a magnitude of the Fast Fourier Transform and has a form of a matrix with absolute complex values [14]. This algorithm is used as a step in the following two methods. The first one is the UNL-Fourier Descriptor [16]. It is based on contour information and uses centroid to transform Cartesian coordinates of a contour into polar coordinates. New coordinate values are then put into a matrix, where rows represent distances from the centroid and columns the corresponding angles. As a result, an image containing unfolded shape contour in polar coordinates is obtained and then the Two-Dimensional Fourier Descriptor is applied. The second Fourier-based description algorithm is the Generic Fourier Descriptor, which is applied to a region shape and uses the transformation of Cartesian points to polar coordinate system. All pixel coordinates from original region shape image are transformed into polar coordinates and new values are put to a rectangular Cartesian image [19]. Row elements correspond to distances from centroid and the columns to 360 angles. The result has a form of an image and as in the previous case the Two-Dimensional Fourier Descriptor is applied.

The Point Distance Histogram (PDH) is a shape descriptor that utilizes information about shape contour [8]. In order to derive a PDH representation, an origin of the polar transform of a contour is firstly selected, usually a centroid. Polar coordinates are stored in two vectors—one for angles and one for radii. In the next step, angle values are converted to the nearest integers. Then the elements in both vectors are rearranged with respect to the increasing angle values. If any equal angles exist, then only the element with the highest radii value is left. Next, only radii vector is selected for further processing. Its elements are normalized and assigned to bins in the histogram. Then values in histogram bins are normalized according to the highest one and the final representation is obtained.

2.2 Step 2—Matching All Shape Descriptors Within One Sequence

For a given sequence, this step 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. Figure 2 shows exemplary plots

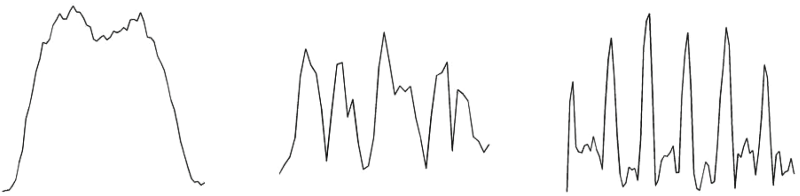

Fig. 2. Exemplary plots of distance vectors for three actions performed by the same actor: bend, walk and run respectively. Low peaks correspond to the silhouettes that are most similar to the first silhouette in a sequence. The exemplary distance vectors were obtained using 2×2 subpart of the Two-Dimensional Fourier Descriptor. For walk and run actions the periodicities are noticeable

of distance vectors of three different actions. The plots reveal action periodicities and the differences between actions.

2.3 Step 3—Converting the Distance Vectors into Sequence Representations

The next step 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 a periodogram. Periodogram is a spectral density estimation of a signal and it can determine hidden periodicities in data [6]. In most cases, the results have improved when fast Fourier transform was firstly applied to a distance vector, the magnitude was extracted and after that the periodogram was used. Due to the fact that distance vectors varied in size, the periodogram helped to equalize final representations' sizes. Ultimately, one periodogram represents one silhouette sequence.

2.4 Step 4—Selection of Matching Procedure and Splitting Data

This step solves the problem of how to split data into templates and test objects. Some initial tests showed that final results were dependent on which part of the data was selected as templates (one template is a single class representative). Therefore, being inspired by the k-fold cross-validation technique [13] we have decided to perform the experiment several times using different set of templates in each iteration. The final recognition effectiveness is then the average of the results from all iterations. For instance, the first iteration used objects with numbers from 1 to k as templates and objects with numbers from $k+1$ to n as test objects, then the second iteration used objects with numbers from $k+1$ to $2 * k$ as templates and the rest as test objects, and so on.

2.5 Step 5—Estimation of the Similarity Between Sequence Descriptors

This step includes the calculation of similarity between sequence descriptors using template matching approach and variable template set as described in the previous step. The correlation coefficient is used. Here template matching, for one iteration, is understood as a process that compares each test object with all templates and selects the most similar one, what simultaneously indicates the probable class of a particular test object. Then the results can be interpreted and analysed in three different ways (considering only the number of correct classifications—'true positive'):

1. Overall recognition effectiveness for each shape descriptor, averaged for all classes and all iterations,
2. Classification accuracy for each iteration, each shape descriptor and all classes,
3. Classification accuracy for each class, each shape descriptor and all iterations.

3 Experiments and Results

Several experiments have been carried out in order to verify the effectiveness and accuracy of the proposed approach. Each experiment consisted of five steps described in Sect. 3, except that for each experiment different shape description algorithm was used. Moreover, various size of shape representation were employed, namely the 2×2, 5×5, 10×10, 25×25 and 50×50 absolute spectrum subparts for the Two-Dimensional Fourier Descriptor, Generic Fourier Descriptor and UNL-Fourier Descriptor, and 2, 5, 10, 25, 50, 75 and 100 histogram bins for the Point Distance Histogram. In the experimental database there were 45 silhouette sequences of 9 actors performing 5 actions—bend, jump, run, walk and one-hand wave—taken from the Weizmann dataset [2]. The number of frames (silhouettes) in a sequence varied from 28 to 125. During the experiment each subgroup of 5 sequences of one person performing these actions was iteratively used as a template set. Percentage experimental results are presented below with respect to the three result analysis manners (see Step 5. of the approach).

Average recognition effectiveness values for each shape descriptor, all classes and all iterations are as follows:

- 49.2 %, 46.1 %, 42.8 %, 41.1 % and 39.4 % for the 2×2, 5×5, 10×10, 25×25 and 50×50 subparts of the Two-Dimensional Fourier Descriptor respectively;
- 51.7 %, 51.4 %, 51.7 %, 52.2 % and 50.0 % for the 2×2, 5×5, 10×10, 25×25 and 50×50 subparts of the Generic Fourier Descriptor respectively;
- 36.7 %, 39.7 %, 37.8 %, 36.9 % and 39.7 % for the 2×2, 5×5, 10×10, 25×25 and 50×50 subparts of the UNL-Fourier Descriptor respectively;
- 39.7 %, 36.7 %, 36.1 %, 34.7 %, 34.4 %, 34.4 % and 34.4 % for the 2, 5, 10, 25, 50, 75 and 100 histogram bins of the Point Distance Histogram respectively.

Table 1. Results for the experiment using Generic Fourier Descriptor—classification accuracy for each iteration, each shape descriptor and averaged for all classes

Iteration	Descriptor size				
No	2×2	5×5	10×10	25×25	50×50
1	45.0 %	45.0 %	45.0 %	50.0 %	47.5 %
2	37.5 %	37.5 %	35.0 %	35.0 %	35.0 %
3	52.5 %	52.5 %	52.5 %	52.5 %	52.5 %
4	52.5 %	52.5 %	50.0 %	52.5 %	37.5 %
5	45.0 %	42.5 %	40.0 %	45.0 %	47.5 %
6	47.5 %	47.5 %	50.0 %	50.0 %	50.0 %
7	57.5 %	57.5 %	60.0 %	57.5 %	55.0 %
8	**67.5 %**	**67.5 %**	**72.5 %**	**70.0 %**	**60.0 %**
9	60.0 %	60.0 %	60.0 %	57.5 %	65.0 %

The average results indicate the Generic Fourier Descriptor as the most effective shape description algorithm for the employed approach and selected data. The percentage recognition effectiveness values are similar for all descriptor sizes. Additional results will enable for a more detailed insight into the classification accuracy using the Generic Fourier Descriptor (see Tables 1, 2 and 3). Table 1 illustrates the results obtained in each iteration 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. 8. 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 averaged results for all iterations can be found. It can be clearly seen that 'bend' action is the most distinctive one, while the 'jump' action is the least recognizable. It is not obvious which shape description size should be indicated as the best to employ for shape representation, because it varies depending on the class. However, if only iteration no. 8 is taken under

Table 2. Results for the experiment using Generic Fourier Descriptor—classification accuracy for each class, each shape descriptor and averaged for all iterations

Class	Descriptor size				
	2×2	5×5	10×10	25×25	50×50
'bend'	87.5 %	87.5 %	88.9 %	88.9 %	87.5 %
'jump'	27.8 %	26.4 %	30.6 %	31.9 %	27.8 %
'run'	48.6 %	50.0 %	51.4 %	54.2 %	55.6 %
'walk'	38.9 %	37.5 %	43.1 %	43.1 %	44.4 %
'wave'	55.6 %	55.6 %	44.4 %	43.1 %	34.7 %

Table 3. Classification accuracy for each class, various size of Generic Fourier Descriptor and iteration no. 8

Class	Descriptor size				
	2×2	5×5	$\mathbf{10 \times 10}$	25×25	50×50
'bend'	100 %	100 %	**100 %**	100 %	100 %
'jump'	25.0 %	25.0 %	**37.5 %**	37.5 %	25.0 %
'run'	87.5 %	87.5 %	**87.5 %**	87.5 %	75.0 %
'walk'	62.5 %	62.5 %	**75.0 %**	75.0 %	75.0 %
'wave'	62.5 %	62.5 %	**62.5 %**	50.0 %	25.0 %

consideration—due to the most distinctive templates—it turns out that the proposed approach is most effective when the 10×10 subpart of the Generic Fourier Descriptor is used. Table 3 depicts percentage classification accuracy values for iteration no. 8.

4 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 shape descriptors within a sequence. Normalized distances create a vector representation of the sequence. All sequence representations are further processed using fast Fourier transform and periodogram, and ultimately are compared using template matching approach and correlation coefficient. Experimental results showed that the developed approach is most effective and accurate when the Generic Fourier Descriptor is used. Generally, the results are promising, however the developed approach needs further improvements and should be examined using more data. Future works involve experimental verification of other shape descriptors and matching measures, that will make the approach more effective.

References

1. Baysal, S., Kurt, M.C., Duygulu, P.: Recognizing human actions using key poses. In: 20th International Conference on Pattern Recognition, pp. 1727–1730, August 2010
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: The Tenth IEEE International Conference on Computer Vision, pp. 1395–1402 (2005)
3. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)
4. Borges, P.V.K., Conci, N., Cavallaro, A.: Video-based human behavior understanding: a survey. IEEE Trans. Circ. Syst. Video Technol. **23**(11), 1993–2008 (2013)

5. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: Silhouette-based human action recognition using sequences of key poses. Pattern Recogn. Lett. **34**(15), 1799–1807 (2013)
6. Chitode, J.: Digital Signal Processing. Technical Publications, Pune (2009)
7. Forczmański, P., Frejlichowski, D.: Robust stamps detection and classification by means of general shape analysis. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2010. LNCS, vol. 6374, pp. 360–367. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15910-7_41
8. Frejlichowski, D.: An experimental comparison of three polar shape descriptors in the general shape analysis problem. In: Swiatek, J., Borzemski, L., Grzech, A., Wilimowska, Z. (eds.) Information Systems Architecture and Technology – System Analysis in Decision Aided Problems, pp. 139–150. Oficyna Wydawnicza Politechniki Wrocławskiej (2010)
9. Frejlichowski, D.: Pre-processing, extraction and recognition of binary erythrocyte shapes for computer-assisted diagnosis based on mgg images. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L., Wojciechowski, K. (eds.) Computer Vision and Graphics, pp. 368–375. Springer, Berlin (2010)
10. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Trans. Pattern Anal. Mach. Intell. **29**(12), 2247–2253 (2007)
11. Goudelis, G., Karpouzis, K., Kollias, S.: Exploring trace transform for robust human action recognition. Pattern Recogn. **46**(12), 3238–3248 (2013)
12. Junejo, I.N., Junejo, K.N., Aghbari, Z.A.: Silhouette-based human action recognition using sax-shapes. Vis. Comput. **30**(3), 259–269 (2014)
13. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, vol. 2, pp. 1137–1143. Morgan Kaufmann Publishers Inc., San Francisco (1995)
14. Kukharev, G.: Digital Image Processing and Analysis (in Polish). SUT Press, Szczecin (1998)
15. Liu, L., Shao, L., Zhen, X., Li, X.: Learning discriminative key poses for action recognition. IEEE Trans. Cybern. **43**(6), 1860–1870 (2013)
16. Rauber, T.W.: Two dimensional shape description. Technical report, Universidade Nova de Lisboa, Lisoba, Portugal (1994)
17. Vaswani, N., Roy-Chowdhury, A.K., Chellappa, R.: Shape activity: a continuous-state hmm for moving/deforming shapes with application to abnormal activity detection. IEEE Trans. Image Process. **14**(10), 1603–1616 (2005)
18. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. Vis. Comput. **29**(10), 983–1009 (2012)
19. Zhang, D., Lu, G.: Shape-based image retrieval using generic fourier descriptor. Signal Process. Image Commun. **17**(10), 825–848 (2002)
20. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. II-819-II-826, June 2004