A Combination of Moment Descriptors, Fourier Transform and Matching Measures for Action Recognition Based on Shape

Katarzyna Gościewska[0000−0002−6726−2174] and Dariusz Frejlichowski[0000−0002−8051−476X]

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

Abstract. This paper presents an approach for human action recognition based on shape analysis. The purpose of the approach is to classify simple actions by applying shape descriptors to sequences of binary silhouettes. The recognition process consists of several main stages: shape representation, action sequence representation and action sequence classification. Firstly, each shape is represented using a selected shape descriptor. Secondly, shape descriptors of each sequence are matched, matching values are put into a vector and transformed into final action representation—we employ Fourier transform-based methods to obtain action representations equal in size. A classification into eight classes is performed using leave-one-out cross-validation and template matching approaches. We present results of the experiments on classification accuracy using moment-based shape descriptors (Zernike Moments, Moment Invariants and Contour Sequence Moments) and three matching measures (Euclidean distance, correlation coefficient and C1 correlation). Different combinations of the above-mentioned algorithms are examined in order to indicate the most effective one. The experiments show that satisfactory results are obtained when low-order Zernike Moments are used for shape representation and absolute values of Fourier transform are applied to represent action sequences. Moreover, the selection of matching technique strongly influences final classification results.

Keywords: Action recognition · Silhouette sequences · Shape descriptors

1 Introduction

An automatic recognition of human movements has gained popularity in recent years due to its wide range of applications, especially related to surveillance systems and human-computer interaction. Other applications include quality-of-life improvement for elderly care, sports analytics, and video retrieval and annotation. This implies a diversity of data and a need for different solutions. Human

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

action can be defined as a sequence of elementary movements that is clearly identifiable by the observer. Combinations of elementary movements can create single (e.g. bending) or periodic (e.g. running) motion patterns [5]. An action is also defined as an activity composed of multiple gestures organized in time, and a gesture is an elementary movement of the body part [20]. To perform action recognition it is common to apply low-level features such as shape which is considered as a distinctive feature supporting accurate classification. Additionally, the order and repeatability of individual silhouettes can help distinguish between actions. Despite a few characteristic elements the recognition process is still a challenging task because of the variations in motion performance, personal differences, speed or duration of individual actions [17].

In this paper we propose an original combination of well-known methods and algorithms aimed to recognize actions based on information contained in a binary foreground masks that were extracted from consecutive video frames representing people performing simple actions. The novelty is accomplished by creating a synthesis of some already approved methods and by testing existing knowledge in a different manner. The proposed approach is applied on coarsely classified sequences. Then the recognition is performed in each subgroup separately using the same procedure composed of three main steps: single shape representation, single action representation and action classification.

The rest of the paper is organized as follows: Section 2 presents several related works on action classification based on shape features. The proposed approach is explained in detail in Section 3 and some methods are presented in Section 4. Section 5 defines experimental conditions and presents the results of the experiments carried out with the use of three moment-based shape descriptors, namely Zernike Moments, Moment Invariants and Contour Sequence Moments. Section 6 summarizes the paper.

2 Related Works

This section describes several methods that are similar to our approach due to the use of shape features and similar input data. We focus on a shape-based action recognition that is classified in [20] as a non-hierarchical approach. This category covers the recognition of short and primitive actions. To recognize such actions, we can use solutions based on space-time volume, like this presented in [4]. The proposed approach generates motion energy images (MEI) to show where the movement is, and motion history images (MHI) to show how the object is moving. Then Hu moments are extracted from MEI and MHI, and the resultant action descriptors are matched using Mahalanobis distance. Hu moments are statistical descriptors which are scale and translation invariant, and allow for good shape discrimination. Another popular space-time volume technique is proposed in [10]. It accumulates silhouettes into space-time cubes (3D representation) and employs a Poisson equation to extract features of human actions, among which are local space-time saliency, action dynamics, shape structure and orientation. Space-time volume is also a global approach—the lo-

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

calized foreground region of interest is encoded as a whole and much of the information is carried. Popular holistic representations are based on silhouettes, edges or optical flow [17]. Among these a silhouette is our most interest. In [12, 14] shape features are calculated for each object separately and objects are not accumulated. The authors of [12] introduced new feature extraction techniques based on Trace transform, namely History Trace Templates and History Triple Features. In the first method, Trace transform is applied to binary silhouettes. The resultant transforms are composed into final history template that represents the whole action sequence and contains much of the spatial-temporal features. In the second method, Trace transform is used to construct a set of features that are invariant to translation, rotation and scaling, as well as robust to noise. Features are calculated for every video frame separately. Ultimately, LDA is applied to reduce dimensionality of final representations. In turn, in [14] every silhouette is transformed into time series and each of these is converted into the symbolic vector—a SAX representation. A set of all vectors represents an action and is called a SAX-Shape.

Action recognition can be performed using only some silhouettes extracted from selected video frames, so called key poses, e.g. [2, 16, 7]. The authors of [2] introduce a shape representation and matching technique that represents each key pose as a collection of line-pairs and can estimate similarity between two frames. A k-medoids clustering algorithm and learning algorithm are used to extract candidate key poses. During the classification process every frame is compared with all key poses in order to assign a label. Then majority voting is used to classify action sequences. Another solution using key poses is presented in [16]. The authors proposed extensive pyramidal features (EPFs) to describe poses. EPFs include Gabor, Gaussian and wavelet pyramids. AdaBoost algorithm is used to learn a subset of discriminative poses. Actions are classified with a new classifier—weighted local naive Bayes nearest neighbour. In [7] the proposed method uses the distance between all contour points and silhouette's centre of gravity to represent a pose. Then, K-means clustering with Euclidean distance is applied to learn key poses and Dynamic Time Warping is used to classify sequences of key poses.

Another solution for action recognition is a fusion of multiple features. The authors of [1] proposed a new algorithm based on Aligned Motion Images (AMIs), where each AMI is a single image that represents the motion of all frames of a single video. Two features are combined—Derivatives of Chord-Distance Signature based on contour and Histogram of Oriented Gradients which capture visual components of a silhouette's region. Action classification is performed using K-Nearest Neighbour and Support Vector Machine (SVM). Another approach based on contour and shape features is presented in [21]. It combines information obtained from the R-transform and averaged energy silhouette images which are used to generate feature vectors based on edge distribution of gradients and directional pixels. Classification is carried out with the use of multi-class SVM.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

3 The Proposed Approach

The proposed approach is composed of selected methods and algorithms, among which are: shape description algorithms based on moments, signal processing algorithms based on Fourier transform as well as distance and correlation-based matching measures. The selection of methods results from the continuation of the works presented in [11], where we have also tested moment-based descriptors using following procedure: each silhouette was represented using selected shape descriptor, shape representations were matched using Euclidean distance and normalized matching values were put into a vector called distance vector. To obtain final sequence representations all distance vectors were transformed using Fast Fourier Transform and periodogram. Classification process was performed iteratively using template matching approach and k-fold cross-validation. In each iteration the database was divided into templates (class representatives) and test objects. Each test object was matched with all templates to indicate the most probable class. Final classification accuracy was an average of all iterations.

In this paper the results of new experiments are given. When compared with previous work there are some significant differences. Firstly, the experimental database consists of eight instead of five classes, four classes for each subgroup. Matching process is performed using three various measures instead of one. This applies to both comparison of shape descriptors and classification of final action representations. Our previous experiments have shown that the accuracy depends on applied matching measure. Secondly, final action representations are prepared using three various methods instead of one. In case of classification process, the leave-one-out cross-validation with template matching approach is applied instead of k-fold cross-validation technique. This is done to avoid a situation in which a set of class representatives affects classification accuracy. Moreover, a coarse classification step has been added.

Here we focus on testing various combinations of several algorithms in order to select relevant features for action description. Therefore, the proposed approach has a form of a general procedure composed of consecutive data processing steps which are:

Step 1. Data preparation

The proposed approach bases on binary silhouettes. We use the Weizmann [3] database which is composed of action sequences—one action sequence is represented by a set of frames from which foreground binary masks are extracted. Each foreground mask contains one silhouette. The dataset is divided into two subgroups based on the centroid trajectory—actions performed in place (a trajectory is very short) and actions with changing location of a silhouette (longer trajectory). Then the approach is applied in each subgroup separately. Let us denote each input action sequence as a set of binary masks $BM_i = \{bm_1, bm_2, ..., bm_n\}$, where n is the number of frames in a particular sequence.

Step 2. Single shape representation

In the next step, we take each bm_i and represent it using selected shape description algorithm. Various methods can be applied and here we examine Zernike Moments, Moments Invariants and Contour Sequence Moments (see Section 4.1).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

In result, we obtain a set of shape descriptors for each action sequence which can be denoted as $SD_i = \{sd_1, sd_2, ..., sd_n\}$. The number of descriptors equals the number of frames. A sd_i can be a matrix or a vector, depending on the applied shape descriptor.

Step 3. Single action representation

Action representation is based on the calculation of similarity or dissimilarity measures for each SD_i separately. We can use various solutions, such as Euclidean distance, correlation coefficient and C1 correlation (see Section 4.2). The shape descriptor of a first frame sd_1 is matched with the rest of descriptors and matching values are put into a vector $MD_i = \{md_1, md_2, ..., md_{n-1}\}$. A sd_1 is not matched with itself therefore we obtain one element less. Here, for instance, md_1 is a matching value calculated using sd_1 and sd_2. All MD vectors are normalized and transformed into frequency domain using periodogram or Fast Fourier Transform algorithm (a magnitude is taken). Each transformed vector creates one-dimensional descriptor of a sequence—a final action representation AR. The transformation into frequency domain makes all representations equal in size—we use a predefined number of elements which exceeds the number of frames in the longest video sequences. Moreover, the resultant transforms reveal some hidden periodicities in the data.

Step 4. Classification

AR vectors are classified based on the leave-one-out cross-validation process and template matching technique. 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 test object, e.g. we take AR_1 and match it with the rest of AR vectors using methods explained in Section 4.2. The percentage of correctly classified actions gives classification accuracy.

4 Shape Description and Matching

4.1 Shape Description Algorithms Based on Moments

The Zernike Moments are derived using Zernike orthogonal polynomials and the formula below [22]:

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

where $R_{nm}(r)$ is the orthogonal radial polynomial [22]:

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

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

The Zernike Moments of order n and repetition m of a region shape $f(x, y)$ are calculated by means of this formula [22]:

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

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

According to [18, 13, 8], to obtain Moment Invariants, the general geometrical moments are firstly calculated using the following formula:

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

The $f(x,y)$ function value is equal to 1 for pixels belonging to an object and 0 for background pixels. The representation is invariant to translation thanks to the use of centroid, which is calculated as follows:

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

Then, Central Moments are calculated using the centroid:

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

In turn, the invariance to scaling is obtained by central normalized moments:

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

Finally, Moment Invariants are derived (seven first values):

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

Based on [19], the calculation of Contour Sequence Moments starts from representing a contour as ordered sequence $z(i)$ which elements are the Euclidean distances from the centroid to N contour points. Then, one-dimensional normalized contour sequence moments are derived as follows:

$$m_r = \frac{1}{N} \sum_{i=1}^{N} [z(i)]^r, \quad \mu_r = \frac{1}{N} \sum_{i=1}^{N} [z(i) - m_1]^r. \tag{9}$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

The r-th normalized contour sequence moment and normalized central sequence moment are calculated using the following formulas:

$$\bar{m_r} = \frac{m_r}{(\mu_2)^{r/2}}, \quad \bar{\mu_r} = \frac{\mu_r}{(\mu_2)^{r/2}}. \tag{10}$$

The final shape description consists of four values:

$$F_1 = \frac{(\mu_2)^{1/2}}{m_1}, \quad F_2 = \frac{\mu_3}{(\mu_2)^{3/2}}, \quad F_3 = \frac{\mu_4}{(\mu_2)^2}, \quad F_4 = \bar{\mu_5}. \tag{11}$$

4.2 Similarity and Dissimilarity Measures

For matching we have selected standard Euclidean distance as a dissimilarity measure and two correlations measuring similarity—correlation coefficient based on Pearson's correlation and C1 correlation based on L1-norm (introduced in [6]).

Let us take two exemplary vectors $V_A(a_1, a_2, \ldots, A_N)$ and $V_B(b_1, b_2, \ldots, B_N)$ which represent object A and object B in a N-dimensional feature space. The Euclidean distance d_E between these two vectors is defined by means of the following formula [15]:

$$d_E(V_A, V_B) = \sqrt{\sum_{i=1}^{N}(a_i - b_i)^2}. \tag{12}$$

The correlation coefficient may be calculated both for the matrix and vector representations of a shape. The correlation between two matrices can be derived using the formula [9]:

$$c_c = \frac{\sum_m \sum_n (A_{nm} - \bar{A})(B_{nm} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{nm} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{nm} - \bar{B})^2\right)}}, \tag{13}$$

where:
A_{mn}, B_{mn}—pixel value with coordinates (m, n), respectively in image A and B,
$\bar{A}$, $\bar{B}$—average value of all pixels, respectively in image A and B.

The C1 correlation is also a similarity measure based on shape correlation. It is obtained by means of the following formula [6]:

$$c_1(A, B) = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}|a_{ij} - b_{ij}|}{\sum_{i=1}^{H}\sum_{j=1}^{W}(|a_{ij}| - |b_{ij}|)}, \tag{14}$$

where:
A, B—matched shape representations,
H, W—height and width of the representation.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

5 Experiments and Results

5.1 Data and Conditions

The experiments were carried out with the use of the Weizmann dataset [3]. The original database consists of 90 video sequences (144×180 px) recorded at 50 fps. The video sequences are very short, each lasting up to several seconds and differing in the number of frames. We have selected eight action types: 'bend', 'jumping jack', 'jump forward on two legs', 'jump in place on two legs', 'run', 'skip', 'walk' and 'wave one hand'. Exemplary frames from selected video sequences representing actions performed in place are depicted in Fig. 1. Fig. 2 shows exemplary frames representing actions with changing location of a silhouette. Binary masks corresponding to all video sequences in the database are available and were used as input data (see Fig. 3 for examples).

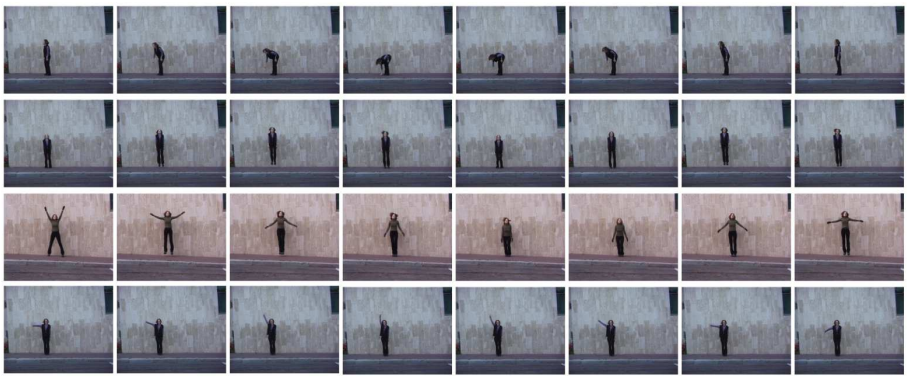

Fig. 1. Exemplary video frames representing actions performed in place (in rows): 'bend', 'jump in place', 'jumping jack' and 'wave one hand' respectively (based on [3]).

The aim of the experiments was to indicate the best result by means of the highest classification accuracy. Action sequences are coarsely classified into two subgroups: actions performed in place ('bend', 'jumping jack', 'jump in place on two legs', 'wave one hand') and actions with changing location of a silhouette ('jump forward on two legs', 'run', 'skip', 'walk'). The following procedure is performed in each subgroup separately. As a single experiment we assume the use of our approach for one shape description algorithm within several tests employing various action representations as well as different matching measures. Thanks to this we can indicate which methods should be used for a specific shape descriptor. Here shapes are represented using Zernike Moments, Moment Invariants or Contour Sequence Moments. Vectors with matching values are transformed into sequence representations using periodogram or Fast Fourier Transform. Both shape descriptors and sequence representations are matched using Euclidean distance, correlation coefficient or C1 correlation. The classification step is based on

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

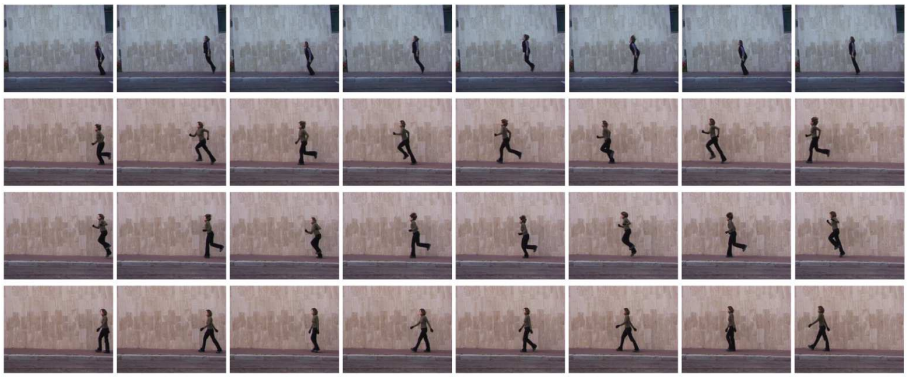

Fig. 2. Exemplary video frames representing actions with changing location of a silhouette (in rows): 'jump', 'run', 'skip' and 'walk' respectively (based on [3]).

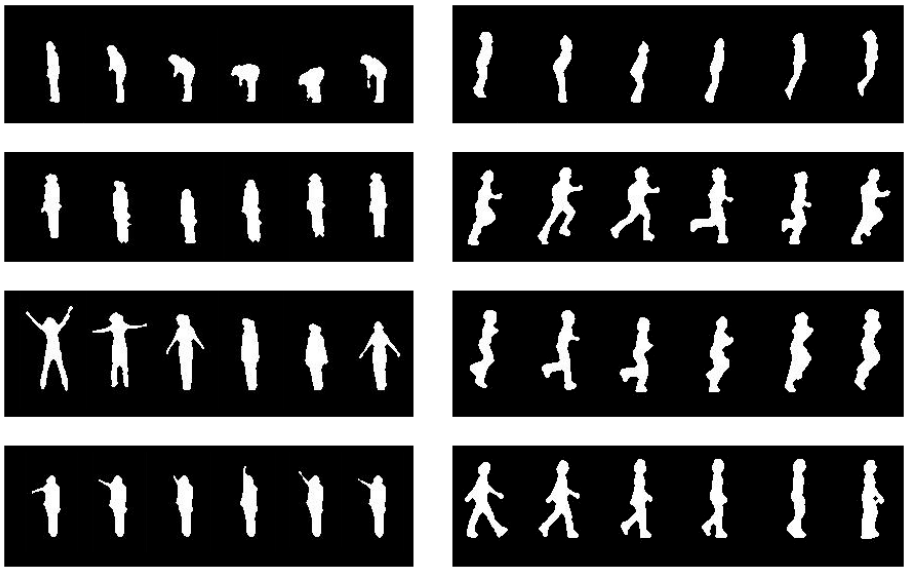

Fig. 3. Exemplary binary masks from the database—left column corresponds to actions presented in Fig. 1 (in rows): 'bend', 'jump in place', 'jumping jack' and 'wave one hand' respectively, and right column corresponds to actions presented in Fig. 2 (in rows): 'jump', 'run', 'skip' and 'walk' respectively (based on [3]).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

the leave-one-out cross-validation and template matching. We take each action representation and match it with the rest of representations. Then we select the most similar one which indicates the probable action class. The percentage of correct classifications, averaged for both subgroups, gives final accuracy.

Additionally, for some shape description algorithms it is possible to calculate shape representations of different size, e.g. Zernike Moments of orders from 1st to 12th with representation size varying from 2 to 49 values. We have performed a set of experiments to select the order of Zernike Moments that gives the highest averaged accuracy. The results for orders from 1st to 12th are as follows: 71%, 71%, 73.04%, 63.31%, 64.97%, 62.59%, 64.06%, 60.03%, 66.25%, 62.22%, 70.10% and 61.84% respectively. The highest averaged accuracy can be obtained using moments of 3rd order. Moreover, Zernike Moments of 1st and 2nd order also give good results. Therefore, only these orders are considered for shape representation during experiments.

5.2 Results

The experimental results were grouped according to the applied shape description algorithm. Therefore, we can indicate which combination of techniques is the most effective when a specific shape representation is employed. Table 1 presents the averaged results for the experiment using Zernike Moments of 3rd order. The highest accuracy is 73.04% and is obtained when silhouette descriptors are matched using Euclidean distance, sequence representation is prepared using Fast Fourier Transform and final representations are matched using C1 correlation.

Table 1. Averaged classification accuracy for Zernike Moments of the 3rd order.

Zernike Moments		*Silhouettes matched by:* Euclidean distance	Correlation Coefficient	C1 correlation
Sequences matched by:				
FFT magnitude and periodogram	+ Correlation Coefficient	46.42%	44.76%	61.84%
	+ C1 correlation	48.83%	37.97%	44.23%
	+ Euclidean distance	44.80%	35.60%	45.51%
periodogram only	+ Correlation Coefficient	36.16%	46.04%	49.36%
	+ C1 correlation	58.18%	23.45%	59.28%
	+ Euclidean distance	44.42%	25.11%	53.39%
FFT magnitude only	+ Correlation Coefficient	53.96%	43.67%	47.89%
	+ C1 correlation	**73.04%**	37.41%	50.64%
	+ Euclidean distance	63.88%	36.12%	52.11%

In case of Moment Invariants (see Table 2) the averaged accuracy reached 65.35% in the experiment using Euclidean distance for shape matching, FFT for

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

action representation and correlation coefficient for action matching. The use of Contour Sequence Moments (see Table 3) did not exceed 52%.

Table 2. Averaged classification accuracy for Moment Invariants.

Moment Invariants		*Silhouettes matched by:* Euclidean distance	Correlation Coefficient	C1 correlation
Sequences matched by:				
FFT magnitude and periodogram	+ Correlation Coefficient	**65.52%**	53.96%	55.43%
	+ C1 correlation	48.45%	35.97%	51.92%
	+ Euclidean distance	45.70%	44.61%	56.33%
periodogram only	+ Correlation Coefficient	50.11%	49.36%	50.83%
	+ C1 correlation	46.98%	44.42%	41.48%
	+ Euclidean distance	36.35%	40.57%	40.20%
FFT magnitude only	+ Correlation Coefficient	65.35%	51.40%	52.49%
	+ C1 correlation	61.31%	51.40%	53.77%
	+ Euclidean distance	59.84%	52.68%	49.74%

Table 3. Averaged classification accuracy for Contour Sequence Moments.

Contour Sequence Moments		*Silhouettes matched by:* Euclidean distance	Correlation Coefficient	C1 correlation
Sequences matched by:				
FFT magnitude and periodogram	+ Correlation Coefficient	44.04%	39.48%	45.32%
	+ C1 correlation	45.51%	33.41%	41.86%
	+ Euclidean distance	34.88%	36.16%	34.50%
periodogram only	+ Correlation Coefficient	41.67%	44.04%	42.95%
	+ C1 correlation	45.51%	42.38%	47.17%
	+ Euclidean distance	45.70%	38.35%	49.92%
FFT magnitude only	+ Correlation Coefficient	**51.58%**	42.57%	45.70%
	+ C1 correlation	40.01%	42.38%	48.64%
	+ Euclidean distance	47.17%	38.35%	47.36%

We should take a closer look at the results in subgroups separately. It turned out that it is advised to apply different algorithms in each subgroup. Table 4 contains results for several experiments in which classification accuracy exceeds 70%. Considering as small shape representation as possible we can indicate the use of Zernike Moments of 3rd order for actions performed in place and Zernike

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

Moments of 1st order for the other subgroup. In addition, we present classification quality measures for these two best experiments (see Table 5), including standard precision and recall for each class.

Table 4. Results for the experiments with accuracy exceeding 70%.

Shape descriptor	Shape matching	Action representation	Action matching	Averaged accuracy	Actions with changing location	Actions performed in place
Moment invariants	Euclidean distance	FFT magnitude only	Correlation Coefficient	65.35%	51.28%	**79.41%**
Zernike Moments (3rd order)	Euclidean distance	FFT magnitude only	C1 Correlation	**73.04%**	66.67%	**79.41%**
Zernike Moments (1st order)	Euclidean distance	FFT magnitude only	C1 Correlation	71.00%	**74.36%**	67.65%

Table 5. Classification quality measures for the best experiments.

Shape descriptor	Subgroup	Class	Precision	Recall
Zernike Moments (1st order)	Actions with changing location	'jump forward'	0.71	0.56
		'run'	0.50	0.70
		'skip'	0.67	0.60
		'walk'	1.00	0.90
Zernike Moments (3rd order)	Actions performed in place	'bend'	0.90	1.00
		'jumping jack'	0.67	0.75
		'jump in place'	0.64	0.78
		'wave one hand'	1.00	0.5

In Section 3 we have listed several changes introduced in our approach, and based on the experimental results we can conclude that the selection of matching measure strongly affects the accuracy. Secondly, an additional coarse classification step increased overall classification quality despite more action classes. Thirdly, the application of a new experimental procedure helped to avoid a situation in which the results depend on the set of templates (class representatives).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

However, there are some classes which are not classified precisely. For instance, running is often confused with jumping or skipping. To improve the results, using a different shape descriptor can be considered or adding another feature that will distinguish between problematic actions.

6 Conclusions

The paper covered the topic of action recognition based on shape features. The presented approach uses binary silhouettes, shape description algorithms and matching techniques to classify action sequences. We represent each silhouette using selected shape descriptor, match all descriptors of a single sequence and put matching values into a vector. Then we transform each vector into frequency domain and classify. We use additional step of coarse classification based on centroid location and perform experiments in each subgroup separately. We have experimentally tested various combinations of the following: shape description algorithms (Zernike Moments, Moment Invariants and Contour Sequence Moments), matching measures (Euclidean distance, correlation coefficient and C1 correlation) and frequency domain techniques (Fast Fourier Transform and periodogram).

The best results are obtained when we use a combination of Zernike Moments for shape representation, Euclidean distance for shape matching, Fast Fourier Transform for action representation and C1 correlation for action classification. The highest averaged accuracy was 73.04% for Zernike Moments of 3rd order—79.41% for actions performed in place and 66.67% for actions with changing location. Moreover, for the second subgroup better results can be obtained by using the Zernike Moments of 1st order instead of 3rd order. Then accuracy equals 74.36% and shape representation is smaller.

References

1. Al-Ali, S., Milanova, M., Al-Rizzo, H., Fox, V.L.: Human Action Recognition: Contour-Based and Silhouette-Based Approaches, pp. 11–47. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-11430-92
2. Baysal, S., Kurt, M.C., Duygulu, P.: Recognizing human actions using key poses. In: 2010 20th International Conference on Pattern Recognition. pp. 1727–1730 (2010). https://doi.org/10.1109/ICPR.2010.427
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2. pp. 1395–1402. ICCV '05, IEEE Computer Society, Washington, DC, USA (2005). https://doi.org/10.1109/ICCV.2005.28
4. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3), 257–267 (2001). https://doi.org/10.1109/34.910878
5. Borges, P.V.K., Conci, N., Cavallaro, A.: Video-based human behavior understanding: A survey. IEEE Transactions on Circuits and Systems for Video Technology **23**(11), 1993–2008 (2013). https://doi.org/10.1109/TCSVT.2013.2270402

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28

6. Brunelli, R., Messelodi, S.: Robust estimation of correlation with applications to computer vision. Pattern Recognition **28**(6), 833–841 (1995). https://doi.org/10.1016/0031-3203(94)00170-Q
7. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters **34**(15), 1799–1807 (2013). https://doi.org/10.1016/j.patrec.2013.01.021
8. Che-Bin Liu, Ahuja, N.: Vision based fire detection. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. vol. 4, pp. 134–137 (2004). https://doi.org/10.1109/ICPR.2004.1333722
9. Chwastek, T., Mikrut, S.: The problem of automatic measurement of fiducial mark on air images (in polish). Archives of Photogrammetry, Cartography and Remote Sensing **16**, 125–133 (2006)
10. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12), 2247–2253 (2007). https://doi.org/10.1109/TPAMI.2007.70711
11. Gościewska, K., Frejlichowski, D.: Moment shape descriptors applied for action recognition in video sequences. In: Nguyen, N.T., Tojo, S., Nguyen, L.M., Trawiński, B. (eds.) Intelligent Information and Database Systems. pp. 197–206. Springer International Publishing, Cham (2017)
12. Goudelis, G., Karpouzis, K., Kollias, S.: Exploring trace transform for robust human action recognition. Pattern Recognition **46**(12), 3238–3248 (2013). https://doi.org/10.1016/j.patcog.2013.06.006
13. Hupkens, T., de Clippeleir, J.: Noise and intensity invariant moments. Pattern Recognition Letters **16**(4), 371–376 (1995). https://doi.org/10.1016/0167-8655(94)00110-O
14. Junejo, I.N., Junejo, K.N., Aghbari, Z.A.: Silhouette-based human action recognition using sax-shapes. The Visual Computer **30**(3), 259–269 (2014). https://doi.org/10.1007/s00371-013-0842-0
15. Kpalma, K., Ronsin, J.: An overview of advances of pattern recognition systems in computer vision. In: Obinata, G., Dutta, A. (eds.) Vision Systems, chap. 10. IntechOpen, Rijeka (2007). https://doi.org/10.5772/4960
16. Liu, L., Shao, L., Zhen, X., Li, X.: Learning discriminative key poses for action recognition. IEEE Transactions on Cybernetics **43**(6), 1860–1870 (2013). https://doi.org/10.1109/TSMCB.2012.2231959
17. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6), 976–990 (2010). https://doi.org/10.1016/j.imavis.2009.11.014
18. Rothe, I., Susse, H., Voss, K.: The method of normalization to determine invariants. IEEE Transactions on Pattern Analysis and Machine Intelligence **18**(4), 366–376 (1996). https://doi.org/10.1109/34.491618
19. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. Thomson-Engineering (2007)
20. Vishwakarma, Sarveshand Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer **29**(10), 983–1009 (2013). https://doi.org/10.1007/s00371-012-0752-6
21. Vishwakarma, D., Dhiman, A., Maheshwari, R., Kapoor, R.: Human motion analysis by fusion of silhouette orientation and shape features. Procedia Computer Science **57**, 438–447 (2015). https://doi.org/10.1016/j.procs.2015.07.515
22. Zhang, D., Lu, G.: Shape-based image retrieval using generic Fourier descriptor. Signal Processing: Image Communication **17**(10), 825–848 (2002). https://doi.org/10.1016/S0923-5965(02)00084-X

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_28