Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 1083–1090

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Recognizing human actions with multiple Fourier transforms

Katarzyna Gościewska*, Dariusz Frejlichowski

West Pomeranian University of Technology, Szczecin, Faculty of Computer Science and Information Technology, Żołnierska 52, 71-210, Szczecin, Poland

Abstract

In this paper we present an approach for action recognition that uses various applications of Fourier transform. The main idea is to classify video sequences based on action representations obtained using shape descriptors. For shape representation we use the Two-Dimensional Fourier Descriptor, Generic Fourier Descriptor and UNL-Fourier Descriptor. For each sequence of binary silhouettes we derive a set of shape descriptors and match the descriptor of the first frame with the rest of descriptors to obtain a vector of similarities (correlation coefficients or C1 correlations) or dissimilarities (Euclidean distances). Then normalized vectors are transformed into action representations using discrete Fourier transform, power spectral density estimate or a combination of both. Classification is performed using leave-one-out cross-validation and various matching measures. Additionally, we incorporate a coarse classification step that distinguish between actions performed in place and actions with changing location of a silhouette. Extensive experiments are carried out to investigate all possible combinations of selected processing steps, and experimental results are promising.

Keywords: Fourier transform; Action recognition; Shape descriptors;

1. Introduction

Action recognition is a part of a human activity recognition system that succeeds motion segmentation, object classification and human tracking, and precedes semantic behaviour description [24]. Based on this structure we assume having foreground objects extracted and located on the consecutive video frames that are input data for action recognition. Object's features may vary from single characteristic point (centre of gravity) or a set of points (a contour) to an entire shape region [27]. An action is defined as a simple and primitive movement performed by a single person during a short period of time [22], such as running or bending. Exemplary frames representing various actions are depicted in Fig. 1.

* Corresponding author. Tel.: +48-91-449-55-63.
E-mail address: kgosciewska@wi.zut.edu.pl

10.1016/j.procs.2020.09.104

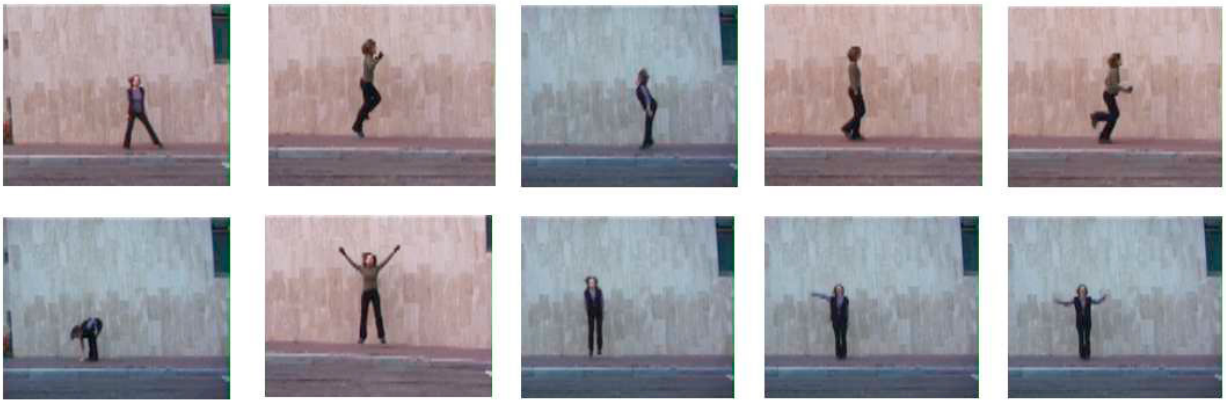

Fig. 1. Exemplary frames of action sequences from the original Weizmann database [3]—'gallop sideways', 'skip', 'jump forward', 'walk' and 'run' in the first row; 'bend', 'jumping jack', 'jump in place', 'wave one hand' and 'wave two hands' in the second row.

In this paper we propose an approach for action recognition that can be classified as a non-hierarchical approach based on holistic and local features. We focus on diverse applications of Fourier transform—in shape description algorithms and calculation of action representations. Therefore, we examine the Two-Dimensional Fourier Descriptor (2DFD) [17], Generic Fourier Descriptor (GFD) [8], UNL-Fourier Descriptor (UNL-F) [19], discrete Fourier transform (DFT) and power spectral density estimate (periodogram) [20]. The Fourier Transform is valued and widely used in the area of object recognition and image retrieval [28, 26]. The use of spectral domain features is popular in human activity recognition as well, e.g. [25, 15].

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 description, single action representation and action classification. The rest of the paper is organized as follows: Section 2 presents some related works on action recognition based on shape, Section 3 explains the proposed approach, Section 4 contains experimental conditions and results, and Section 5 summarizes the paper.

2. Related works

The recognition of simple and primitive actions based on shape has been investigated using features of various complexity and dimensionality. Among the most popular ones are solutions that accumulate all action silhouettes into single 2D [4] or 3D [12] representations. The authors of [4] proposed an approach, based on binary masks, that represents movement as images, so called motion energy images and motion history images. Then matching is performed using Hu moments and Mahalanobis distance. The approach presented in [12] combines silhouettes into 3D space-time cubes and extracts some characteristic features for classification, such as action dynamics, shape structure and orientation. The authors of [13] proposed to accumulate skeletons instead of silhouettes, extract salient regions and classify them using SVM classifier. The opposite manner of treating action data is to base solely on selected key poses (e.g. [18, 6]) or selected features calculated for particular silhouettes (e.g. a combination of contour-based and silhouette-based features [1] or pose orientation combined with shape features [23]).

According to [14], a good representation of action features should provide a description for a sufficiently large number of classes, reflect the similarity of two similar actions as well as be easy to calculate and invariant. It is believed that Fourier transform-based features can provide all of that, and some researchers have already applied these features for action classification. For instance, the authors of [25] extract some specific features from hand/face region in time domain, and then apply Fourier transform to generate new set of features that separates low and high frequencies. Each video sequence is represented using predefined number of selected Fourier features. In [15] the discrete Fourier transform is applied to images. During a training stage, motion image features are selected and a focus is put on high amplitude Fourier coefficients. Feature dimensionality is reduced using the Principal Component Analysis and action classification is performed using a Support Vector Machine-based approach. The authors of [2] use the trajectories of spatio-temporal points as motion features and represent them using Fourier transform coefficients. A combination of optical flow and Fourier features is also presented in [21].

3. Proposed approach

In this section we present our approach to action recognition. It has a form of a general procedure composed of several processing steps. In the first step all shapes are represented using shape descriptors (Section 3.1). In the second step those descriptors are combined into action representations (Section 3.2), which are used for action classification in the third step (Section 3.3). In each step there is a group of methods to choose from which makes it possible to adapt the approach to different applications. We can distinguish two main groups of methods. In the first group there are Fourier transform-based methods applied to calculate shape and action representations, and the second group consists of matching measures used to calculate similarity or dissimilarity between shape or action representations. From a computational point of view, we have a Discrete Fourier Transform that can be obtained using Fast Fourier Transform algorithm. Nowadays DFT implementations, such as FFTW library [11], can handle arbitrary size transforms and vectors being transformed do not need to have a length equal to the power of 2. The application of DFT, to the image or signal, is simple and efficient. Spectral analysis allows the disclosure of additional information and the reduction of data by selecting only part of the Fourier coefficients, which are easy to match as well.

For our approach we assume that an action sequence is a set of foreground binary masks, each containing a single human silhouette (see Fig. 2 for exemplary masks), which can be denoted as $BM_i = \{bm_1, bm_2, ..., bm_n\}$, where n is the number of frames in a particular sequence. All sequences in the database are divided into two general classes—actions performed in place and actions with changing location of a silhouette. This division is made based on the length of centroid trajectory. Thank to coarse classification the recognition is performed in smaller subgroups separately.

Fig. 2. Exemplary foreground masks from the Weizmann database [3].

3.1. Single shape representation

Firstly, each binary mask bm_i is represented using selected shape description algorithm. Therefore, for each action sequence we obtain a set of shape descriptors which number is equal to the number of frames. A set of descriptors can be denoted as $SD_i = \{sd_1, sd_2, ..., sd_n\}$ and each sd_i is a matrix or a vector.

In this paper we investigate three shape description algorithms (2DFD, GFD and UNL-F) that utilize DFT for images (the formula is given below [17]):

$$C(k,l) = \frac{1}{HW}\left|\sum_{h=1}^{H}\sum_{w=1}^{W} P(h,w)\cdot \exp^{(-i\frac{2\pi}{H}(k-1)(h-1))}\cdot \exp^{(-i\frac{2\pi}{W}(l-1)(w-1))}\right|, \tag{1}$$

where:
H, W—height and width of the image in pixels,
k—sampling rate in vertical direction ($k \geq 1$ and $k \leq H$),
l—sampling rate in horizontal direction ($l \geq 1$ and $l \leq W$),
$C(k,l)$—value of the coefficient of discrete Fourier transform in the coefficient matrix in k row and l column,
$P(h,w)$—value in the image plane with coordinates h, w.

The above transform is applied to binary images containing different amount of information about the foreground object. In case of the 2DFD [17], it is applied to a region shape in Cartesian coordinates. The resultant spectrum has a size equal to the image size. The GFD [8] employs Fourier transform after a binary image with a region shape is transformed into polar coordinates. In turn, the UNL-F [19, 9] is a contour-based shape descriptor, which utilizes complex representation of contour Cartesian coordinates. These are then transformed into polar coordinates and projected onto Cartesian image to allow the application of the two-dimensional Fourier transform. The two-dimensional DFT produces a matrix with Fourier spectrum, containing low frequencies on the corners and high frequencies in the middle. For further processing we use only a part of the absolute spectrum—2×2, 5×5, 10×10, 25×25 or 50×50 sub-part taken from the upper left corner—and we concatenate these sub-matrices into vectors. Low frequency components carry the largest part of the original information about the shape.

3.2. Single action representation

For each action sequence from the previous step we have a set of shape descriptors but we do not use them directly. Instead, we match the descriptor of the first frame sd_1 with the rest of the descriptors within a particular sequence using matching measures explained in Section 3.3. The resultant similarity or dissimilarity values are put into a vector $MD_i = \{md_1, md_2, ..., md_{n-1}\}$ and normalized. For instance, md_1 is a matching value calculated using sd_1 and sd_2 (a sd_1 is not matched with itself). Due to various number of frames in the database, all vectors are transformed into frequency domain, and then each transformed vector is a one-dimensional action representation AR.

Such transformation makes all representations equal in size. The following methods are applied: periodogram, DFT with zero padding or a combination of both. Zero padding is a method of extending a signal with zeros what is equivalent to appending zeros to the end of the original signal. The DFT of the zero-padded signal corresponds to spectral interpolation [20]. By using FFT algorithm with a predefined signal length we can truncate an input vector as well. In our experiments we declare the number of Fourier coefficients equal to the nearest power of 2, and we perform tests for other lengths as well. Periodogram is a type of the DFT that estimates the spectral density of a signal. It can help to reveal some periodicities in the data. The number of resultant coefficients is the larger of 256 or the nearest power of 2, only if periodogram is applied for complex numbers (double-sided spectrum). If it is not, and we use real values as an input instead, we obtain one-sided spectrum and the length of the resultant vector is equal to half the double-sided spectrum increased by 1.

3.3. Classification

All action representations (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 [10]. The percentage of correctly classified actions gives classification accuracy.

During the classification process we examine three matching measures in order to indicate the solution which appropriately captures differences between classes. The first is a standard Euclidean distance for two vectors, $V_A(a_1, a_2, \ldots, A_N)$ and $V_B(b_1, b_2, \ldots, B_N)$, that is obtained using a following formula [16]:

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

The second technique is correlation coefficient based on Pearson's correlation, which for two exemplary objects, A and B, can be derived using the formula [7]:

$$c_c = \frac{\sum_m \sum_n (A_{nm} - \bar{A})(B_{nm} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{nm} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{nm} - \bar{B})^2\right)}}, \tag{3}$$

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

Another correlation is a C1 correlation based on L1-norm (introduced in [5]). It is obtained by means of the following formula [5]:

$$c_1(A,B) = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\left|a_{ij} - b_{ij}\right|}{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(\left|a_{ij}\right| - \left|b_{ij}\right|\right)}, \tag{4}$$

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

4. Experimental results

The experiments were performed with the use of the Weizmann database [3], which consists of video sequences (144×180 px) recorded at 50 fps, and the corresponding foreground masks. Video sequences last up to several seconds and the number of frames varies from 26 to 127. We have selected four unique actions performed in place ('bend', 'jumping jack', 'jump in place on two legs', 'wave one hand') and four actions with changing location of a silhouette ('jump forward on two legs', 'run', 'skip', 'walk'). Selected frames representing these actions are depicted in Fig. 3 and Fig. 4. During an initial coarse classification the database is divided into two subgroups using centroid trajectory and the following steps of the approach are performed in each subgroup separately.

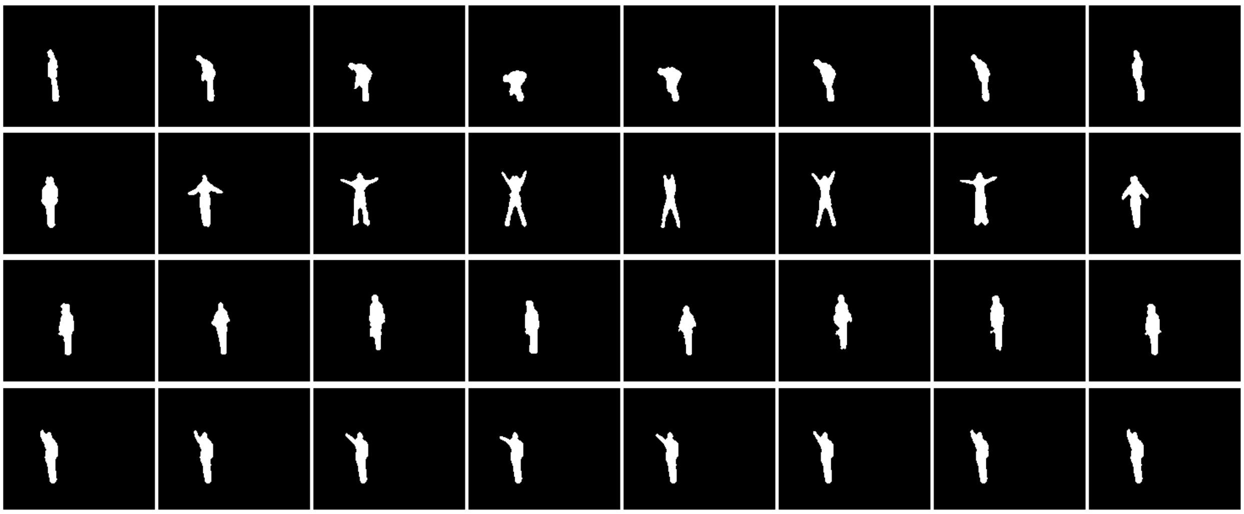

Fig. 3. Selected frames representing actions performed in places by one actor. In rows are: 'bend', 'jumping jack', 'jump in place on two legs', 'wave one hand' classes [3].

For each shape description algorithm, 135 tests were carried out which are a combination of a shape representation, action representation and matching measures applied both for shape matching and action classification. Following elements were used to prepare tests (the abbreviations used to tabulate the results are given in brackets):

1. Three shape descriptors: the Two-Dimensional Fourier Descriptor (2DFD), the Generic Fourier Descriptior (GFD) and the UNL-Fourier Descriptor (UNLF);
2. Five sizes of shape representations: 2×2, 5×5, 10×10, 25×25, 50×50;
3. Three types of action representations:

   (a) absolute spectrum of Fourier transform for zero-padded vectors (absFFT)—128 elements what equals to a nearest power of 2 for 127 frames;

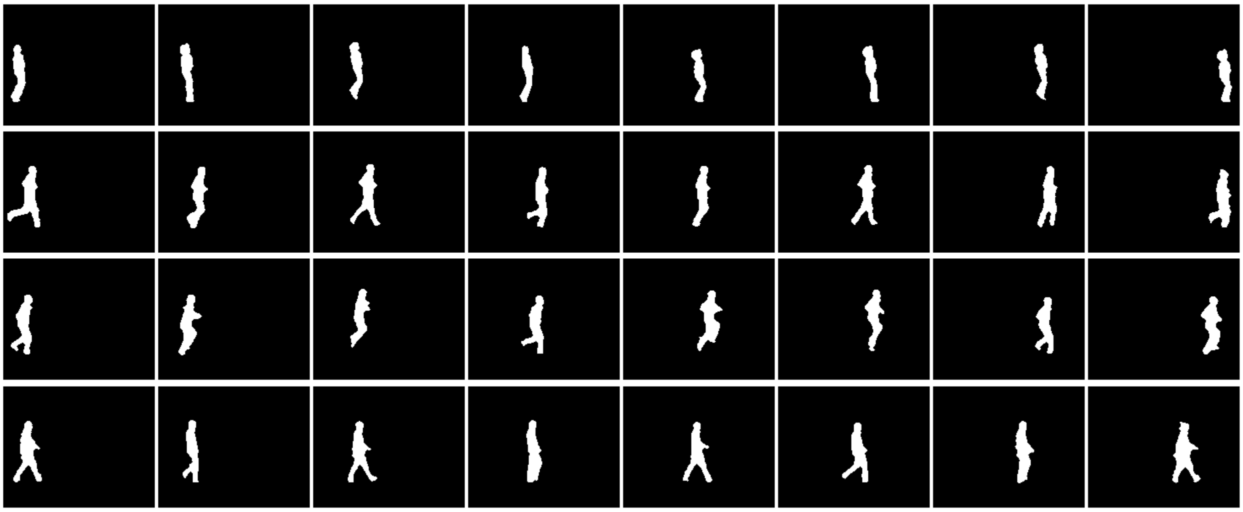

Fig. 4. Selected frames representing actions with changing location of a silhouette performed by one actor. In rows are: 'jump forward on two legs', 'run', 'skip', 'walk' classes [3].

(b) power spectral density (PSD) estimate—129 elements for one-sided spectrum;
(c) both combined (absFFT+PSD)—129 elements what is a result of obtaining absolute spectrum of Fourier transform for original vectors and applying a periodogram on it.

4. Three matching measures applied both for shape matching and for action representation matching—C1 correlation (C1), correlation coefficient (CC) and Euclidean distance (EU).

The results obtained for all shape descriptors are combined into three rankings. The first one contains results for the entire database (see Table 1), the second one for actions performed in place (see Table 2) and the third one for actions with changing location of a silhouette (see Table 3). Different combinations of methods and algorithms are sorted in the descending order based on accuracies and five best results are given. In case of the results for actions with changing location of a silhouette presented in Table 3, five top combinations yielded the same accuracy.

Table 1. Five best results for the entire database.

Shape descriptor	Classification accuracy	Descriptor size	Shape matching	Action representation	Sequence matching
GFD	67.72%	10×10	C1	absFFT	CC
GFD	67.35%	5×5	CC	absFFT	CC
2DFD	66.97%	10×10	CC	absFFT	CC
2DFD	66.06%	25×25	C1	absFFT	EU
GFD	65.87%	10×10	CC	absFFT	CC

The results show that similar accuracies were obtained in several tests, and some dependencies can be concluded:

1. Various combinations of methods can result in the same accuracy;
2. It is more effective to use different versions of the approach in each subgroup;
3. Correlations give better results than distance measure;
4. DFT with zero-padding is the most common representation.

Table 2. Five best results for the subgroup of actions performed in place.

Shape descriptor	Classification accuracy	Descriptor size	Shape matching	Action representation	Sequence matching
2DFD	76.47%	50×50	C1	absFFT	C1
GFD	76.47%	5×5	C1	absFFT	C1
GFD	76.47%	10×10	C1	absFFT	CC
2DFD	73.53%	5×5	EU	absFFT	C1
2DFD	73.53%	25×25	C1	absFFT	C1

Table 3. Five best results for the subgroup of actions with changing location of a silhouette.

Shape descriptor	Classification accuracy	Descriptor size	Shape matching	Action representation	Sequence matching
2DFD	69.23%	10×10	CC	absFFT	CC
2DFD	69.23%	10×10	C1	PSD	CC
GFD	69.23%	5×5	EU	absFFT+PSD	CC
UNLF	69.23%	10×10	CC	absFFT	CC
UNLF	69.23%	10×10	C1	PSD	CC

Due to the possibility of applying DFT to an arbitrary input size, we performed an additional experiment for the predefined vector lengths varying from 2 to 128. The experiment was performed for three best versions of the approach obtained for the entire database (see Table 1). In Table 4 we present the results indicating the size of the DFT vector with the highest accuracy. We can see from the Table that in most cases smaller number of Fourier coefficients is sufficient (less than 128) and the results are improved. In order to indicate the best results, the classification accuracy and the number of DFT coefficients should be taken into account. Referring to that, the best combination of methods for the entire database (accuracy equal to 71.75%) consists of 5×5 GFD for shape description, 73 DFT coefficients for action representation and correlation coefficient for matching. For actions with changing location of a silhouette we can apply the same combination but the number of DFT coefficients can be reduced to 55 (71.79% accuracy). The highest accuracy for actions performed in place amounted to 82.35% and was obtained for a combination of 10×10 GFD matched by C1 correlation. Action representation was composed of 49 DFT coefficients and actions were classified using correlation coefficient.

Table 4. Selected results for varying number of discrete Fourier transforms coefficients in action representation.

Shape descriptor	GFD	GFD	2DFD
Descriptor size	10×10	5×5	10×10
Shape matching	C1	CC	CC
Action representation	absFFT	absFFT	absFFT
Sequence matching	CC	CC	CC
Accuracy—entire database	67.72%	71.75%	69.91%
Number of DFT coefficients	128	73	101
Accuracy—actions with changing location	67.67%	71.79%	71.79%
Number of DFT coefficients	93	55	80
Accuracy—actions performed in place	82.35%	79.41%	70.59%
Number of DFT coefficients	49	73	44

5. Summary and Conclusions

In the paper, an approach for action recognition based on silhouette sequences is presented. It is a combination of several processing steps that utilize such methods as shape descriptors, spectral transforms and matching measures.

Extensive experiments were carried out, both for the entire database and two subgroups separately. Some accuracies were very similar for different tests, but there are common recommendations that indicates the use of 2DFD or GFD for shape representation, DFT for action representation and correlations for matching. Moreover, an additional experiment was carried out which examined various number of Fourier coefficients for action representation. It turned out that it is sufficient to process vectors smaller than 128 to obtain an increase in accuracy.

References

[1] Al-Ali, S., Milanova, M., Al-Rizzo, H., Fox, V.L., 2015. Human Action Recognition: Contour-Based and Silhouette-Based Approaches. Springer International Publishing, Cham. pp. 11–47. doi:10.1007/978-3-319-11430-92.

[2] Beaudry, C., Péteri, R., Mascarilla, L., 2014. Action recognition in videos using frequency analysis of critical point trajectories, in: IEEE International Conference on Image Processing (ICIP 2014), Paris, France. pp. pages 1445–1449.

[3] Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., 2005. Actions as space-time shapes, in: Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2, IEEE Computer Society, Washington, DC, USA. pp. 1395–1402. doi:10.1109/ICCV.2005.28.

[4] Bobick, A.F., Davis, J.W., 2001. The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 257–267. doi:10.1109/34.910878.

[5] Brunelli, R., Messelodi, S., 1995. Robust estimation of correlation with applications to computer vision. Pattern Recognition 28, 833–841. doi:10.1016/0031-3203(94)00170-Q.

[6] Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F., 2013. Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters 34, 1799–1807. doi:10.1016/j.patrec.2013.01.021.

[7] Chwastek, T., Mikrut, S., 2006. The problem of automatic measurement of fiducial mark on air images (in polish). Archives of Photogrammetry, Cartography and Remote Sensing 16, 125–133.

[8] Dengsheng Zhang, Guojun Lu, 2002. Generic Fourier descriptor for shape-based image retrieval, in: Proceedings. IEEE International Conference on Multimedia and Expo, pp. 425–428. doi:10.1109/ICME.2002.1035809.

[9] Frejlichowski, D., 2008. An algorithm for binary contour objects representation and recognition, in: Campilho, A., Kamel, M. (Eds.), Image Analysis and Recognition, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 537–546.

[10] Frejlichowski, D., 2010. An experimental comparison of seven shape descriptors in the General Shape Analysis problem, in: Campilho, A., Kamel, M. (Eds.), Image Analysis and Recognition, Springer, Berlin, Heidelberg. pp. 294–305.

[11] Frigo, M., Johnson, S.G., 2005. The design and implementation of FFTW3. Proceedings of the IEEE 93, 216–231.

[12] Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R., 2007. Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 2247–2253. doi:10.1109/TPAMI.2007.70711.

[13] Hazar, M., Zaafouri, R., Hammami, M., 2018. Human action recognition based on discriminant body regions selection. Signal, Image and Video Processing doi:10.1007/s11760-017-1227-z.

[14] Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: A survey. Image and Vision Computing 60, 4 – 21. doi:https://doi.org/10.1016/j.imavis.2017.01.010. regularization Techniques for High-Dimensional Data Analysis.

[15] Imtiaz, H., Mahbub, U., Schaefer, G., Zhu, S.Y., Ahad, M.A.R., 2015. Human action recognition based on spectral domain features. Procedia Computer Science 60, 430–437. doi:10.1016/j.procs.2015.08.161.

[16] Kpalma, K., Ronsin, J., 2007. An overview of advances of pattern recognition systems in computer vision, in: Obinata, G., Dutta, A. (Eds.), Vision Systems. IntechOpen, Rijeka. chapter 10. doi:10.5772/4960.

[17] Kukharev, G., 1998. Digital Image Processing and Analysis. SUT Press, Stettin, Poland.

[18] Liu, L., Shao, L., Zhen, X., Li, X., 2013. Learning discriminative key poses for action recognition. IEEE Transactions on Cybernetics 43, 1860–1870. doi:10.1109/TSMCB.2012.2231959.

[19] Rauber, T., 1994. Two-dimensional shape description. Technical Report. University Nova de Lisboa, Portugal.

[20] Smith, J., 2007. Mathematics of the Discrete Fourier Transform (DFT): With Audio Applications. W3K Publishing, BookSurge Publishing.

[21] Tran, A., Guan, J., Pilantanakitti, T., Cohen, P., 2014. Action recognition in the frequency domain .

[22] Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O., 2008. Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology 18, 1473–1488. doi:10.1109/TCSVT.2008.2005594.

[23] Vishwakarma, D., Dhiman, A., Maheshwari, R., Kapoor, R., 2015. Human motion analysis by fusion of silhouette orientation and shape features. Procedia Computer Science 57, 438–447. doi:10.1016/j.procs.2015.07.515.

[24] Vishwakarma, S., Agrawal, A., 2012. A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer 29. doi:10.1007/s00371-012-0752-6.

[25] Wang, J.J., Singh, S., 2005. Spatial feature based recognition of human dynamics in video sequences, in: Singh, S., Singh, M., Apte, C., Perner, P. (Eds.), Pattern Recognition and Image Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 84–93.

[26] Yadav, R.B., Nishchal, N.K., Gupta, A.K., Rastogi, V.K., 2007. Retrieval and classification of shape-based objects using Fourier, generic Fourier, and wavelet-Fourier descriptors technique: A comparative study. Optics and Lasers in Engineering 45, 695–708. doi:10.1016/j.optlaseng.2006.11.001.

[27] Yilmaz, A., Javed, O., Shah, M., 2006. Object tracking: A survey. ACM Computing Surveys 38, 1–45.

[28] Zhang, D., Lu, G., 2004. Review of shape representation and description techniques. Pattern Recognition 37, 1–19. doi:10.1016/j.patcog.2003.07.008.