Combining Shape and Trajectory Features for Human Action Classification using a Neural Network and Synthetic Data

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

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

Abstract. Human activity recognition systems using visual content analysis algorithms use data collected from a variety of sensors, the most popular of which are RGB cameras. Highly accurate motion information can be recorded using motion capture and then used to generate synthetic human body models. The advantage of such data is the absence of other objects in the background, the visualization accuracy and the anonymity of a person. This paper proposes a modified action recognition approach which creates action representations using simple features observed over time, including shape measurements, ratios and centroid trajectory. A feature vector consists of shape descriptors calculated separately for each video frame. These are then normalised, transformed to the frequency domain and supplemented with trajectory information. Action representations are classified using a feed-forward neural network with one hidden layer and varying number of hidden neurons. The high effectiveness values obtained in the experiments show that the appropriate composition of elementary features of moving objects brings considerable benefits.

Keywords: action classification · shape features · centroid trajectory · feed-forward neural network

1 Introduction

Human Activity Recognition (HAR) systems incorporate algorithms related to artificial intelligence, both classical machine learning algorithms and deep learning solutions. They benefit from many approaches, including pattern recognition, signal processing, video content analysis, neural networks and visual computing. HAR systems replace traditional observer-guarded surveillance systems, and simultaneously offer accelerated task execution and new functionalities, thanks to the availability of increased computational capabilities. Human movements vary in complexity — from elementary, single gestures and primitive actions to longer activities and behaviours [7]. A gesture can be, for example, a raise of the hand or a tilt of the head, while an action consists of multiple gestures performed over

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

a short period of time. Actions are characterised by simple movement patterns and certain poses of the human silhouette may occur cyclically [6]. If a silhouette shape is used, human movement is seen as a continuous change of human pose. Shapes extracted from video frames can be described by shape descriptors and used to create action representations. Classes of actions include, but are not limited to, walking, bending, waving or jumping.

Numerous attempts to classify action recognition methods have been presented, but despite a broad view of the state of the art, it is difficult to include all existing solutions. Many recent taxonomies distinguished between learning-based methods and traditional machine learning hand-crafted approaches, e.g. [20, 17]. On the other hand, taking the type of input data as a criterion, the area of interest includes visual data extracted from video sequences, depicting human movement. Video sequences provide a lot of information, including the poses and location of the person over a period of time. It is possible to use low-cost, off-the-shelf RGB cameras (laptop, smartphone) that allow the video to be recorded without the need to carry any additional equipment such as wearable sensors. This is related to the proposed use case scenario, which matches current trends in human activity recognition, including automatic physical exercise recognition based on video sequences. For experimental purposes, when real data is limited, synthetic data may be used — namely visualizations of the human model based on points recorded using motion capture. A very large number of calculations is required to simulate experimental conditions, including visual data. However, the problem of classifying physical exercises can be solved without any direct human involvement.

Our previous publications take into account the use of the Weizmann dataset and the method of action recognition based on the combination of shape features of individual silhouettes and two-stage classification of actions in subgroups — this method is described in [12] and [11], among others. Here a modification of this approach is proposed. Object trajectories are used in a different way and the manner of constructing action descriptors is changed. The effectiveness of the modified approach is verified on a larger number of data than before, extracted from the AMASS dataset [16] — the classes correspond to selected physical exercises. The main idea of the experiments is to test numerous versions of the modified method in order to find its most effective variant for the action recognition procedure in the assumed application. A physical exercise is characterized more in terms of the type of human movement than the meaning or the purpose of a person's behaviour. As an example, the following physical exercises are recommended for the elderly [24]: aerobic exercise, e.g. walking, jumping jack (activities in which the body's large muscles move in a rhythmic manner); balance training, e.g. sideways, vertical jumping, jumping forward (activities increasing lower body strength); flexibility exercise, e.g., hand waving, bending (activities preserving or extending motion range around joints).

The rest of the paper is organised as follows: Section 2 presents selected related works concerning action recognition based on binary shapes and neural

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

networks. Section 3 describes the modified approach for action recognition. Section 4 presents experiments and their results. Section 5 summarizes the paper.

2 Related Works

In recent years, interest in action recognition methods has shifted towards learning-based approaches. Convolutional neural network is one of the more commonly used model based on supervised learning. It is a hierarchical architecture with multiple hidden layers of different types, which process the input data into output categories (classes) [27]. The entire recognition process is carried out without the knowledge of an expert and features are extracted directly from data [25]. On the other hand, there are hand-crafted techniques, including traditional descriptors and classifiers prepared manually [1]. Examples of taxonomies separating hand-crafted methods from learning-based algorithms can be found in [20] or [17], among others. There are also hybrid methods combining hand-crafted features and neural network-based classifiers [15]. These approaches can achieve similar activity recognition rates as deep neural networks used alone, especially in tasks where a small amount of data is available [23].

One of the simpler examples of a unidirectional neural network is the multi-layer perceptron, the use of which for activity recognition is presented in [5]. The input data are binary human silhouettes, extracted from a video sequence and described using a set of values based on the bounding rectangle, area and centre of gravity of the object. In [21] an action recognition method using two convolutional neural networks is proposed. The task of the first network is to evaluate each frame from the sequence and create a normalized histogram showing how often a certain class is indicated. In turn, the second network estimates the dense optical flow for consecutive frames, and the results of the analysis are stored in the form of a second histogram. After summing the histograms, the final classification result is obtained. In [9], a binary silhouette sequence is represented by a single image that contains a combination of binary masks extracted from the video frames. The masks are collected based on a weighted function that assigns progressively higher weights to more recent frames. Some areas of the silhouette are made lighter, and this is a way of representing the flow of movement.

In [23], three different shallow neural networks with one hidden layer and one output layer are experimentally tested, each differing in the type of cost function and activation function in the hidden layer. It is concluded that the relevant parameters of the most effective solution can be randomly searched for, one of the most important being the number of neurons in the hidden layer. This demonstrates that the use of a shallow neural network in the activity recognition task is effective when the availability of data is limited, such as when a relatively small dataset is involved. In [4], the results of a study are presented, that aims to create and evaluate the potential for automated, unsupervised monitoring of lower back and shoulder physiotherapy exercises performed at home. Key points corresponding to the positions of the joints are extracted from the video sequences (recorded using a smartphone camera). This positions on consecutive

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

frames form a time series, the segments of which are used to train a convolutional neural network. In [8], a new action recognition scheme based on deep learning is proposed. Each detected human silhouette is described by eleven features extracted from its bounding box. An LSTM-type recurrent neural network is used for classification.

3 Modified Method for Action Recognition

In this paper, human action recognition is performed using a modification of the approach presented in [12] and [11]. It is a combination of image processing operations that use binary shape descriptors to create video sequence descriptors (see Fig. 1). The current version of the approach includes video preprocessing, therefore the input data can be RGB video sequences.

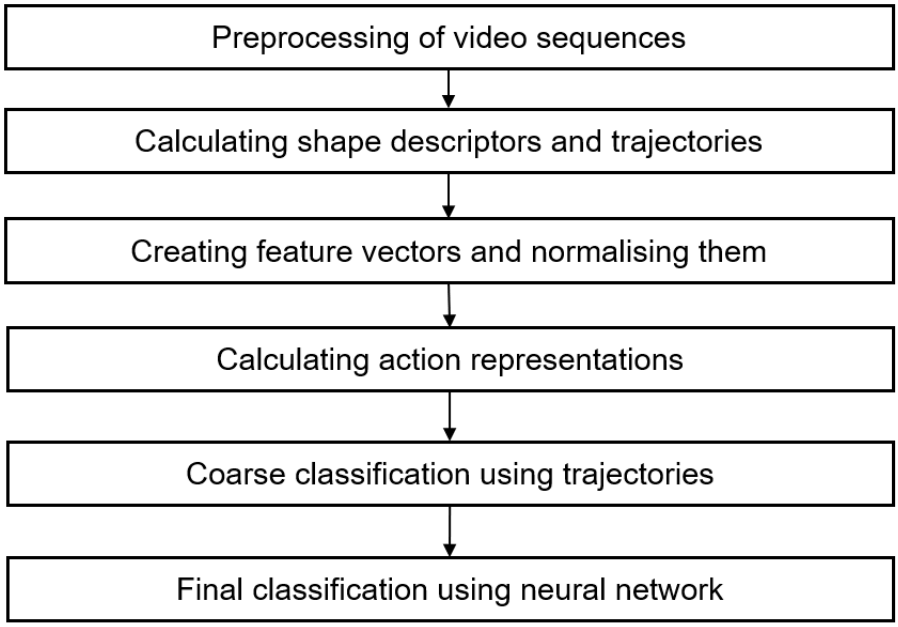

Fig. 1. Main steps of the modified action recognition approach.

3.1 Preprocessing of video sequences

Regardless of the type of video frames, the aim of preprocessing is to extract the area corresponding to the moving person in the foreground. Static background subtraction methods are applied if a background is simple. Resulting foreground

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

masks are thresholded and converted to binary images — a foreground shape is white, and the background is black. A video sequence is stored as a set of binary foreground masks and denoted as $BM_i = \{bm_1, bm_2, ..., bm_n\}$, where $i = 1, 2, ..., n$ and n is the number of video frames.

3.2 Calculating shape descriptors and trajectories

Each binary mask bm_i is individually represented using selected shape descriptor, sd_i. Various shape measurements and shape ratios are used for shape description. Since they are scalar values, a reduction of the two-dimensional binary image to a single number is obtained. Shape descriptors can be calculated based on any binary shape — a human silhouette, its convex hull or a bounding box enclosing all shape points. Four subgroups of simple shape descriptors are considered in this study, and these corresponds to the following shapes (shape descriptors are given in brackets):

1. A silhouette (area, perimeter, eccentricity [14], elongation [2], ellipticity ratio and circularity ratio [26], compactness [19], and Feret diameters including X Feret, Y Feret, XY Feret and Max Feret);
2. A convex hull, CH (area, perimeter, convexity [19], solidity [26], and the difference between the area of convex hull and the shape);
3. An axis-aligned minimum bounding rectangle, AABR (area, perimeter, horizontal side and vertical side as number of pixels, elongation [26], rectangularity [26]);
4. An arbitrarily oriented minimum bounding rectangle, OBR (area, perimeter, shorter side and longer side as number of pixels, elongation [26], rectangularity [26]).

At the same time, the coordinates of the centre of gravity of the foreground object are determined and the trajectory of the object is constructed. The trajectory is analysed within the field of view of the camera.

3.3 Creating feature vectors and normalising them

Shape descriptor values are concatenated into feature vectors. Each set of binary masks BM is now represented by a set of descriptors $SD_i = \{sd_1, sd_2, ..., sd_n\}$. Each feature vector is normalized using min-max normalization [13]. Values are scaled to a range — the smallest feature value is represented by 0 and the highest feature value corresponds to 1. This facilitates vector comparison and reveals the characteristics of feature variation over time. In addition, the influence of shape size, especially for basic shape measurements, is eliminated.

3.4 Deriving action representations

At this stage, the dataset consists of vectors of different lengths, and these vectors can be treated as signals. Therefore, the one-dimensional Fourier transform is applied for two main reasons — the number of resulting coefficients can

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

be declared in advance and periodicity in the data is highlighted. The number of predefined coefficients equals 128 or 256. It is the closest power of 2 compared to the number of frames in the longest video sequence. A m-point one-dimensional discrete Fourier transform is computed to obtain action representations $AR = \{ar_1, ar_2, ..., ar_m\}$, where m is the predefined number of resultant Fourier coefficients. If m is greater than n, the feature vectors are zero-padded in the time domain, which corresponds to interpolation in the frequency domain [22]. Otherwise, the feature vectors are first truncated. The result of the transform is a Fourier spectrum with complex numbers — only the magnitude of the spectrum is used, and its values are normalised by dividing all the coefficients by the value of the first coefficient. For real signals, the Fourier spectrum is a two-sided spectrum, therefore the use of half of the spectrum is also tested.

3.5 Coarse classification using trajectories

Each AR vector corresponds to a trajectory of the centre of gravity. The length of the trajectory is used to determine coarse classifications. If the length of the trajectory exceeds 20% of the video frame width, the corresponding AR vector is assigned to the group of actions performed with changing location, and is marked by appending the value 1. Otherwise, AR vectors belong to the group of actions performed in place and are appended with the value 0. Updated action representations are not divided into subgroups as previously [11], instead they are just labelled accordingly.

3.6 Final action classification using neural network

Updated action representations are classified using a pattern recognition network, which is a feed-forward network with a single hidden layer and multiple hidden neurons. The dataset is divided into training, validation and testing sets. The layer initialisation function uses the Nguyen-Widrow algorithm, which generates a different weight and bias each time the function is called. By eliminating this randomness it is possible to focus on adjusting other parameters of the proposed approach — the shape descriptor and the number of neurons. Multiple initial experiments were carried out to check the validity of individual variables, which showed that high effectiveness is repeated for random initialisation parameters. Therefore, it is possible to use a stored set of random values in the experiments in order to provide as many fixed parameters as possible, and to ensure the reproducibility of the results. In addition to this, other training functions and different proportions of data partitioning were tested. In the end, the use of the conjugate gradient method and a data partitioning ratio of 70/15/15 worked best. Due to the large number of possible parameter combinations, the main criterion for evaluating the results is the percentage effectiveness, which reflects the correspondence between the network's indications and the original action classes.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

4 Experiments and Results

4.1 Dataset

The test data set is constructed based on examples selected from the AMASS dataset [16] — Archive of Motion Capture as Surface Shapes — which refers to a unified synthetic dataset generated from other datasets containing motion information captured by optical markers (motion capture, mocap). The mocap data are converted into realistic 3D meshes representing a model of the human body. Synthetic visualisations for more than 20 datasets have already been created within the AMASS framework and are available for research purposes [3]. Their main advantage is the availability of renderings in the form of video sequences containing individuals performing various activities.

For the purpose of this research, sequences rendered using the mocap data from the MoVi — Motion and Video dataset [10] — are selected. This dataset is a collection of 21 activity classes (20 predefined and one arbitrary) performed by 90 actors. The activities in the MoVi dataset [18] were recorded in a continuous manner, one after the other, however, in the AMASS dataset [3] they are already divided into video sequences of a few seconds. From the available classes, five were selected, and they include hand waving, walking, sideways, vertical jumping and jumping jacks — there are 75, 77, 71, 70 and 77 video sequences, respectively, and 370 in total. Each sequence depicts one person performing an action — the silhouette of this person is rendered based on the previously recorded mocap points. Video sequences used in the experiments are characterized by: a resolution of 2048×1600 pixels (scaled to 256×200), 24 frames per second, a duration of less than 10 seconds, a black background and a camera following a moving person. Examples of video frames for five selected action classes from the AMASS dataset are provided in Fig. 2.

Coordinates of the centre of gravity of the moving object are stored separately. Fig. 3 shows example trajectories for five different classes — in this case, the values range from almost zero to around 3.5. The length of the trajectory is calculated as a distance between the most distant points (horizontal axis) using Euclidean distance. The differences are evident — actions with short trajectory (less than 1) are referred to as actions performed in place, while the long trajectories (more than 1) correspond to actions with changing location of a silhouette.

Video sequences from the dataset are converted into binary masks, as specified in the preprocessing step. These masks (see examples in Fig. 4) are used to calculate shape features. A single video sequence is represented as a cell array with dimensions $256 \times 200 \times n$, where n corresponds to the number of frames.

4.2 Description of the Experiments

The purpose of the experiments is to test the effectiveness of the modified action recognition method and to select its parameters in the proposed use case scenario. The research is carried out in the Matlab R2022b environment and using a laptop

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

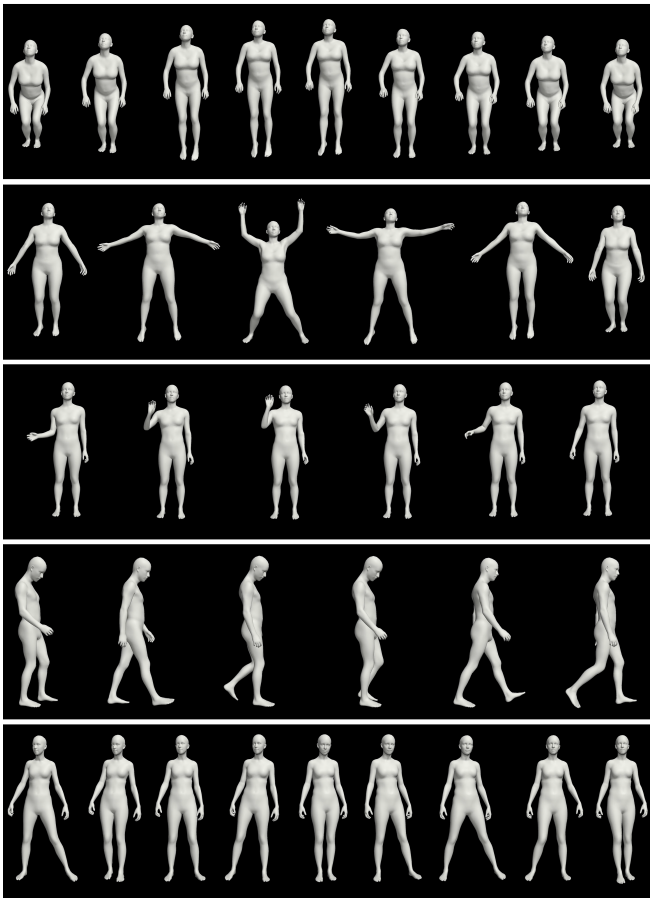

Fig. 2. Example frames corresponding to: vertical jumping, jumping jacks, hand waving, walking and sideways [16].

computer equipped with 32 GB of RAM. A set of binary images is used to prepare representations of actions, which are then classified using a neural network-based classifier. A number of experiments were carried out and each was repeated for all descriptors and varying number of neurons in a hidden layer.

The main experimental guidelines for a single experiment are summarised in the list below:

1. Binary masks are represented using shape descriptors and trajectories are calculated;
2. Feature vectors are created and values in each feature vector are normalised to a range [0-1];
3. Feature vectors are transformed using the one-dimensional Fourier transform — the predefined number of the absolute spectrum coefficients equals 128

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

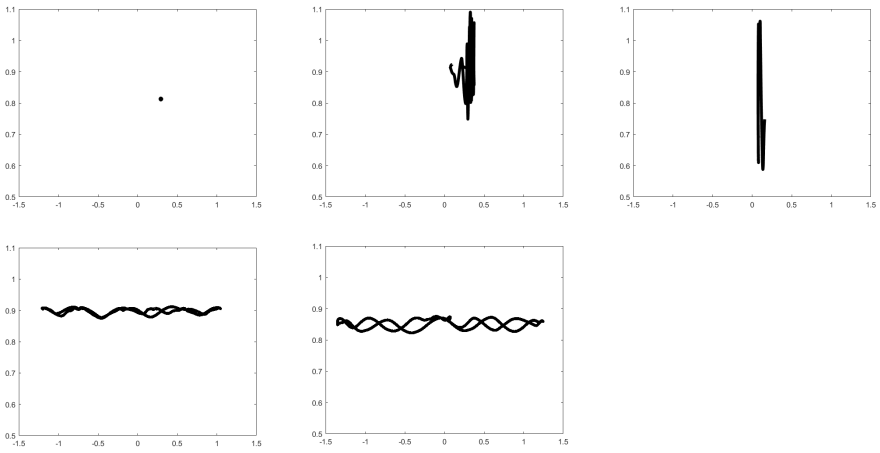

Fig. 3. Trajectories calculated for five example actions from the AMASS dataset, given in Fig. 2. The first row corresponds to hand waving, jumping jacks and vertical jumping, while the second row illustrates the trajectory of sideways and walking actions.

or 256, and all values in the vector are divided by the value of the first coefficient;

4. Action representations are created using resultant coefficients (whole or half of the spectrum), and a number is appended which determines the type of trajectory;
5. A simple feed-forward neural network is used for classification:
   (a) neural network input consists of 370 vectors of action representations;
   (b) for a single neural network input there is a single output representing the degree of similarity to classes under recognition;
   (c) there is one hidden layer with a variable number of neurons (1 to 50);
   (d) the activation function is based on the conjugate gradient method;
   (e) the input data is divided into training, validation and testing sets in a 70/15/15 ratio;
   (f) target classes are encoded into one-hot vectors and there are five unique classes;
   (g) the initialization parameters of a network are a set of random values that is usually generated again each time the network is trained. In order to be able to repeat the results of the experiments, a separate set of parameters is retained and used for all tests.

4.3 Results and Analysis

Through preliminary experiments, some parameters of the neural network-based classifier have been established. However, the number of possible versions of the modified action recognition approach is very large. With 28 descriptors and 50

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

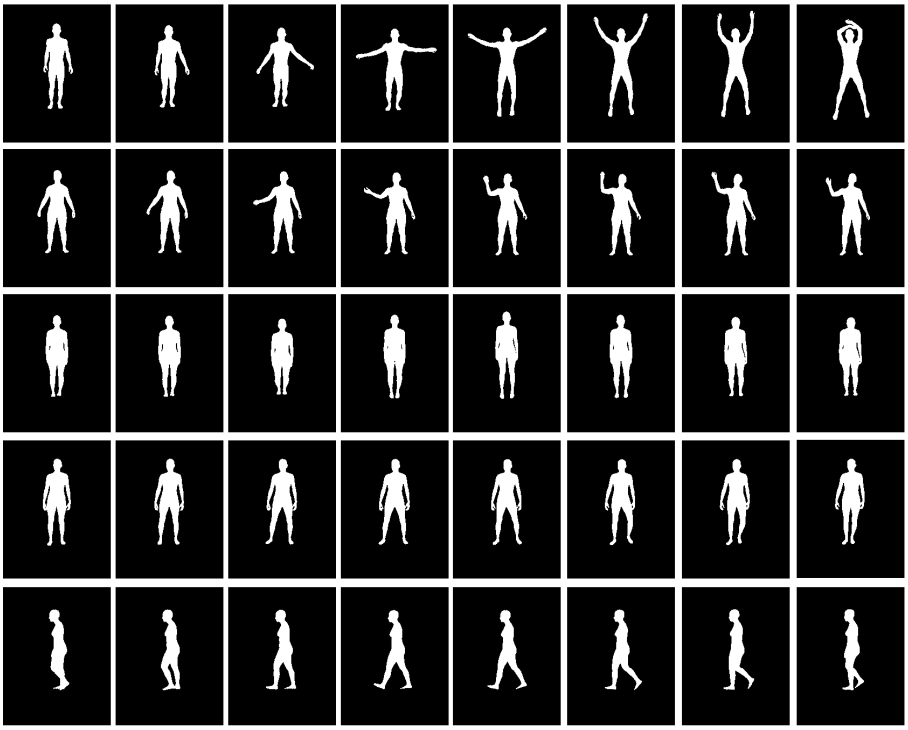

Fig. 4. Examples of binary masks corresponding to the following actions: walking, hand waving, vertical jumping, sideways and walking.

hidden layer sizes, this gives 1,400 combinations. Any change in action representation vectors will multiply this value. Hence, the effectiveness is taken as the main measure for evaluating classification results — it is a percentage of predicted classes that coincide with real class labels. Table 1 presents the experimental results for which the highest effectiveness is obtained. In all experiments the same action representation type is applied. For each shape descriptor, a series of tests is performed for a different number of hidden neurons, and then the value that corresponds to the highest effectiveness is selected (see Table 1).

According to the results provided in Table 1, the highest classification effectiveness is 99.72% in the experiment that used the perimeter of the convex hull. Such a result indicates that only one action is misclassified — in this case, an action from the class hand waving is incorrectly labelled as vertical jumping. In some other tests, the effectiveness exceeded 97% — this includes the area of the convex hull and the difference of the areas of the convex hull and the shape, as well as the area and perimeter of the axis aligned minimum bounding rectangle. This allows us to conclude that the calculation of shape descriptors on the basis of their bounding shapes is sufficient and improves the results. An

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

important aspect that determines the quality of the results is the size of the action descriptor and the way it is prepared — the results discussed here were obtained thanks to action representations with the following properties:

1. The normalized feature vectors are transformed using the Fourier transform with a declared number of resulting coefficients equal to 256;
2. The real part of the spectrum is selected and normalized with respect to its first coefficient;
3. Action representation uses half of the spectrum and an indication regarding coarse classification is appended at the end of the vector, therefore each final action representation consists of 129 elements.

Table 1. The results of the best experiments with an indication of the effectiveness values and the number of hidden neurons.

Shape descriptors		**Classification effectiveness**	**Number of neurons**
Silhouette	Area	95,68 %	25
	Perimeter	92,97 %	48
	Eccentricity	91,62 %	46
	Elongation	94,32 %	32
	Ellipticity ratio	92,43 %	17
	Circularity ratio	94,86 %	47
	Compactness	94,59 %	35
	X Feret	94,32 %	47
	Y Feret	91,89 %	24
	XY Feret	91,35 %	38
	Max Feret	92,70 %	39
Convex hull (CH)	CH area	97,57 %	46
	CH perimeter	**99,73 %**	34
	CH convexity	95,95 %	19
	CH solidity	96,76 %	26
	CH area difference	97,84 %	21
Axis-aligned minimum bounding rectangle (AABR)	AABR area	97,57 %	31
	AABR perimeter	97,84 %	31
	AABR horizontal side	94,32 %	47
	AABR vertical side	91,89 %	24
	AABR elongation	92,70 %	42
	AABR rectangularity	94,59 %	44
Arbitrarily-oriented minimum bounding rectangle (OBR)	OBR area	95,68 %	21
	OBR perimeter	96,76 %	47
	OBR shorter side	94,05 %	41
	OBR longer side	88,11 %	43
	OBR elongation	94,59 %	31
	OBR rectangularity	95,14 %	44

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

5 Summary

This paper presents a modified approach to action recognition that combines simple shape features and a classifier based on a unidirectional neural network. Video sequences are converted into binary images and then into vectors of shape descriptors, whose values are scaled to the interval [0-1]. These vectors are then transformed using the one-dimensional Fourier transform. The magnitude of the spectrum is taken and then all spectrum coefficients are normalized with respect to the first coefficient. The object trajectory is used to encode a coarse classification labels, and final classification is performed using simple feed-forward neural network. Experimental studies prove that it is possible to obtain high classification effectiveness using small action representations (129 elements) that are simple to compute. In addition, the approach is adaptable and can be used in other applications, using both real and synthetic video data. Future work includes comparative studies using other neural network architectures, such as Recurrent Neural Networks, while using more activity classes and video sequence examples.

References

1. Aggarwal, C.C.: Machine learning with shallow neural networks. In: Neural Networks and Deep Learning: A Textbook, pp. 53–104. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-94463-0_2
2. Aktaş, M.A., Žunić, J.: Measuring shape ellipticity. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) Computer Analysis of Images and Patterns. pp. 170–177. Springer Berlin Heidelberg (2011)
3. AMASS: AMASS: Archive of Motion Capture As Surface Shapes. https://amass.is.tue.mpg.de/download.php (2019), accessed: 20.05.2024
4. Arrowsmith, C., Burns, D., Mak, T., Hardisty, M., Whyne, C.: Physiotherapy exercise classification with single-camera pose detection and machine learning. Sensors **23**(1) (2023). https://doi.org/10.3390/s23010363
5. Babiker, M., Khalifa, O.O., Htike, K.K., Hassan, A., Zaharadeen, M.: Automated daily human activity recognition for video surveillance using neural network. In: 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA). pp. 1–5 (2017). https://doi.org/10.1109/ICSIMA.2017.8312024
6. 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)
7. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Expert Systems with Applications **39**(12), 10873–10888 (2012). https://doi.org/10.1016/j.eswa.2012.03.005
8. Cob-Parro, A.C., Losada-Gutiérrez, C., Marrón-Romera, M., Gardel-Vicente, A., Bravo-Muñoz, I.: A new framework for deep learning video based human action recognition on the edge. Expert Systems with Applications **238**, 122220 (2024). https://doi.org/10.1016/j.eswa.2023.122220

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

9. Dobhal, T., Shitole, V., Thomas, G., Navada, G.: Human activity recognition using binary motion image and deep learning. Procedia Computer Science **58**, 178–185 (2015). https://doi.org/10.1016/j.procs.2015.08.050, second International Symposium on Computer Vision and the Internet (VisionNet'15)
10. Ghorbani, S., Mahdaviani, K., Thaler, A., Kording, K., Cook, D.J., Blohm, G., Troje, N.F.: Movi: A large multi-purpose human motion and video dataset. Plos one **16**(6), e0253157 (2021)
11. Gościewska, K., Frejlichowski, D.: A combination of moment descriptors, fourier transform and matching measures for action recognition based on shape. In: Krzhizhanovskaya, V.V., Závodszky, G., Lees, M.H., Dongarra, J.J., Sloot, P.M.A., Brissos, S., Teixeira, J. (eds.) Computational Science – ICCS 2020. pp. 372–386. Springer International Publishing, Cham (2020)
12. Gościewska, K., Frejlichowski, D.: The analysis of shape features for the purpose of exercise types classification using silhouette sequences. Applied Sciences **10**(19) (2020). https://doi.org/10.3390/app10196728
13. Han, J., Kamber, M., Pei, J.: 3: Data preprocessing. In: Han, J., Kamber, M., Pei, J. (eds.) Data Mining (Third Edition), pp. 83–124. The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Boston (2012). https://doi.org/10.1016/B978-0-12-381479-1.00003-4
14. Haralick, R.M., Shapiro, L.G.: Computer and Robot Vision. Addison-Wesley Longman Publishing Co., Inc., USA (1992)
15. Khan, M.A., Sharif, M., Akram, T., Raza, M., Saba, T., Rehman, A.: Hand-crafted and deep convolutional neural network features fusion and selection strategy: An application to intelligent human action recognition. Applied Soft Computing **87**, 105986 (2020). https://doi.org/10.1016/j.asoc.2019.105986
16. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
17. Morshed, M.G., Sultana, T., Alam, A., Lee, Y.K.: Human action recognition: A taxonomy-based survey, updates, and opportunities. Sensors **23**(4) (2023). https://doi.org/10.3390/s23042182
18. MoVi: MoVi: A Large Multipurpose Motion and Video Dataset. https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP2/JRHDRN (2020), accessed 20.05.2024
19. Peura, M., Iivarinen, J.: Efficiency of simple shape descriptors. In: Arcelli, C., Cordella, L.P., di Baja, G.S. (eds.) Advances in Visual Form Analysis. pp. 443–451. World Scientific (1997)
20. Saif, S., Tehseen, S., Kausar, S.: A survey of the techniques for the identification and classification of human actions from visual data. Sensors **18**(11) (2018). https://doi.org/10.3390/s18113979
21. Silva, V., Vidal, F., Romariz, A.: Human action recognition based on a two-stream convolutional network classifier. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 774–778 (2017). https://doi.org/10.1109/ICMLA.2017.00-64
22. Smith, J.: Mathematics of the Fourier Transform (DFT): With Audio Applications. W3K Publishing, BookSurge Publishing (2007)
23. Suto, J., Oniga, S.: Efficiency investigation from shallow to deep neural network techniques in human activity recognition. Cognitive Systems Research **54**, 37–49 (2019). https://doi.org/10.1016/j.cogsys.2018.11.009

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10

24. Thaxter-Nesbeth, K., Facey, A.: Exercise for healthy, active ageing: A physiological perspective and review of international recommendations. West Indian Medical Journal **67**(5), 351–356 (2018). https://doi.org/10.7727/wimj.2018.177
25. Ullah, H.A., Letchmunan, S., Zia, M.S., Butt, U.M., Hassan, F.H.: Analysis of deep neural networks for human activity recognition in videos—a systematic literature review. IEEE Access **9**, 126366–126387 (2021). https://doi.org/10.1109/ACCESS.2021.3110610
26. Yang, M., Kpalma, K., Ronsin, J.: A survey of shape feature extraction techniques. Pattern Recognition pp. 43–90 (2008)
27. Yao, G., Lei, T., Zhong, J.: A review of Convolutional-Neural-Network-based action recognition. Pattern Recognition Letters **118**, 14–22 (2019). https://doi.org/10.1016/j.patrec.2018.05.018

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_10