Human action recognition using dimensionality reduction and support vector machine

L V Shiripova[1], E V Myasnikov[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

e-mail: shiripova.lubov@yandex.ru

Abstract. The paper is devoted to the problem of recognizing human actions in videos recorded in the optical range of wavelengths. An approach proposed in this paper consists in the detection of a moving person on a video sequence with the subsequent size normalization, generation of subsequences and dimensionality reduction using the principal component analysis technique. The classification of human actions is carried out using a support vector machine classifier. Experimental studies performed on the Weizmann dataset allowed us to determine the best values of the method parameters. The results showed that with a small number of action classes, high classification accuracy can be achieved.

1. Introduction

Human action recognition is actively used in various fields: in creating human-machine interfaces, in entertainment, in ensuring public safety, etc.

Human actions recognition involves solving two problems [1]:

1. The extraction of some feature information, i.e., converting a video stream or image sequence into a form suitable for subsequent classification.

2. Actually classification of the feature information obtained at the first stage.

To solve these problems, many approaches have been proposed, described in detail in [1]. Let us consider some of them.

To obtain the feature information, the authors of the paper [6] proposed to extract a silhouette from each frame, calculate images of the difference between adjacent frames and build the final image, superimposing the obtained images on each other. The resulting image was called Motion Energy Image (MEI). In addition, the authors introduce the concept of Motion History Image (MHI), i.e., the image, in which the intensity of each pixel depends on the time of action occurrence at a given point. The proposed approach has shown good results, but it has drawbacks when the angle of observation changes [6].

To eliminate this problem, a generalizing approach related to the use of 3D motion history volume (MHV) was proposed in [7]. MHV is based on 3D voxels obtained for various viewing angles. Further, the Fourier transform is used to acquire features that are invariant to position and rotation.

Another approach to obtaining features is associated with the extraction of space-time interest points (STIPs). Thus, the authors of [2,8] extended the Harris angle detector to the space-time domain.

The Gaussian function is then used to determine changes in movement in the spatial and temporal domains. In papers [9, 10, 11], a histogram of oriented gradients (HOG) and a histogram of optical flow (HOF) are used to obtain features. However, points of interest help to get information only for a short period of time. The authors of paper [12] proposed to use the Kanade – Lucas – Tomasi (KLT) feature tracker to track changes in points of interest.

In [13], simple parameters of convex figures are used as features.

For classification of the obtained features, various approaches are used, namely, the support vector machine (SVM) [14, 15, 9], k-nearest neighbors algorithm (k-NN) [16, 17, 18], as well as Hidden Markov Models (HMM) [19, 20, 21], etc.

In this paper, an approach based on the dimensionality reduction using the principal component analysis and subsequent classification using the support vector machine is used to solve the problem of human action recognition. A similar approach was successfully used by us earlier [22, 23] in solving the problem of person recognition by gait.

The paper has the following structure. Section 2 describes the developed method for human action recognition. Section 3 describes the results of experimental studies performed on the Weizmann dataset. The conclusions and the list of literature is given at the end of the paper.

2. Methods

The method proposed previously [22, 23] consists of the following steps:

- detection of a moving person in the video sequence,
- normalization of the frame size of the selected video sequence fragment,
- generation of subsequences,
- dimensionality reduction of the generated subsequences,
- classification of video sequences.

2.1. Detection of a moving person on a video sequence

At the first stage of the developed method, the moving person is detected in the video sequence. When the video sequence source is a video surveillance camera, background subtraction methods are used most frequently. The main idea of the methods of this class is to use a certain background model and to decide whether the particular pixel belongs to the background or a moving object, based on its correspondence to the background model. The background model is gradually refined over time. Although the time-averaged observation image can be used as a background model in the simplest applications, better results to this problem are given by more complex models, for example, [24-26].

In this paper, we use the background subtraction algorithm based on the mixture of Gaussian distributions (Gaussian mixture model, GMM) [25] to extract a moving person in a video sequence. According to this method, each background pixel is modeled by a weighted sum (mixture) of Gaussians. The weights of Gaussians correspond to the periods of time during which the corresponding Gaussian color is present on the video sequence.

We note that when choosing a method based on a mixture of Gaussian distributions, both our preliminary experiments and the experience by other researchers in solving the problem under consideration, were taken into account [27, 28].

As a result of the first stage, the set of masks corresponding to individual frames of the video sequence is formed. Each mask reflects the result of the segmentation of a frame into the foreground area corresponding to a moving person and the background.

2.2. Normalization of the size of detected fragments

At the second stage of the method, obtained masks are processed as follows. First, the center of mass for each foreground region is calculated, then the linear sizes of the region are determined, and a framing (a truncation of the mask image) is performed. After that, the cropped image is resized (compressed) to the specified size.

Taking into account the time coordinate, the dimensionality of the sequence of masks, which describes the movement of a person, remains high even after the size normalization (framing and

compression). In this regard, the fourth stage reduces the dimensionality of data describing the movement of a person.

2.3. Generation of subsequences

For each sequence of frames containing motion, a set of subsequences of a given length is allocated. Generation of subsequences is carried out with some specified step, starting from the beginning of the original sequence. A detailed description of the allocation of subsequences is given in previous papers [22, 23].

For each selected subsequence, the vector of features is formed as follows: each normalized frame of the subsequence is expanded into a row, and the rows obtained for individual frames are concatenated to each other.

The feature vectors of the subsequences of all sequences form the input matrix for the dimensionality reduction stage.

2.4. Dimensionality reduction using the principal component analysis technique

Both linear and nonlinear methods are used to reduce the dimensionality of multidimensional data. Linear methods such as principal component analysis (PCA) [29] and independent component analysis (ICA) are most commonly used. Nonlinear dimensionality reduction methods (for example, nonlinear mapping, ISOMAP, LLE) are used less often due to the high computational complexity of such methods. It should be noted that recent attempts have been made to accelerate such methods [30].

In this paper, we use the principal component analysis technique, as the most often used in similar cases and in other tasks (for example, see our previous papers [22, 23, 31]). This technique searches for a linear projection into the subspace of a smaller dimension that maximizes the variance of data. The PCA is often considered as a linear dimensionality reduction technique, minimizing the loss of information.

When principal components are found, the projection of feature vectors onto the first N principal components is taken as a feature description.

2.5. Classification of video sequences

The features obtained as a result of the principal component analysis are used to train the classifier Support Vector Machine (SVM) [33]. In the considered case, the classes correspond to individual actions, and feature vectors obtained for all subsequences correspond to individual observations (examples).

Note that the description given above is valid for the training mode in which the parameters of the dimensionality reduction and classifier are configured. In the testing mode, the data is processed in the same way, except that the parameters of the linear transformation are fixed to the values obtained in the training mode, and the trained SVM classifier performs the classification.

3. Experiments

The proposed method was implemented in C ++ using the OpenCV library. A PC based on the Intel Core i5-3470 CPU 3.2 GHz was used to perform experimental studies.

For the experimental study of the proposed method, the video sequences from the open Weizmann dataset (Figure 1) were used. This dataset contains sequences of binary images corresponding to individual frames of the video sequence, on which moving objects have already been extracted (foreground and background segmentation). The dataset contains video sequences for 9 people performing 10 different actions. The total dataset contains 90 sequences. Thus, there were 9 sequences in each class. The minimum sequence length was 28 frames.

The sequences were divided into the training and test sets containing four and five sequences correspondingly for each class. The sequences were pre-processed using the algorithm described in Section 2.2. Further, subsequences were generated according to Section 2.3. Then, the dimensionality reduction using the method described in Section 2.4 and classification using the algorithm described in section 2.5 were performed. To estimate the quality of the method, we used the classification accuracy, defined as the proportion of correctly classified objects.

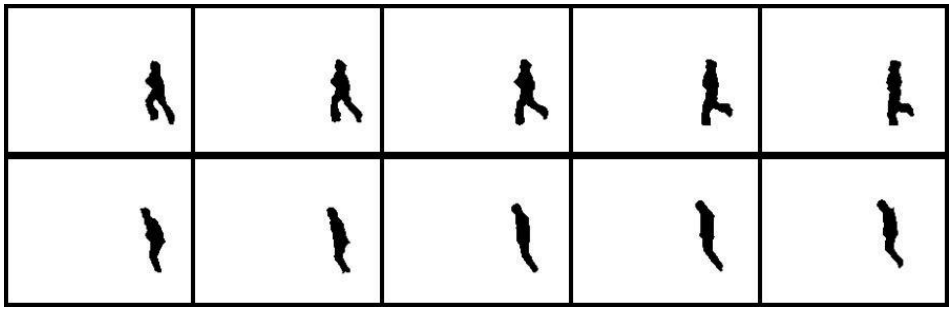

Figure 1. Examples of sequences from the Weizmann database: running and long jumps.

In the first experiment, we studied the dependence of the classification accuracy on the step size used in the generation of subsequences. The length of the subsequence was 28 frames. The output dimension of the feature vectors formed in step 2.4 of the considered method was equal to 8.

The experimental results are shown in figure 2. It was experimentally determined that the best classification accuracy is achieved with small step values. In further experiments, a step equal to 2 was used.

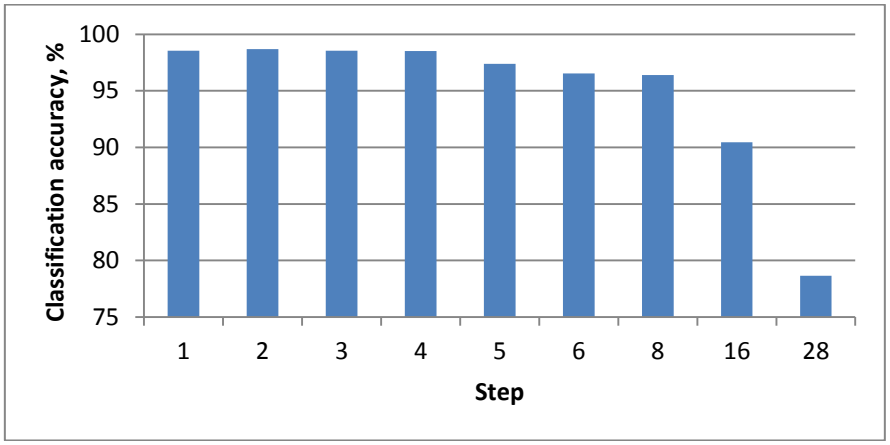

Figure 2. Dependence of the classification accuracy on step size.

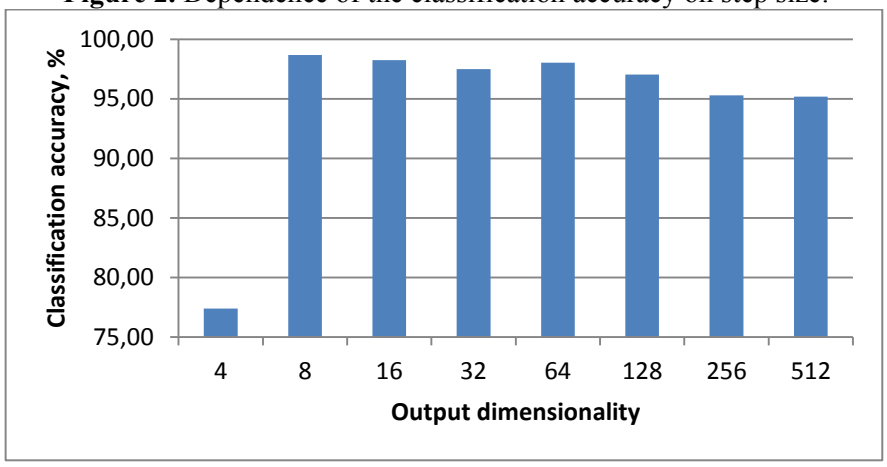

Figure 3. Dependence of the classification accuracy on the dimensionality.

In the second experiment, we investigated the dependence of the classification accuracy on the output dimensionality of the feature vectors, formed in step 2.4 of the proposed method. The output dimensionality varied from 4 to 512, while other parameters remained fixed. In particular, the step used in the allocation of subsequences was 2 frames.

The results of the experiment are shown in figure 3. As can be seen from the above results, the best values of the classification accuracy are achieved for dimensionality 8, 16 and 64.

In the third experiment, we investigated the dependence of classification accuracy on the length of subsequences. The length of the subsequence varied from 4 to 28 frames, while other parameters remained fixed. Thus, the step used in the allocation of subsequences was 2 frames; the output dimension was 64. The results of the experiment are shown in figure 4. As it can be seen from the results, the best values of the classification accuracy are achieved with a subsequence of 28 frames.

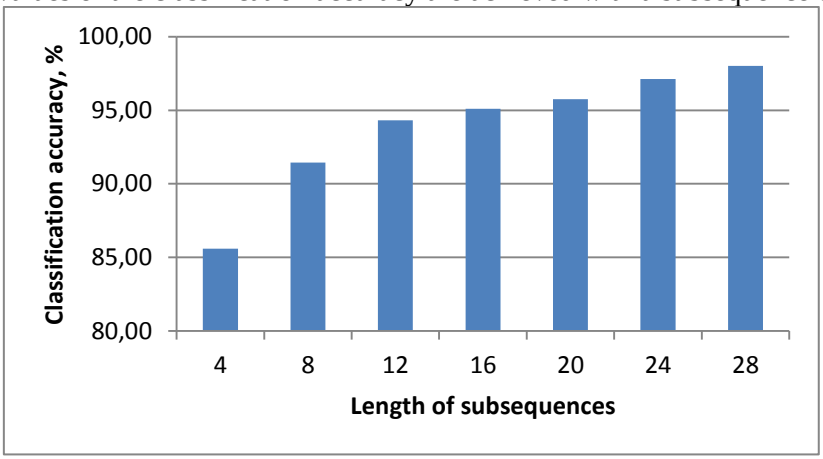

Figure 4. Dependence of classification accuracy on the length of subsequences.

As can be seen from the results of these experiments, with a relatively small number of classes (10 classes), a high (not less than 95%) classification accuracy can be achieved.

4. Conclusion

The proposed method for human actions recognition consists in the detection of a moving person in a video sequence, normalization of the size, generation of subsequences, dimensionality reduction using the principal component analysis and classification using the support vector machine.

The experiments performed on the Weizmann dataset allowed us to determine the best values of the parameters of the developed method. It was shown that with a small number of classes (10 classes), the proposed method provides on this dataset a high (with a wide range of parameters - at least 95%, and using the best values - up to 98%) accuracy of classification.

In the future, it is planned to expand the list of algorithms used to form a feature description and the list of classification methods. Another possible direction of further research is the detection of abnormal behavior (see, for example, [32]).

5. References

[1] Kong Y and Fu Y 2018 Human action recognition and prediction: a survey *J. of Latex class files* **19**
[2] Laptev I 2005 On space-time interest points *IJCV* **64** 107-123
[3] Raptis M and Sigal L 2013 Poselet key-framing: a model for human activity recognition *CVPR* 2650-2657
[4] Ji S, Xu W, Yang M and Yu K 2013 3d convolutional neural networks for human action recognition *IEEE Trans. Pattern Analysis and Machine Intelligence* **35** 221-231
[5] Carreira J and Zisserman A 2017 Quo vadis, action recognition? a new model and the kinetics dataset *CVPR* 6299-6308
[6] Bobick A F and Davis J W 2001 The recognition of human movement using temporal templates *IEEE Trans Pattern Analysis and Machine Intelligence* **23** 257-267
[7] Weinland D, Ronfard R and Boyer E 2006 Free viewpoint action recognition using motion history volumes *Computer Vision and Image Understanding* **104** 249-257
[8] Laptev I and Lindeberg T 2003 Space-time interest points *ICCV* 432-439

[9] Laptev I, Marszalek M, Schmid C and Rozenfeld B 2008 Learning realistic human actions from movies *CVPR*
[10] Klaser A, Marszalek M and Schmid C 2008 A spatio-temporal descriptor based on 3d-gradients *BMVC*
[11] Dalal N and Triggs B 2005 Histograms of oriented gradients for human detection *CVPR*
[12] Messing R, Pal C and Kautz H 2009 Activity recognition using the velocity histories of tracked keypoints *ICCV*
[13] Gosciewska K and Frejlichowski D 2018 Sillhouette-based action recognition using simple shape descriptors *Springer*
[14] Laptev I, Schuldt C and Caputo B Recognizing human actions: a local SVM approach *Proc. ICPR'04* (Cambridge, UK)
[15] Marszalek M, Laptev I and Schmid C 2009 Actions in context *CVPR*
[16] Blank M, Gorelick L, Shechtman E, Irani M and Basri R 2005 Actions as space-time shapes *Proc. ICCV*
[17] Laptev I and Perez P 2007 Retrieving actions in movies *ICCV*
[18] Tran D and Sorokin A 2008 Human activity recognition with metric learning *ECCV*
[19] Duong T V, Bui H H, Phung D Q and Venkatesh S 2005 Activity recognition and abnormality detection with the switching hidden semi-markov model *CVPR*
[20] Rajko S, Qian G, Ingalls T and James J 2007 Real-time gesture recognition with minimal training requirements and on-line learning *CVPR*
[21] Ikizler N and Forsyth D 2007 Searching video for complex activities with finite state models *CVPR*
[22] Shiripova L, Strukova O and Myasnikov E 2018 Gait analysis for person recognition using principal component analysis and support vector machines *CEUR Workshop Proceedings* **2210** 170-176
[23] Shiripova L and Myasnikov E 2018 Comparative analysis of classification methods for human identification by gait *CEUR Workshop Proceedings* **2268** 118-128
[24] KadewTraKuPong P and Bowden R 2001 An improved adaptive background mixture model for real-time tracking with shadow detection
[25] Zivkovic Z 2004 Improved adaptive Gausian mixture model for background subtraction
[26] Andrew B, Matsukawa A and Goldberg K 2012 Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation
[27] Murukesh C, Thanushkodi K, Padmanabhan P and Mohamed D 2014 Secured authentication through integration of gait and footprint for human identification *Journal of Electrical Engineering and Technology*
[28] Wang L, Tan T, Hu W and Ning H 2003 Automatic gait recognition based on statistical shape analysis *Transactions on image processing* **12**
[29] Fukunaga K 2003 *Introduction to statistical pattern recognition* (London: Academic Press)
[30] Myasnikov E V 2017 Fast techniques for nonlinear mapping of hyperspectral data *Proc. SPIE* **10341** 103411D.
[31] Myasnikov E V 2017 Hyperspectral image segmentation using dimensionality reduction and classical segmentation approaches *Computer Optics* **41(4)** 564-572 DOI: 10.18287/2412-6179-2017-41-4-564-572
[32] Shatalin R A, Fidelman V R and Ovchinnikov P E 2017 Abnormal behavior detection method for video surveillance applications *Computer Optics* **41(1)** 37-45 DOI: 10.18287/2412-6179-2017-41-1-37-45
[33] Cortes C and Vapnik V 1995 Support-vector networks *Machine Learning* **20(3)** 273-297

Acknowledgments
The work was partly funded by RFBR according to the research project 17-29-03190 in parts of «1. Introduction» – «2. Methods» and by the Russian Federation Ministry of Science and Higher Education within a state contract with the "Crystallography and Photonics" Research Center of the RAS under agreement 007-ГЗ/Ч3363/26 in part of «3. Experiments».