Classification of a two-dimensional pose using a human skeleton

Marina Pismenskova[1,*], *Oxana* Balabaeva[1], *Viacheslav* Voronin[1], and *Valentin* Fedosov[2]

[1]DSTU, Department of Radio-electronic and electrotechnical systems and complexes, 346500 Shakhty, Russia
[2]SFEDU, Department of Theoretical Foundations of Radio Engineering, 347922 Taganrog, Russia

Abstract. This article proposes an approach for the human pose recognition based on to preliminary prepared high-level data of the human skeleton. The coordinates of the feature points on the human skeleton is constructed by using a neural network. In the presented method, the coordinates of the feature points are a normalized relative of the human body height and the centre of gravity. We use these data for training the neural network.

1 Introduction

The task of human actions recognition is a classic task of computer vision and has a high relevance in such industries as intelligent video surveillance systems, analysis of monitoring cameras data, robotic vision systems, medical services, video analysis of sports games, etc. In recent decades, significant progress has been made in this area through the use of neural networks. The main current trend in the development of neural networks is the architecture of a convolutional neural network, proposed by Jan LeCun [1]. Methods based on convolutional neural networks have made a significant contribution to solving the problems of analyzing the video data of intelligent surveillance systems [2], but there are difficulties in their implementation in practice. Either recognition system requires an individual set of training, corresponding to a specific task. Preparation large sets of video data are required huge human resources, and training a neural network is an expensive process, both in terms of computation and time. Training sets consist of thousands of video sequences and has a significant amount of unnecessary information, which introduces error into the learning process. While images contain discriminative action poses, it improves of efficient to classify the pose [3].

2 Related works

Recently, many works have been prepared and methods have been implemented for recognizing actions and classification of the posture without using optical markers or any other motion sensors attached to parts of the human body. In [4], the authors use a continuous sequence of depth maps of the video sequence. The proposed online system extracts space-time information about human movement, and also information about the skeleton based on the detection of joints of the human body. For training, the authors use the hidden Markov model (HMM). In [5], methods for recognizing human actions based on a data stream of deep information are presented, which improves the performance of intelligent analysis systems Zanfir et al. [6] developed a descriptor of a moving pose, which takes into account both the information on the pose and the differential values (speed and acceleration) of the joints of the human body. Then the proposed descriptor is used in conjunction with the modified k-nearest neighbors (KNN) classifier. X. Yang and Y. Tian [7] proposed a function based on differences in the position of the three-dimensional joints and their own joints (Eigenjoints), which include information such as static posture, motion and displacement. They use the naive Bayesian classifier (NBNN) to classify. Xia et al. [8] used a compact representation of poses called HOJ3D, which characterizes human poses as histograms of the locations of three-dimensional joints in a modified coordinate system. Then they taught a discrete hidden Markov model to classify successive poses. In [9], the authors focused on understanding complex human activities. The activity consists of a set of actions with simple semantic information - elementary (atomic) actions (for example, a triple jump includes the following atomic actions: takeoff, jump and landing in the sand). As an interpretation of such a video sequence, which contains complex activities, a hierarchical description of the data containing the following information is assumed: the type of activity that takes place in the video clip, what atomic actions it includes and when these actions are performed. This method, in addition to obtaining specific recognition results on databases containing complex activities, also informatively describes the actions taking place on the video, captures the semantic concept and the temporal structure of elementary actions. In [10], proposed a modified global video descriptor for classification of realistic videos. This method detection of interest points, the extraction of local video descriptors and the quantization of descriptors into a codebook; it represents each video sequence as a single feature vector. The global descriptor is computed by applying a bank of 3-D spatiotemporal filters on the frequency spectrum of a

* Corresponding author: mpismenskova@mail.ru

video sequence, hence it integrates the information about the motion and scene structure.

3 The proposed approach

The purpose of the presented work is to classify a two-dimensional human pose from one image (frame). In this article, a method for recognizing a person's pose according to previously prepared high-level data of a human skeleton is proposed, which makes it possible to reduce the error of excessive information.

3.1 Representation of key point

As a method of key points representing of a human body, the recurrent neural network V. Belagiannis and A. Zisserman [11] is used. The method detects 16 key points on the skeleton of the human body, and namely: right ankle, right knee, right hip, left hip, left knee, left ankle, torso, neck, chin, top of the head, right wrist, right elbow, right shoulder, left wrist, left elbow, left shoulder.

A training set of images are frames obtained on the basis of a video data set "UCF101-Action Recognition Data Set" [13]. Three types of actions were used: Pull-ups, Squatting and push-ups from the wall (Wall Push Ups), in Figure 2 examples of frames of different actions are presented. Each video is divided into frames. As a result, the total training set was 8061 images.

3.2 Preparing high-level data

For each image, the coordinates of the singular points i_n were obtained, where n is the joint number (Figure 1). To prepare the data for the input of the neural network, the coordinates were normalized relative to the body length and relative to the center of gravity.

The length of the body (6) is calculated as the sum of the lengths of individual parts: the head (1), the torso (2), the leg from the hip to the knee and from the knee to the ankle (3,4), the maximal length of the right or left leg (5):

$$Length_{head} = \sqrt{(j_{10}(x) - j_9(x))^2 + (j_{10}(y) - j_9(y))^2} \quad (1)$$

$$Length_{torso} = \sqrt{(j_8(x) - j_7(x))^2 + (j_8(y) - j_7(y))^2} \quad (2)$$

$$Length_{leg\_right} = \sqrt{(j_3(x) - j_2(x))^2 + (j_3(y) - j_2(y))^2} + \sqrt{(j_2(x) - j_1(x))^2 + (j_2(y) - j_1(y))^2} \quad (3)$$

$$Length_{leg_{left}} = \sqrt{(j_4(x) - j_5(x))^2 + (j_4(y) - j_5(y))^2} + \sqrt{(j_5(x) - j_6(x))^2 + (j_5(y) - j_6(y))^2} \quad (4)$$

Fig. 1. An example of building a skeleton for an image from the dataset «Leeds Sports Pose Dataset» [12]

$$Length_{leg} = \max(Length_{leg\_right}, Length_{leg\_left}) \quad (5)$$

$$Length_{body} = Length_{head} + Length_{torso} + Length_{leg} \quad (6)$$

where $j_n(x)$ – x - the x coordinate of the n-joint, $j_n(y)$ – the y coordinate of the n – joint, $Length_{head}$ - is the head length, $Length_{torso}$ – is the length of the torso, $Length_{leg\_right}$ – is the length right leg, $Length_{leg\_left}$ – is the length left leg, $Length_{leg}$ – maximum leg length, $Length_{body}$ – total body length.

The center of gravity is calculated as:

$$Centr_x = \sum_n \frac{j_n(x)}{16} \quad (7)$$

$$Centr_y = \sum_n \frac{j_n(y)}{16} \quad (8)$$

At the next stage, the coordinates are normalized relative to the length of the body and the center of gravity:

$$i_n(x) = \frac{j_n(x) - Centr_x}{Length_{body}} \quad (9)$$

$$i_n(y) = \frac{j_n(y) - Centr_y}{Length_{body}} \quad (10)$$

where $i_n(x)$ – x – normalized coordinate n – joint, $i_n(y)$ – y – normalized coordinate n – joint.

Information on the distance between certain joints, which can characterize the features of the presented posture, namely the distance from the wrist to the shoulder

(11,12) and from the ankle to the hip (13,14), is also fed to the input.

$$Dist_{shoulder-wrist-r} = \sqrt{(j_{13}(x) - j_{11}(x))^2 + (j_{13}(y) - j_{11}(y))^2} \quad (11)$$

neurons in the output layer (linear activation function). Figure 3 shows the described architecture of the neural network.

The validation dataset contains 1201 images corresponding to the types of actions. The presented implementation of the neural network and training data allow obtaining the correct recognition result in 85.5% of

a) b) c)

Fig. 2. Examples of actions from the training set: a - pull-ups; b - squatting; c - push-ups from the wall

$$Dist_{shoulder-wrist-l} = \sqrt{(j_{14}(x) - j_{16}(x))^2 + (j_{14}(y) - j_{16}(y))^2} \quad (12)$$

$$Dist_{hip-ankle-r} = \sqrt{(j_3(x) - j_1(x))^2 + (j_3(y) - j_1(y))^2} \quad (13)$$

$$Dist_{hip-ankle-l} = \sqrt{(j_4(x) - j_6(x))^2 + (j_4(y) - j_6(y))^2} \quad (14)$$

3.3 Configuration and learning the neural network

As the final stage of the recognition of actions, a neural network is used. In the input of the network is fed the prepared high-level data. Using descriptors, rather than the whole image, as input, allows you to reduce the excess of information and reduce learning time.

The data presented above are fed to the input of a neural network, which has 36 inputs. Because three types of actions were used to test the network, the output layer is represented by three neurons. During the experiments, various variants of neural network configuration were considered. The best result was achieved by using a perceptron consisting of 36 neurons on the input layer (logsig activation function), 64, 32 and 16 neurons on three hidden layers (logsig activation function) and three

cases.

Table 1 shows the percentage of correct recognition of individual actions

Table 1. correct recognition of individual actions.

Action	Correctly (%)
Pull Ups	97
Squatting	86
Wall Push Ups	78
All set	85

Errors in the recognition and construction of the skeleton take place in cases where an object is partially hidden (occluded), flipped over, there are several foreground objects.

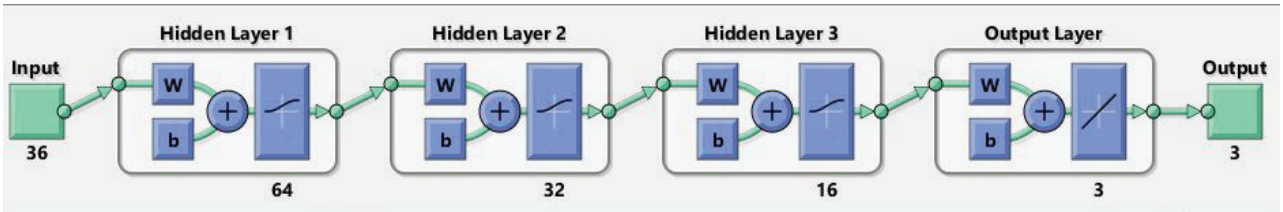

Fig. 3. The architecture of the neural network used

4 Conclusion and future work

We proposed an approach for the human pose recognition based using a neural network. In the future, there is

planned to apply the presented approach to the video sequence in real time, to make the comparisons with state-of-the-art methods on public databases. There is supposed to conduct research into the effect of the visual data pre-processing step on the quality of action recognition, because various defects can take place during the creation and transfer of video information: incorrect focusing, lens shake, low resolution of images due to physical limits of cameras, scratches and dust on devices and etc. [14, 15].

The reported study was supported by the Russian Foundation for Basic research (RFBR), research projects №16-37-00386, №16-37-00391 and № 17-57-53192.

References

1. Y. LeCun. et al., The handbook of brain theory and neural networks (MIT Press, Cambridge 1995)
2. M. Pismenskova, N. Gapon, P. Sizykin, *Materials of the international scientific-practical conference. The Scientific and Publishing Center "The World of Science.",* 161-167 (2016)
3. S. Ma et al. *In Proc. CVPR*, 334-345 (2017)
4. A. Jalal et al. *In Proc. CVPR,* 295-308 (2017)
5. M. Pismenskova, V. Voronin, N. Gapon, L. Levina, Information Technology. Radioelectronics. Telecommunications, no. 6, 164-171 (2016)
6. M. Zanfir, M. Leordeanu, C. Sminchisescu. *Proceedings of the IEEE International Conference on Computer Vision*, 2752-2759 (2013)
7. X. Yang, Y. Tian. *In Proc. CVPRW*, 14-19 (2012)
8. Xia L., Chen C. C., Aggarwal J. K. *In Proc. CVPRW*, 20-27 (2012)
9. C. Liu, X. Wu., Y. Jia. International Journal of Computer Vision, 1-16 (2016)
10. T. Silkina, V. Voronin, V. Marchuk, M. Pismenskova, *Innovations, ecology and resource-saving technologies materials of the XI International Scientific and Technical Forum,* 1311-1315 (2014)
11. V. Belagiannis, A. Zisserman. *In Proc. 12th IEEE International Conference on Automatic Face & Gesture Recognition,* 468-475 (2017)
12. S. Johnson, M. Everingham, *In Proc. BMVC* (2010)
13. Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, *In Proc. CRCV-TR* (2012)
14. S. Kundu, M. R. Raksha. 2016.
15. M. Ponomarenko, M. Pismenskova, K. Egiazarian, *In Proc. 25th European Signal Processing Conference (EUSIPCO)* (2017)