Improvement of the Human Action Recognition Algorithm by the Pre-processing of Input Data

M M Zhdanova¹, O S Balabaeva², V V Voronin¹, E A Semenishchev¹, A A Zelensky¹

¹ Center for Cognitive Technology and Machine Vision, Moscow State University of Technology «STANKIN», Moscow, Russia
² Lab. «Mathematical methods of image processing and intelligent computer vision systems», Don State Technical University, Rostov-on-Don, Russia

Abstract. The paper presents an approach to recognizing human actions using an additional pre-processing stage of input data. The growing volumes of video information do not always allow support the quality of data at a high level; this can cause limitations in the further processing of digital data. In this regard, it becomes urgent to introduce an additional stage of image enhancement into the algorithm for recognizing actions in video. The proposed method includes three main steps: image enhancement, constructing a descriptor, and classification. The presented image enhancement stage is based on the combined local and global image processing in the frequency domain. The basic idea in using local alfa-rooting method is to apply it to different disjoint blocks with different sizes. To solve the problem of constructing a descriptor, a three-dimensional microblock dense difference (3D DMD) algorithm is used, which provides a highly oriented representation of image regions by tightly capturing microblocks within each region in several orientations and scales. 3D DMD has several advantages over other methods: higher efficiency compared to existing methods; minimal computational costs when using an integrated image; low dimension; ease of implementation; does not require settings. The presented modification allows to increase productivity by 2-4%.

1. Introduction
The problem of human action recognition is a classic example of computer vision task and is highly relevant today, including various applications in video surveillance, robotic vision, medical equipment development and research, sports games annotation, etc. Ever increasing demand on the opportunity to recognize human actions leads to the situation when many researchers in the computer vision field are focused on this problem. In recent years, plenty of works on this topic have been published. Particularly interesting are approaches for action recognition and pose classification without any need for motion capture markers and various motion sensors attached to the human body parts [1].

Below are a few popular methods for constructing local descriptors for object recognition and matching. These descriptors are the basis for many modern effective human action recognition algorithms.

The SIFT (Scale Invariant Feature Transform) [2] is the method detects and describes local features of the image. The descriptors obtained on the basis of this method are scale-invariant and rotation-invariant, robust to affine transformations, and noise. A modification of the SIFT method for a video sequence is presented in [3,4]. Algorithms are able to detect and describe the spatio-temporal position of feature points. They are based on three main stages: detection of points of interest, building a descriptor for these points, and creating a classifier.
Optical flow methods are methods that record movement in a scene by bias surfaces or edges of an image. The optical flow is used for recognizing objects, tracking objects, detecting, and recognition action. The optical flow methods calculate the motion between two frames taken at time t in each pixel [5]. In [6] presented a method for detecting and recognizing actions on a video sequence, which combines the motion descriptors HOG [7] (Histogram of Oriented Gradients), HOF [7] (Histogram of optical flow) and MBH [8] (Motion boundary histograms).

One of the methods for constructing descriptors for a video sequence is the global video descriptor (Global Video Descriptor) [9]. The idea of a GIST descriptor [10], used to describe the configuration of a scene based on histograms of gradient directions, is used in the work. Because the GIST does not describe information about spatio-temporal changes in the data, the authors calculate the spectral-frequency components that are useful for extracting information about the movement, and the GIST is used to describe the configuration of the scene. In the original method for training and classification of the obtained descriptors, the support vector machine (SVM) method is used.

As part of this work, an action recognition method based on the construction of a global descriptor for describing human actions on complexly structured video sequences and background is proposed.

2. Proposed algorithm

The block diagram of the proposed method of human action recognition with the preprocessing stage is shown in Figure 1.

![Figure 1. Block-scheme of the proposed method](image)

The structure of the proposed algorithm is as follows:

1. Two data streams are input: RGB video and a video sequence of depth maps;
2. Video sequences enhancement based on combined local and global image processing;
3. Fusing data obtained from sensors of various modalities and pre-processed on previously stage;
4. Recognition of human actions occurring in the frame based on the method of the 3-D dense difference of microblocks (3D-DMD);
5. Assigning a sequence of frames to a specific class of actions;
6. The end.

2.1. The pre-processing step

Large video streams are often an obstacle to fast transfer and are computationally intensive and storage-intensive. Therefore, the quality of video information is not always at a high level. In addition, many other factors affect the quality: lighting, weather conditions, the speed of movement of objects in the frame, etc. Therefore, for tasks of digital signal processing, preprocessing plays an important role and can qualitatively affect the further processing result.

Image enhancement is the processing of an image to improve quality so that the results are more suitable for display or further analysis of the image.

Image enhancement means:
- Improving the interpretability or perception of information in images for viewers.
- Providing "better" input for other automatic image processing methods.

The principal objective of image enhancement is to modify attributes of an image to make it more suitable for a given task and a specific observer [11, 12].

We use image enhancement algorithm based on combined local and global image processing. Figure 2 shows the structure of the used image enhancement algorithm.

![Figure 2. Flowchart of the image enhancement step [1, 2]](image)

The main idea of the image enhancement algorithm is using extraction blocks with different sizes from the input image. The sizes of the blocks are 8 by 8, 16 by 16, 32 by 32, and, i.e., For this purpose, we move the square block from up left corner to down right corner with step equal one. For every block, we use the transform-based enhancement algorithm base on the $\alpha$-rooting and magnitude reduction method [13]:

$$
\hat{X}(p,s) = X(p,s) \cdot |X(p,s)|^{\alpha-1} = |X(p,s)|^\alpha \cdot e^{i\theta(p,s)}
$$

where $X(p,s)$ is the transform coefficients of the image, $\alpha$ is a user defined operating parameter, $\theta(p,s)$ is the phase of the transform coefficients.

The $\alpha$-rooting transform depends on the parameter $\alpha$. We are choosing the best (optimal) enhancement image through optimization of measure enhancement (EME) introduced by Agaian [14].

The image enhancement results obtained by the proposed algorithm is illustrated in Figure 3.
2.2. Fusion step

The goal of image fusion (IF) is to integrate complementary multisensor, multitemporal and/or multiview information into one new image containing information the quality of which cannot be achieved otherwise.

The basis of the image fusion model is a technology similar to the human vision system (HVS). Because the human visual system processes the light logarithmically, the transmitted images must be combined according to logarithmic laws. The PLIP model [15-16] is a parameterized LIP (logarithmic image processing) model [17]. PLIP considers grayscale images and processes them using new arithmetic operators that replace the standard ones, as shown in [16]. The grayscale function $\Psi$ for a given image is generated using:

$$\Psi = \alpha - 1$$

where $\alpha$ is a parameter of the model. The $\alpha$ value is the maximum value of the image intensity $I$.

Fusing images are done using the following formula:

$$fusedImage = \Psi_1 \oplus \Psi_2 \oplus \Psi_3$$

where $\Psi_1$ is the image of the visible spectrum; $\Psi_2$ – depth image; $\Psi_3$ – the maximum of the combined image obtained by selecting the maximum value of the intensity of the depth or visible image at each pixel location; $\oplus$ is the addition operator of the parameterized model of logarithmic image processing, $\tilde{\Psi}_i$ is the result of scalar multiplication of the parameterized model of logarithmic image processing, as shown below:

$$\Psi_1 \oplus \Psi_2 = \Psi_1 + \Psi_2 - \left( \frac{\Psi_1 \Psi_2}{\max(\Psi_1, \Psi_2)} \right)$$

$$\tilde{\Psi}_i = (\Omega_i \ominus \Psi_i) = \max(\Psi_i) - \max(\Psi_i) \left( 1 - \frac{\Psi_i}{\max(\Psi_i)} \right)^{\Omega_i}$$

The constants $\Omega_i$ presented in [16] have the following values $\sum \Omega_i = 1$, $\Omega_1 = 0.2989$, $\Omega_2 = 0.5870$ и $\Omega_3 = 0.1141$.

The examples of combining images are presented in Figure 4. Video frames were taken from the video database [18].
Figure 4. The examples of fusion images obtained from sensors of various modality:
a) RGB images; b) the images obtained from the depth sensor; c) the fusion images

2.3. The action recognition step
The action recognition stage consists of the following steps:
- The 3D discrete Fourier transform (3D DFT);
- Convolution with three-dimensional space-time Gabor filters;
- The inverse 3D discrete Fourier transform;
- The 3-D dense difference of microblocks;
- Action classification.

Motion is an important element that represents the action being performed on the scene. The frequency spectrum of a two-dimensional image lies on a plane whose orientation depends on the speed of the model. Having a two-dimensional function \( f_0(x, y) \), possibly to create a three-dimensional space-time sequence of images by moving \( f_0(x, y) \) with a rate \( \bar{u} = [u_1, u_2] \) in time. This volume is expressed as [19]:

\[
f(x, y, t) = f_0(x - u_1 t, y - u_2 t)
\]  
(6)

The three-dimensional Fourier transform \( F(x, y, t) \) in space and time is calculated by the following formula [19]:

\[
F(f_x, f_y, f_t) = \frac{1}{MNT} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{t=0}^{T-1} f(x, y, t) e^{-j2\pi \left( \frac{fx}{M} + \frac{fy}{N} + \frac{ft}{T} \right)}
\]  
(7)

where \( M, N, T \) are the width, height and length of the video, respectively, \( x, y, t \) is the spatio-temporal position of each point in the created space. The three-dimensional discrete Fourier transform has the same volume as the created space.

In order to reflect the various characteristics of movements in the scene, a bank of 3D Gabor filters is used, the components of which differ in orientation and scale. Then, each generated 3D filter is applied to the frequency spectrum of the clip.

As a result, for each filter, from the 3D Gabor filter bank, a function vector is obtained that represents a separate clip characterizing a certain direction of actions. The resulting array has a dimension of 4 (frames width \( \times \) frames height \( \times \) number of frames \( \times \) number of filters). After reducing the dimension, the array retains spatial information in itself, as an answer to each filter.

Then a modified DMD [20] algorithm is used extended for 3D space. This feature provides highly oriented 3D representation of image areas by tightly capturing microblocks within each area in multiple orientations and scales. DMD has several advantages over existing methods. The presented method is based on the idea that small patches (areas) of the image have a characteristic structure and, if they are captured effectively, it is possible to obtain discriminatory information.

In order to encode the local structure of a patch, the difference in intensity of a pair of blocks ("microblocks") in the image patch is taken. An image patch is usually 9 \( \times \) 9 to 15 \( \times \) 15 pixels in size and a smaller microblock inside the patch.
The final vector is formed by sequentially recording the difference in intensity of each microblock into a single vector. As far as we consider the video sequence, the intensity and intensity difference of microblocks is calculated in 3D dimension, taking into account the changes that occur in subsequent frames [21]. At the final stage, descriptor is fed to the classifier to categorize the actions performed on the video sequence.

3. Experimental results
The multimodal human action data set (UTD Multimodal Human Action Dataset, UTD-MHAD) [18] was used as a database for the experiments. The UTD-MHAD dataset consists of 27 different actions performed by 8 subjects and contains data received from sensors of various modalities: an RGB camera and a depth sensor.

Analysis of the results obtained shows that the efficiency of the presented method with an additional stage of image improvement leads to an increase in accuracy by 2-4%, in comparison with the efficiency of this algorithm without using the image improvement stage.

Acknowledgments
This work was supported by the Russian Ministry of Education and Science in accordance to the Government Decree № 218 from April 9, 2010 (project number № 074-11-2018-013 from May 31, 2018 (03.G25.31.0284)).

References
[1] Voronin V, Pismenskova M, Zelensky A, Cen Y, Nadykto A, Egiazarian K 2018 Proc. SPIE Security + Defence (Berlin) vol 10802
[2] Lowe D G 2004 Int. j. of computer vision 60 (2) p 91- 110
[3] Chen M, Hauptmann A 2009 Mosift: Recognizing human actions in surveillance videos
[4] Ji X F, Wu Q Q, Ju Z J, Wang Y Y 2014 Int. J. of Automation and Computing 11(5) p 500-509
[5] Franz V A, Voronin V V, Marchuk V I, Fisunov A V, Pismenskova M M 2013 J. Engineering Bulletin of the Don 3 (26)
[6] Wang H, Kläser A, Schmid C, Liu C L 2013 Int. j. of computer vision 103(1) p 60-79
[7] Laptev I, Marszalek M, Schmid C, Rozenfeld B 2008 Proc. CVPR p 1-8
[8] Dalal N, Triggs B, Schmid C 2006 Proc. European Conference on Computer Vision
[9] Solmaz B, Assari S M, Shah M 2013 J. Machine vision and applications 24(7) p 1473-1485
[10] Oliva A, Torralba A B, Guérin-Dugué A, Hérault J 1999 Proc. Int. conf. on Challenge of Image Retrieval p 9-9
[11] Voronin V, Semenishchev E, Ponomarenko M, Agaian S 2018 Proc. Electronic Imaging vol 2018 (13)
[12] Voronin V, Tokareva S, Semenishchev E, Agaian S 2018 Proc. 2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI) p 5-8
[13] Grigoryan A and Agaian S 2004 Image enhancement. Advances in Imaging and Electron Physics (New York: Academic) p 165–243
[14] Agaian S S, Silver B, Panetta K 2007 Proc IEEE transactions on image processing 16(3) p 741-758
[15] Panetta K, Wharton E, Agaian S 2007 Proc IEEE Tran. Systems, Man, and Cybernetics, Part A: Systems and Humans
[16] Wan Q, Rao S P, Kaszowska A, Voronin V, Panetta K, Taylor H A, Agaian S 2018 Proc. SPIE 10668
[17] Jourlin M, Pinoli J C 2001 J Advances in Imaging and Electron Physics vol 115 p 129 - 196
[18] Chen C, Jafari R, Kehtarnavaz N 2015 Proc. of IEEE International Conference on Image Processing
[19] Solmaz B, Assari S M, Shah M 2013 J Machine vision and applications 24(7) p 1473-1485
[20] Mehta R, Egiazarian K E 2016 Proc. IEEE Transactions on Image Processing 25(4) p 1604-1616
[21] Voronin V, Zhidanova M, Semenishchev E, Zelensky A, Tokareva O 2020 Proc. Int. conf. Signal Processing, Sensor/Information Fusion, and Target Recognition XXIX vol 11423