Advances in Intelligent Systems Research, volume 147
International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Human Action Recognition Based on Deep Images and Dense Trajectories

Xiaopeng Cui[1, a)], Binwen Fan[2, b)] and Jingyu Yi [1]

[1] *Harbin Institute of Technology, Shenzhen, China*
[2.] *Harbin Institute of Technology, Harbin, China*

[a)] hitsz_cuixiaopeng@163.com
[b)] Corresponding author: 472299471@qq.com

Abstract. The main content of this paper is to implement a human action recognition method based on depth image and dense trajectories. Firstly, the binocular RGB camera is used to collect images, and then the depth image is obtained through stereo matching algorithm. We use depth images to extract human action sequences. Then we choose dense optical flow field to calculate the trajectory of human sequence. After that, we compare the HOG method with MBH method and HOF method. Finally, we use SVM to complete the recognition of human actions.

Key words: Human Action Recognition; Depth Map; Dense Trajectories; SVM.

INTRODUCTION

The main problem to be solved in human motion recognition based on vision is to process and analyze the original image or image sequence data collected by a computer, and to learn and understand the action and behavior of the person [1]. As the development of society, people increasingly want to analyze themselves and their surroundings through a more intelligent and convenient way. With the rapid development of computer technology in recent decades, people can identify human actions in a completely new way. Human action recognition based on visual images can be applied in many areas, such as VR games, elderly monitoring, sports competition discrimination and so on. Most video records are people's activities as the main body of social activities. It is of great academic and application value to study human motion recognition in video, whether it is from the point of view of security, monitoring, entertainment, or personal archiving [2]. The depth map we use can use depth information to improve the accuracy of action recognition and reduce the probability of erroneous recognition.

OBTAINING A DEEP IMAGE FROM BINOCULAR CAMERA

The depth map is obtained by using two parallel RGB cameras and a series of arithmetic operations. They include calibration of binocular cameras, stereo matching of the results after calibration, and finally get the depth map we need.

Calibration of Binocular Camera

The first step to get the depth map is to calibrate binocular cameras. The purpose of the single target is to get the internal parameters of a single camera, and the subsequent binocular correction is to use the relative position of the two cameras to complete the full camera calibration task.

First, we get a pinhole camera model with an image point as a point, as shown in Fig. 1. We can get the relationship between camera coordinate system and world coordinate system:

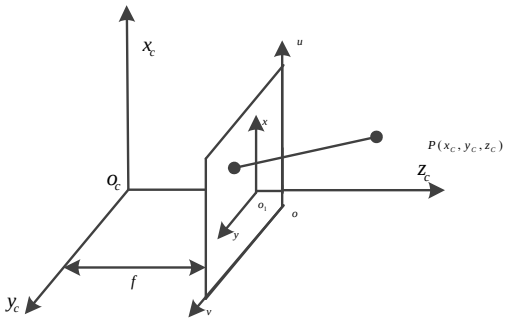

FIGURE 1. Pinhole camera model

First, the camera coordinate of point P is (X_C, Y_C, Z_C), its world coordinate is (X_W, Y_W, Z_W). $\mathbf{R}$ is the rotation matrix of the camera, and $\mathbf{t}$ is the translation vector of the camera [3] By $x = f/Z_C \, X_C$, $y = f/Z_C \, Y_C$ we can get.

$$Z_C \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} = \mathbf{M_1 M_2 X_W} = \mathbf{M X_W} \tag{1}$$

In (1), f_x and f_y are the focal length of the camera. $\mathbf{X_W}$ is the homogeneous coordinate of the spatial coordinates in the world coordinate system $\mathbf{M_1}, \mathbf{M_2}$ are the internal and external parameters of the camera, respectively $\mathbf{M}$ is called the projection matrix, which is composed of the internal and external camera parameters determined [4].

Then we calibrate the monocular camera by Zhang Ding your calibration method. The binocular correction is based on the internal parameters of the single target, the rotation matrix and the translation vector between the binocular and distinguish the aberration and line alignment of the left and right images, so that the coordinates of the original point coordinates of the left and right views are consistent.

Stereo Matching

After binocular correction, the two images have been aligned, and they have been strictly matched. A point on the image can be found on the corresponding image, and the two points must be on the same line. The opposite pole line of the two images is on the same horizontal line, which simplifies the process of finding the same pixels in the stereo matching. Stereo matching is to find matching points with the best line, and row alignment makes two-dimensional search of pixels become one-dimensional search.

We finally select the SGBM (Semi Global Block Matching) algorithm, SGBM algorithm based on mutual information, approximate global matching, and complete the two-dimensional smoothing constraint by multiple one-dimensional constraints.

Mutual information $MI_{1,2}$ is defined by the information threshold H_1 and H_2 of two images and their joint information threshold $H_{1,2}$, which are defined as follows:

$$MI_{1,2}=H_1+H_2+H_{1,2} \tag{2}$$

The information threshold H of a single image is calculated by the probability distribution of histogram description, and the information threshold $H_{1,2}$ is calculated by the joint probability distribution of the gray level of the cross-matching image [5].

The core step of the SGBM algorithm is to select the matched primitives, construct the cost energy and function of the scanning line based on multiple directions, and obtain the optimal solution of the energy cost and function.

Finally, we get the depth map generated by SGBM algorithm.

EXTRACTION OF ACTION SEQUENCE

After the completion of the deep image generation, we need to deal with the depth image. We need to separate the foreground of motion from the depth image we have obtained. Here we use the Vibe algorithm.

The idea of Vibe algorithm is to set up a collection belonging to every pixel. The origin of the set midpoint is composed of two parts, one is the pixel value of the pixel, and the other is the pixel value in the surrounding area. Then, each pixel of the detected image is compared with the set. We determine whether the point to be detected belongs to this set, and identify the foreground and background based on this.

We make $v(x)$ a pixel value of x in the image, and we set up a sample set with a sample size of N. As shown in Fig.2, we also get a range $S_R(v(x))$ with R as the radius and the pixel x as the center of the circle.

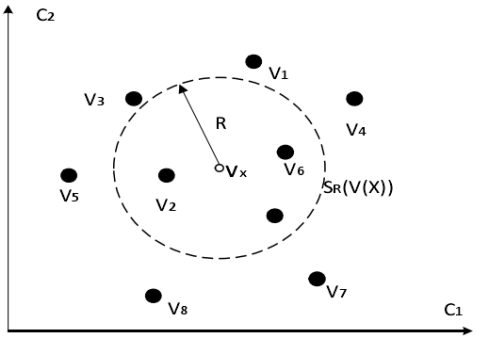

FIGURE 2. A schematic diagram for judging the degree of correlation

$$M(x)=\{v_1,v_2,\cdots,v_N\} \tag{3}$$

$$\#\{S_R(v(x))\cap M(x)\} \tag{4}$$

From formula (3) and (4), if $\#$ is greater than or equal to a given threshold $\#_{\min}$, it shows that their correlation is relatively high, so we regard it as a background point.

Finally, the vibe algorithm is used to extract the depth map, and the result is shown in the following figure.

（a） （b）

FIGURE 3. (a) is the original depth map. (b) is the extracted human action.

CHARACTERISTIC CALCULATION ALONG DENSE TRAJECTORIES

After getting the action sequence, we begin to extract their features. We proceed in two steps. First, we obtain the dense optical flow trajectories of pixels. In the second part, we construct spatiotemporal grids along the trajectory, compute descriptors from the grid, and finally get the features.

Obtain Dense Trajectories

The sample set in this paper is a unit video with ten frames. First, an 8-scale image of Pyramid is built for each image. The ratio factor of the adjacent layer is $\sqrt{2}$, then the dense sampling is carried out in each scale, and the sampling interval is 5.

Considering that there is no structural information in the flat area, we remove the point where the eigenvalue of the autocorrelation matrix is less than the threshold T.

$$T = 0.001 \times \max_{i \in I} \min\left(\lambda^1{}_i, \lambda^2{}_i\right) \tag{5}$$

$\lambda^1{}_i, \lambda^2{}_i$ is the eigenvalue of the autocorrelation matrix M of the i points in the image I. Max and min denote the maximum and minimum values respectively, and the effective feature points can be obtained by setting the threshold

After obtaining the feature points, we use Frame back method to build dense optical flow field. As the formula (6), we set up the optical flow field in every frame of the video [6].

$$w_t = \left(u_t, v_t\right) \tag{6}$$

u_t and v_t represent horizontal and vertical components.

We use the corresponding points set in the video as trajectories, and they are combined from the same point in different positions in the continuous frame, as follows:

$$\left(P_t, P_{t+1}, P_{t+1}, \cdots\right) \tag{7}$$

Where $P_t\left(x, y\right)$ is the pixel P of t at any time, and its point is at t+1 time, we use the median filter M to get the position of P_{t+1}. The specific process is shown by the following formula:

$$P_{t+1} = \left(x_{t+1}, y_{t+1}\right) = \left(x_t, y_t\right) + \left(M * w_t\right)\big|_{(x_t, y_t)} \tag{8}$$

The Calculation of Descriptors

As shown in Fig.4., after obtaining the trajectory, we follow the trajectory in a continuous L frame to track the $N \times N$ neighborhood with the feature point as the center to form the $N \times N \times L$ space-time pipeline, and further divide the space-time pipeline into the time and space grid of the $n_\sigma \times n_\sigma \times n_\tau$ [7].

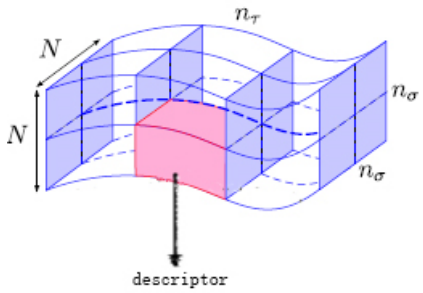

FIGURE 4. Space-time pipeline and space-time grid

Then we use the local features of depth map to calculate HOG (Histogram of Oriented Gradient) features in the grid. The gradient direction histogram describes the local contour characteristics by using the statistical information of the local gradient direction of the image, and has a certain adaptability for rotation, translation and illumination change.

At the end of each grid we get 8-dimensional hog features. We will analyze it in the next chapter

HOG features only use depth information of depth map instead of using optical flow information, so we add HOF (Histogram of Optical Flow) and MBH (Motion Boundary Histogram) descriptors.

HOF is based on statistics of optical flow information of local pixels. It first divides the image into different cell, then calculates the statistical information of each cell and combines it into the information of the block, and finally makes the normalization operation to form the feature vector. The MBH descriptor obtains the motion information representing acceleration by calculating the gradient of the optical flow image.

We combine these three descriptors to get a descriptor which is different from using HOG only and analyze their effects in the next chapter.

EXPERIMENT RESULT

In this paper, 7 different human behaviors under the laboratory environment were collected, including 5 individual behaviors and two interactive behaviors, which were completed by people with different sex and height. The five kind of single person behavior is eaten, call, sleep, walk and jump. two kinds of interactive behaviors, fight and handshake. A total of 700 videos were collected and ten frames per video. We selected 595 segments as training samples and 105 as test samples.

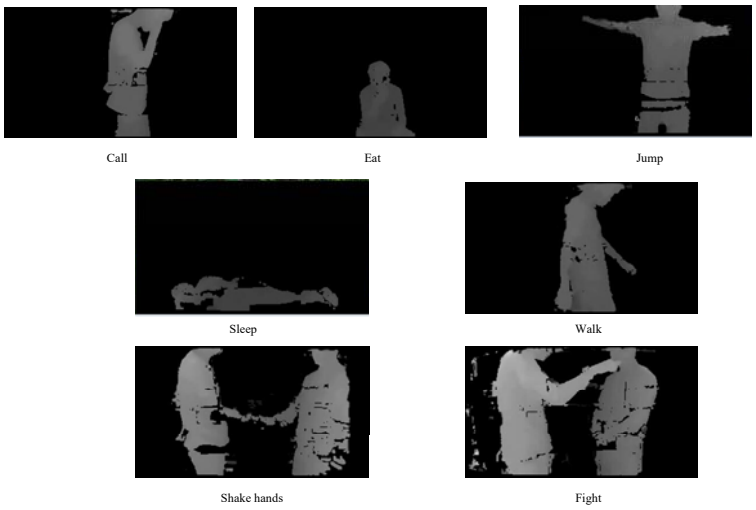

FIGURE 5. Schematic diagram of human movement

We use SVM (Support Vector Machine) to train these samples, and SVM maps the sample space into a feature space with a high dimension to infinite dimension through a nonlinear mapping. The nonlinear separable problem in the original sample space is transformed into a linear separable problem in the feature space. SVM can be used in our human action recognition very well.

We get a feature that contains the direction gradient information of the depth map when we only use the HOG feature to calculate the feature in the space-time pipeline. We use SVM for training. At the same time, we also used the HOG, HOF and MBH descriptors to train, and their training process is as follows:

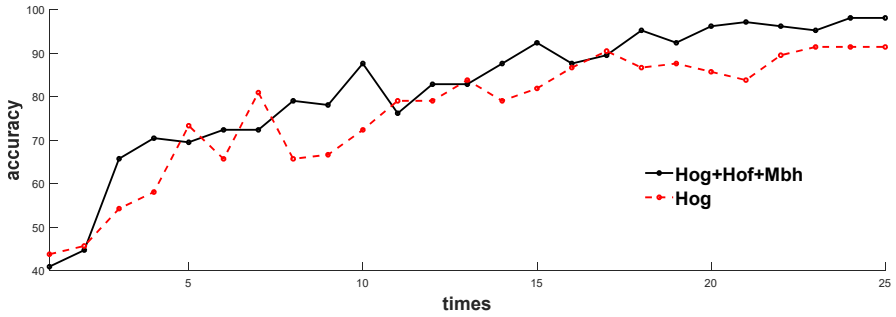

FIGURE 6. The training process of SVM

We can see that using HOG alone can achieve a 91.4285 recognition rate after 23 times and achieve a better recognition effect. With the combination of three kinds of descriptors, the recognition rate has increased to 98.0952, achieving a very good recognition effect. For the difference between the two characteristics, we analyze his confusion matrix as shown below:

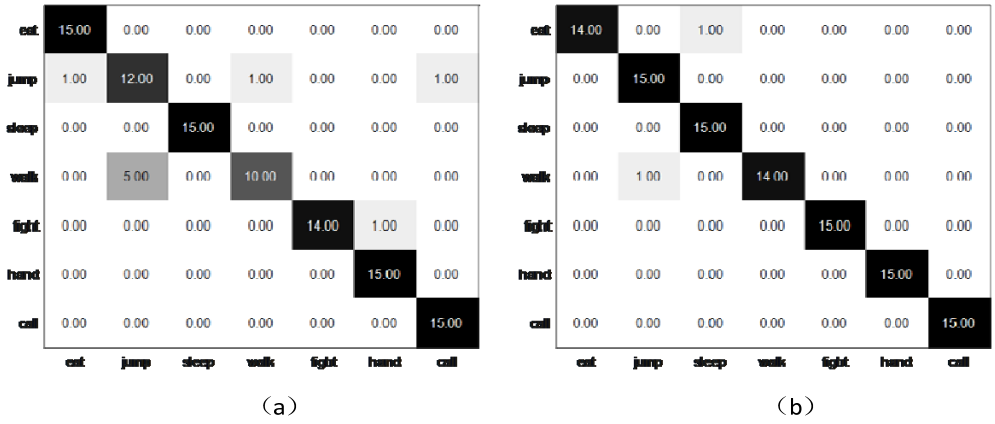

FIGURE 7. The confusion matrix

From Fig.7(a)., it is not difficult to see that the parts of the error recognition are mainly concentrated on the two categories of jump and walk. The two categories are difficult to distinguish between the two types of behavior before the features of the optical flow are added, because their movements vary greatly, and the HOG is not enough to describe the difference between the two.

After the combination of the three features, we can clearly see in Fig.7(b). that some recognition rates for jump and walk have been greatly improved, and only two samples have been wrong. This shows that the newly introduced optical flow characteristics HOF and MBH play a great role in the discrimination between them.

Finally, we collect the characteristics of HOG and HOF and MBH in the spatiotemporal grid along the path of dense optical flow and combine them to achieve satisfactory results.

CONCLUSION

In this paper, we first use the imaging principle of binocular camera to calibrate the camera, so that we can prepare for the stereo matching later. Then we use the SGBM algorithm to successfully get the matching depth map, then we extract the depth map of the human action sequence, and divide it into 7 categories, a total of 700 samples, to carry out our human action recognition test. When we extract the characteristics, we first obtain dense optical flow field, then we get the trajectory of the light flow, and carry out the characteristic calculation in the grid it formed. We use SVM to compare the results of simple use of hog and the comprehensive use of hog, MBH, Hof, and finally get more than 98 percent knowledge under the three combinations. It achieves ideal human motion effect.

ACKNOWLEDGMENTS

This work was supported by Special funds for the development of strategic emerging industries in Shenzhen (no. JCYJ20150625142543448).

REFERENCES

1. Hu Q, Qin L, Huang Q. A Survey on Visual Human Action Recognition[J]. Chinese Journal of Computers, 2013, 36(36):2512-2524.
2. Aggarwal J K, Ryoo M S. Human activity analysis: A review[J]. Acm Computing Surveys, 2014, 43(3):1-43.
3. Cheng Tian, Yang Sisi, Feng Rong, et al. Based on binocular calibration for the fall detection algorithm of solitary elderly, [J]. sensor and microsystem, 2014, 33 (10): 100-103.
4. Cai Yao, Yu Tao, leaf kumyoung. Binocular stereo vision, camera calibration and distortion correction software and application of [J]. computer, 2012 (19): 9-10.
5. Zhang Huan, Amway, Zhang Qiang, et al. SGBM algorithm and BM algorithm analysis and research [J]. mapping and spatial geographic information, 2016 (10): 214-216.
6. Farnebäck G. Two-Frame Motion Estimation Based on Polynomial Expansion[C]// Scandinavian Conference on Image Analysis. Springer-Verlag, 2003:363-370.
7. Wang H, Kläser A, Schmid C, et al. Dense Trajectories and Motion Boundary Descriptors for Action Recognition[J]. International Journal of Computer Vision, 2013, 103(1):60-79.