Journal of Robotics, Networking and Artificial Life, Vol. 4, No. 1 (June 2017) 5–9

# Action recognition based on binocular vision

**Yiwei Ru[1,2*], Hongyue Du[1],Shuxiao Li[2], Hongxing Chang[2]**

[1] *Harbin University of Science and Technology, Harbin, China*
[2] *Institute of Automation Chinese Academy of Sciences, Beijing, China*

*E-mail: *ruyiwei2014@ia.ac.cn*

**Abstract**

Aimed at the problem that the recognition accuracy of the monocular camera is low, we propose a binocular vision recognition algorithm for action recognition based on HART-Net(Human action recognition networks).Firstly, the left and right views obtained by the binocular camera are matched to obtain the depth map of the human body .Then, the depth information is projected onto the three planes, the projection images of three directions are used to construct MHI (motion history image), and are combined into a new image. Finally, we use HART-Net to train a classifier for action recognition. Experimental results show that the binocular recognition algorithm is 18% more accurate than the monocular recognition algorithm.

*Keywords*: action recognition; binocular version; convolutional neural networks; motion history image;

## 1. Introduction

Since 1980s, the concept of search and rescue robot has been presented. Recognizing human activity is one of the important areas of computer vision research today[1].Its applications include video surveillance, virtual reality, human-computer interaction and others. In recent years, many human motion recognition research is based on monocular vision. Although the recognition algorithm based on monocular vision has achieved good results, but due to the monocular camera's own limitations, this approach is very sensitive to complex backgrounds, so when the background changes significantly, the accuracy of recognition will decline. In order to improve the detection accuracy of human action recognition algorithm in complex background, people use time-of-flight or structured light technology to obtain the depth map of the object. However, for some applications such active sensors are not suitable. For example, in outdoor setup or in a scenario with multiple autonomous robots whose active sensors would interfere to each other[2].

In order to solve the problem that the human action recognition algorithm is sensitive to the background under the monocular visual and cannot obtain the object depth map based on the structured light or time-of-flight technology in the outdoor environment[3], we propose an action detection algorithm based on binocular vision. In order to improve the accuracy of motion recognition in binocular vision, a novel convolutional neural networks architecture named HART-Net (human action recognition net) is presented. Firstly, the left and right views obtained by the binocular camera are matched to obtain the depth map of the human body under the camera coordinate system (O-XYZ). Then, the depth information is projected onto the three planes of O-XY, O-YZ and O-ZX respectively in the camera coordinate system[4]. In order to reflect the motion of the timing information, the projection images of three directions are used to construct MHI (motion history image), and then the three MHIs are regarded as the three channels of the image to construct a new image. We finally use CNN to train the classifier for action recognition.

## 2. Related Work

### *2.1. Stereo Matching Algorithm in Binocular*

Binocular stereo matching algorithm can be divided into: local matching algorithm and global matching algorithm. Local matching algorithm mainly compare the matching point within a certain range of loca characteristics to match. Bigone et al.[5] used the edge information of the image as a matching feature. Marr6 et al. Proposed to use zerocrossing as the basis, plus and continuous constraint iterations obtained after the disparity map. Nevatia and Medina[7] use line segments as matching primitives.

(a) left image (b) right image (c) depth map

Fig1. Left-right image after line alignment, and depth map

(a)Outline of the human body (b) MHI in O-XY (c) MHI in O-YZ (d)MHI in O-ZX

Fig2. Outline and MHI image

Global matching algorithm is generally used to scan the line or the overall consideration of the image information to be matched to solve the disparity. Boykov[8] et al. first introduced the graph-cut theory into the stereo matching algorithm. For the first time, Kolmogorov[9] incorporated the solution of the occlusion problem into the construction of the energy function. Confidence propagation was proposed by Sun Jian[10] of Microsoft Research Asia. Ohta[11] proposed an edge-characterized DP algorithm.

### *2.2. Action Recognition Algorithm Based on RGB and RGBD*

As the RGB image is different from the RGBD image[16] which can be directly segmented out of the body region using depth information, RGB-based action recognition generally requires the use of segmentation methods to extract the contours of the human body. Ali and Aggarwal[12] define an action boundary by using a feature vector that contains three angles to the main part of the human body. Hanjalic[13] et al. used logical tale units (each of which is represented by one or several events not related to timing) to detect the motion bounds. Zhai and Shah[14] used the Markov chain Monte Carlo technique to segment the temporal scenes in various videos. Shi et al. [15] used semi-Markov model to achieve the action segmentation. For the RGBD method, the depth of the human body can be obtained by face detection or head and shoulder detection, and then the human body position can be extracted using the depth information. After the outline of the human body is obtained, the MHI[17] is used to reflect the time when the human action takes place and the time to change. The obtained MHI is used as the input picture of the classifier.

## 3. Proposed Method

### *3.1. Overall framework*

The proposed method mainly include the following four modules:

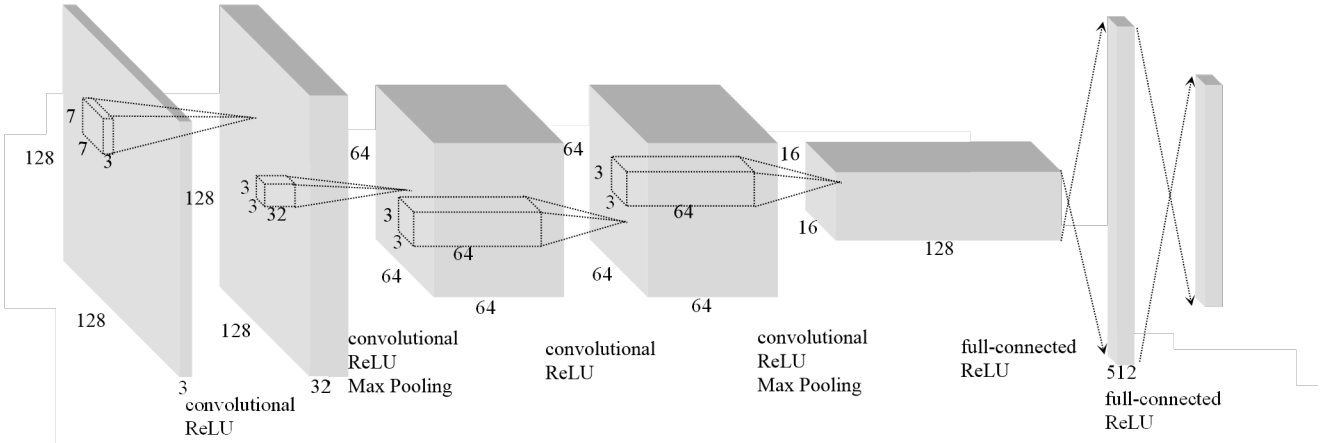


Fig3. HART-Net architecture

1. Human body region extraction. After using the camera's internal and external parameters for line alignment between the left and right images, we can get the depth image by seed pixel propagation matching algorithm, as shown in Fig 1. In order to get the location of the human body, we first use the seetaface detection algorithm to detect the face position. Then, the average depth value of the face is computed in the depth map, denoted as *D*(mm). Finally, [D-300, D+300] is used to extract the contours of the human body.

2. Human body region refinement. As the ground and the human body is connected, the acquired human body region will contain part of the ground information. In order to remove the ground, we use RANSAC algorithm to detect plane in the picture, and then remove it from the body outline image, as shown in Fig2 (a).

3. Get MHI images. After the human body depth map is obtained, the depth map is cropped and cut into 480 * 480. According to the position information of the human face, the human body profile is located at the middle position of the image, and the depth map is projected onto the O-XY, O-YZ, O-ZX three planes, and then the three planes were accumulated to get MHI images, as shown in Fig 2. In order to meet the different speed of movement of different people, we selected 15,20, 25, 30 images in the same sequence of motion for each action to construct MHI. Finally, the MHI of the three projection planes is merged into one image as an RGB image as shown in Fig 4.

4. Action recognition by HART-Net. By using the MHI image as the input image of the classifier A novel convolutional neural networks architecture named HARTNet (human action recognition net) is presented in the following subsections.

### *3.2.HART-Net architecture*

In order to meet the training needs, we propose our own convolution neural network training framework: HARTNet, as shown in Fig3. This convolutional neural network has four convolutions and two fully-connected layers. The input of the first convolution layer is 64 * 64 * 3, this layer contains 32 convolution kernel, and the convolution kernel size is 7 * 7 * 3. The second layer contains a pooling layer with max pooling. The output of the first convolution layer is the input of the pooling layer. The pooling layer contains 64 convolution kernels. The size of the convolution kernel is 3 * 3 * 32. The third layer and the fourth layer are similar to the first layer and the second layer, and the specific parameters are shown in Fig3. The output of the last fullyconnected layer is fed to a soft-max layer, and the output of the soft-max layer is used as the confidence of the classifier for each class to be categorized.

Notice that, the ReLU non-linearity is applied to the output of every convolutional and fully-connected layer.

### *3.3.Training process*

Our classification task has 8 kinds of action: Drinking, eating, making phone call, reading book, walking, waving, hand clapping, boxing. For each category, there are 1000 MHI images. For these images, we will first of all resize them to 130 * 130.For each resized picture, we randomly cut out 128 × 128 windows on these images as training samples, and then horizontally rotated all the training samples. After these transformations, our training sample is 50 times the original. Although the samples obtained from these transformations are highly dependent, training with these samples still greatly improves the training results. Because our training sample

size is relatively large, limited to GPU performance, so in the course of training batch_size is general set to 16.Because the difference between the samples is relatively small, we set base_lr to 0.01 and lr_policy to "inv" where gamma is set to 0.0001 and power is set to 0.75.

## 4. Experiment

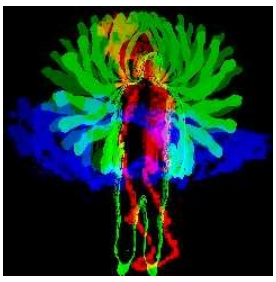

Fig4. MHI with three channels

In order to evaluate the performance of the classifier accurately, we use the self-made set, using crossvalidation to test the classification of the results. The MHI is divided into five parts, one fifth is test set, three fifth is training set, and the remaining one fifth is verification set. The verification set and training set are mainly used to test the classification performance of HART-Net during training. The test set is used to test the performance of a classifier trained on RGB and RGBD samples in a real sample. For comparison, we use the projection of RGBD in the O-XY direction as a contour in the RGB.

Fig5.outline of the human body in RGB image

From the classification result of HART-Net, we can get the corresponding classification confusion matrix. Through the confusion matrix, we can see: For actions of drinking, eating, making phone call and reading book, the classifier performance trained by the RGBD sample is higher than the classifier trained by RGB by about 8%. But for actions of walking, waving, hand clapping and boxing, the classifier performance trained by RGBD is significantly better than the classifier trained by the RGB, because these three-dimensional characteristics of the action are more obvious. For RGBD, classifier training based on RGBD can get more information, so the classification performance will be better. Overall, the classifier performance training by the RGBD classification is higher than the classifier training by RGB classification by about 18%.

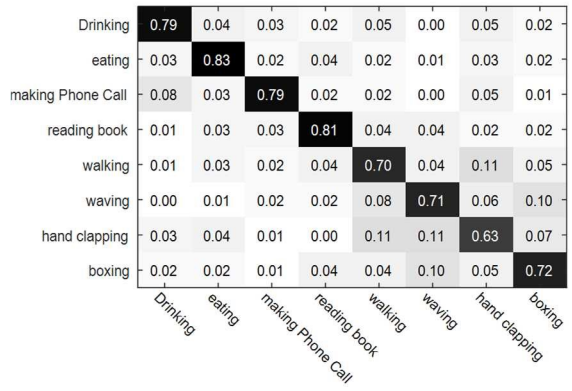


Fig6. Confusion matrix - RGB

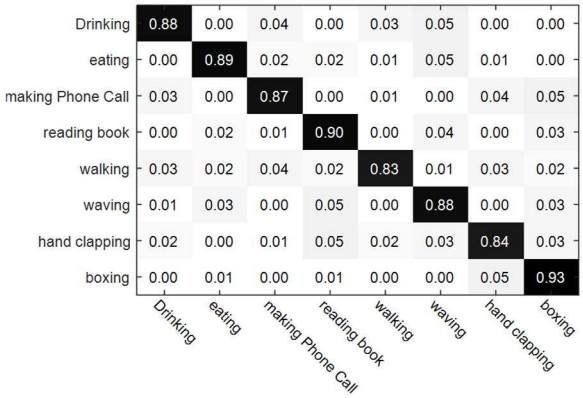


Fig7. Confusion matrix - RGBD

## 5. Conclusion

In this wok, we propose a human motion detection fork based on binocular vision. In speed, we can easily estimate the location of the human body according to face position or head and shoulder position .However, based on monocular human action recognition algorithm need to use motion information and graph cut algorithm to get the outline of human motion, so compared with monocular vision algorithm our algorithm has certain advantages in speed. In terms of applicability, our algorithm can be used outdoors, while algorithms based on structured light and time of flight can only be used indoors, and there is interference between devices, so our algorithm has the same advantage in the applicability.

**Acknowledgements**

This work is jointly supported by the Natural Science Foundation of Heilongjiang Province of China (F2015043) and the Special Research Program for Basic Science and Advanced Technology of Chongqing of China (cstc2016jcyjA0362).

**References**

1. S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR'98*, pages 232–237, Santa Barbara, CA, June 23-25, 1998.
2. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV'05, volume 2*, pages 1395–1402, Beijing, China, Oct. 17-21, 2005.
3. A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(3):257–267, Mar. 2001.
4. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS'05*, pages 65–72, Beijing, Oct. 15-16, 2005.
5. Frank Bignone, Olof Henricsson, Pascal Fua, et al. Automatic extraction of generichouse roofs fromhigh resolution aerial imagery. Proceedings of the 4th European Conference on Computer Vision: *Cambridge*, United Kingdom, Springer Verlag, 1996: 85-96.
6. D. Marr, T. Poggio. Cooperative Computation of Stereo Disparity. Science，1976, 194: 209-236.
7. Cordelia Schmid, Andrew Zisserman. The geometry and matching of curves in multiple views. *Proceedings of the 5th European Conference on Computer Vision*: Freiburg, Germany, Springer Verlag, 1998: 104-118.
8. Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on*, 2001, 23(11): 1222-1239.
9. Kolmogorov V, Zabih R. Computing visual correspondence with occlusions using graph cuts. Computer Vision, 2001 *ICCV 2001 Proceedings Eighth IEEE International Conference on. IEEE*, 2001, 2: 508-515.
10. X. Sun, X. Mei, S.Jiao, et al. Stereo matching with reliable disparity propagation. *In Proc: 3DIMPVT*, 2011: 132-139.
11. Ohta Y, Kanade T. Stereo by intra-and inter-scan line search using dynamic programming. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1985, (2): 139-154 .
12. J. Cech, J. Sanchez-Riera, and R. Horaud. Scene flow estimation by growing correspondence seeds. In *IEEE Conference on Computer Vision and Pattern Recognition, 2011.*
13. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. *European Conference on Computer Vision, 2012.*
14. P.Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision, 57:137–154, 2004.*
15. W G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *European Conference on Computer Vision, 2008.*
16. D. Wu, F. Zhu, and L. Shao. One shot learning gesture recognition from rgbd images. In *Computer Vision and Pattern Recognition Workshops, 2012.*
17. A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers. Stereoscopic scene flow computation for 3d motion understanding.*International Journal of Computer Vision, 2010.*
18. Bobick A F, Wilson A D. A State-Based Approach to the Representation and Recognition of Gesture. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 1997, 19( 12) : 1325−1337.