Proceeding of the IEEE
International Conference on Robotics and Biomimetics (ROBIO)
Shenzhen, China, December 2013

A New Approach for Hand-Waving Detection in Crowds

Nannan Li, Xinyu Wu, Dan Xu, Ruiqing Fu and Wei Feng

Abstract— **In this paper, we propose a new approach to detect hand-waving motion in crowds. Different from previous approaches which are often based on segmentation and motion detection, our method can be seen as a complexity reduction process from the problem of 3D motion detection to 2D object detection. Through arranging the same row of per video frame along the time sequence, we obtain 2D images composed of traces of moving object, and in these images hand-waving motion are represented as 2D patterns which have strong periodicity. Considering the similarity between these patterns, we first use Chamfer Distance Matching to locate the patterns' position roughly. Then we convert 2D patterns to 1D signals and apply Fourier Transformation to these signals. Finally, we identify a region as a hand-waving one, if its intensity spectrum of Fourier Transformation of 1D signals has a peak response between specified locations, and the value of the peak is greater than a threshold. Experiments on two self-built hand-waving datasets demonstrate that the proposed method can detect hand-waving motion effectively.**

I. INTRODUCTION

In recent years, with the amount of surveillance video in use increasing, intelligent video analysis has been becoming a research hotspot. Among numerous studies upon the topic, the ability to automatically detect actions performed by people in crowds has particular value for surveillance applications. Just as object recognition is a key problem in image understanding, action recognition is a fundamental challenge for interpreting video [1]. The same action has different expression due to different actors, varying environment conditions and camera viewpoints. Many methods have been proposed for detection of a set of human actions, such as falling, hands waving, running. Those approaches can be broadly classified as two categories. One is based on volumetric analysis of video, the other one is employing video features of varying types. Generally speaking, there are three leading representations of video features, which are shape based methods, interest-point based methods and optical-flow based methods [2]. Up to now, both of these types of methods have achieved promising detection results, but also have significant limitations.

In shape-based methods, silhouette-based approaches attempt to recognize actions by characterizing the shape of the actors silhouette through space-time, and thus are robust to variations in clothing and lighting [1]. Although these approaches have showed robust detection performance on a number of actions, there exist some drawbacks: firstly, they need the background clean enough to obtain complete silhouette; secondly, the silhouette blob covers the actions which take place within it. Optical-flow based methods calculate optical-flow between adjacent frames and use it as basis for action recognition. For example, in [3], the authors propose the conception of motion history image (MHI), which is a description of the cumulative spatial distribution of motion energy in a video sequence. The authors in [4], [5], [6] directly use the optical flow to derive a representation which can be used for recognition. Saad Ali et al. [2] propose a set of kinematic features that are derived from the optical flow for human action recognition in videos. Compared with shape-based method, these kinds of approach require no background subtraction, so they can be applied in complex environments, even endure limited camera motion. Apart from these approached mentioned above, another important direction of research is the use of space-time interest points and their trajectories for action and activity analysis. These low-level visual features, such as Gist, SIFT, STIP and so on, are used for action recognition. Works by Laptev and coworkers [7], Nibbles et al. [8] and Dollar et al. [9] belong to this class. The main advantage of these means lies in its robustness to occlusion, since it is not necessary to track or detect the whole human body.

In this paper, we pay our attention on hand-waving detection, which has practical significance, for example, when people say hello to friends far away or call for help in case of an emergence, they often wave hands. Being different from previous approaches which are often based on segmentation or motion detection, our method needs neither human body detection nor optical flow calculation which may be inaccuracy and time consuming in complex environment. It can be seen as a complexity reduction process from 3D motion detection to 2D object detection. Our method is mainly based on template matching and periodicity analysis of patterns in 2D images, which are concise and efficient. This property makes it more robust in complex environment.

Firstly, we convert a 2D video clip to 2D images by arranging the same row of per video frame along the time sequence. Traces of hand-waving movement are reflected in 2D patterns which have strong periodicity in these 2D images. In consideration of the similarity of those 2D patterns, we make use of Chamfer Distance Matching to obtain approximate regions where hand-waving motion occurs. Then we convert the 2D patterns in candidate regions into 1D signals, and

This work is partially supported by the National Science Foundational of China (61005012), Guangdong Innovative Research Team Program (201001D0104648280).

Nannan Li, Xinyu Wu, Dan Xu, Ruiqing Fu and Wei Feng are with Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen Key Lab for Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. Email:{nn.li, xy.wu, dan.xu, rq.fu, wei.feng@siat.ac.cn}

Xinyu Wu is also with the Chinese University of Hong Kong.

978-1-4799-2744-9/13/$31.00 ©2013 IEEE

perform Fourier Transformation on these signals. Through observing from the results of Fourier Transformation, we wipe out candidate regions which are wrongly identified, because traces caused by hand-waving motion have strong periodicity, and the results of intensity spectrum of Fourier Transformation have intense peak response at specified location. Lastly, we identify a candidate region as a target region with hand-waving motion occuring, if its intensity spectrum of Fourier Transformation of 1D signal has peak response between specified locations and the value of the peak is greater than a threshold. Fig. 1 shows the flow chart of our algorithm. We continue the paper as follows: in section II the proposed algorithm is described in detail. Performance evaluation is given in section III. Conclusion and Analysis are presented in section IV.

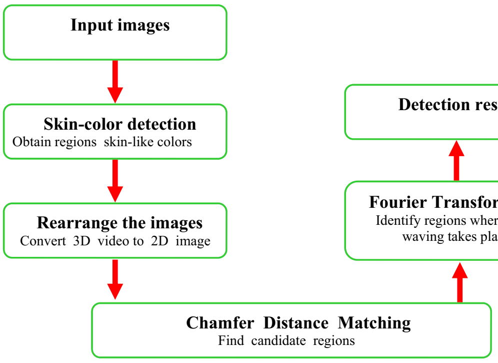

Input images
Skin-color detection
Obtain regions skin-like colors
Rearrange the images
Convert 3D video to 2D image
Chamfer Distance Matching
Find candidate regions
Fourier Transformation
Identify regions where hand-waving takes place
Detection result

Fig. 1: The flow chart of our algorithm

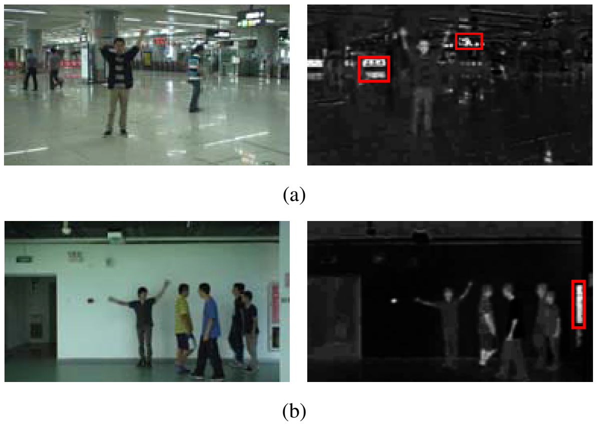
II. Approach

A. Skin-Color Detection

Complicated background in the dynamic environment brings a great change on hands localization and tracking. Skin-color detection based methods can deal with the problem of distinguishing objects with skin-like colors in the background. Referring to [10], regions with skin-like colors can be segmented in color images through a statistical elliptical boundary of skin-color model. We model skin-color in YCbCr color space, in which luminance component Y and chrominance components Cb and Cr can be considered separately. From the distribution of skin-color in CbCr color plane, the elliptical boundary can be derived from a Single Gaussian Model (SGM) [10]. The probability of a pixel's color belonging to skin-color is defined as:

$$p(x) = \frac{1}{(2\pi)|\Sigma|^{1/2}} \exp\{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}. \quad (1)$$

Where x denotes the color vector of a pixel, μ and Σ are skin-color mean vector and covariance in CbCr color plane respectively. For every frame in the video, we obtain a probability image which consists of probabilities of color belonging to skin-color at each pixel. Then we convert the probability image to gray image for subsequent processing. As showed in Fig. 2, more white regions have colors more similar to skin color. From Fig. 2, we can see that some non skin regions show strong similarity to skin regions, for example the regions occupied by red lights.


(a)

(b)

Fig. 2: Examples of skin-color detection. (a) is at a bus station;(b) is at the corridor of a building.Regions confined by bold red line exhibite strong similarity to skin-color.

Since hand-waving motion is a kind of periodic movement, we consider detecting such motion making use of periodicity. In order to make the periodicity of such motion more explicit and more conductive, we process gray images as follows: supposing that the original video has T frames and each frame is composed of M by N pixels, firstly, we specify a row in the first frame, then pick up all the corresponding rows from the entire remaining frames, and at last arrange the rows to form a image according to the time order. We perform the same operation to all the rows in the first frame, then we can get an image sequence constituted by T by N pixels, M frames, which is referred to as ”row image”. Examples of the row image are showed as Fig. 3.

From Fig. 3, we can note that if the row of gray image contains motion of periodicity, there are obvious periodic patterns in its corresponding row image, for example, as shown in Fig. 3. Fig. 3(a) shows the row image of row including hand-waving motion, Fig. 3(b) shows the row image of row including walking. On the other hand, rows containing none cyclical movements present in patterns having none regularity, for example, Fig. 3(c) shows the row image of row including none evident cyclical motion. From the manifestation of patterns, we note that periodic patterns produced from hand-waving motion share similar shape. So we consider to use the template matching method to locate hand-waving region in row image roughly. By performing the process, we can eliminate regions containing irregular motions which may cause false detection in the following steps and then focus on regions of interesting (ROI).

B. Candidate Region Selection By Chamfer Distance Matching

In this paper, we use chamfer distance matching to find the patterns produced by hand-waving with several predefined masks. Chamfer distance matching [11] is a popular technology for matching two edge maps, which has

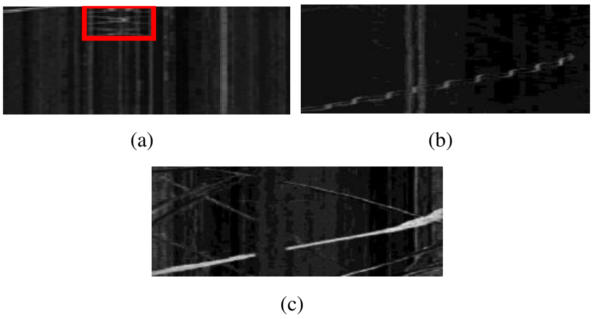


Fig. 3: Examples of 'row image'. (a) is the image of hand-waving, and the region confined by bold red line contains pattern with strong periodicity; (b) is the image of walking; (c) is the image containing irregular motion.

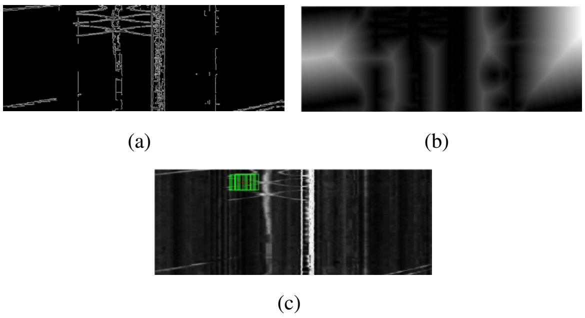


Fig. 5: An example of Chamfer Distance Matching. (a) is a edge map; (b) is the distance transformation map; (c) is the result of Chamfer Distance Matching, and regions confined by green boxes are the ones whose matching score are over the specified threshold.

low computational overhead and high robustness to clutter. Supposing $U_T(u_i \in U_T, i = 1, 2, ..., n)$ and $V_Q(v_j \in V_Q, j = 1, 2, ..., m)$ represent the point set of template edge image and the edge image of querying window, the chamfer distance between them can be computed as follows:

$$d_{chamf}(U_T, V_Q) = \frac{1}{n} \sum_{u_i \in U_T} \min_{v_j \in V_Q} \|u_i - v_j\|. \quad (2)$$

Actually, d_{chamf} is the mean of distances between edge point $u_i \in U_T$ and its nearest edge point $v_j \in V_Q$. To reduce the matching cost, firstly doing the distance transformation (DT) before calculating mean distance, which converts the query edge image into gray image by assigning edge pixel with zero and non edge pixel with the distance value to its nearest edge point. In our application, considering waving hands at varying speed and extent, we use a set of masks to carry out chamfer distance matching, as showed in Fig. 4. When the template slides through the querying image, every pixel get a chamfer matching score which represents how similar the edge map covered by the template centered at that pixel resembles the template. If a region's highest matching score among the templates is over a specified threshold, it means that the region has patterns with shape similar to templates and is identified as a candidate. Fig. 5 shows an example of the edge map, the distance transformation map and the matching result of one frame of row image.

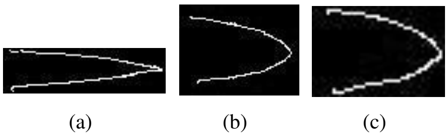


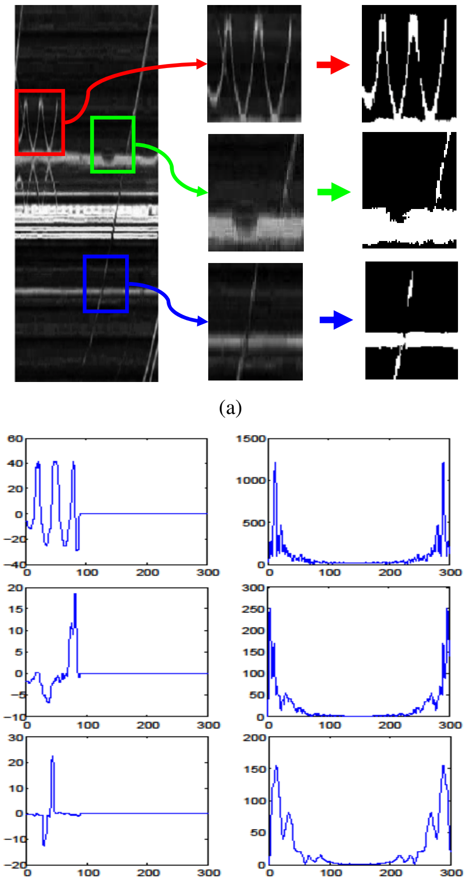
Fig. 4: Template sets

C. Hand-Waving Detection By Fourier Transformation

The matching result shows that although hand-waving pattern can be detected by Chamfer Distance Matching, there are false positive matchings. It is because that our templates have an angle like shape, every region having edges with shape of this kind will get a high matching score. But when observing those regions where hand-waving take place from a larger scope, we find out that edge curves within such regions exhibit strong periodicity. So we perform Fourier Transformation to edge curves in order to find a way of distinguishing true regions from false ones. At first, We choose a wider range around candidate regions, and the extent of the range is large enough to include at least two complete cycles. For example, the region confined by bold red lines in Fig. 6(a) has two and a half cycles. We rotate the edge map with 90 degree in order to process conveniently in next step. Then we convert the two dimensional edge maps to one dimensional signals for Fourier Transformation as follows: within the range, firstly, we convert the gray image into binary image, and then calculate the mean value of y coordinate location per column as value of the corresponding column for the one dimensional signal. After Fourier Transformation, we observe that the results from hand-waving regions have a maximum peak response between specific locations in the intensity spectrum of Fourier Transformation. Actually, edge curves of patterns from hand-waving performed with slight-change speed will have a maximum peak response at nearby locations of their Fourier Transformation spectrum. The entire process is showed in Fig. 6.

III. EXPERIMENT

To evaluate the performance of the proposed hand-waving detection algorithm, we test it on two self-built video clips. One of them is captured in a scene of the subway station. In the video one person stands close to the camera and there are other people going by behind him. The video length is 300 frames and the size of frame is 408 by 720. We select three templates with different shape, the size of which are 30 by 105, 55 by 85 and 60 by 80 respectively, as showed in Fig.4. The size of candidate region for Fourier Transformation is chosen as 101 by 101 in order to include at least two complete cycles. Before carrying out Fourier Transformation, we extend the length of one dimensional signals to 300 points by adding zero to the end, in order to make the location of the peak response of true hand-waving region be clearly differentiated from regions of non hand-waving or irregular motion. We identify a true positive detection,



(a)

(b)

Fig. 6: The illustration of the process of Fourier Transformation. (a) shows three different regions from a row image as well as their binary images. The one confined by bold red line contains hand-waving motion, the others include irregular motion. (b) shows their one dimensional signals and results of Fourier Transformation in intensity spectrum, from top to down they belong to regions confined by bold red, green and blue line.

if the location of maximum peak response in its intensity spectrum of Fourier Transformation is between 9 and 12, and the value of the peak is greater than 800. Examples of the detection results are showed in Fig. 7(a).

The other video is taken in the scene of a corridor in a building. This video has a cleaner background than video one; one person waves his hand in front of a white wall and several other people pass in front of him. The length of the video is 600 frames. The detection process is similar to that of video clip one, except that we consider the detection result of a region to be positive, if its intensity spectrum of Fourier Transformation has the maximum peak response between 11 and 13 location, and the value of the peak is more than 350 at the same time. Examples of detection results are showed in Fig. 7(b).

In our algorithm, there are two processes in which thresholds are needed. One process is Chamfer Distance Matching, and the other is Fourier Transformation. Only when a region is identified as a candidate according to the result of Chamfer Distance Matching, then we carry out Fourier Transformation on it. So the threshold of Chamfer Distance Matching is more influential on the detection result than that of Fourier Transformation. To describe how threshold affects the performance of our algorithm, we plot ROC of threshold of Chamfer Distance Matching, which is showed in Fig. 8. The True Positive Ratio(TPR) and False Positive Ratio(FPR) are calculated based on the frame-level groundtruth, which means that if there is hand-waving motion occurring in the frame, and the detection result shows that there is hand-waving, we say it is a true positive detection; if there is not hand-waving taking place in the frame, but the detection result shows that there is hand-waving occurring, we say it is a false positive detection. The Equal Error Rate(EER) of our algorithm are 4.2% for dataset one and 8.5% for dataset two. We also calculate the Area Under Curve(AUC), the value of which are 98% for dataset one and 94% for dataset two respectively. The proposed algorithm is implemented by Matlab, and we use OPENCV function to operate Chamfer Distance Matching and mex the code to be called in Matlab for improvement of computation efficiency. The developing environment of our PC platform is 2 GHz CPU and 2 GB memory. Our algorithm is running offline, and the most time consuming operation is Chamfer Distance Matching, and the average time needed for doing this per frame is 15.466s for video clip one, and 26.837s for video clip two.

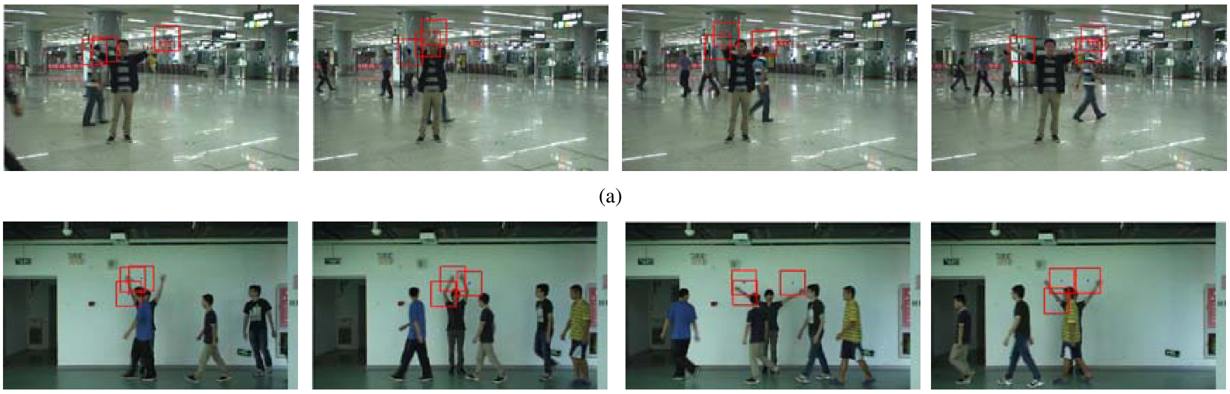
IV. Analysis and Conclusion

A. Analysis

The experiments demonstrate that the proposed algorithm can effectively detect hand-waving motion in crowds. If a person waves hands without moving, there is a high rate(approximating to 95%) to detect such movement successfully, even in the case that there is slight occlusion. But there are also false positive detections, because our method is based on detection of periodic motion. When an edge curve exhibits periodicity in row image, it will be detected as a trace of hand-waving, although it may be caused by irregular motion. For example, when two person cross over, the intersection of traces of their arms movement may form periodic curve. Besides our approach relies on skin color detection, so some objects having skin-like color will bring adverse influence towards detection, like regions with red color in Fig. 2. In future work, we can reduce such influence through combination of skin-color detection and moving object detection to pay our attention on moving human body.

B. Conclusion

We proposed a new approach for detecting hand-waving motions in crowds. Our method can be deemed as a complexity reduction process from 3D action detection to the problem of 2D object detection. We converted video clip to 2D images consisted of tracks of movement, and the hand-waving action is transformed into 2D patterns which have strong periodicity. Then we use Chamfer Distance Matching and Fourier Transformation to detect such patterns.



(a)

(b)

Fig. 7: Examples of detection results of hand-waving motion on two datasets. The scene of images in the first row is at a subway station, the scene of images in the second row is at a corridor in a building. The red rectangle indicates the region where hand-waving motion are occurring.

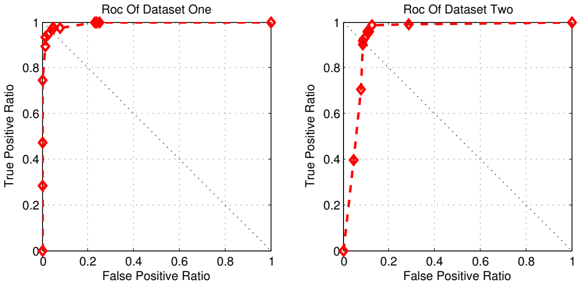


Fig. 8: The ROC of two datasets. The left is for dataset one, the right is for dataset two.

Experiments on two self-built video clips demonstrate that our algorithm can achieve promising detection results.

In future work, we can reduce influence from objects having skin-like color through combination of skin-color detection and moving object detection to pay our attention on moving human body.

References

[1] Y. Ke, Sukthankar, Rahul and H. Martial "Spatio-temporal shape and flow correlation for action recognition," *in IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-8.

[2] S. Ali and M. Shah "Human action recognition in videos using kinematic features and multiple instance learning," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, vol. 32(2), pp. 288-303.

[3] Bobick, F. Araron, Davis and W. James "The recognition of human movement using temporal templates," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, vol. 23(3), pp. 288-303.

[4] J. Little and J. Boyd "Recognizing people by their gait: the shape of motion," *in Videre:Journal of Computer Vision Research*, 1998, vol. 1(2), pp. 1-32.

[5] J. Little and J. Boyd "Describing motion for recognition," *in IEEE International Symposium on Computer Vision*, 1995, pp. 235-240.

[6] Y. Yacoob and M. Black "Parameterized modeling and recognition of activities," *in IEEE International Conference on Computer Vision*, 1998, pp. 120-127.

[7] Laptev and Ivan "On space-time interest points," *in International Journal of Computer vision*, 2005, vol. 64(2-3), pp. 107-123.

[8] J. Niebles and H. Wang "Unsupervised learning of human action categories using spatial-temporal words," *in International Journal of Computer Vision*, 2008, vol. 79(3), pp. 299-318.

[9] P. Dollar, V. Rabaud and G. Cottrell "Behavior recognition via sparse spatio-temporal features," *in IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65-72.

[10] J. Wang, W. Lu and Waibel "skin-color modeling and adaption," *in Computer Vision ACCV*, 1997, pp. 687-694.

[11] M. Liu, O. Tuzel and A. Veeraraghavan "Fast directional chamfer matching," *in IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1696-1703.