2015 The 5th International Workshop on Computer Science and Engineering
doi: 10.18178/wcse.2015.04.072

Hand Gesture Recognition Using Superpixel Tracking and RBF Classifier

Hong-Min Zhu and Chi-Man Pun

Department of Computer and Information Science, University of Macau, Macau SAR, China

Keywords: Hand detection, motion tracking, superpixel, template matching.

Abstract. A hand gesture recognition system using superpixel tracking and radial basis function (RBF) classifier is proposed in this paper. Firstly we employed the motion detection of superpixels and unsupervised image segmentation to detect the target hand in the first few video frames. Then the hand appearance model is constructed from its surrounding superpixels. By incorporating the failure recovery and template matching in the tracking process, the target hand is tracked by the proposed adaptive superpixel based tracking algorithm. Experimental results show that the extracted motion trajectories recognized with a trained RBF classifier can achieve better performance compared to the existing state of the art methods.

Introduction

Being a significant part in interaction of communication in our daily life, hand gestures provide us a natural and user friendly way of interaction. The computer vision field has experienced a new opportunity of applying a practical solution for building a variety of systems [1, 2] such as surveillance, smart home and sign language recognition. However, locating the hands and segment them from the background usually encounter with difficulties when there are occlusions, lighting variances, fast motion or other objects present with similar appearance.

There are many vision-based hand gesture recognition algorithms proposed in past several decades attempted to provide a robust and reliable systems, as reviewed in [3, 4]. The hand tracking solution can be benefit from visual object tracking solutions [5-7]. The PROST method [5] used multiple modules to reduce the drifts and object deformation, however the tracker is easily distracted by object with similar appearance. The visual tracking decomposition approach (VTD) [6] gets the tracking result with significant amount of noise from the background patches which combined particle filter with multiple observation and motion models, the tracker encounter failures when distinguish the target object and its background. Another potential solution is super-pixel tracking (SPT) [7], which used mid-level clustering of histogram information captured in superpixels and a discriminative appearance model formulated with target-background confidence map, which tried to find proper appearance models that distinguish one object with all other targets or background. However, this approach is very unreliable when severe deformation or background confusion exists. In the area of hand gesture recognition, there are less works relayed on hand's motion trajectories. Alon's work [1] proposed a classifier-based pruning framework and a subgesture reasoning algorithm to identify falsely matched parts in longer gestures, however they detect the hand location in each frame independently with color and motion information and the appearance changes are not adaptively learnt, the multiple hand region candidates cause confusion between the palm and the arm. Recently, many works [10-16] have been also proposed to address the problem of object segment and hand gesture recognition.

In this paper an adaptive superpixel based hand gesture tracking and recognition system was proposed, in which hand gestures drawn in free air are recognized from the extracted motion trajectory. With the given input video sequence, the moving target hand is first detected to construct its appearance model by the proposed Initial Hand Detection and Model Construction algorithm using the first few video frames. Then the hand gesture motion trajectory is tracked by the proposed Adaptive Hand Gesture Tracking algorithm. The rest of the paper is organized as follows. In Section 2 we describe the details of our proposed Initial Hand Detection and Model Construction

algorithm. In section 3, the proposed Adaptive Hand Gesture Tracking algorithm will be described, experimental results are given and discussed in section 4, and finally we conclude the work in section 5.

Initial Hand Detection and Model Construction

The first step of our proposed hand gesture recognition system is to detect the moving target hand and construct its appearance model. We employed the motion detection of superpixels and unsupervised image segmentation on the first few video frames. With the widely used image segmentation method Simple Linear Iterative Clustering (SLIC) superpixels [8], we detect the slight hand motion from corresponding superpixels changes in between adjacent frames. The first frame I_1 is segmented into P superpixels S_p (Fig.1a), from which the object boundaries are approximated. The accumulated intensity changes D_p of each superpixel S_p between I_1 and I_i can be computed as:

$$D_p = \sum_{i=2}^{M} |I_i(S_p) - I_1(S_p)|, \; p = 1, \ldots, P \tag{1}$$

And the slight motion of a superpixel is detected (Fig.1b) if:

$$\frac{D_p}{|S_p|} > T_0 \tag{2}$$

Where T_0 is a threshold of the normalized distance and $|S_p|$ is the size of p-th superpixel. After we merged neighbored superpixels with intensity changes as R candidate regions of the hand(Fig.1c), we used the Compression-based Texture Merging (CTM) [9] based image segmentation to select the hand region from candidates. We get the K CTM object regions O_k with areas A_k ($k=1,\ldots,K$) on the surrounding of candidate hand region (twice of the area size) in the first frame (Fig.1d), and the region with maximum percentage of the region area overlapped with hand candidates R is stated as detected hand R_H (Fig.1e):

$$R_H = O_k \mid A_k = \max\left(\frac{size(R \cap O_i)}{size(R \cup O_i)}\right), i = 1, \ldots, K \tag{3}$$

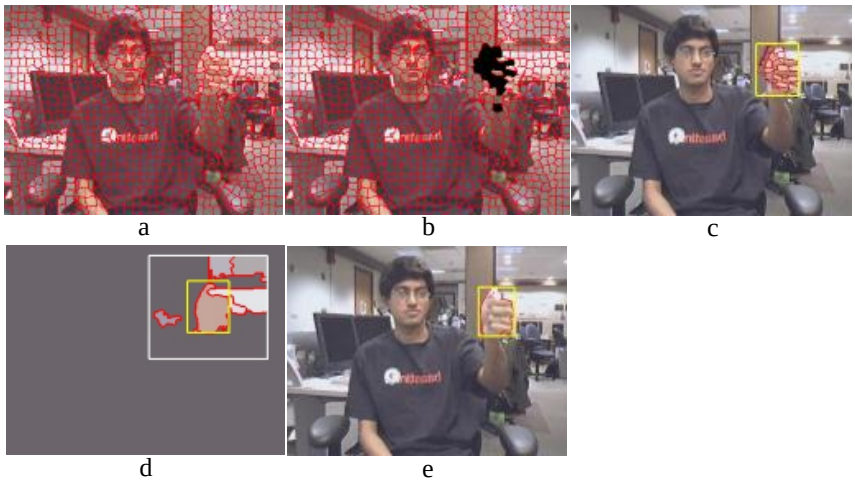

Figure 1. SLIC and CTM hand detection. a). SLIC superpixels on the first frame. b). superpixels with slight motions. c). candidate hand region on the connected superpixels. d). CTM objects on the surrounding of candidate hand region. e). refined hand region.

As we can see from the example, motion detection based on SLIC superpixels locate the hand region as shown in Fig.1c, which include the region besides the left side of the hand since the hand moves from right to left in this case. The result is then refined by CTM segmentation to exclude the false region part. The initial hand detection is represented by a bounding box of the hand region in the first frame, although the motion information with changed intensity is accumulated from the first M frames.

Superpixel Hand Gesture Tracking

After the initial hand appearance model is constructed from the first few frames. We employed an adaptive superpixel based hand gesture tracking approach to detect the hand position in following frames. The existing superpixel tracking (SPT) method [7] proposed for general object tracking, frequently encounters failures in our hand gesture tracking task. The first row of Fig 2 gives some typical examples that SPT fails to track the gesturing hand. We state the *occlusion* in Fig.2a occurred when the match score between the candidate hand region and the hand model bellows a threshold, which may be caused by hand deformation and blur of fast motion. Fig.2b and Fig.2c shows the example that if the target hand disappeared in the scene for a long period, the model will be updated with false information and the subsequent hand tracking will fail. And Fig.2d gives an example that SPT detect the background as the hand region when it is skin-color liked. If the problem continuously appears, the appearance model will eventually updated with features extracted from the background. We term this problem as *background confusion*.

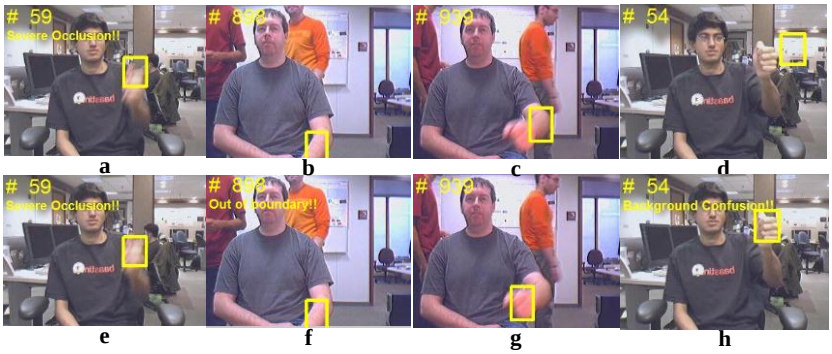

Figure 2. Typical example results in SPT (first row) and in our solution (second row). Columns from left to right are: *occlusion* occured; hand region disappeared in the scene; hand region tracked after it reappeared; and *background confusion* occurred.

Our proposed adaptive hand tracking solution recovers from these failures to provide reliable tracking results. Hand region candidates are pre-refined by incorporating domain specific knowledge, after which re-tracking with template matching gives the hand detection more accurately. Firstly we select hand detection from candidates by matching to the initial/updated model, in case if any failure occurred as introduced in Fig.2, we recover and re-tracking the hand with template matching to give positive detections. The detected hand will be continuously and periodically sampled and used to update the hand appearance model.

Figure 3. Gesture hand templates

Experimental Results

In this section, our proposed hand gesture recognition system using adaptive and robust superpixel based tracking was evaluated on the hand signed digit gesture dataset provided by Alon's work [1], the dataset defined 10 classes of gesture from digit 0 to 9. With the gesture motion trajectories tracked, a radial basis function (RBF) network is used for our multi-class hand gesture trajectories classification task. The test set contains 30 video sequences, three from each of 10 users which are

captured in office enviroment. The user signed each of 10 gestures once and wore short sleeves, totally there are 300 gesture instances in this set.

The RBF network involves three different layers, namely, input layer, hidden layer, and output layer. The input layer is made up of a number of source / input nodes, one node for one energy signature from the reduced feature vector of a given query image. The goal of the hidden layer is to cluster the data and to further reduce its dimensionality. The output layer supplies the responses of the network to the reduced feature vector applied to the input layer during classification. The responses correspond to the distances between the input image and the different database image classes. The proposed adaptive image classification algorithm can be divided into two stages. The first stage is for training, which is done only once. Its main objective is to construct an RBF network based on the number of features in the feature vectors and the number of classes involved, and to compute the corresponding weights of the hidden layer in the RBF network using a number of training images. The inputs to the RBF network include the feature vectors of the training image samples and their corresponding image classes. The output of the training would be the weights of the hidden layer of the network. The network starts with some initial weights which would be adjusted incrementally by the network as each feature vector and its class data are input. Therefore, the objective of the training is to produce the weights to represent the image classes of the training samples for achieving good classification results. Such weights would be used to classify query images during the classification stage. For efficiency sake, the training can be performed offline and the trained network information, including the weights, be saved for future use. The second stage is for online classification. Its main objective is to find the best match of any given query image to one of the predefined classes captured in the trained RBF network.

Firstly we use one sequences from each user for RBF network training, and test on the remaining sequences. By switching the training/test video sequences, there are three tests. Table 1 gives the confusion matrix of the recognition results. The number of correctly and falsely recognized gestures for each class are accmulated from the three tests. For first row is ground truth labels of gesture calsses, and the firt column is the recognized class labels. We see that totally 10 gestures are falsely classified out of 600 gestures from three tests, the recognition accuracy is 590/600=98.3%.

Table 1: Confusion matrix of recognition result on easy set, using 1/3 data for training and 2/3 for testing. Gestures counts are accmulated from three tests by switch training/test data.

	0	1	2	3	4	5	6	7	8	9
0	60	0	0	0	0	0	2	0	0	0
1	0	60	0	0	0	0	0	0	0	0
2	0	0	60	0	0	0	0	0	0	0
3	0	0	0	58	0	0	0	0	2	1
4	0	0	0	0	59	0	0	0	0	0
5	0	0	0	0	0	60	0	0	0	1
6	0	0	0	0	0	0	58	0	0	0
7	0	0	0	0	0	0	0	59	0	0
8	0	0	0	2	0	0	0	0	58	0
9	0	0	0	0	1	0	0	1	0	58
false	0	0	0	2	1	0	2	1	2	2

We also compared our approach with the state of arts as shown in Table 2. The experiments show that our hand gesture recognition approach outperforms the other solutions with significant improvement, which benefit mainly from our reliable hand motion tracking solution in long sequences.

Table 2: Comparsion with state of the art methods on the same datasets.

Approach	**Accuracy (%)**
Correa M. *et al.* [10]	75.00
Malgireddy M. *et al.*[11]	93.33
Y. Yao *et al.* [13]	95.67
Hanson, D.A [14]	100
Alon *et al.* [1]	94.60
Our proposed method	**98.3**

Conclusion

We proposed an adaptive superpixel based hand gesture tracking and recognition system in this paper to address the gestures expressed by human hand motion trajectories. With the target hand detected in first few frames using SLIC segmentation and motion subtraction and then refined by segmented object regions of CTM, our adaptive hand motion tracking well handles the occlusion and background confusion problem. Experimental results of trajectory classification using RBF networks on hand signed digit gestures show that our proposed system can achieve better performance compared to the existing state of the art methods with the recognition accuracy 98.3%. Future works may focus on multi-objects or two hand gesture tracking.

Acknowledgment

This research was supported in part by Research Committee of the University of Macau (MYRG134-FST11-PCM, MYRG181-FST11-PCM) and the Science and Technology Development Fund of Macau SAR (Project No. 008/2013/A1).

References

[1] J. Alon, V. Athitsos, Y. Quan, and S. Sclaroff, "A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, pp. 1685-1699, 2009.

[2] E. Sato, T. Yamaguchi, and F. Harashima, "Natural Interface Using Pointing Behavior for Human–Robot Gestural Interaction," *Industrial Electronics, IEEE Transactions on,* vol. 54, pp. 1105-1112, 2007.

[3] R. S. J. G. R. S. Murthy, "A review of vision based hand gesture recognition," *Int. J. Inf. Technol. Knowl. Manage.,* vol. 2, pp. 405-410, 2009.

[4] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 37, pp. 311-324, 2007.

[5] C. L. J. Santner, A. Saffari, T. Pock, and H. Bischof, "PROST: Parallel robust online simple tracking," *Computer Vision and Pattern Recognition,* pp. 723-730, 2010.

[6] J. K. a. K. M. Lee, "Visual tracking decomposition," *Computer Vision and Pattern Recognition,* pp. 1269-1276, 2010.

[7] W. Shu, L. Huchuan, Y. Fan, and Y. Ming-Hsuan, "Superpixel tracking," in *Computer Vision (ICCV), 2011 IEEE International Conference on,* 2011, pp. 1323-1330.

[8] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, Su, *et al.,* "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 34, pp. 2274-2282, 2012.

[9] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Comput. Vis. Image Underst.,* vol. 110, pp. 212-225, 2008.

[10] M. Correa, J. Ruiz-del-Solar, R. Verschae, J. Lee-Ferng, and N. Castillo, "Real-time hand gesture recognition for human robot interaction," in *RoboCup 2009*, B. Jacky, G. L. Michail, N. Tadashi, and G. Saeed Shiry, Eds., ed: Springer-Verlag, 2010, pp. 46-57.

[11] M. Malgireddy, I. Nwogu, S. Ghosh, and V. Govindaraju, "A Shared Parameter Model for Gesture and Sub-gesture Analysis," in *Combinatorial Image Analysis*. vol. 6636, J. Aggarwal, R. Barneva, V. Brimkov, K. Koroutchev, and E. Korutcheva, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 483-493.

[12] N.-Y. An and C.-M. Pun, “Color Image Segmentation Using Adaptive Color Quantization and Multiresolution Texture Charaterziation," Signal Image and Video Processing, 8(5), pp. 943-954, 2014.

[13] H.-M. Zhu and C.-M. Pun, “An Adaptive Superpixel Based Hand Gesture Tracking and Recognition System,” The Scientific World Journal, Volume 2014 (2014), Article ID 849069, 12 pages, 2014.

[14] C.-M. Pun and P. Ng, “Skin Color Segmentation by Gaussian Mixture Models Classifier and Wavelet Texture Feature Generation,” Information Journal, 17(10), pp. 4937-4942, 2014

[15] G. Huang and C.-M. Pun, “Robust Interactive Segmentation Using Color Histogram and Contourlet Transform,” International Journal of Computer Theory and Engineering, 7(6), pp. 489-494, 2014

[16] C.-M. Pun and P. Ng, “Skin Segmentation Using GMM Classifier and Texture Feature Extraction,” International Journal of Machine Learning and Computing, 4(1), pp.57-62, 2014.