Robust Hand Gesture Recognition Using Machine Learning With Positive and Negative Samples

Hong-Min Zhu, Chi-Man Pun and Cong Lin
Department of Computer and Information Science, University of Macau, Macau SAR, China
{yb07422, cmpun, yb17403}@umac.mo

Abstract—**Human action understanding is one of the most attractive research areas in computer vision. In this paper, we focus on a subset of human action which is the gesture performed by hand motion. To track the trajectory of motion, we adopt efficient and robust object detection and tracking schemes, which used Randomized Forest and Online Appearance model. Multiple hand templates are leaned using positive and negative samples (P-N learning). According to robust hand tracking and trajectory enhancement, we recognize the gesture with the baseline SVM tool. The effectiveness of the approach is demonstrated by experiments on the dataset of hand signed digit gestures.**

Index Terms—**motion tracking, P-N learning, trajectory, classification**

I. Introduction

Human action understanding is one of the most active area of research in the video content analysis and computer vision community, there are many potential applications such as video surveillance, human computer interfaces, sports video analysis, video retrieval, and sign language recognition. After the progress achieved in several decades, even the best existing systems still exhibit limitations due to complexities and variances of video applications. As a most recent example of many potential applications using human action understanding techniques, Microsoft developed the Xbox games which use Kinect to capturing the pose and motion of whole human body, it provides a nature interface[1] between the user and game without any additional devices wore on users.

Automatic categorization and localization of human action videos, however, remains a challenging problem due to cluttered background, camera motion, occlusion, viewpoint changes, and geometric and lighting conditions variances. Among related computer vision approaches to face these difficulties, object recognition that rely on sparsely detected features in a particular arrangement [1] tend to be robust to pose, image clutter, occlusion, object variation, and the imprecise nature of the feature detectors. The idea is extended to 3D video volumes for 3D interest points detecting in [2]. Other than more general whole body actions, we focus on hand gestures which are performed by hand motion. The authors in [3] introduced a unified framework for hand gesture segmentation and recognition, with multiple candidate hand detections in every frame from a spatiotemporal matching algorithm, a classifier-based pruning framework rejects poor matches to gesture models in early stage, and a subgesture reasoning module learns gestures which can falsely match parts of other longer gestures. In our approach, the hand trajectory is tracked by learning of hand templates, which is a bootstrapping binary classifier with structural constraints on positive (hand object) and negative (background) samples (namely P-N learning). This object tracking, learning and detection approach was introduced in [4] and is shown to be reliable in long sequence tracking.

The rest of the paper is organized as follows. We review some related works in section 2. In Section 3 we describe our proposed hand gesture recognition system, and motion trajectory tracking based on P-N learning is introduced. In section 4, we present the experimental results on hand signed digit gesture dataset. Finally we conclude in section 5.

II. Related works

There exists a wealth of approaches on topic related to human action recognition. In this section we review only some of closely related work.

Color cues may be the first attempt which is used to detect human body based on skin classification, the motion can then be tracked once the human/hand is located in a frame. The main drawback of such approaches is the confusion introduced by skin-color-liked object in the background, the local lighting changes and skin color variance will also make the task challenging. Gaussian Mixture Model (GMM) can be used over YCbCr color space[5] to segment the skin regions, and depth information is utilized to filter out unexpected regions to enhance the hand detection. Motion cues and color cues can also be combined to detect the hand location in each frame [3], on the skin likelihood image computed from a generic histogram, a motion mask is applied to obtain the hand likelihood image. *K* subwindows are considered as candidate hand regions are maintained and are used to match with gesture models in the spatial-temporal matching module. In our previews work[6], we proposed a similar hand tracking method where the hand is detected by

Manuscript received September 3, 2012; revised November 18, 2012, accepted November 11, 2012.
Corresponding author: Chi-Man, Pun (cmpun@umac.mo)

doi:10.4304/jcp.8.7.1831-1835

inter-frame difference of skin-color-liked images with candidate hand regions.

Video representation based on spatial-temporal (ST) interest points has been studied in several literatures recently, which is extended from approaches for static images. The authors in [7] presents a space-time interest point detector based on the idea of the Harris and Forstner interest point operators. They detect local structures in space-time where the image values have significant local variations in both dimensions. In addition, [2] propose a detector based on a set of separable linear filters which generally produces a high number of detections. This method responds to local regions which exhibit complex motion patterns, including space-time corners. Also, a number of descriptors are proposed for cuboids which are the resulting video patches around each interest point. Combined with discriminative classifiers to learn and recognize human actions, the performance on KTH dataset [8]was improved to 81.2%. The best performance achieved on the same dataset by combining the cuboids descriptor with more advanced classification solutions, is the approach [9] which introduces a discriminative Semi-Markov Models (SMM) to recognize the descriptor and get the accuracy of 95%. Although the cuboids feature descriptor has been explored in many works and showed good potentials on KTH dataset as well as other human actions, it is unsatisfactory for human actions that the pattern is greatly rely on relative spatial position and temporal occurrence, since the feature descriptor as cuboids prototypes ignores the positional arrangement, in space and time, of the ST interest points.

More recently, another potential solution for our targeted task was proposed by Kalal et. al. [10], which takes advantage of a sequential process of a tracker, a discriminative classifier, and a generative template-based model. It is showed robust on object detection and online tracking against complex background and suit well to appearance changes of tracked target. We will review this P-N learning based object tracking mechanism in more detail in next section.

III. Proposed Hand Gesture Recognition System

We introduce our proposed hand gesture recognition system in this section as four main stages: initial hand localization which served as target to be tracked; hand motion tracking using P-N learning; trajectory enhancement and feature representation; and gesture classification by Support Vector Machine (SVM).

A. Initial Hand Localization

As the first step, we identify the hand location at the beginning of gesturing, which will be tracked to form the motion trajectory. As we have showed in [6], the solution benefit from both color cue and motion cue provided reasonable hand detection. We review the procedures of initial hand detection as follows:

Algorithm I –Gesturing hand detection

Input: a reference frame F_0 and a gesture frame F
Step 1: calculate the binary images I_0 and I for F_0 and F respectively. ex.: $I(j, k) = 1$ if $F(j,k) \in R_{skin}$.
Step 2: I=*exclusive-or*(I, I_0), remove static skin objects in F_0.
Step 3: I=*AND*(I, *NOT* (I_0)), remove region of gesturing hand from F_0.
Step 4: I=*closing* (*opening* (I)).
Output: Bounding box of region in I.

Fig.1 shows an example of hand detection. Step 1 classifies each pixel in the reference frame (Fig.1a) and gesture frame (Fig.1c) as a skin pixel if it falls into the skin range which will result in two corresponding binary skin images (Fig.1b&d). We adopt the skin color range in YCbCr color space (77<=Cb<=127 & 133<=Cr<=173) for skin classification, as YCbCr is more robust for object detection than other color spaces when the lighting condition changes. While some static skin-liked object in the background as well as non-gesturing parts of the body are detected, Step 2 performs logical exclusive-or between two skin images, which retains only two main clusters (Fig.1e) as the gesturing hands from two frames. And the gesturing hand from the reference frame can further be eliminated by step 3, as shown in Fig.1f. Finally the gesturing hand was detected after de-noise process, and the bounding box of the hand region will be used as the input of motion tracking module.

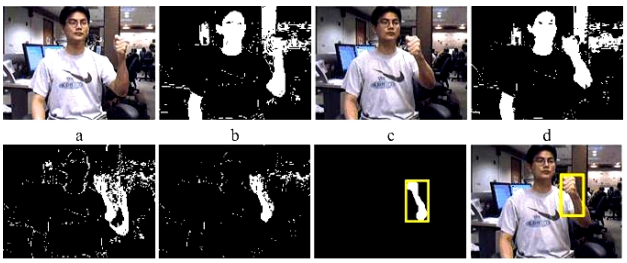


Figure 1. Example of hand detection. (a)&(c): gesture frames; (b)&(d): skin images; (e)&(f): active hand filtering; (g): result of de-noise; (h) resulting bounding box.

B. Motion Tracking with P-N Learning

To extract trajectory of gesture, we track the hand motion in a video sequence with approach in [10]. The process of P-N learning is guided by two constraints which are positive (P) and negative (N) ones, corresponding to the gesturing hand and its complement in our case. The tracking strategy involves three main components.

1). A short-term Kanade Lucas Tracker (KLT): the KLT features inside the bounding box of gesturing hand (Fig.1h) are extracted in each frame and tracked in next. The next hand location is computed based on the median of the features.

2). An object detector based on randomized forest: a scanning window technique is applied for exhaustive search in the current frame. The KLT and the randomized forest based object detector work together. The KLT predicts the location of the hand in following frames while the object detector verifies the KLT's output. A

local feature called 2bitBP[10], which is similar to the Cascade feature[11], is used in the detection. The rectangle patches (sub-window) contains that target are divided into four parts with same size. To generate the local feature, the patch is firstly integrated by:

$$I(i,j)=\sum_{i<M,j<N}^{n} P(i,j) \tag{1}$$

Where

$$[M,N]=Size(P) \tag{2}$$

P is the patch from the video frame. The 2bitBP feature is then generated by comparing the integral from each part.

$$2bitBP=[f_t(i,j)] \tag{3}$$

Where

$$(i,j)\in\{(M/2,N/2),(M,N/2),(M/2,N),(M,N)\}$$

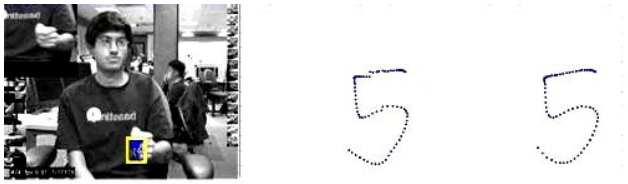
The $f_t(i,j)$ is defined as:

$$ft(i,j)=\begin{cases}1 & \text{if } I(i,j)>I(i-M/2,j) \text{ AND } I(i,j)>I(i,j-N/2)\\ 0 & \text{if } I(i,j)<I(i-M/2,j) \text{ AND } I(i,j)<I(i,j-N/2)\\ 0.5 & \text{Otherwise}\end{cases} \tag{4}$$

Therefore the local features generator outputs 4 codes for a patch. Each output sample of sub-window is evaluated by the binary randomized forest classifier. All candidates are tested and the best samples in a tracked region are transferred to the hand model for final decision.

3). An online model based on template matching: The distance to model for all candidates are computed and the best one is selected as positive sample, while others far from the result location are treated as negative samples. Those P-N samples are used to train the randomized forest classifier and to update online hand model.

As the result of P-N learning based tracking of the hand motion, a bounding box indicates the location of hand in each frame with varied confidence is fired. We use the center of the area to form the trajectory of a gesture. Fig.2 gives an example of hand motion tracking with P-N learning, (a) is the result of hand detection on the last frame, and (b) shows the tracked trajectory with all centers of hand bounding box.



a b c

Figure 2. Example of hand motion tracking

C. Trajectory Feature Representation

To fit with feature representation used in SVM, we normalize all extracted gesture trajectories to be with a unified size. As the gesture length is proportional to the complexity of gesture, we take the mean length L of all gesture categories for normalization. Two dimensions of a trajectory with T points are approximated and interpolated with B-form Spline polynomial function.

A Spline is any smoothed piecewise polynomial function that an interval $[a, b]$ (*e.g.* $[1, T]$) is subdivided into sufficiently small intervals $[\xi_i, \xi_{i+1}]$ with $a=\xi_1<\ldots<\xi_{i+1}=b$ (e.g. all intervals are with equal length = 1). On each such interval, a polynomial p_i of relatively low degree can provide a good approximation. The B-form Spline describes the piecewise polynomial function as a weighted sum of the required order k, with their number $n>k-1+i$:

$$f(t)=\sum_{i=1}^{n} B_{i,k}(t).a_i \tag{5}$$

Each $B_{j,k}$ is defined on an interval $[\xi_i, \xi_{i+1}]$ and is zero elsewhere, t is called knots and is provided based on the smoothness required. B-splines are functions that:

$$\sum_{i=1}^{n} B_{j,k}(x)=1, x\in[t_k,t_{n+1}] \tag{6}$$

We use the order k=4 and knots t= 8 to formalize our approximation function g, and $g(x)$ is interpolated with $x = [1, 1+s, 1+2s,\ldots, T]$ where $s=(T-1)/L$. Fig.2c shows the normalized and smoothed result of trajectory in Fig.2b.

D. Gesture Classification with SVM

According to the reliability of trajectory extraction based on motion tracking with P-N learning, we simplify the recognition task by using the baseline SVM solution. SVM was firstly introduced in[12] which simultaneously minimizes the empirical classification error and maximizes the geometric margin, which has many applications related to classification tasks.

In the scenario of binary classification on a linearly separable training set $\{x_1, x2,..., xn\}$ with class labels $\{y_1, y_2,..., y_n\}$, $y_i\in\{-1,1\}$, an optimal hyperplane as classifier is defined by solving the optimization problem

$$\text{Minimise: } \phi(\omega)=\frac{1}{2}\|\omega\|^2 \tag{7}$$

$$\text{Subject to: } y_i(\omega\bullet x_i+b)\geq 1$$

Where

$$\omega=\sum\alpha_i y_i x_i \tag{8}$$

And a new input x is then classified by the hyperplane according to its side related to the hyperplane as:

$$f(x)=sign(\sum_i \alpha_i y_i(x_i\bullet x)+b) \tag{9}$$

An important feature of SVM is that non-linearly separable features can be mapped to higher dimension space by a kernel function K which is then linearly separable by the classifier:

$$f(x) = sign(\sum_i \alpha_i y_i K(x_i \bullet x) + b) \quad (10)$$

Due to the discriminating ability on non-linearly separable data of SVM, our gesture trajectory in points' representation can be recognized. For the trajectory's points' representation, we normalize the coordinates into the range of [0, 1].

We use the SVM library provided by [13] to do our classification task. The library used "one-against-one" approach to do multi-class classification, which means *k(k-1)/2* classifiers are constructed and each one trains data from two classes. In testing, a simple voting strategy is used to recognize a new feature vector. We adopt RBF (Radial Basis Function) as kernel function since it usually performs better than other kernels. Two parameters, the cost *c* of quadratic problem and gamma *g* used in RBF kernel, are optimized by grid search in the range of 2^{-8} to 2^{8}. In each run to adjust the *c* and *g*, 3-fold cross validation is used to get the averaged accuracy for comparison on training feature set.

IV. Experimental results

A. Dataset and Methodology

As an example given in Fig.2, we demonstrate the performance of our proposed hand gesture recognition system by testing on a set of hand signed digit from 0 to 9, where the pattern is presented by the trajectory of single hand motion in front of a static camera in the laboratory environment.

The dataset provided in [3] defines two types of test sets. The easy test set contains 30 short sleeve video sequences three from each of 10 users. The gestures in hard test set are defined in the same way, with two sequences from each of seven users. In hard set there are distractors present beside the gesturing user moving back and forth in the background, which makes the tracking of gesturing hand challenging. In each sequence for both sets the user signed each of 10 digits once, we use the ground truth to segment the sequence into 10 clips that we assume the gesture segmentation is known. Fig.3 shows an instance for each class of gesture with indication of starting point.

Figure 3. Hand signed digit gestures [3].

We use the training/testing strategy motivated from [2] that the extracted trajectories from the easy test set are grouped for each user. There are 10 groups and each group contains 30 trajectories with three for each of 10 gesture classes. The experiment is carried out by leave one (group) out cross-validation, each time we train the SVM model on 9 groups and test on the remaining group, the overall recognition rate is the averaged on the result of 10 runs by changing test group.

B. Recognition results

Fig.4 shows the confusion matrix on the easy test set, the recognition rates for each gesture class in each group are averaged. The overall accuracy rate is 96.67%, that 290 out of 300 trajectories are correctly recognized. We can see that the gesture "6" cannot be well classified which has a high probability to be classified as gesture "0". This is due to the similarity of spatial appearance and temporal occurrence between these two gestures. The result is comparable to the original experiment in [3] which atomically segment the gesture with sub-gesture reasoning and achieved the best accuracy of 94.6%.

	0	1	2	3	4	5	6	7	8	9
0	.97	.00	.00	.00	.03	.00	.00	.00	.00	.00
1	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00
2	.00	.00	1.0	.00	.00	.00	.00	.00	.00	.00
3	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00
4	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00
5	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00
6	.23	.00	.00	.00	.00	.00	.77	.00	.00	.00
7	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00
8	.00	.00	.00	.00	.03	.00	.00	.00	.97	.00
9	.00	.00	.00	.00	.00	.03	.00	.00	.00	.97

Figure. 4. confusion metrix of recognition on easy test set.

	0	1	2	3	4	5	6	7	8	9
0	.86	.00	.00	.00	.14	.00	.00	.00	.00	.00
1	.00	.93	.00	.00	.07	.00	.00	.00	.00	.00
2	.00	.00	.93	.00	.07	.00	.00	.00	.00	.00
3	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00
4	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00
5	.00	.00	.00	.00	.07	.86	.00	.00	.00	.07
6	.29	.00	.00	.00	.07	.00	.59	.00	.00	.06
7	.00	.07	.00	.00	.07	.00	.00	.86	.00	.00
8	.00	.00	.00	.00	.07	.00	.00	.00	.93	.00
9	.00	.01	.00	.00	.14	.00	.00	.00	.00	.86

Figure. 5. confusion matrix of recognition on hard test set

For the hard test set, we state that the initial hand localization described in section III.A is not reliable to detect only the gesturing hand, as the background is not as relatively static as in the easy test set. While multiple hand region candidates may be detected with some of them from the distractors, we track each hand candidate independently, and the best trajectory is chosen according its longest duration. The casual hand motions from distrac-

tors are likely to be occluded by the gesturing user or their own orientation to the camera, which cause the tracked trajectory much shorter than the gesturing user's. We use the SVM models generated from the easy test set to classify the trajectories in the hard test set, where all trajectories are classified by one of ten models and the accuracy is computed. The overall recognition rate is the averaged result of classification with all models, which is given in Fig.5. There are totally 123 out of 140 trajectories be correctly recognized with the accuracy of 88%, which also outperforms the best result in [3] for the hard test set with 85%.

V. CONCLUSION

In this paper, we proposed a hand recognition system. The initial hand location is automatically detected, whose motion is then tracked with P-N learning strategy. The extracted motion trajectory is normalized and smoothed by B-form Spline polynomial function, and classification with SVM on the hand signed digit gestures shows good performance of the system. While P-N learning successes in dealing with gesturing video taken from relatively static background, it can also prevent the gesturing user from the corruption of distractor's motion in a more challenged setting.

The future work may utilize the depth information to enhance the initial hand localization when a complex background is present.

REFERENCES

[1] Agarwal, S., A. Awan, and D. Roth, *Learning to Detect Objects in Images via a Sparse, Part-Based Representation.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. 26(11): p. 1475-1490.

[2] Dollar, P., et al., *Behavior Recognition via Sparse Spatio-Temporal Features.* Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, 2005: p. 65-72.

[3] Alon, J., et al., *A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009. 31(9): p. 1685-1699.

[4] Kalal, Z., J. Matas, and K. Mikolajczyk. *P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints*. in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. 2010.

[5] Elmezain, M., et al., *A Hidden Markov Model-Based Isolated and Meaningful Hand Gesture Recognition.* International Journal of Electrical, Computer, and Systems Engineering, 2009. 3(3): p. 156-163.

[6] Pun, C.-M., H.-M. Zhu, and W. Feng, *Real-Time Hand Gesture Recognition using Motion Tracking.* International Journal of Computational Intelligence Systems (IJCIS), 2011. 4(2): p. 277-286.

[7] Laptev, I., *On Space-Time Interest Points.* International Journal of Computer Vision, 2005. 64(2): p. 107-123.

[8] Schuldt, C., I. Laptev, and B. Caputo. *Recognizing Human Actions: A Local SVM Approach*. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. 2004.

[9] Shi, Q., et al., *Human Action Segmentation and Recognition Using Discriminative Semi-Markov Models.* International Journal of Computer Vision, 2011. 93(1): p. 22-32.

[10] Kalal, Z., J. Matas, and K. Mikolajczyk. *Online Learning of Robust Object Detectors During Unstable Tracking*. in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. 2009.

[11] Viola, P. and M. Jones. *Rapid Object Detection using A Boosted Cascade of Simple Features*. in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. 2001.

[12] Vapnik, V., *The Nature of Statistical Learning Theory.* NewYork: Springer, 1995.

[13] Chang, C.-C. and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines.* ACM Trans. Intell. Syst. Technol., 2011. 2(3): p. 1-27.

Hong Min Zhu was born in Zhejiang province, China on September 1985. Hong Min Zhu received the B.Sc degree in software development and application at Macau University of Science and Technology, Macau, China in 2008, and received the M.Sc degree in software engineering at University of Macau, Macau, China in 2010. He is currently a PhD candidate in software engineering at University of Macau.

He had carried out some projects and published several conference and journal papers. His major is software engineering and his research interests include computer vision, image processing and pattern recognition.

Chi Man Pun was born in Macau, china. He received his B.Sc. and M.Sc. degrees in Software Engineering from the University of Macau in 1995 and 1998 respectively, and Ph.D. degree in Computer Science and Engineering from the Chinese University of Hong Kong in 2002.

He is currently an Associate Professor at the Department of Computer and Information Science of the University of Macau. He is now teaching several bachelor and master courses and supervising severalmaster and PhD theses. He has investigated several funded research projects and published more than fifty refereed papers in international journals, books and conference proceedings. His research interests include Content-Based Multimedia Indexing and Retrieval; Digital Watermarking; Multimedia Databases; Image/Video Compression, Analysis and Processing; Pattern Recognition and Computer Vision, Intelligent Multimedia Systems and Applications.

Dr. Pun has also been invited to serve as referee/reviewer and/or committee member for international journals and conferences. He is also a Senior Member of the IEEE.