Hand Posture Recognition Using Adaboost with SIFT for Human Robot Interaction

Chieh-Chih Wang[†‡] and Ko-Chih Wang[‡]
[†]Department of Computer Science and Information Engineering
[‡]Graduate Institute of Networking and Multimedia
National Taiwan University, Taipei, Taiwan
Email: bobwang@ntu.edu.tw,casey@robotics.csie.ntu.edu.tw

***Abstract*— Hand posture understanding is essential to human robot interaction. The existing hand detection approaches using a Viola-Jones detector have two fundamental issues, the degraded performance due to background noise in training images and the in-plane rotation variant detection. In this paper, a hand posture recognition system using the discrete Adaboost learning algorithm with Lowe's scale invariant feature transform (SIFT) features is proposed to tackle these issues simultaneously. In addition, we apply a sharing feature concept to increase the accuracy of multi-class hand posture recognition. The experimental results demonstrate that the proposed approach successfully recognizes three hand posture classes and can deal with the background noise issues. Our detector is in-plane rotation invariant, and achieves satisfactory multi-view hand detection.**

I. Introduction

When robots are moved out of factories and introduced into our daily lives, they have to face many challenges such as cooperating with humans in complex and uncertain environments or maintaining long-term human-robot relationships. Communication between human and robots instinctively and directly is still a challenging task. As using hand postures/gestures is natural and intuitive for human-to-human interaction and communication, hand detection and hand posture recognition could be essential to human-robot interaction. Figure 1 illustrates an example of human robot interaction through hand posture in which our NTU PAL1 robot and an image from an onboard camera are shown. In this paper, the issues of hand detection and posture recognition are addressed and the corresponding solutions are proposed and verified.

As the Viola-Jones face detector based on an Adaboost learning algorithm and Harr-like features [1] has been successfully demonstrated to accomplish face detection in real time, these approaches are also applied to detect other objects. Unfortunately, it failed to accomplish the hand detection task because of its limited representability on articulated and non-rigid hands [2]. In addition, hand detection with the Viola-Jones detector can be accomplished with about 15° in-plane rotations compared to 30° on faces [3]. Although rotation invariant hand detection can be accomplished using the same Adaboost framework in a way of treating the problem as a multi-class classification problem, the training process needs much more training images and more computational power is needed for both training and testing. In this paper,

(a) A person interacts with the NTU PAL1 robot via hand posture.

(b) An image from the onboard camera.

Fig. 1. Hand posture based human robot interaction

a discrete Adaboost learning algorithm with Lowe's SIFT features [4] is proposed and applied to achieve in-plane rotation invariant hand detection. Multi-view hand detection is also accomplished straightforwardly with the proposed approach.

It is well understood that background noise of training images degrades detection accuracy significantly in the Adaboost learning algorithm. In the face detection applications, the training images seldom contain background noise. However, it is unlikely to show an articulated hand without any background information. Generating more training data with randomly augmented backgrounds can solve this background noise issue with a highly computational cost [5]. With the use of the SIFT features, the effects of background noise in the training stage are reduced significantly and the experimental results will demonstrate that the proposed approach performs

with high accuracy.

Given that hand detection is successfully accomplished, hand posture recognition can be done in a way that one classifier/detector is trained for each hand posture class [6]. An *one versus all* strategy is often used where the results from all classifiers are computed and the class with the highest score is labeled as the class of the test image. The computational cost of these sequential binary detectors increases linearly with the number of the classes. The one versus all strategy do not always generate correct recognition. In this paper, we apply a sharing feature concept proposed by Torralba *et al.* [7] to separate sharing and non-sharing features between different hand posture classes. Sharing features of different hand posture classes are used for detecting hand robustly. As non-sharing features represent the discrimination among classes, these non-sharing features are used to increase recognition accuracy and to speed up the recognition process.

The remainder of this paper is organized as follows. Related work is reviewed in Section II. The details of hand detection using the Adaboost learning algorithm with SIFT features are described in Section III. Hand posture recognition based the sharing feature concept is described in Section III-E. Ample experimental results and comparisons are demonstrated in Section IV. Conclusions and future work are in Section V.

II. Related Work

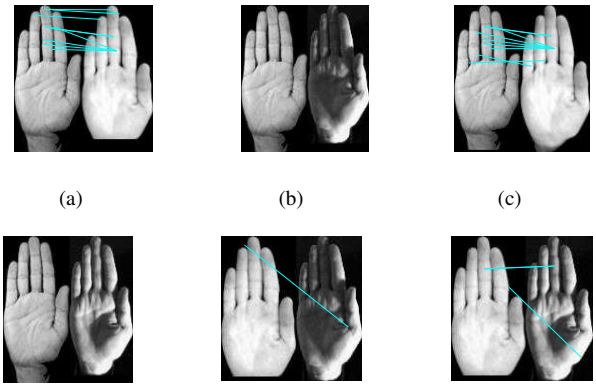
The Adaboost learning algorithms are currently one of the fastest and most accurate approaches for object classification. Kölsch and Turk [2] exploited the limitations of hand detection using the Viola-Jones detector. A new rectangle feature type was proposed to have more feature combinations than the basic Haar-like features proposed by Viola and Jones. As the feature pool for learning contains about 10^7 features, a highly computational cost is needed for training. Ong and Bowden [8] applied the Viola-Jones detector to localize/detect human hands, and then exploited shape context to classify differences between hand posture classes. Anton-Canalis and Sanshez-Nielsen [5] proposed to collect more training images for reducing the background noise effects. Their approach is to collect images under several controlled illuminations and to randomly augment the training images with various backgrounds to increase the robustness of the detectors. Just *et al.* [6] integrate a variant of Adaboost with a modified censes transform to accomplish illumination invariant hand posture recognition.

In addition to the Adaboost-based approaches, Athitsos and Scalaroff [9] formulated the hand posture recognition problem as an image database index problem. A database contains 26 hand shape prototypes, and each prototype has 86 difference viewpoint images. A probabilistic line matching algorithm was applied to measure the similarity between the test image and the database for recognizing hand posture class and estimating hand pose.

In this paper, the discrete Adaboot learning algorithm is integrated with SIFT features for accomplishing in-plane rotation invariant, scale invariant and multi-view hand detection. Hand posture recognition is accomplished with the sharing feature concept to speed up the testing process and increase the recognition accuracy.

(a) (b) (c) (d) (e)

Fig. 2. The SIFT features are extracted and shown. From left to right, 67, 57, 63, 70 and 85 SIFT features are extracted from the hand images respectively.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Hand Detection using the SIFT matching algorithm. Most of the pair images only contain less than five SIFT keypoint matches.

III. Hand Detection and Posture Recognition

In this section, the SIFT keypoint detector and the Adaboost learning algorithm are briefly reviewed. The modification to integrated Adaboost with SIFT and the sharing feature concept are described in detail.

A. SIFT

The Scale Invariant Feature Transform (SIFT) feature introduced by Lowe [10] consists of a histogram representing gradient orientation and magnitude information within a small image patch. SIFT is a rotation and scale invariant feature and is robust to some variations of illuminations, viewpoints and noise. Figure 2 shows the extracted SIFT features from five hand images. Lowe also provided a matching algorithm for recognize the same object in different images. However, this approach is not able to recognize *a category of the objects*. Figure 3 shows some examples of hand detection using the SIFT matching algorithm in which most of the pair images only contain less than five SIFT keypoint matches.

B. The Adaboost Learning Algorithm

The Adaboost learning algorithms provide an excellent way to integrate the information of a category of objects. As a single weak classifier can not provide a satisfactory result, Adaboost combines many weak classifiers to form a strong classifier in which a weak classifier can be slightly better than randomly guess to separate two classes. Given a set of positive and negative images, the Adaboost learning algorithm chooses the best weak classifier from the large pool. After choosing the best weak classifier, Adaboost adjusts the weights of the training images. The weights of misclassified training images of this round are increased and the weight of correct ones are decreased. In the next round, the Adaboost will focus more on the misclassified images and try to correctly classify the misclassified images in this round. The whole procedures are iterated until a predefined performance requirement is satisfied.

C. Adaboost with SIFT

Our hand detection approach applies Adaboost with SIFT features. Compared to the existing Adaboost-based hand detection approaches, the proposed approach has three advantages. First, thanks to the scale-invariant characteristic of SIFT, it is unnecessary to scale the training images to a fixed resolution size to adapt the characteristic of the Harr-like features in the original Viola-Jones approach. Second, rotation-invariant and multi-view detection is straightforwardly accomplished because of the rotation-invariant characteristic of SIFT features. Finally, the background noise issue is taken care easily. During the training data collection stage, the background of positive training images is set to a single color without any texture. Therefore, the extracted SIFT features from the positive training images exist only in the hand areas of the images. The classification performance is achieved without increasing the number of training samples.

Here the modifications to integrate the discrete Adaboost with SIFT features are described in detail. Let $\{I_i,\ i = 1,\ \ldots,\ N\}$ be the training image set where N is the number of the training images. Every image is associated with a label, $\{l_i,\ i = 1,\ \ldots,\ N\}$ and $l_i = 1$ if the image contains a hand, otherwise $l_i = 0$. Each image is represented by a set of SIFT features $\{f_{i,j}, j = 1,\ldots,n_i\}$, where n_i is the number of SIFT features in image I_i.

The weights, $\frac{1}{2N_p}$, $\frac{1}{2N_n}$, are initially set to positive training samples and negative training samples respectively where N_p is number of positive sample and N_n is number of negative sample. Each weak classifier, h_m, consists of a SIFT keypoint (f), a threshold (t) and a polarity (p).

A weak classifier, h_m, is defined as:

$$h_m(I_i) = \begin{cases} 1, & \text{if} \quad p * f_m(I_i) < p * t \\ 0, & otherwise \end{cases} \tag{1}$$

The next step is to choose m weak classifiers and combine them into a strong one. Our detector uses the function $F(f, I_i) = min_{1 \le j \le n_i} D(f, f_{i,j})$, where D is Euclidean distance, to define the distance between an image and a feature. Algorithm 1 shows the details of classification using the discrete Adaboost learning algorithm with SIFT features.

Algorithm 1 Training using the Discrete Adaboost Algorithm with SIFT features

Require: Given training images$(I_1, l_1) \ldots (I_n, l_n)$ where l_i = 0, 1, for negative and positive examples respectively.

1: Initialize weights $w_{1,i} = \frac{1}{2N_p}, \frac{1}{2N_n}$ for l_i = 0, 1 respectively, where N_p and N_n are the number of negatives and positives respectively.
2: **for** m = 1, ..., T **do**
3: Normalize weight of all training samples such that $\sum_{i=1}^{N} w_{m,i} = 1$;
4: Choose a SIFT keypoint feature (f_m), a threshold (t_m) and a polarity (p_m) to form a weak classifier such that the error is minimize. We define the error is;

$$e_m = \sum_{i=1}^{N} w_{m,i} |h_m(I_i) - l_i| \tag{2}$$

5: Define $h_t(x) = h(x, f_m, p_m, t_m)$ where f_m, p_m, t_m are the minimizers of e_m
6: Update the weights:

$$w_{m+1,i} = w_{m,i} \beta_m^{1-e_i} \tag{3}$$

where e_i = 0 if example I_i is classified correctly, e_i = 1 otherwise, and $\beta_m = \frac{e_m}{1-e_m}$
7: **end for**
8: The final strong classifier is:

$$H = \begin{cases} 1, & \text{if} \quad \sum_{t=1}^{m} \alpha_t * h_t > T \\ 0, & otherwise \end{cases} \tag{4}$$

where $\alpha_t = log \frac{1}{\beta_t}$
9: **return** strong classifier $H(x)$

D. Hand Detection

With the learned SIFT features, hand detection is accomplished as follows. The SIFT features are firstly extracted from the test image. For each weak classifier, the distance between its associated SIFT feature and the extracted SIFT features from the test image are computed. The best match with the shortest distance shorter than a threshold t_m is treated as a valid result from this weak classifier. Then the weight factors α_m of all valid weak classifiers are summed. If this summed value is greater than the threshold of the strong classifier, T, the test image is classified as a hand image. Algorithm 2 shows the details of the hand detection algorithm.

E. Multi-Class Recognition

With the use of the proposed hand detection method, multi-class hand posture recognition is done using the sharing feature concept. As different object classes still have sharing and non-sharing features, our method use non-sharing features to speed up the recognition process with a higher accuracy than the one versus all approaches.

Algorithm 2 Detection

Require: Given a strong classifier (T, W), T is the threshold of strong classifier. $W : (h_1, ..., h_m)$ is a set of weak classifiers. h_i consists of $(\alpha_i, f_i, t_i, p_i)$. α_i, f_i, t_i, p_i are the weight, SIFT feature, threshold and polarity of h_i, respectively. An image: I

Initialize $WeightSum = 0$
S = Extracting SIFT features from I
for $i = 1, \ldots, m$ **do**
  S_x = Find the nearest SIFT feature of f_i in S
  if EuclideanDistance(S_x, f_i)* $p_i < t_i$ * p_i **then**
    $WeightSum + \alpha_i$;
  end if
end for
if $WeightSum > T$ **then**
  return 1
else
  return 0
end if

In the detection phase, the sharing feature set is used to detect hand robustly. If the test image does not exist any sharing feature, the image is discarded. In the posture recognition stage, only non-sharing features are used in the sequential classification process. The class with the highest score is labeled as the image class. It should be noted that the current system trains each classifier independently. All classifiers could be trained jointly to get a better performance.

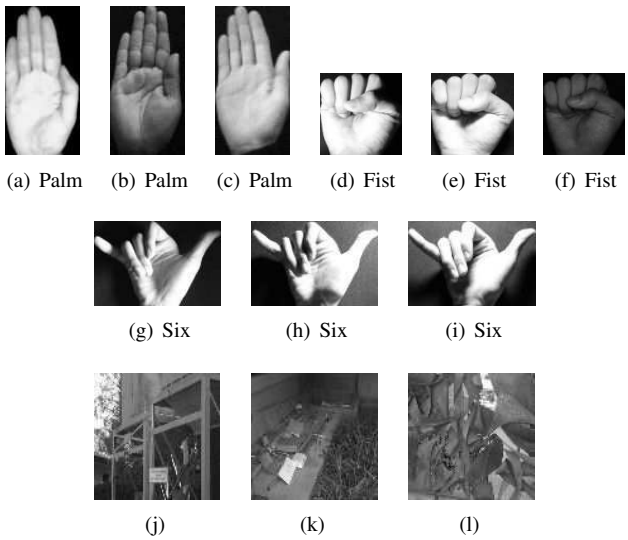
IV. Experimental Results

In this paper, three targeted hand posture classes, "palm", "fist" and "six", are trained and recognized. 642 images of the "palm" posture class from the Massey hand gesture database provided by Farhad Dadgostar *et al.*[1] are used as positive samples. As the Massey hand gesture database does not contain images of "fist" and "six", 450 "fist" images and 531 "six" images were collected by ourself under different lighting conditions. The negative/background images are consist of 830 images from the internet and 149 images collected in the building of our department. Figure 5 shows examples of the training data.

For testing, 275 images were collected using the onboard Logitech QuickCam Pro 5000 with a resolution of 320x240. Figure 6 shows the sharing and non-sharing features determined by our algorithm.

Figure 7 shows samples of correct hand detection and posture recognition using the proposed algorithms. Figure 8 shows some of correct hand detection but incorrect posture recognition. Tables I and II show the performances of multi-class hand posture recognition using our proposed detection algorithms without and with using the sharing feature concepts. The experimental results show that the approach using the sharing feature concept is superior.

[1]http://www.massey.ac.nz/fdadgost/xview.php?page=farhad

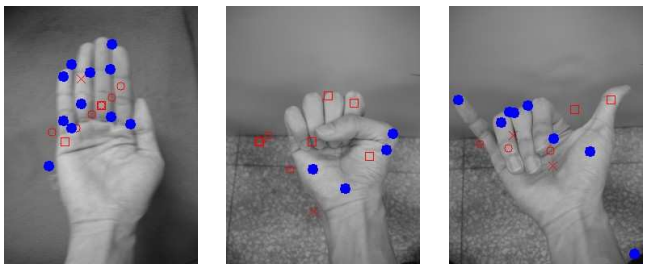


(a) Palm (b) Palm (c) Palm (d) Fist (e) Fist (f) Fist

(g) Six (h) Six (i) Six

(j) (k) (l)

Fig. 5. The training images of the "palm", "fist" , "six" and the backgrounds.



(a) (b) (c)

Fig. 6. Sharing and non-sharing features. Blue solid circles indicate sharing features. Red circles indicates non-sharing features detected by the "palm" detector. Red rectangles are non-sharing features detected by the "fist" detector. Red 'X' indicates a non-sharing feature detected by the six detector. Note that the weights of the detected features are different.

			Result		
Truth	PALM	FIST	SIX	Total	Accuracy
PALM	80	3	8	91	87.9%
FIST	0	94	4	98	95.9%
SIX	3	7	76	86	88.3%
Total				275	90.9%

TABLE I
Hand Posture Recognition without using the sharing feature concept.

			Results		
Truth	PALM	FIST	SIX	Total	Accuracy
PALM	89	2	0	91	97.8%
FIST	1	97	0	98	98.9%
SIX	3	6	77	86	89.5%
Total				275	95.6%

TABLE II
Hand Posture Recognition using the sharing feature concept.

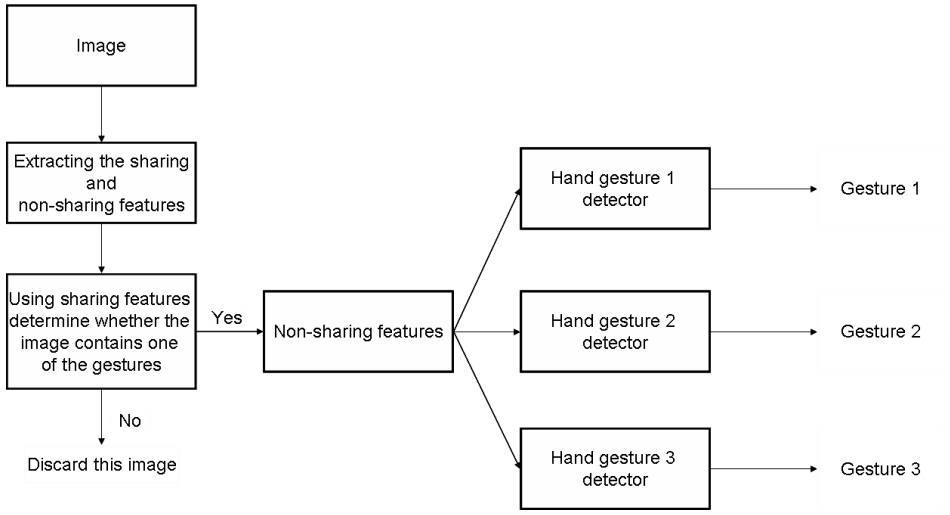
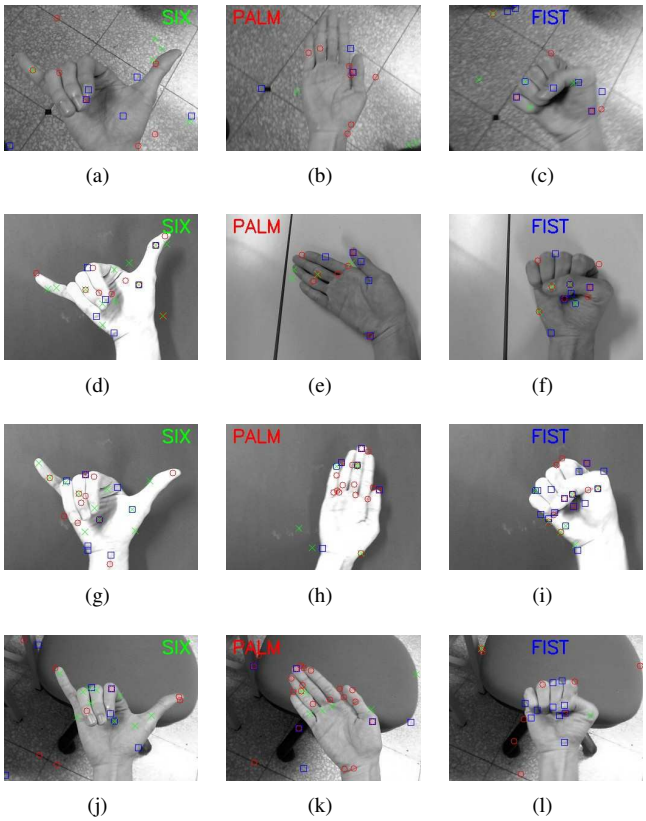


Fig. 4. Multi-class hand posture recognition using the sharing feature concept.



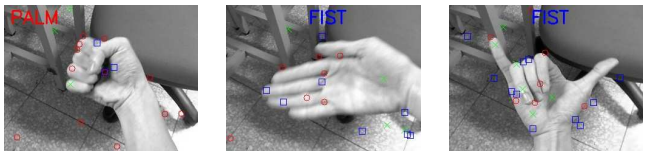
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 7. Correct hand gestures recognition results.



(a) (b) (c)

Fig. 8. Incorrect hand gestures recognition results. In (a), the palm detector may detect too many features between fingers and the background. In (b), the test image could be too blurred. In (c), both the six detector and the fist detector found many features.

We will show the quantitative results in terms of background noise, in-plane rotation variant recognition and multi-view recognition.

A. Background Noise

The performances of hand detection using the Viola-Jones detector and the proposed approach are compared.

The training results are shown by the ROC curve of the detector in Figure 9. Given the 90% detection rate, hand detection using the Viola-Jones detector generates 3 - 4% false positive rate. This result is unsatisfactory and is worse than the same approach applied on face detection. As the face training images almost do not contain any background. The experimental results show that background noise in the training images could degrade the performance of the Adaboost learning algorithm significantly. On the contrary, the results of hand detection using Adaboost with SIFT are satisfactory.

B. Rotation Invariant Recognition

Figure 10 shows the ROC curve in which the in-plane rotation invariant recognition is demonstrated. Three data sets were collected to test 0°, 90° and -90° of in-plane rotations. It is clearly shown that the performances between different in-plane rotations are very similar. The proposed approach use only one detector to accomplish in-plane rotation invariant hand detection.

C. Multi-View Recognition

Here we further verify if the proposed approach can achieve multi-view hand posture detection. Although more data from different viewpoints can be collected and trained to achieve multi-view hand detection, only data with a fixed viewpoint are used in this experiment. Figure 11 shows that

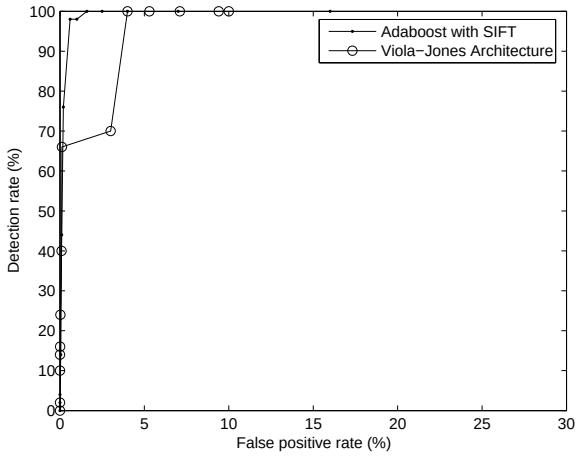


Fig. 9. The ROC curves of hand detection using the Viola-Jones detector and the proposed approach.

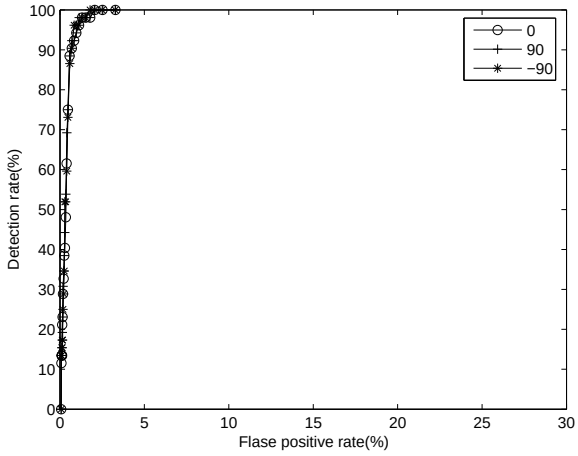


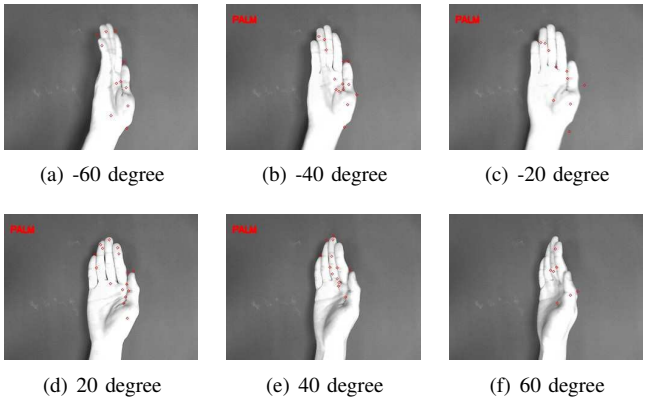
Fig. 10. the ROC curve shows that the proposed approach accomplishes in-plane rotation invariant recognition.

the detector can still work in the situation of the 40 degree viewpoint .

V. Conclusion and Future Work

In this paper, we presented a robust hand detection and posture recognition system using Adaboost with SIFT features. The accuracy of multi-class hand posture recognition is improved using the sharing feature concept. The experimental results demonstrated that the proposed hand detector can deal with the background noise issues. Our detector is in-plane rotation invariant, and achieves satisfactory multi-view hand detection.

The future work is to add more hand posture classes for analyzing the performances and limitations of the proposed approaches. Different features such as contrast context histogram [11] will be studied and applied to accomplish hand posture recognition in real time. The system will be integrated with the NTU PAL1 robot for performing human-robot interaction. It should be of interest to study the methodology of jointly training and testing multiple hand posture classes.



(a) -60 degree (b) -40 degree (c) -20 degree

(d) 20 degree (e) 40 degree (f) 60 degree

Fig. 11. Multi-view hand detection. The detector works until that the viewpoint is larger than 40 degree. The images with red "PALM" texts are the correct recognition results.

VI. Acknowledgments

We acknowledge the helpful suggestions by an anonymous reviewer. This work was partially supported by grants from Taiwan NSC (#95-2218-E-002-039, #95-2221-E-002-433); Excellent Research Projects of National Taiwan University (#95R0062-AE00-05); Taiwan DOIT TDPA Program (#95-EC-17-A-04-S1-054); and Intel.

References

[1] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57(2), pp. 137–154, 2004.

[2] M. Kölsch and M. Turk, "Robust hand detection," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[3] ——, "Analysis of rotational robustness of hand detection with a viola-jones detector," in *the 17th International Conference on Pattern Recognition(ICPR'04)*, 2004.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] L. Anton-Canalis and E. Sanchez-Nielsen, "Hand posture dataset creation for gesture recognition," in *International Conference on Computer Vision Theory and Applications (VISAPP'06)*, Setúbal, Portugal, February 2006.

[6] A. Just, Y. Rodriguez, and S. Marcel., "Hand posture classification and recognition using the modified census transform," in *IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, 2006.

[7] A. Torralba, K. Murphy, and W. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[8] E. J. Ong and R. Bowden., "A boosted classifier tree for hand shape detection," in *IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, 2004.

[9] V. Athitsos and S. Sclaroff., "Estimating 3d hand pose from a cluttered image." in *Computer Vision and Pattern Recognition*, 2003.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, 1999.

[11] C.-R. Huang, C.-S. Chen, and P.-C. Chung, "Contrast context histogram - a discriminating local descriptor for image matching," in *International Conference of Pattern Recognition (ICPR)*, 2006.