

# Hand Posture Recognition Using Hidden Conditional Random Fields

Te-Cheng Liu, Ko-Chih Wang, Augustine Tsai and Chieh-Chih Wang

**Abstract**—Body-language understanding is essential to human robot interaction, and hand posture recognition is one of the most important components in a body-language recognition system. The existing hand posture recognition approaches based on robust local features such as SIFT can be invariant to background noise and in-plane rotation. However the ignorance of the relationships among local features is a fundamental issue. The part-based models argue that objects of the same category share the same part-structure which consists of parts and relationships among parts. In this paper, a discriminative part-based model, Hidden Conditional Random Fields (HCRFs), is used to recognize hand postures. Although the existing global locations of features have been used to consider large scale dependency among parts in the HCRFs framework, the results are not invariant to in-plane rotation. New features by the distance to the image center are proposed to encode the global relationship as well as to perform in-plane rotation-invariant recognition. The experimental results demonstrate that the proposed approach is in-plane rotation-invariant and outperforms the approach using AdaBoost with SIFT.

## I. INTRODUCTION

Body language is an important part of communication between people. Understanding body language could play a key role in human robot interaction as illustrated in Figure 1. A body-language recognition system could consist of face detection and analysis, arm gesture recognition and hand posture recognition. In our body-language recognition system, the Viola-Jones face detector [1] and a skin color model [2] are integrated into a robust frontal face detector. False alarms from the Viola-Jones face detector can be rejected by the skin color constraint. A 3D Active Appearance Model [3] is used to align 3D faces in 2D images for estimating the orientation of the user's head. Figure 2(a) shows the positive face detection result in which a person is facing toward the robot. Given the location of the face, the upper body of the detected person is extracted using the disparity map from the stereo cameras as shown in Figure 2(b). Kernel Sliced Inverse Regression [4] and Support Vector Machine [5] are used to train a raising-arm detector. The disparity map of the upper body is binarized as depicted in Figure 2(c) and used as the



Fig. 1. A person uses his body language to interact with the NTU-PAL2 robot.



(a) Face detection using single camera (b) Upper body localization using stereo cameras



(c) The input of the raising-arm detector (d) The extracted hand posture image

Fig. 2. The results of face detection, raising-arm detection and hand posture image extraction in our body-language recognition system.

Te-Cheng Liu was with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan [atwood@robotics.csie.ntu.edu.tw](mailto:atwood@robotics.csie.ntu.edu.tw)

Ko-Chih Wang is with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan [casey@robotics.csie.ntu.edu.tw](mailto:casey@robotics.csie.ntu.edu.tw)

Augustine Tsai is with the Innovative DigiTech-Enabled Applications and Services Institute, Institute for Information Industry, Taipei, Taiwan [atsai@iii.org.tw](mailto:atsai@iii.org.tw)

Chieh-Chih Wang is with the Department of Computer Science and Information Engineering, and the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan [bobwang@ntu.edu.tw](mailto:bobwang@ntu.edu.tw)

input of the raising-arm detector. The hand posture classifier is called only when both the face detector and the raising-arm detector provide positive results. In addition, the face detector and the raising-hand detector serve as the localizer of hand posture for further classification as shown in Figure 2(d). As the current human robot interaction design of our NTU-PAL2 robot depends on hand postures of users, a robust hand posture classifier is critical.

Wang and Wang [6] proposed to recognize hand posture using AdaBoost with Scale Invariant Feature Transform (SIFT) features [7]. Their model is invariant to background noise, illumination and in-plane rotation. In their model an image is represented only by a set of SIFT features.

The relationship among SIFT features are not considered. However, Lazebnik et al. [8] proposed a matching algorithm which considers spatial information in conjunction with spatial location to match objects. But they did not consider in-plane rotation problem. In additions, the part-based models argue that objects of the same category share the same *part-structure* which consists of parts and relationships among parts. Under the part-based model, relationships among features are expected to be helpful for recognition. In this paper, we propose to recognize hand posture by a part-based model in which relationships among features are taken into account.

Recently, the part-based models have significant progress in modeling the structure of an object. Fergus et al. [9] proposed a part-based model to overcome the variance of illumination and background noise in training images without labeling parts. Quattoni et al. [10] proposed a discriminative part-based model without labeling parts called Hidden Conditional Random Fields (HCRFs). The part-based models are expected to more accurately describe objects of which most instances share similar spatial configuration. Our experiments show that part-based models considering the global position features have higher accuracy, and this can be explained by two reasons. First, the relationship among parts is one of the necessary relationships for an instance of this kind to belong to such object class. Second, the within-class variance of local features of an object class could be large. By dividing an object into sub-classes with lower variance, the original problem is divided into easier sub-problems. Moreover, HCRFs obtain higher accuracy mainly due to the abandonment of the assumption on the form of the distribution of data.

In this paper, Global relationships among features are incorporated into HCRFs to consider large scale dependency among parts. Our experiments show that HCRFs are successfully applied on the upright hand posture dataset, but not on the cases with in-plane rotation, since the global relationship used by Quattoni et al. [10] is not invariant to in-plane rotation. In human robot interaction scenarios, hand postures are frequently in the rotated cases. In this paper, we propose to encode the global relationship of features by the distance to the image center to perform in-plane rotation-invariant hand posture recognition. The difference of part-structures with different global relationships is further discussed. The ample experiments demonstrate that the proposed approach has an advance in accuracy from 2.5% to 3.5% compared with the model of Wang and Wang [6], which supports our statements.

The rest of this paper is organized as follows. HCRFs are briefly reviewed in Section II and the cognitive meanings of HCRFs and part-structures are addressed and discussed in Section III. In Section IV, the ample experimental results and the comparisons are shown. Finally, we conclude the paper in Section V.

## II. HIDDEN CONDITIONAL RANDOM FIELDS

In order to investigate the part-structure representation under HCRFs in Section III, HCRFs are briefly reviewed

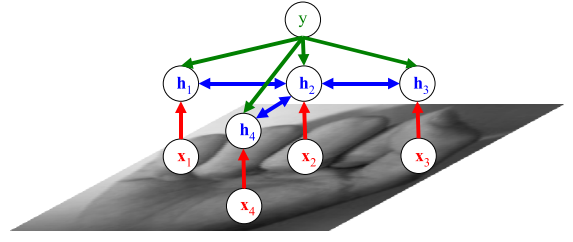


Fig. 3. An illustration of a double-layer CRF

in this section. The motivation of HCRFs is to construct a part-based model by Conditional Random Fields (CRFs) [11] without the efforts to label parts. Let  $\mathbf{X}$  be a vector of node features in an image,  $Y$  the set of class labels for images and  $\mathbf{Y}$  the set of assignments of  $Y$  to  $\mathbf{X}$ . Instead of modeling  $P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y})$  by some forms of distributions such as Gaussian distributions, the conditional probability  $P(\mathbf{Y}|\mathbf{X})$  is computed directly from  $\mathbf{X}$  in CRFs.

### A. Formulation

Quattoni et al. [10] defined a double-layer CRF with a random variable of parts  $\mathbf{h}$  as illustrated in Figure 3. Let  $H$  be the set of part labels and  $\mathbf{H}$  the set of all possible assignments of  $H$  to  $\mathbf{X}$ . The probability of a double-layer CRFs conditioned on observation  $\mathbf{x}$  is :

$$P(Y=y, \mathbf{H}=\mathbf{h}|\mathbf{X}=\mathbf{x}; \theta) = \frac{1}{Z_{y', \mathbf{h}'(\mathbf{x})}} \exp \left\{ \sum_i \theta_i f_i(y, \mathbf{h}, \mathbf{x}) \right\} \quad (1)$$

where  $\theta$  is the vector of parameters and  $Z_{y', \mathbf{h}'(\mathbf{x})}$  is a normalization factor over  $Y \times \mathbf{H}$ . As  $\mathbf{h}$  is unobserved and one method to deal with hidden variables is to marginalize out  $\mathbf{h}$  and work with  $P(Y=y|\mathbf{X}=\mathbf{x}; \theta)$ :

$$\begin{aligned} P(y|\mathbf{x}; \theta) &= \sum_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{x}; \theta) \\ &= \frac{1}{Z_{y', \mathbf{h}(\mathbf{x})}} \sum_{\mathbf{h}} \exp \left\{ \sum_i \theta_i f_i(y, \mathbf{h}, \mathbf{x}) \right\} \end{aligned} \quad (2)$$

Let  $N$  be the number of node features in an image,  $D$  the dimension of a node feature and  $E$  the edge set. Three specific forms of feature functions  $f_i(y, \mathbf{h}, \mathbf{x})$  are defined:

$$\begin{aligned} \sum_i \theta_i f_i(y, \mathbf{h}, \mathbf{x}) &= \sum_{j=1}^D \sum_{t=1}^N \theta_{\mathbf{h}, j}^{node} f_j^{node}(\mathbf{x}_t) + \\ &\sum_{k=1}^{|H|} \sum_{(s,t) \in E} \theta_{y_k}^{edge} f_{y_k}^{edge}(\mathbf{h}_s, \mathbf{h}_t) + \\ &\sum_{l=1}^{|H \times H|} \sum_{t=1}^N \theta_{y_l}^{node} f_{y_l}^{node}(\mathbf{h}_t) \end{aligned} \quad (3)$$

Let  $(u, v)$  be the normalized image location of a SIFT feature and  $s$  be the normalized scale at which the SIFT feature is found. A feature in the specific form of HCRFs is  $(u, v, s, S)$  where  $S$  is a SIFT descriptor of 128 dimensions.

Thus,  $D$  is 131 in this case.  $\theta_{\mathbf{h}_t, j}^{node}$  is used to evaluate the score for a node to be labeled as  $\mathbf{h}_t$ , and is shared by all classes.  $\theta_{yk}^{edge}$  and  $\theta_{yl}^{node}$  are used to evaluate the score for a part assignment to be of a class. Moreover,  $\theta_{yk}^{edge}$  is for the histogram of edges in the part layer.  $\theta_{yl}^{node}$  is for the histogram of nodes in the part layer.  $f_{yk}^{edge}$  and  $f_{yl}^{node}$  are designed as binary counting features:

$$f_{yl}^{node}(\mathbf{h}_t) = \begin{cases} 1, & \text{if } \mathbf{h}_t = l \in H \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

$$f_{yk}^{edge}(\mathbf{h}_s, \mathbf{h}_t) = \begin{cases} 1, & \text{if } (\mathbf{h}_s, \mathbf{h}_t) = edge_k \in H \times H \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The underlying expectation is that each category has its own distribution of parts different from other categories, and the scores from  $f_{yk}^{edge}$  and  $f_{yl}^{node}$  will discriminate the cases in which the scores from node features are not discriminative enough.

### B. Learning and Inference

The maximum-likelihood method is applied to estimate  $\theta$  of  $P(y|\mathbf{x})$ . Since the result of optimization remains and the computation is easier to deal with in the log domain, the log-likelihood function  $L(\theta)$  is to be maximized:

$$L(\theta) = \log P(\theta; y, \mathbf{x}) = \log Z_{\mathbf{h}}(y, \mathbf{x}) - \log Z_{y', \mathbf{h}}(\mathbf{x}) \quad (6)$$

$Z_{\mathbf{h}}(y, \mathbf{x})$  is the summation of  $P(y, \mathbf{h}|\mathbf{x})$  over  $\mathbf{H}$ . For  $n$  training instances, the objective function is  $\sum_{k=1}^n \log(\theta; y_k, \mathbf{x}_k)$ . The functions of the log-sum-exp form is convex [12]. However this log-likelihood function is the difference of two functions of this form and thus might not be concave [13]. The multiple initial cases are needed.

With learned  $\hat{\theta}$ , which class  $\mathbf{x}$  belongs to and what the most possible part assignment is for  $\mathbf{x}$  are two important questions which are solved simultaneously by  $\arg\max_{y', \mathbf{h}} P(y', \mathbf{h}|\mathbf{x})$ . However it is intractable and thus approximated by two tractable stages using belief propagation:

$$\hat{y} = \arg\max_y P(y|\mathbf{x}; \hat{\theta}) \quad (7)$$

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} P(\mathbf{h}|\hat{y}, \mathbf{x}; \hat{\theta}) \quad (8)$$

### C. Multi-class Recognition

For the multi-class recognition problems, there are two main approaches. One is the combination of multiple binary classifiers and the other is to solve a multi-class problem directly. The formulation of HCRFs can solve the multi-class recognition problem directly. In order to compare our approaches with others, we also implement the one-against-other approach. Let  $c_i(\mathbf{x})$  be a classifier for class  $i$ .  $c_i(\mathbf{x})_1$  is the probability that  $\mathbf{x}$  belongs to class  $i$  and  $c_i(\mathbf{x})_0$  is the probability that  $\mathbf{x}$  do not belong to class  $i$ . As our system is currently implemented to recognize three hand postures, we have a three dimensional classifier  $t_i(\mathbf{x})$ :

$$\begin{cases} t_i(\mathbf{x})_p & = c_p(\mathbf{x})_1 \\ t_i(\mathbf{x})_{q \neq p} & = \frac{1}{2} c_p(\mathbf{x})_0 \end{cases} \quad (9)$$

Given the assumption that the prior probability of each classifier is uniform, we have a combined one-against-other classifier  $O(\mathbf{x})$ :

$$O(\mathbf{x}) = \frac{1}{3} \sum_{i=0}^2 t_i(\mathbf{x}) \quad (10)$$

Experiments on these two approaches are shown in Section IV-C.

## III. PART-STRUCTURE

In this section the cognitive meaning of HCRFs is introduced. In HCRFs, a part-structure is not characterized by a part assignment. It is expressed in terms of  $\theta_{yk}^{edge}$  and  $\theta_{yl}^{node}$ .  $\theta_{yk}^{edge}$  is the histogram of the edges of parts and  $\theta_{yl}^{node}$  is the histogram of the nodes of parts in the objective function.  $\theta_{yk}^{edge}$  and  $\theta_{yl}^{node}$  does not directly correspond to a part assignment, although the most possible part assignment can be inferred from  $\theta_{yk}^{edge}$  and  $\theta_{yl}^{node}$ . Such part-structure representations are more abstract than part assignments. Therefore, the setting of these parameters is obscure to human beings. One advantage of HCRFs is that these parameters for part-structures are learned in a semi-supervised way.

### A. Semi-supervised Learning

A part-based model with supervised learning requires both of class labels and part labels. The definition and annotation of parts are two difficult tasks. In HCRFs, the part labels are hidden and the selection of part-structures is guided merely by class labels. In the parameter estimation of HCRFs by maximum-likelihood, the optimization converges if the gradient of  $L(\theta)$  reaches  $\vec{0}$ . Assume the derivative of  $L(\theta)$  with respect to  $\theta_{yl}^{node}$  and the one with respect to  $\theta_{yk}^{edge}$  are 0. Then,

$$\sum_t P(\mathbf{h}_t = j|y, \mathbf{x}; \hat{\theta}) = \sum_{y', t} P(\mathbf{h}_t = j|y', \mathbf{x}; \hat{\theta}) P(y'|\mathbf{x}; \hat{\theta}) \quad (11)$$

$$\sum_{(s,t) \in E} P(\mathbf{h}_s = a, \mathbf{h}_t = b|y, \mathbf{x}; \hat{\theta}) = \sum_{y', (s,t) \in E, a, b} P(\mathbf{h}_s = a, \mathbf{h}_t = b|y', \mathbf{x}; \hat{\theta}) P(y'|\mathbf{x}; \hat{\theta}) \quad (12)$$

, in which  $H$  is the set of part labels.

The left side of Equation 11 is the sum of the marginal probabilities for all nodes of class  $y$  to be labeled as part  $j$ . The left side of Equation 12 is the sum of marginal probabilities for all edges of class  $y$  to be labeled as  $(a, b)$ . The right side of each equation is the same sum weighted by class probabilities  $P(y'|\mathbf{x}; \hat{\theta})$ . That is, before convergence  $\theta_{yk}^{edge}$  and  $\theta_{yl}^{node}$  are adjusted according to the difference between the sum of the marginal probabilities in data and

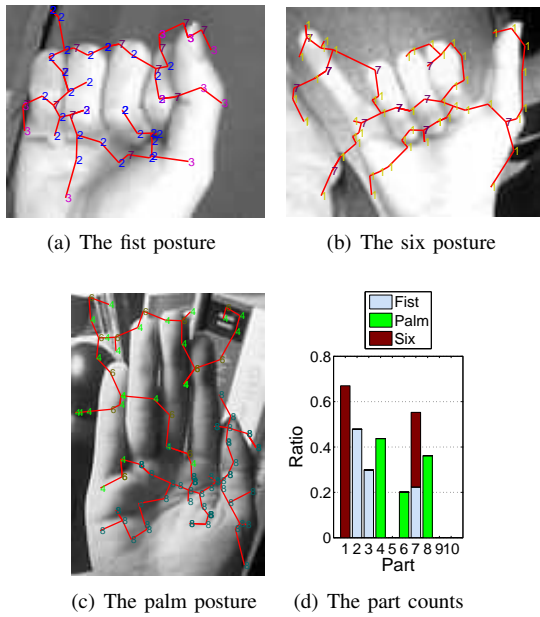


Fig. 4. The most possible part assignments from a three-class model with 86.7% accuracy.

the same sum weighted in current model during optimization. With convergence, with respect to the sum of marginal probabilities, the empirical expected value equals the expected value under the model. That is a variant of the equation between the empirical expected value and the expected value under the model for each feature function in CRFs [14].

It is observed that part-structures selected by the maximum likelihood method do not conform to our intuition on hand posture, but are more similar to a result of clustering in the defined feature space. This could be an essential property of a discriminative classifier. Part assignments annotated by people are filled with prior knowledge which may not be apparent in the defined feature space. Without an analysis of the similarity among features, the annotation are highly likely to be unreasonable in the sense that the within-part variance is unexpectedly high.

### B. Sharing and Non-Sharing Parts for Multi-class Recognition

A HCRF with higher accuracy tends to have a part-structure capable of indicating sharing and non-sharing parts. As depicted in Figure 4, each hand posture class has its own distribution of part counts in the most possible part assignment. For instance, Part 1 is only labeled in the six posture class, Parts 2 and 3 are only labeled in the fist posture class. These parts are non-sharing. Part 7 is a part shared by the fist and six posture classes. It conforms to our intuition on the similarity between Fist and Six. On the other hand, Part 1, 2 and 4 seem to have large potential to respectively indicate the existence of the classes six, fist and palm

### C. Global Relationship

In HCRFs, any two nodes are not assumed to be independent and thus may be overlapped as shown in Figure 5(a). In

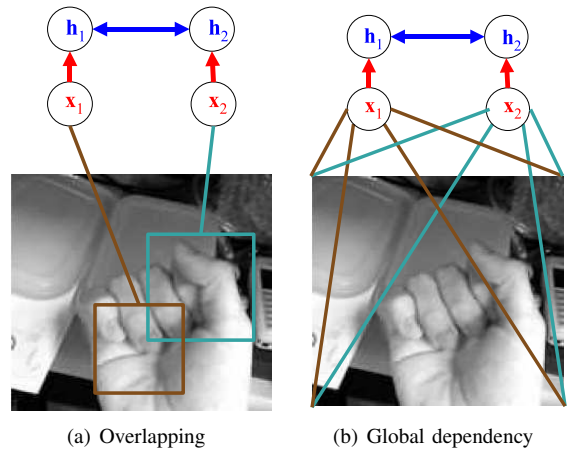


Fig. 5. The data relationship of nodes in HCRFs.

addition, global relationship of nodes may be incorporated into HCRFs so as to represent large scale dependency among data as shown in Figure 5(b).

In the specific form of Quattoni et al.[10], the scales and global positions of features are taken into account. The scales of features are used to discriminate the cases in which the thumb of an adult is of the size of a baby. The global positions of features are designed to discriminate the cases where a thumb is not at the proper position. However, the global locations of features extracted from 90°-rotated images do not work well using the model trained with upright images. Such models with significant weights on global locations are not invariant to in-plane rotation. We propose to encode the global relationship by the distance of each feature to the image center. Let the normalized distance of a feature in an image to the image center be  $d$ , then our feature form in HCRFs is  $(d, s, S)$ . Our experiments show that models with the proposed global relationship is invariant to in-plane rotation.

It could be argued that encoding the global relationship by the distance to the image center ignores the angular information. Thus, the distances to the image center are less informative than the global locations. However, our experiments show that a model considering global location is better than a model considering global distance to a small extent. HCRFs with our feature form performs in-plane rotation-invariant recognition with a quite small cost of accuracy. Additionally, the proposed models are better than the models considering no global relationship.

The differences of the part-structures trained with different global relationships can be easier observed via the most possible part assignment evaluated by Equation 8. The global positions could be a spatial cue to make features of a part compact in space. As depicted in Figure 6(a), Part 5 are mostly distributed in the top half of the image and Part 1 are mostly distributed in the bottom half of the image. The distances to the image center could be also a spatial cue but make features of a part less compact in space. Figure 6(b) shows that the distribution of Part 7 is mainly restricted in the

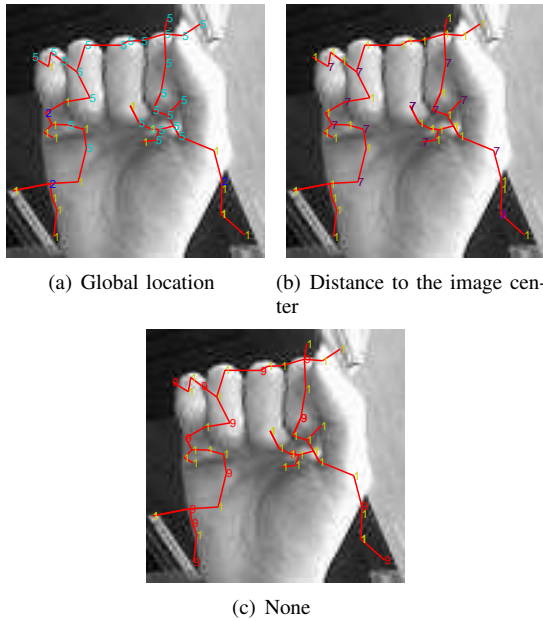


Fig. 6. The most possible part assignments from the models with and without considering the global relationships of features.

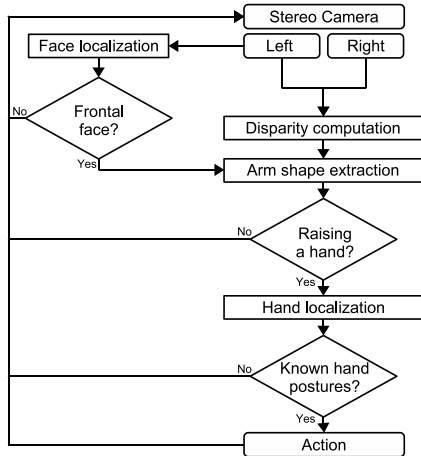


Fig. 7. The flowchart of the body-language recognition system. The NTU-PAL2 robot greets, head to or leave a user according to his/her hand postures.

central region. Without taking the global relationships into account, parts could be distributed over the whole image as shown in Figure 6(c).

#### IV. EXPERIMENTS AND DISCUSSION

##### A. System Overview and Dataset

The body-language recognition system is developed for the NTU-PAL2 robot to perform human robot interaction. The NTU-PAL2 robot greets, heads to or leaves a user according to his/her hand postures. Figure 7 is the flowchart of our body-language recognition system. A laptop computer with 1.66G CPU and 1.5G RAM is used in our system. The VIDERE Design STOC stereo camera with an image resolution of 640x480 is used to collect visual images. The average executing time of the face detector for each frame

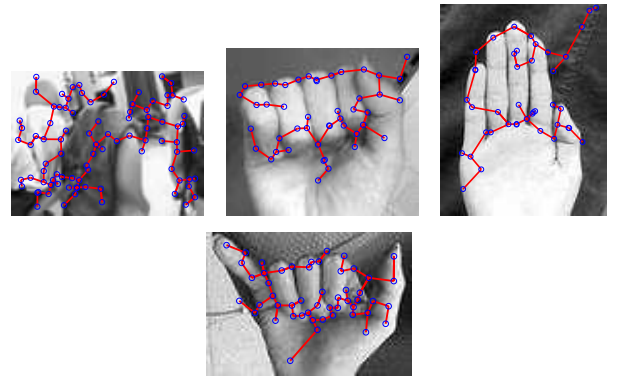


Fig. 8. Four sample images of the hand posture dataset. The corresponding trees of SIFT features for HCRFs are shown.

is 230 ms, the raising-hand detector is 140 ms, and the hand posture recognizer is 385 ms.

Figure 8 shows four sample images of our upright hand posture dataset and the corresponding trees of SIFT features. There are four classes: Background, Fist, Palm and Six. For each class, there are 200 images for training and 100 images for testing. For each hand posture, the variance of lighting condition and deformation are collected as well as the slight variance of 3D rotation. The 90° cases are obtained by flipping the upright cases. The size of each image is around 100x100 pixels.

##### B. Single-class Recognition

The performances of HCRF-based recognition with different settings are shown in Table I. The UIUC car side dataset and our hand posture dataset were used for evaluation. All the test images were also in-plane rotated 90° to test if the approaches are in-plane rotation-invariant. The algorithm none indicates that no global relationship of features is used. The algorithm global indicates that the global locations of features are used, and distance indicates that the proposed distances to the image center of features are used.

It is shown that the accuracy of our feature form of HCRFs has no obvious decrease in the 90°-rotated cases. thus, our proposal is invariant to in-plane rotation. On the other hand, models considering no global relationship are also invariant to in-plane rotation. However, the proposed global relationships can enhance the accuracy to the maximum of 4.5% in the class of Six. In terms of the upright cases, a model considering global location is better than a model considering global distance by 1.375% in average. That is, our feature form of HCRFs obtains invariance of in-plane rotation with a quite small cost of accuracy.

It should be noted that the accuracy of the algorithm Location for the classes of Fist and Palm decrease much less than the classes of Car Side and Six in the 90°-rotated cases. In addition, the algorithms Location and Distance make less enhancements in accuracy in the classes of Fist and Palm than Car Side and Six. A possible explanation could be that the contribution of the global relationship information is in proportion to its distribution to the rotated case.

TABLE I

THE ACCURACY OF THE MODEL TRAINED WITH ONLY UPRIGHT IMAGES

	None	Location	Distance
UIUC Car Side	89.0%	94.0%	92.0%
UIUC Car Side 90°	88.0%	78.0%	92.0%
Fist	93.0%	94.5%	94.0%
Fist 90°	93.0%	91.0%	93.5%
Palm	87.0%	89.5%	89.5%
Palm 90°	86.0%	86.0%	88.5%
Six	83.0%	90.5%	87.5%
Six 90°	82.5%	75.0%	87.5%

The recognition results using HCRFs with the proposed global relationships are also compared with the model on hand posture recognition [6] which uses AdaBoost with SIFT and is invariant to in-plane rotation. Table II shows that the proposed approach has an advance from 2.5% to 3.5%.

TABLE II

THE COMPARISON BETWEEN THE PROPOSED APPROACH AND THE APPROACH USING ADABOOST WITH SIFT

	Fist	Palm	Six
HCRFs+Distance+SIFT	94.0%	89.5%	87.5%
AdaBoost+SIFT	91.0%	87.0%	84.0%

### C. Multi-class Recognition

In this section, the one-against-other approach is compared with the multi-class approach in a three-class hand posture recognition problem. Table III and Table IV show the confusion matrices from the one-against-other approach and the multi-class approach, respectively. The accuracies of two approaches do not differ evidently. It may require experiments on more object classes to confirm if the multi-class approach is better or not than the one-against-other approach.

TABLE III

THE CONFUSION MATRIX OF THE ONE-AGAINST-OTHER APPROACH

	Fist	Palm	Six	Total	Accuracy
Fist	85	3	12	100	85%
Palm	3	92	5	100	92%
Six	8	6	86	100	86%
Total	96	101	103	300	87.6%

TABLE IV

THE CONFUSION MATRIX OF THE MULTI-CLASS APPROACH

	Fist	Palm	Six	Total	Accuracy
Fist	83	4	13	100	83.0%
Palm	1	92	7	100	92.0%
Six	7	8	85	100	85.0%
Total	91	104	105	300	86.7%

## V. CONCLUSION AND FUTURE WORK

In this paper, HCRFs were successfully applied to solve the multi-class hand posture recognition problem. HCRFs

with the proposed global relationships are in-plane rotation-invariant and outperform the approach using Adaboost with SIFT.

The future work is to test if the proposed approach is also multiview invariant. A dataset with multiple viewpoints as well as the theoretical foundations to tackle the variance of viewpoint under the part-based models will be established. In addition, it should be of our interest to integrate temporal information with the temporal extension of HCRFs to increase hand posture recognition performance.

## ACKNOWLEDGMENTS

This work was also partially supported by grants from Taiwan NSC (#96-2628-E-002-251-MY3, #97-2218-E-002-017, #96-2218-E-002-008); Excellent Research Projects of National Taiwan University (#95R0062-AE00-05); Taiwan DOIT TDDPA Program (#95-EC-17-A-04-S1-054); Taiwan ITRI, CCI, MSI; and Intel. This work was also partially conducted under the III Innovative and Prospective Technologies Project of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

## 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] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, pp. 264–277, 1999.
- [3] C.-W. Chen and C.-C. Wang, "3d active appearance model for aligning faces in 2d images," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2008.
- [4] H. M. Wu, "Kernel sliced inverse regression with applications to classification," *Journal of Computational and Graphical Statistics*, vol. 17, no. 3, pp. 590–610, 2008.
- [5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.
- [6] C.-C. Wang and K.-C. Wang, "Hand posture recognition using adaboost with sift for human robot interaction," *Springer Lecture Notes in Control and Information Sciences*, vol. 370, pp. 317–329, 2008.
- [7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60(2), pp. 91–110, 2004.
- [8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.
- [9] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of models for visual recognition," *International Journal of Computer Vision*, vol. 71, no. 3, pp. 273–303, 2007.
- [10] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 10, pp. 1848–1852, 2007.
- [11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning*, 2001.
- [12] S. D. Pietra, V. D. Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19(4), pp. 380–393, 1997.
- [13] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2007, pp. 93–128.
- [14] Y. Gong and W. Xu, *Machine Learning for Multimedia Content Analysis*, ser. Multimedia Systems and Applications. LLC, 233 Spring Street, New York, NY 10013, USA: Springer, 2007, vol. 30.