Image Set-based Hand Shape Recognition Using Camera Selection Driven by Multi-class AdaBoosting

Yasuhiro Ohkawa, Chendra Hadi Suryanto, and Kazuhiro Fukui

Graduate School of Systems and Information Engineering,
University of Tsukuba, Japan
{ohkawa@cvlab.cs,chendra@cvlab.cs,kfukui@cs}.tsukuba.ac.jp

Abstract. We propose a method for image set-based hand shape recognition that uses the multi-class AdaBoost framework. The recognition of hand shape is a difficult problem, as a hand's appearance depends greatly on view point and individual characteristics. Using multiple images from a video camera or a multiple-camera system is known to be an effective solution to this problem. In our proposed method, a simple linear mutual subspace method is considered as a weak classifier. Finally, strong classifiers are constructed by integrating the weak classifiers. The effectiveness of the proposed method is demonstrated through experiments using a dataset of 27 types of hand shapes. Our method achieves comparable performance to the kernel orthogonal mutual subspace method, but at a smaller computational cost.

1 INTRODUCTION

In this paper, we propose a hand shape recognition method that uses sets of image patterns captured by a multiple-camera system. By introducing camera selection based on the multi-class AdaBoost framework, the proposed method can classify the nonlinear distributions of input images effectively. Computational complexity is reduced as the method is based only on linear classifiers.

Hand gestures are often used in our daily life to facilitate communications with another person. Therefore, it is also expected that hand gestures can be used to achieve a more natural interaction between humans and computer systems. To recognize hand gestures automatically, the recognition of the three-dimensional shape of a hand is the most elementary requirement. Many types of hand shape recognition methods have been proposed. They can be divided into two categories: model-based methods and appearance-based methods[1].

Model-based methods use a three-dimensional hand model for recognition[2, 3]. They extract feature points such as edges and corners of hand images and match them to a three-dimensional hand model. For example, Imai has proposed a method for estimating hand posture in three dimensions by matching the edges extracted from a hand image to the silhouette generated from a typical hand model[3]. Although the model-based methods are widely used in various trial systems, they often suffer from unstable matching and high computational complexity, since a hand is a complex three-dimensional object with 20 degrees of freedom [1].

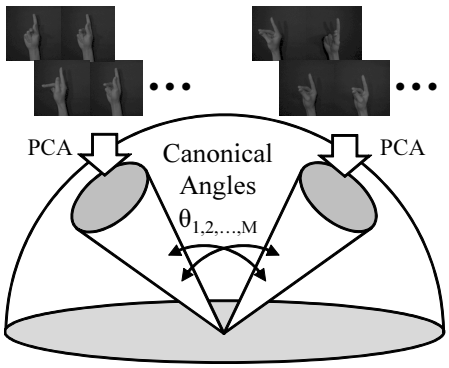


Fig. 1. Conceptual diagram of MSM. The distributions of multiple-viewpoint image sets of hands are represented by linear subspaces, which are generated by PCA. The canonical angles between two subspaces are used as a measure of the similarity between the distributions.

On the other hand, appearance-based methods [4–7] classify a hand shape from its appearance, where an $n \times n$ pixel pattern is treated as vector $\mathbf{x}$ in a n^2-dimensional space. These methods can deal with the variation of appearances due to changes of viewpoint, illumination and differences between individuals by preparing a static model representing these variations.

The mutual subspace method (MSM)[8] is one of the most suitable and efficient appearance-based methods for recognizing hand shape. The novelty of MSM is its ability to handle multiple sets of images effectively. MSM represents a set of patterns $\{\mathbf{x}\}$ of each class by a low-dimensional linear subspace in high-dimensional vector space using the Karhunen-Loève (KL) expansion, which is also known as principal component analysis (PCA). By introducing the subspace-based representation, the similarity between two sets of patterns can be easily obtained from canonical angles θ_i between subspaces, as shown in Fig.1.

MSM is better able to deal with variations of appearance due to changes of view point than are conventional methods using a single input image, such as the k-NN method. However, the classification ability of MSM declines considerable when the distribution of patterns has a nonlinear structure, such as that captured through a multiple-camera system. To overcome this problem, MSM has been extended to a nonlinear method, called the kernel mutual subspace method (KMSM) [9, 10]. Further, to boost classification ability, KMSM has been extended to the kernel orthogonal MSM (KOMSM) by adding the orthogonal transformation of class subspaces [11]. The ability of KOMSM to classify multiple sets of image patterns is as good as or better than other extensions of MSM [12–16]. KOMSM has also been demonstrated to be effective for hand recognition[7].

However, KOMSM has the serious problem that its computational cost and memory size requirements increase in proportion to the number of learning patterns and classes. In particular, the generation of the orthogonal transformation matrix, which is an essential component of KOMSM, is almost impossible when these numbers are large. The

problem is difficult, even for the distribution of patterns from a single camera. Thus, we can hardly apply KOMSM to a multiple-camera system, although the distribution of patterns obtained from the multiple-camera system contains more fruitful information about hand shape.

This problem of computational complexity cannot be completely solved even if the method of reduction [7] by k-means clustering or the incremental method [17] is applied. Accordingly, we propose an alternative approach based on the framework of ensemble learning[18] without using the kernel trick. In the proposed method, we regard a classifier based on the MSM as a weak classifier.

When applying the framework of ensemble learning to our problem, the method of generating various types of weak MSM-based classifiers is an important issue to be considered. We are able to achieve better performance than that of the original MSM method by generating the classifiers of ensemble learning from each camera, but performance is still far below that of nonlinear methods, such as KOMSM. This is because the classifiers generated from each camera hold an information obtained from a local viewpoint, but they do not hold the combination which contains richer information about the distribution. In contrast, KOMSM is able to encode the complete appearance of the image patterns in the nonlinear subspace. Therefore, we consider generating classifiers from all possible combinations of the multiple cameras so that we can obtain classifiers with a richer pattern distribution by combinations of camera selection. Thus, the number of classifiers increases from n to $2^n - 1$, where n is the number of cameras installed for ensemble learning.

It is difficult to determine suitable dimensions for an input subspace and reference subspaces. Thus, we add the dimension selection to the camera selection in the above framework. This additional process increases the number of the combinations substantially. However, such combinations may include ineffective classifiers and the computational cost with all the combinations is very high. Therefore, we select the best combinations from these using multi-class AdaBoost[19].

The rest of this paper is organized as follows. In Section 2, we explain the method for camera selection based on the multi-class AdaBoost. In Section 3, we explain the process flow of the proposed method. In Section 4, the effectiveness of our method is demonstrated through evaluation experiments with actual multiple-image sequences. Section 5 presents our conclusions.

2 Proposed method based on multi-class AdaBoost

In this section, we first explain the construction of weak classifiers that are effective for image set-based recognition using multiple cameras. Then, we explain the recognition of multiple-view images based on the MSM. Finally, we propose a method for selecting valid weak MSM-based classifiers from all the possible classifiers using the multi-class AdaBoost.

2.1 Generating Weak Classifiers

Figure 2 shows the concept of the proposed method for generating weak classifiers from a combination of selections from five cameras. First, the hand shape images are captured

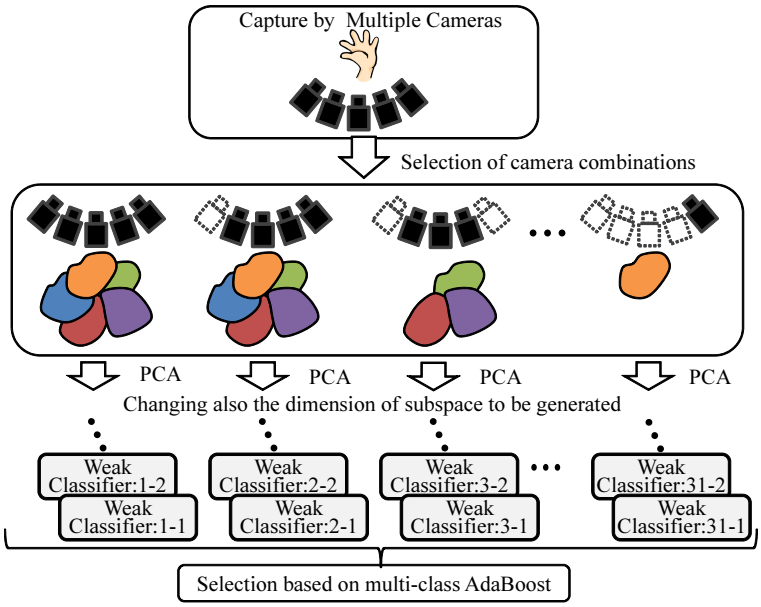


Fig. 2. Conceptual diagram of proposed method: Various weak classifiers are generated by changing the combinations of the cameras used for recognition and the dimensions of reference subspaces. Valid weak classifiers are selected using multi-class AdaBoost learning.

by the multiple-camera system. Next, we construct sets of combined images from the five cameras. Since we employ five cameras, the number of the camera combinations is $31(=2^5-1)$. Finally, weak classifiers are constructed by employing MSM to classify the sets of the combined images with various combinations of subspace dimensions. As not all of the weak classifiers are capable of constructing strong classifiers, we use multi-class AdaBoost to select the valid ones.

2.2 Mutual Subspace Method

In MSM, the distributions of reference patterns and input patterns are represented by linear subspaces, which are generated by principal component analysis (PCA). Then, the canonical angles between the two subspaces are used as a measure of the similarity between the distributions.

Definition of canonical angles between two subspaces The canonical angles can be calculated as follows. Given the M_1-dimensional subspace $\mathcal{P}_1$ and the M_2-dimensional subspace $\mathcal{P}_2$ in D-dimensional feature space, the M_1 canonical angles $\{0 \leq \theta_1, \ldots, \theta_{M_1} \leq \frac{\pi}{2}\}$ between $\mathcal{P}_1$ and $\mathcal{P}_2$ (for convenience $M_1 \leq M_2$) are uniquely defined by

$$\cos^2 \theta_i = \max_{\substack{\mathbf{u}_i \perp \mathbf{u}_j,\ \mathbf{v}_i \perp \mathbf{v}_j \\ 1 \leq i,j \leq M,\ i \neq j}} \frac{(\mathbf{u}_i \cdot \mathbf{v}_i)^2}{||\mathbf{u}_i||^2 ||\mathbf{v}_i||^2} , \tag{1}$$

Algorithm 1 Selection of weak MSM classifier based on multi-class AdaBoost

1: Given example input-subspaces $\mathcal{P}_1, \ldots, \mathcal{P}_N$, and class-labels $c_1, \ldots, c_N$ where $c_n = 1, \ldots, K$. $F^{(l)}$ indicates l-th weak classifier, which outputs a value of $1, \ldots, K$.
2: Initialize the weights $w_n = 1/N, n = 1, 2, \ldots, N$.
3: **for** $m = 1$ to M **do**
4: (a) Compute the weighted error of each weak classifier $err^{(l)} = \sum_{n=1}^{N} w_n(c_n \neq F^{(l)}(\mathcal{P}_n)), l = 1, \ldots L$.
5: (b) Select the weak classifier with the minimum error as mth weak classifier $T^{(m)}$ $T^{(m)} \leftarrow F^{(\arg\min_l err)}$.
6: (c) Compute the reliability α^m from the weighted error of mth weak classifier $T^{(m)}$ by $\alpha^m = \log\frac{1-err}{err} + \log(K-1)$.
7: (d) Update the weights: $w_n = w_n \exp(\alpha^{(m)}(c_n \neq T^{(m)}(\mathcal{P}_n))), n = 1, \ldots N$.
8: (e) Normalize the weights: $w_n \leftarrow \frac{w_n}{\sum_i^N w_n}$.
9: **end for**
10: output $C(\mathcal{P}) = \arg\max_k \sum_{m=1}^{M} \alpha^m \cdot (T^{(m)}(\mathcal{P}) = k)$.

where $(\cdot)$ denotes inner product and $||\cdot||$ denotes the norm of a vector.

A practical method of finding the canonical angles is by computing the $M_1 \times M_2$ matrix

$$\begin{aligned} \mathbf{C} &= \mathbf{V}_1{}^{\top}\mathbf{V}_2, \ , \\ \mathbf{V_1} &= [\boldsymbol{v}_1^1, \ldots, \boldsymbol{v}_{M_1}^1] \ , \\ \mathbf{V_2} &= [\boldsymbol{v}_1^2, \ldots, \boldsymbol{v}_{M_2}^2] \ , \end{aligned} \tag{2}$$

where $\boldsymbol{v}_s^1$ and $\boldsymbol{v}_s^2$ denote the s-th D-dimensional orthonormal basis vectors of the subspaces $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. The canonical angles $\{\theta_1, \ldots, \theta_{M_1}\}$ are the arccosine $\{\arccos(\kappa_1), \ldots, \arccos(\kappa_{M_1})\}$ of the singular values$\{\kappa_1, \ldots, \kappa_{M_1}\}$ of the matrix $\mathbf{C}$.

Similarity between two subspaces From these canonical angles, we calculate the similarity between two subspaces as $S = \frac{1}{M_1}\sum_{m=1}^{M_1}\cos^2\theta_m$. If the two subspaces coincide completely, S is 1.0, since all canonical angles are 0. The similarity S becomes smaller as the two subspaces separate. Finally, the similarity S is zero when the two subspaces are orthogonal to each other.

2.3 Selection of valid MSM-based weak classifiers by multi-class AdaBoost

Five cameras were used for the recognition. Therefore, the number of multiple-camera combinations is $31(= 2^5 - 1)$. Among those combinations, unnecessary weak classifiers are discarded and valid weak classifiers are selected by multi-class AdaBoost to generate the strong classifier. Algorithm 1 shows the detailed process of MSM-based weak classifier selection based on multi-class AdaBoost.

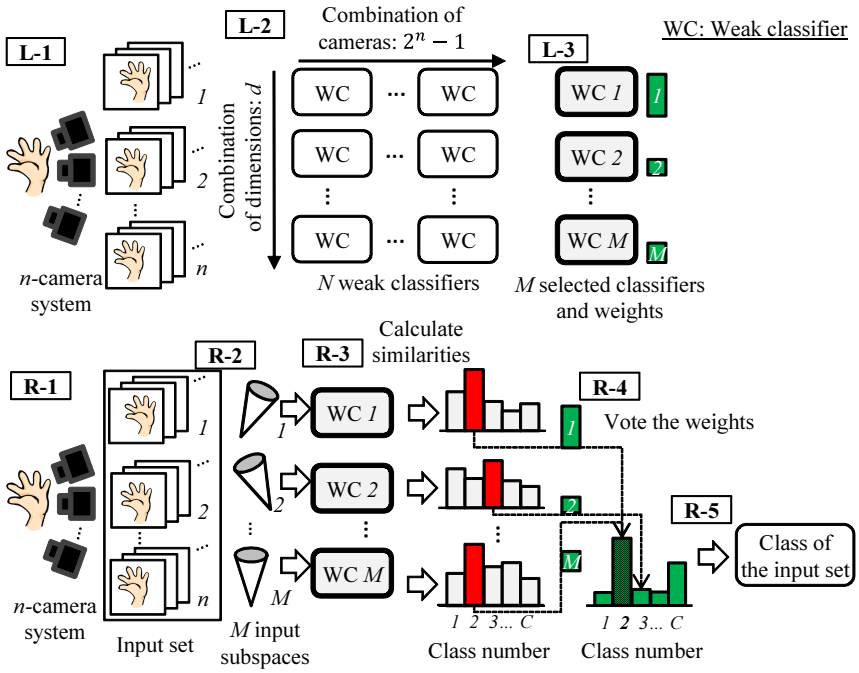


Fig. 3. The flow of the hand shape recognition process based on the proposed method.

3 Flow of hand shape recognition based on the proposed method

Figure 3 shows the flow of the recognition process based on the proposed framework. The whole process is divided into a learning phase and a recognition phase.

Learning phase

L-1: Collect n image sequences of each hand shape using a n-camera system, where n is the number of cameras installed and the number of class is C.

L-2: Generate $N(=(2^n-1)d)$ weak classifiers while changing both the combination of cameras used for inputting the image sequence and the dimensions of the input subspace and reference subspaces, where d is the number of the combinations of the dimensions.

L-3: Select the $M(\ll N)$-weak MSM classifiers and determine their weights by using the multi-class AdaBoost shown in Algorithm.1.

Recognition phase

R-1: Input n image sequences of an unknown hand shape using the n-camera system.

R-2: Generate $M(\ll N)$ input subspaces with the M combinations of the cameras and the dimensions of input and reference subspaces, which are corresponding to the selected weak classifiers in **L-3**.

R-3: Calculate the similarities of the input subspace and all class subspaces using a weak MSM classifier for each the combination.

R-4: Vote the weight to the class with the highest similarity of all the that obtained. Do this vote for all the combinations.

R-5: Classify the input set into the class with the highest voting.

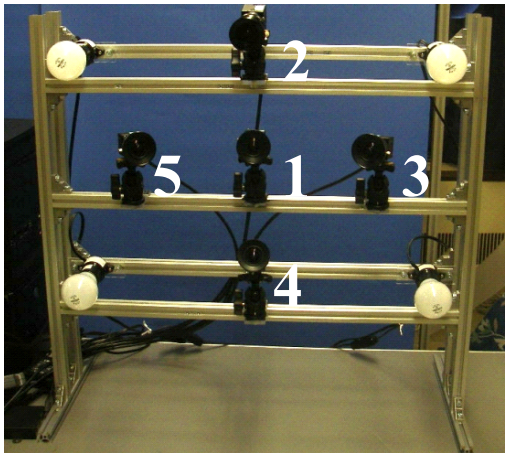


Fig. 4. Multiple-camera system.

Fig. 5. 27 types of hand shapes.

4 Experiments

4.1 Evaluation Data

We constructed a multiple-camera system to collect the evaluation images from seventeen subjects. The multiple-camera system consists of five IEEE1394 Point Grey Flea 2 cameras, as shown in Fig.4. The position of each camera is adjusted in such a manner that various view points of hand shape can be captured. The angle between the optical axes of the center camera and the other cameras was set to 18 degrees. The distance between the center camera and the other cameras was set to 21cm, and the distance from the center camera to the hand of a subject was about 40cm. To obtain more heterogeneous view of the hand shape, during the capture process the subjects rotated their hands to the left and right at a constant slow speed.

Fig. 6. Images of the same hand shape collected from 17 subjects.

Methods	ER [%]	EER[%]
Naive MSM	10.54	4.91
Voting-5	9.47	3.24
Voting-31	8.82	2.42

Table 1. Results of Experiment-I.

Using this multiple-camera setup, we collected the 27 types of hand shapes shown in Fig.5. The total number of collected images is 123000 (=90 frames $\times$ 5 cameras $\times$ 27 shapes $\times$ 17 subjects). Figure 6 shows the various appearances of the same type of hand shape collected from 17 subjects. Figure 7 shows the sequential images captured by the five cameras.

We cropped the hand shapes using skin color information and reduced the size to 32×32 pixels. Next, we extracted 140-dimensional feature vectors using higher-order local auto-correlation[20] from the four-level pyramid structure of the input image.

Sixteen subjects were used for learning, and one subject was used for evaluation. We repeated the experiment 17 times (once for each of the 17 subjects) and the average was taken as the experimental result.

We divided the 90 test images into 15 sets, each containing 6 images. Classification is done 6885 (=15 sets $\times$ 27 shapes $\times$ 17 subjects) times for each experiment. We adopt error rate (ER) and equal error rate (EER) for performance evaluation.

4.2 Experiment-I

This experiment evaluated the effectiveness of the multiple-camera selection by using naive MSM and voting classification. The dimension of the reference subspaces are set from 1 to 15, and the input subspaces from 1 to 5. Since each method in this experiment requires a different optimized subspace dimension, we use the optimized subspace dimension in the recognition phase.

Fig. 7. Example of sequential images captured by five cameras.

Num. Patterns	Method	ER[%]	EER[%]	Recog.Time[ms]
194400	5-Camera Ranking	9.47	3.24	20
	Proposed	8.73	2.16	132
	KMSM	7.7	4.48	2044
	KOMSM	7.33	2.11	2728
648000	5-Camera Ranking	9.21	2.74	21
	Proposed	7.79	2.03	134
	KMSM	-	-	-
	KOMSM	-	-	-

Table 2. Results of Experiment-III.

The experimental results are shown in table 1. In the Ranking-5 method, the voting is done using the five cameras only. While, the Ranking-31 method uses the 31 selections from the five camera. The experimental results show that by utilizing all of the possible camera combinations, the recognition performance is notably improved.

4.3 Experiment-II

In this experiment, we evaluated the relationship between the number of weak classifiers and the classification performance and computational cost of the recognition process. Various weak classifiers are generated not only by changing the camera selection, but also by changing the dimensions of input and reference subspaces. The dimension of the input subspace is set to 1, 2, or 3. The dimensions of reference subspaces are set from 5 to 90 in increments of 5. Thus, the total number of weak classifiers is 1674 ($=3 \times 18 \times 31$).

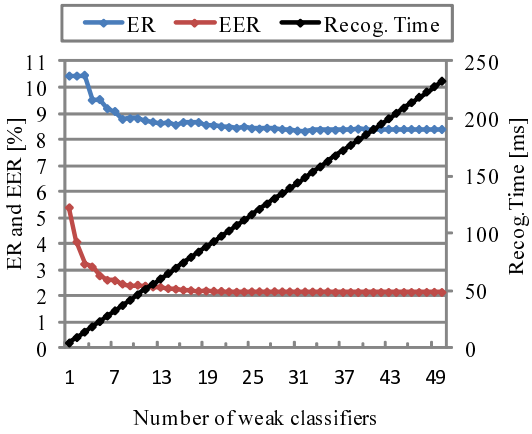


Fig. 8. Results of Experiment-II.

The experimental results are shown in Fig.8. The figure shows that the performance is notably improved by increasing the number of multi-class AdaBoost weak classifiers. When the number of weak classifiers reaches 30, both the ER and EER converge. The number of weak classifiers and the recognition time are linearly related.

4.4 Experiment-III

In this experiment, we compare the proposed method with the MSM, KMSM, and KOMSM classifiers. Since KMSM and KOMSM use the kernel trick, calculation becomes impossible when the number of learning patterns is substantially increased due to the complexity and the large memory requirement for the kernel trick computation. In fact, in our experimental setup, we are unable to add more learning patterns for KOMSM on a PC with 16GB of memory. On the other hand, the proposed method does not have this limitation on the number of learning patterns. To show the advantages of the proposed method, we performed another experiment in which the number of learning patterns is substantially increased. We collected new 481950 hand shape images (= 210 frames $\times$ 5 cameras $\times$ 27 shapes $\times$ 17 subjects) to be used as additional learning patterns.

The experimental results are shown in Table 2. As with Experiment-I, the 5-Camera Ranking methods shown in Table 2 do not use combinations of the five cameras. The proposed method uses all of the possible five-camera combinations as weak classifiers and employs multi-class AdaBoost to generate strong classifiers from them. This experiment demonstrates that the proposed method is about 20 times faster than KOMSM while having comparable performance. In the experiment in which the number of learning patterns is substantially increased, the EER of the proposed method is better than that of KOMSM while the recognition time is still much shorter than that of KOMSM.

○ : Selected camera ● : Unselected camera

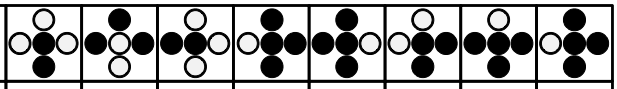
Camera Combination								
Weight	5.4	2.8	2.4	2.0	1.8	1.7	1.4	1.1
Ref. Dim	45	85	5	80	60	10	25	45
Input Dim	2	2	1	1	1	1	1	1

Fig. 9. Top eight weak classifiers selected by multi-class AdaBoost.

Next, we show the detail of the kinds of weak classifiers that are selected by multi-class AdaBoost. As explained previously, the weak classifiers are generated from all possible camera combinations and various input and reference subspace dimensions. Figure 9 shows the top eight selected weak classifiers arranged from the highest weight 5.4 (leftmost) to the 1.1 (rightmost). As an example from the figure, the first weak classifier selected by the multi-class AdaBoost chooses the upper, left, and right cameras with reference subspace dimension 45 and input dimension 2. It is worth noting that in the total of 510 selections, weak classifiers using all of the five cameras are never selected by the multi-class AdaBoost. Another interesting fact is that the center camera is less likely to be selected.

5 Conclusion

This paper proposes an image set-based hand shape recognition method using camera selection driven by the multi-class AdaBoost. In the proposed method, we consider a simple linear mutual subspace method as a weak classifier, and construct a strong classifier by integrating these weak classifiers. The obtained strong classifier could outperform one of the state-of-the-art nonlinear kernel methods, KOMSM, without using the kernel trick technique and with smaller computational cost.

Acknowledgment

This work was supported by KAKENHI (22300195).

References

1. Erol, A., Bebis, G., Nicolescu, M., Boyle, R., Twombly, X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding **108** (2007) 52–73
2. Stenger, B., Thayananthan, A., Torr, P., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. IEEE transactions on pattern analysis and machine intelligence (2006) 1372–1384
3. Imai, A., Shimada, N., Shirai, Y.: Hand posture estimation in complex backgrounds by considering mis-match of model. Asian Conference on Computer Vision (2007) 596–607

4. Martin, J., Crowley, J.: An appearance-based approach to gesture-recognition. Image Analysis and Processing (1997) 340–347
5. Birk, H., Moeslund, T., Madsen, C.: Real-time recognition of hand alphabet gestures using principal component analysis. Scandinavian Conference on Image Analysis **1** (1997) 261–268
6. Cui, Y., Weng, J.: Appearance-based hand sign recognition from intensity image sequences. Computer Vision and Image Understanding **78** (2000) 157–176
7. Ohkawa, Y., Fukui, K.: Hand shape recognition based on kernel orthogonal mutual subspace method. IAPR Conference on Machine Vision Applications (2009) 222–225
8. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. IEEE International Conference on Automatic Face and Gesture Recognition (1998) 318–323
9. Sakano, H., Mukawa, N., Nakamura, T.: Kernel mutual subspace method and its application for object recognition. Electronics and Communications in Japan **88** (2005) 45–53
10. Wolf, L., Shashua, A.: Learning over sets using kernel principal angles. The Journal of Machine Learning Research **4** (2003) 913–931
11. Fukui, K., Yamaguchi, O.: The kernel orthogonal mutual subspace method and its application to 3d object recognition. Asian Conference on Computer Vision (2007) 467–476
12. Fukui, K., Yamaguchi, O.: Face recognition using multi-viewpoint patterns for robot vision. 11th International Symposium of Robotics Research (2003) 192–201
13. Kawahara, T., Nishiyama, M., Kozakaya, T., Yamaguchi, O.: Face recognition based on whitening transformation of distribution of subspaces. Workshop on Asian Conference on Computer Vision, Subspace2007 (2007) 97–103
14. Li, X., Fukui, K., Zheng, N.: Image-set based face recognition using boosted global and local principal angles. Asian Conference on Computer Vision (2009) 323–332
15. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Transactions on Pattern Analysis and Machine Intelligence (2007) 1005–1018
16. Kim, T., Arandjelovic, O., Cipolla, R.: Boosted manifold principal angles for image set-based recognition. Pattern Recognition **40** (2007) 2475–2484
17. Chin, T., Suter, D.: Incremental kernel principal component analysis. IEEE Transactions on Image Processing **16** (2007) 1662–1674
18. Bishop, C.: Pattern recognition and machine learning. Springer New York (2006)
19. Zhu, J., Rosset, S., Zou, H., Hastie, T.: Multi-class adaboost. Technical report, Department of Statistics, University of Michigan **1001** (2006)
20. Otsu, N., Kurita, T.: A new scheme for practical flexible and intelligent vision systems. IAPR Workshop on Computer Vision (1988) 467–476