Classification of similar 3D objects with different types of features from multi-view images
– An approach to classify 100 Apples –

Hitoshi Niigaki[1] and Kazuhiro Fukui[2]

University of Tsukuba, Graduate School of Systems and Information Engineering,
1-1-1 Tennoudai, Tsukuba, Ibaraki, Japan
[1]niigaki@cvlab.cs.tsukuba.ac.jp,[2]kfukui@cs.tsukuba.ac.jp

Abstract. This paper proposes a method for classifying 3D objects with similar appearances using different types of features from multi-view images. We can find this type of task in various practical applications, such as flaw inspection of industrial components, quality checking, ans screening of fruits and vegetables. In this paper, as an example such a concrete task, we will deal with the problem of classifying apples, a task that is difficult even for human vision. To tackle this task, we will introduce the mutual subspace method (MSM)-based methods as a weak classifiers in an ensemble learning framework. In addition, we will consider three types of features: shape, texture and color in the terms of invariants of position and scale, as input vectors of each MSM-based classifier. The effectiveness of the proposed method will be demonstrated through the results of evaluation experiments using 100 apples.

1 introduction

Many view-based methods have been proposed for 3D object recognition, which is one of active research areas in computer vision[1]. Several investigations into the issue suggest the effectiveness of utilizing rich information obtained from multi-view images to achieve high-performance[2]-[6].

The mutual subspace method (MSM) has the ability to handle multiple images, including sequential images or multi-view images, and so is suitable and efficient for recognizing 3D objects. Let an $n \times n$ pixel pattern be treated as a vector $\mathbf{x}$ in n^2-dimensional space. In MSM, the set of patterns of each class is represented by a low-dimensional linear subspace using Karhunen-Loève(KL) expansion, also known as principal component analysis (PCA). The classification of a set of patterns is executed based on the canonical angles θ between subspaces, $\mathcal{P}$ and $\mathcal{Q}$, where smaller angles indicate higher similarity between the two subspaces as shown in Fig. 1.

MSM and its extensions, CMSM[4] and OMSM[5], have been successfully applied to various practical applications, such as face recognition[6], ISAR image analysis[7], and lip reading[8]. The successes in face recognition are especially noteworthy. However, face recognition may be a relatively easy classification

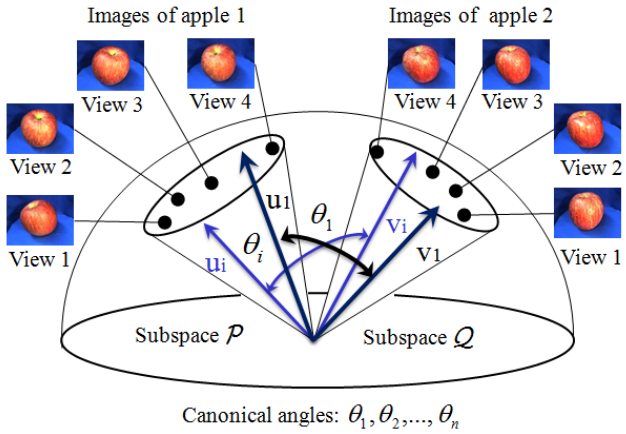

Fig. 1. Similarity between two distributions of multi-view image patterns.

problem, considering that it can be easily executed by human vision. This leads us to the question, how well the MSM-based methods perform in classifying 3D objects with appearances so similar that even human vision has difficulty classifying them? There are many types of 3D objects with such characteristics in various practical applications; flaw inspection of industrial component, quality checking, and screening of fruits and vegetables[9][10].

In this paper, we consider the challenging problem of classifying 100 apples as a representative task. As shown in Fig. 2, even human vision has trouble classifying them as the number of apples increases. If MSM-based methods achieve high performance in this task, it should indicate a potential applicability to the classification of other objects with very similar appearances.

Face recognition systems based on MSM-based methods have achieved high performance by using only single appearance feature. However, to perform more difficult tasks such as those described above, it is necessary to use multiple feature types obtained from multi-view images, and use them by considering their characteristics. We consider three types of features: shape-type(P-type Fourier Transform descriptor[11]), texture-type(2D FFT power spectrum, view, HOG[12], HLAC[13]) and color-type(color histogram) in the terms of invariants of position and scale. These are used as input vectors for each MSM-based methods. Then the multiple classification results from all the MSM-based methods are combined in a ensemble learning framework.

The rest of the paper is organized as follows. Section 2 outlines an our method, including the algorithms of the MSM-based methods. Section 3 demonstrates the effectiveness of our method through evaluation experiments using 100 apples. Section 4 presents our conclusions.

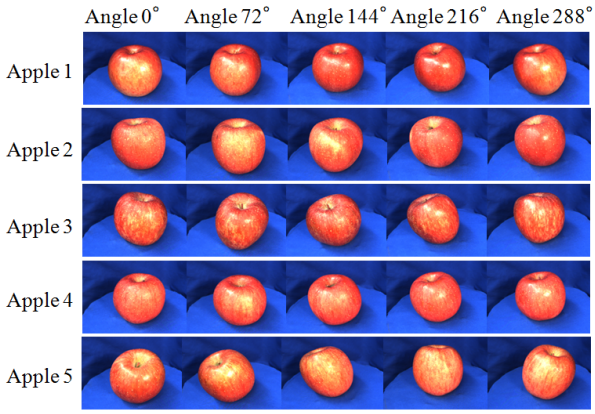

Fig. 2. Multi-view image patterns of apples.

2 The proposed method

In this section, we first describe the algorithm of the MSM-based methods, MSM, CMSM and OMSM before discussing various feature types extracted from multi-view images in the term of invariant of position and scale. Then, we combine the results of all the MSM-based classifiers using different feature types in the framework of the ensemble learning.

2.1 Mutual Subspace Method (MSM)

As mentioned in Sec. 1, MSM measures the similarity between the distributions of reference patterns and input patterns by using canonical angles between two subspaces. These canonical angles can be calculated by the following procedure.

Given an m-dimensional subspace $\mathcal{P}$ and an n-dimensional subspace $\mathcal{Q}$ (as a matter of convenience $m \geq n$), the n canonical angles $\{0 \leq \theta_1, \theta_2, .., \theta_n \leq \pi/2\}$ are determined as shown in Fig. 1. The first canonical angle is the smallest angle between the two subspaces and the second canonical angle is the smallest angle along the direction orthogonal to the first canonical angle. $cos^2\theta_i$ for $i = 3, \ldots, n$ are calculated similarly. The canonical angles θ_i between $\mathcal{P}$ and $\mathcal{Q}$ are uniquely defined as :

$$\cos^2 \theta_i = \max_{\substack{u_i \perp u_j (j=1,2,\ldots,i-1) \\ v_i \perp v_j (j=1,2,\ldots,i-1)}} \frac{(\mathbf{u}_i \cdot \mathbf{v}_i)^2}{\| \mathbf{u}_i \|^2 \| \mathbf{v}_i \|^2} \quad , \tag{1}$$

where $\mathbf{u}_i \in \mathcal{P}, \mathbf{v}_i \in \mathcal{Q}$.
Let $\mathbf{\Phi}_i$ and $\mathbf{\Psi}_i$ denote the i-th n-dimensional orthonormal basis vectors of the subspaces $\mathcal{P}$ and $\mathcal{Q}$, respectively. These orthonormal basis vectors can be obtained as the eigenvectors of the autocorrelation matrix $\sum_{i=1}^{l} \mathbf{x}_i \mathbf{x}_i^T$ calculated from the l learning patterns $\{\mathbf{x}\}$ of each class.

A practical method of finding the canonical angles is by computing the matrix $\mathbf{X}$=$\mathbf{A}^T\mathbf{B}$, where $\mathbf{A} = [\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_m]$ and $\mathbf{B}$=$[\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_n]$. Let $\{\kappa_1, \ldots, \kappa_n\}$ $(\kappa_1 \geq, \ldots, \geq \kappa_n)$ be the singular values of the matrix $\mathbf{X}$. The canonical angles $\{\theta_1, \ldots, \theta_n\}$ can be obtained as $\{cos^{-1}(\kappa_1), \ldots, cos^{-1}(\kappa_n)\}$.

2.2 Definition of similarity

The similarity between two subspaces is defined as

$$S[n'] = \frac{1}{n'} \sum_{i=1}^{n'} \cos^2 \theta_i \quad , \tag{2}$$

where n' is the number of canonical angles used for calculating the similarity. The value $S[n']$ reflects the structural similarity between two subspaces. In cases in which two subspaces coincide completely with each other, $S[n']$ is 1.0, since all canonical angles are zero. The similarity $S[n']$ becomes smaller as the two subspaces separate. The similarity $S[n']$ is zero, only when the two subspaces are orthogonal to each other.

2.3 Extensions of MSM: CMSM and OMSM

In order to improve the performance of MSM, it has been extended to the constrained mutual subspace method (CMSM[4]) and the Orthogonal Mutual Subspace Method (OMSM[5]).

In CMSM, each class subspace is projected onto a discriminant space referred to as the constraint subspace $\mathcal{D}$. This projection extracts a common subspace of all the class subspaces from each class subspace, such that the canonical angles between class subspaces are enlarged to approach orthogonal relation. Given the projection matrices $\mathbf{P}_i (i = 1, 2, ..., k)$ of k classes, the constraint subspace $\mathcal{D}$ is spanned by the eigenvectors corresponding to the N_d-th smallest eigenvalue of the following matrix:

$$\mathbf{G} = \sum_{i=1}^{k} \mathbf{P}_i \quad . \tag{3}$$

The dimension N_d is set experimentally.

OMSM also realizes the orthogonalization by Fukunaga and Koontz's method [14], so that it improves the performance of MSM. In their method, orthogonalization is achieved by applying the whitening transformation matrix $\mathbf{O}$ to the training patterns or orthonormal basis vectors of each class subspace. The whitening matrix $\mathbf{O}$ is defined as

$$\mathbf{O} = \mathbf{\Lambda}^{-1/2}\mathbf{B}^T \quad , \tag{4}$$

where $\mathbf{\Lambda}^{-1/2}$ is the diagonal matrix whose i-th component is the reciprocal of the square root of the i-th highest eigenvalue of $\mathbf{G}$. $\mathbf{B}$ is the matrix whose i-th column vector is the eigenvector of the matrix $\mathbf{G}$ corresponding to the i-th highest eigenvalue.

2.4 Valid features extracted from multi-view images

The conventional MSM-based methods have mainly used the "view (appearance) feature" that is obtained by raster scan of an image. However, the view feature varies largely depending on the position and scale of an object. Thus, we can predict that using only view feature is inadequate to classify similar 3D objects with a high degree of accuracy. There are numerous features that we can extract from multi-view images.

- **Shape-type**: P-type Fourier Transform descriptor(P-FT)
  P-FT is based on the boundary-based shape descriptor, which is position and scale invariant feature. However, it is not so robust with regards to variations caused by shadows and illumination. When the number of objects is large, its classification ability may decrease.
- **Texture-type**: HOG, HLAC, 2D-DFT descriptor(2D-DFT), view feature
  HOG is based on the magnitude information of edges in local region. HLAC is based on the autocorrelation between pixels in a local region. 2D-DFT is obtained as low-frequency Fourier components of an image. HLAC and 2D-DFT are invariant to position. In contrast, HOG and view feature are not invariant.
- **Color-type**: Color histogram
  This type is widely used in image retrieval, as it is position and scale invariant. However, when the number of apples increases, its classification ability may drop.

As shown in Table 1, these features are classified into three types: shape, texture and color. From this table, we can see that these characteristics are mutually complementary. Therefore, it should be effective to utilize multiple different feature types obtained from multi-view images in order to realize the classification of many apples that have very similar appearances.

2.5 Ensemble learning

In this section, we construct the ensemble classification based on multiple MSM-based classifiers with different kinds of features. Fig. 3 shows the flow chart of

Table 1. Characteristics of each feature

	Shape type	Texture type	Color type	Position invariant	Scale invariant
View		o		x	x
HOG		o		x	x
P-FT	o			o	o
2D-DFT		o		o	x
HLAC		o		o	x
Color histogram			o	o	o

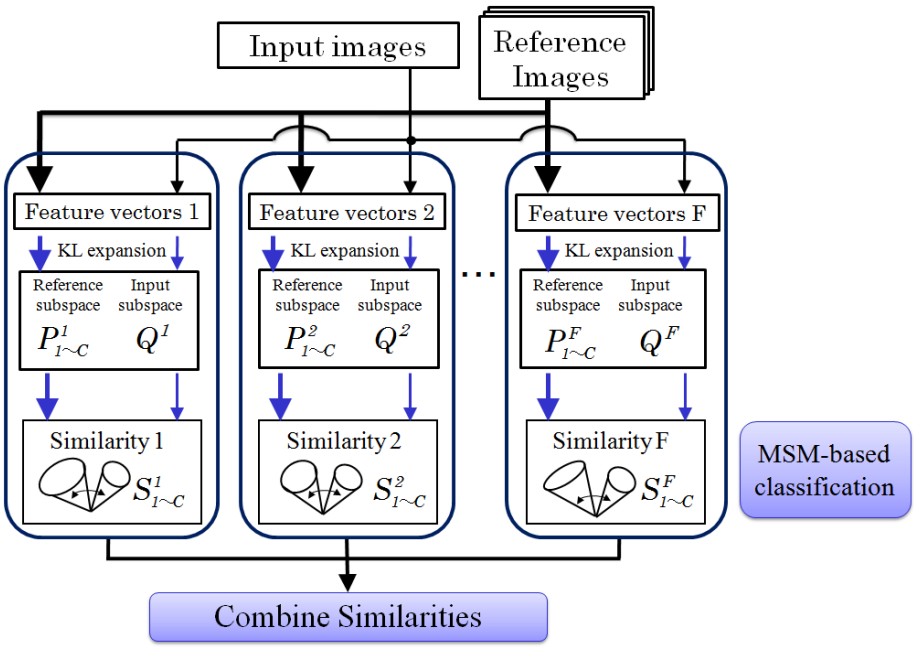

Fig. 3. Ensemble classification

the ensemble classification, which consists of a learning stage and a classification stage.

Learning stage:

1. The multiple kinds of feature vectors $\{\mathbf{x}_i^f\}(i = 1, \ldots, C, f = 1, \ldots, F)$ are extracted from learning multi-view images, where F is the number of feature types and C is the number of classes.
2. The reference subspaces P_i^f are generated from learning set $\{\mathbf{x}_i^f\}$ using KL expansion.
3. The above operations are executed for all the C classes.

Classification stage:

1. An input subspace Q^f is generated from feature vectors extracted from input multi-view images using KL expansion.
2. The similarity S_i^f between an input subspace Q^f and reference subspace P_i^f is measured using MSM-based method.
3. The similarity S_i^f from feature f is normalized to $S_i^{f'}$ by dividing it by the max similarity among $\{S_1^f, \ldots, S_C^f\}$.
4. The normalized similarities $\{ S_i^{1'}, ..., S_i^{F'} \}$ are combined by using $\frac{1}{F}\sum_{f'=1}^{F} S_i^{f'}$. This operation is executed for all the C classes.
5. The input set is classified as the class with the highest similarity.

3 Experiments

In this section, we first describe the details of each feature extraction. Next, we execute a preliminary experiment to evaluate the effectiveness of several features. We then show the results of classification of 100 apples using a single feature are shown. Finally, we demonstrate the effectiveness of using multiple different types of features.

3.1 Details of each feature extraction

- **P-type Fourier Transform descriptor (P-FT)**[11]
  The P-type Fourier descriptor is a representation of the object boundary. In this experiment, we used 40 low-frequency components extracted from two images, which were an original 320×240 pixel image and a half size 160×120 pixel image. The dimension of feature vector was set to 80(= 40 × 2).
- **Color histogram feature**
  Color histogram was made by partitioning the red, green and blue axes into 16 regions. Histogram elements were divided by the number of pixels in the object for normalization. A 96-dimensional feature vector was extracted from the two images described in the explanation of P-FT.
- **View base feature**
  A 768-dimensional feature vector was extracted from the 32 × 24 pixels monochrome image converted from an input image.
- **2D Discrete Fourier Transform descriptor (2D-DFT)**
  The feature vector was obtained as the low-frequency components by applying the Fourier transform to an input image. A 198-dimenional feature vector was extracted from the two images described in P-FT.
- **Histograms of Oriented Gradients descriptor (HOG)**[12]
  A 32×24 pixel image converted from an input image was divided into cells of 8 × 8 pixels, and each group of 2 × 2 cells was integrated into a block. Each cell consisted of an 8-bin histogram of oriented gradients and each block consisted of a vector of combined histograms of its cells.
  The vector was normalized with respect to each block, and its dimension was 32(= 8 × 4). The dimension of HOG was set to 192(= 32 × 6), as an image had 6 blocks,
- **Higher-order Local Auto-Correlation feature (HLAC)**[13]
  HLAC is defined as :

$$X(a_1, ...a_n) = \int I(r)I(r + a_1)....I(r + a_n)dr \quad , \tag{5}$$

  where the nth-order autocorrelation functions with n displacements $\{a_1, ..., a_n\}$. $I(r)$ denotes the pixel value of the image. Here, we restrict the order n up to the second and restrict the range of displacements within a local 3×3 window.
  The feature vectors (HLAC1) were extracted from the two images described

in P-FT. The feature vectors (HLAC2) were extracted from an x-orient differential image and a y-orient differential image, and the image size was 320×240. The dimensions of HLAC1 and HLAC2 feature were set to 70.

3.2 Preliminary experiment using five apples

Experimental conditions:
We collected multi-view images of apples by using a gathering system with an IEEE1394 camera and a turntable as shown in Fig. 4. We captured multiple images of an apple while rotating the turntable. This operation was repeated three times for each apple. Note that the positions of the apples were different for the three gatherings.

In the first gathering, two hundred images were captured at 1.8 degree intervals around the entire circumference. These images were used as training data. The reference subspaces were generated from these 200 images. In the second and third gatherings, the multi-view images were captured at the 3.6 degree intervals, while changing the range θ of the view angles from 36 degrees to 360 degrees, as in Fig. 4. We thus obtained the six test sets. The numbers of images in the sets were 10, 13, 20, 25, 50, and 100. These images were used as testing data. The dimensions of all the reference subspace were set to 10. The dimension of the input subspace was set to 3.

Experimental results and discussion:
Fig. 5 shows the changes of the recognition rate against the range θ of the view angles. We can see that the recognition rate increases as the range of view angle widens. When the range θ was set to more than 90 degrees, the classification rates of P-FT, 2D-DFT, color histogram and HLAC2 were over 90%. The rate of HLAC2 was higher than that of HLAC1 when the range θ was small. On the other hand, the rates of view feature and HOG were not particularly high even when the range θ was set to 360 degrees. This difference indicates that position invariant features are effective for MSM-based method in classification

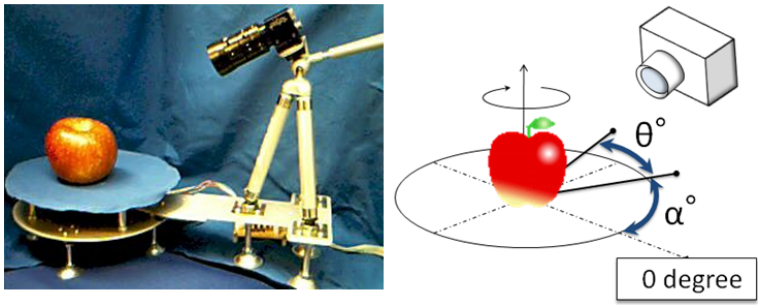

Fig. 4. Gathering system of multi-view images of apples.

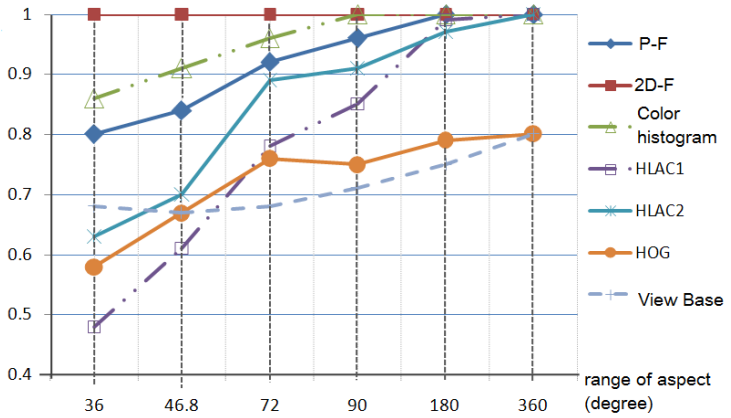

Fig. 5. Recognition rate against the range θ of view angles

of 3D objects. From these results, for the next experiment using a large data set we chose some position invariant features: P-FT, 2D-DFT, color histogram and HLAC2 and the range θ of the view angles to 90 degrees.

3.3 Classification of 100 apples using a single feature

Experimental conditions:
We used the MSM-based methods (MSM, CMSM, OMSM) and k-NN method as classifiers and compared their performances. An input subspace was generated from 25 images obtained in the range θ from α to $\alpha+90$ degrees. We changed the start angle α at steps of 36 degrees. Such image gathering was executed twice, and so the total number of the evaluation trials was 2000 (=10(=360/36) $\times$ 2 (two rotations) $\times$ 100 (number of apples)).

The reference subspaces were generated from 200 learning images. The dimensions of reference subspaces were set to 10. The dimension of the input subspace varied from 1 to 5. The constraint subspace $\mathcal{D}$ and the whitening matrix $\mathbf{O}$ were generated from learning data obtained from 20 apples that were different from the 100 apples used in the above learning phase.

In k-NN method, we registered 10 vectors created by clustering of 200 training data for each class. Given m input patterns, the output is determined by the $k \times m$ voting result.

Results and discussion:
Table 2 and Table 3 show the experimental results. In the tables, the best recognition rates are shown when the number of canonical angles, n', in Equation 2, varied from 1 to 5. The notation between brackets () indicates the value of n. In the "classifier" column, the figures for "CMSM " item indicate the dimension of the principal component subspace $\mathcal{M}$ [4].

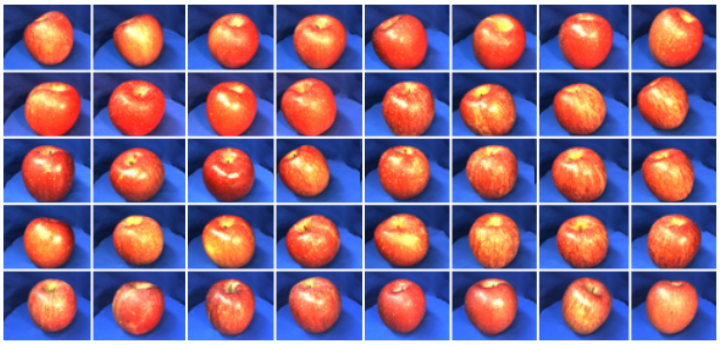

Fig. 6. 40 of all the apples used in the experiments

Classifier	P-FT	2D-FT	Color histogram	HLAC
MSM	76.9(3)	89.75(4)	89.6(3)	54.9(5)
CMSM-1	85.6(3)	92.2(4)	94.55(1)	76.95(4)
CMSM-2	88.25(3)	93.4(4)	94.25(1)	94.9(2)
CMSM-5	88.25(2)	92.15(4)	94(1)	98.75(1)
CMSM-10	80.3(1)	90.65(4)	93(1)	99(1)
OMSM	82.2(5)	86.2(4)	98.45(1)	94.45(1)
NN	87.75	72.7	91.95	34.9
3-NN	79.4	69.35	83.4	28.85
5-NN	70.1	61.75	72.95	23.8

Table 2. Recognition rate of each classifier (%).

We compared the performances of all the classifiers. The performance of k-NN was worst among them. We can see that the recognition rates of CMSM and OMSM were largely improved in comparison with that of MSM. For example, the performance of MSM using HLAC was extremely low, whereas even when HLAC was used, the performance of CMSM and OMSM were extremely superior. This result suggests that MSM lacks in the classification ability as compared with CMSM and OMSM. None of the classifiers using P-FT achieved the rate of 90%. The reason may be that the stable extraction of the contour of the apples was affected by shadows. Color histogram feature had better performance for all the methods in comparison with the other types of features. From the above results, we can confirm that these four kinds of features have different characters for classification of apples.

Classifier	P-FT	2D-FT	Color histogram	HLAC
MSM	11.7(3)	6.2(4)	5.8(3)	21.7(5)
CMSM-1	8.4(3)	4.5(4)	4.1(1)	15.4(4)
CMSM-2	7.3(3)	3.9(4)	4.1(1)	17.9(2)
CMSM-5	7.7(2)	4.0(4)	5.1(1)	2.8(1)
CMSM-10	9.7(1)	4.4(4)	6.6(1)	1.6(1)
OMSM	10.8(5)	6.4(4)	8.9(1)	24.4(1)

Table 3. Equal error rate of each classifier (%)

Classifier	S[1]	S[2]	S[3]	S[4]	S[5]
MSM	94.3	90.75	97.9	98.75	98.9
CMSM-1	88.5	96.1	98.7	99.25	99.1
CMSM-2	91.5	98.05	99.1	99.4	99.25
CMSM-5	92.4	98.05	99.15	99.2	99.05
CMSM-10	91.95	97.95	99.1	99.05	98.75
OMSM	94.9	96.35	98.85	99.15	99.35

Table 4. Recognition performance(%). Similarity S[n'] is defined in Sec. 2.2

Classifier	S[1]	S[2]	S[3]	S[4]	S[5]
MSM	45.9	2.8	0.8	0.3	0.4
CMSM-1	4.5	1.1	0.3	0.1	0.1
CMSM-2	3.2	0.7	0.2	0.2	0.2
CMSM-5	2.9	0.6	0.3	0.2	0.3
CMSM-10	3.1	0.6	0.3	0.3	0.4
OMSM	2.1	1.1	0.4	0.3	0.3

Table 5. Equal Error Rate (%)

3.4 Classification of 100 apples using ensemble learning

We evaluated the effectiveness of the ensemble classification using four kinds of features. In this experiment, we used MSM,CMSM and OMSM. The experimental conditions were as described above.

Results and discussion:
Table 4 and Table 5 show the performances of all the methods. In the ensemble classification framework, all the performances were largely improved in comparison with using only a single feature. The best recognition rates were 98.9% by MSM, 99.4% by CMSM and 99.35% by OMSM. Especially the improvements of EER were particularly notable.

In addition, the performances were better as the number n of the canonical angles increases. This tendency is not seen in the experiment using a single feature. This implies that the ensemble classification well derived the effectiveness of using multiple canonical angles.

4 Conclusion

In this paper we have proposed a method to classify 3D objects with similar appearances. We considered the classification of one hundred apples as a concrete task. To tackle this challenging task, we used three types of feature of shape,

texture and color type as input vectors for each of the MSM-based classifiers. The results of classification from all the MSM-based methods were combined in the the ensemble learning framework. The effectiveness of the proposed method was demonstrated through the results of the evaluation experiments using 100 apples. In future works, we will evaluate the performance of our method by using larger data sets to estimate the limitations of classification ability of MSM-based methods.

References

1. Murase, H., Nayar, S. K.: Visual Learning and Recognition of 3-D Objects from Appearance. International Journal of Computer Vision, Vol. 14, pp. 5-24 (1995)
2. Li, Y., Gong, S., Liddell, H.: Video-Based Online Face Recognition Using Identity Surfaces. ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (2001)
3. Satoh, S., Katayama, N.: An Efficient Implementation and Evaluation of Robust Face Sequence Matching. International Conference on Image Analysis and Processing, pp. 266-271 (1999)
4. Fukui, K., Yamaguchi,O.: Face Recognition Using Multi-viewpoint Patterns for Robot Vision. International Symposium of Robotics Research, pp. 192-201 (2003)
5. Kawahara, T., Nishiyama, M., Kozakaya, T., Yamaguchi, O.: Face Recognition based on Whitening Transformation of Distribution of Subspaces. Workshop on ACCV2007, Subspace2007, pp. 97-103 (2007)
6. Maeda, K., Yamaguchi, O., Fukui, K.: Towards 3-Dimensional Pattern Recognition. SSPR2004 & SPR2004, pp. 1061-1068 (2004)
7. Maki, A., Fukui, K.: Ship identification in sequential ISAR imagery. Machine Vision and Applications, Vol. 15, pp. 149-155 (2004)
8. Ichino, M., Sakano, H., Komatsu, N.: Speaker recognition using Kernel Mutual Subspace Method. In Proc. of International Conference on Control, Automation, Robotics and Vision, Vol. 1, pp. 397-402 (2004)
9. Diaz, R., Gil, L., Serrano, C., Blasco, M., Moltó, E., Blasco, J.: Comparison of three algorithms in the classification of table olives by means of computer vision. Journal of Food Engineering, Vol.61, pp. 101-107 (2004)
10. Rocha, A., Hauagge, D. C., Wainer, J., Goldenstein, S.: Automatic produce classification from images using color, texture and appearance cues. In Proc. of the Brazilian Symposium of Computer Graphics and Image Processing (2008)
11. Zheng, Z., Iwata, I., Hirata, Y., Tamura, Y.: Quantitative evaluation of the degree of sprout leaf bending of rice cultivars using P-type Fourier descriptors and principal component analysis. Euphytica, Vol. 163, No. 2, pp. 259-266 (2008)
12. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. CVPR, pp. 886-893 (2005)
13. Otsu, N., Kurita, T.: A New Scheme for Practical Flexible and Intelligent Vision Systems. In Proc. of IAPR Workshop on Computer Vision, pp. 431-435 (1988)
14. Fukunaga, K., Koontz, W. L. G.: Application of the Karhunen-Loève Expansion to Feature Selection and Ordering. IEEE Trans. on Computers, Vol. C-19, No. 4, pp. 311-318 (1970)