Vol.2013-CVIM-187 No.33
2013/5/30

Face Model from Local Features: Image Clustering and Common Local Feature Extraction based on Diverse Density

TAKAYUKI FUKUI[†1] TOSHIKAZU WADA[†1]
HIROSHI OIKE[†1]

Face image retrieval based on local features has advantages of short elapsed time and robustness against the occlusions. However, the keypoint detection, beforehand with the feature description, may fail due to illumination changes. For solving this problem, top-down model-based keypoint detection can be applied, where man-made face model does not fit this task. This report addresses the problem of bottom-up face model construction from example, which can be formalized as common local features extraction among face images. For this purpose, a measure called Diverse Density (DD) can be applied. DD at a point in a feature space represents how the point is close to other positive example while keeping enough distance from negative examples. Because of this property, DD is defined as product of metrics, which can easily be affected by exceptional data, i.e., if one negative data leaps into the neighbour of a positive example, the DD around there becomes lower. Actually, face images have wide variations of face organs' positions, beard, moustache, glasses, and so on. Under these variations, DD for wide varieties of face images will be low at any point in the feature space. For solving this problem, we propose a method performing hierarchical clustering and common local feature extraction simultaneously. In this method, we define a measure representing the affinity of two face image sets, and cluster the face images by iteratively merging the cluster pair having the maximum score. Through experiments on 1021 CAS-PEAL face images, we confirmed that multiple face models are successfully constructed.

1. Introduction

Local image feature based image retrieval has the following advantages over pixel-wise comparison for the similar image search.

1. Short elapsed time.
2. Image alignment free.
3. Robust against occlusion.

However, the bottom-up extraction of local features [1, 2] has a drawback that the keypoint detection can easily be affected by illumination changes as shown in Figure 1. Specifically, the disappearance of keypoints is a fatal problem, because insufficient keypoints make the image comparison unstable and unreliable.

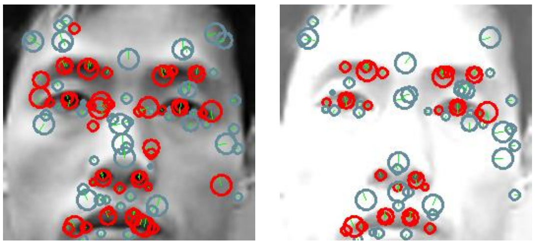


Figure 1: Effect of illumination changes in keypoint detection.

When a face model consisting of local features is given, top-down keypoint detection can be applied by matching the model to the face image. That is, by finding local feature correspondence, the geometric transformation from the model to the image can be estimated, and then, the missing keypoints' location, orientation, and scale in the image can be estimated from the transferred keypoints in the model. Feature descriptor can be applied by using these parameters to restore the missing local features.

For constructing the face model consisting of local features, manual model design does not work, because the parameters of the model have to fit to the real image and the human intervention may destroy the natural arrangement of them. Then, the bottom-up model construction has to be used for the face model. This bottom-up model construction can be formalized as a common local feature extraction problem from local feature instances.

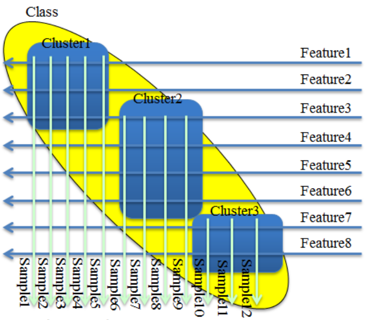
For this purpose, Diverse Density (DD) [3, 4] proposed in the field of Multiple Instance Learning (MIL) [5] can be applied. When we compute positive local features from face images and providing some negative features extracted from non-face images, we can compute DD in the feature space. In this case, DD at certain point in the feature space represents the situation how the point is close to common positive features while keeping enough distance from any negative features. Thus, we can find common local features from face and some non-face images by simply finding the maxima of DD in the feature space.

However, this strategy fails when the face images have wide variations. That is, face organs' positions and shapes have wide variations, some faces have beard, moustache, glasses, and so on. For the local features extracted from these face images, very few maxima of DD, i.e., common local features, are extracted. This phenomenon can also be understood from the fact that DD is defined as a product of terms and only one exceptional feature damages DD around there. One may think that modifying the DD to be robust against the exceptional data can be a solution of this problem. However, this is logically incorrect, because such highly sensitive property to exceptional data is essential for the task of common local feature extraction.

We believe that image clustering is necessary for the common local feature extraction, because the common local feature extraction is a signal-level processing and the coherency of feature should be guaranteed. From this viewpoint, clusters

[†1] Wakayama University

having coherent features are the building block of the semantic class, i.e., “face” as shown in Figure 2.



$$\begin{aligned} Class &= Cluster1 \cup Cluster2 \cup Cluster3 \\ &= (Feature1 \cap Feature2 \cap Feature3) \cup (Feature3 \cap Feature4 \cap Feature5 \cap Feature6) \cup (Feature7 \cap Feature8) \end{aligned}$$

Figure 2: Common local features based clustering.

In this context, the criterion for the clustering should be the coherence of features, which can be represented as the number or the strength of the extracted common local features. In this report, we propose an affinity measure between two image sets based on DD. This affinity measure represents how many or strong common local features are preserved after merging them. Starting from small image sets, each of which consists of a single face image, face images can be clustered just by iteratively merging the two image sets having the maximum affinity. Note that this clustering is conducted so as to preserve maximum common local features in each cluster.

In the following sections, we present related works and our method in Section 2 and 3, respectively. Some experimental results are shown in Section 4.

2. Related Works

Local feature, such as SIFT [1] and SURF [2], are widely used in the field of image correspondence, retrieval, and so on. Bag-of-Features (BoF) [6, 7] is a single vector representation of an image, which is essentially a histogram vector representing the frequency of Visual Words in the image. Nister and Stewenius [8] proposed vocabulary tree for accelerating similar image search. Vocabulary tree is a hierarchical code book, which is obtained by performing hierarchical k-means clustering of local features. The advantage of this approach is BoF representation and similar image search are accelerated. On the other hand, search algorithm without BoF representation has also been investigated. For example, Kise et al. [9] have proposed hashing based high-speed image search algorithm based on local features.

These local features are often regarded as robust against rotation, scale change, slight affine transformation, slight blurring, and illumination change within the dynamic range. However, the keypoint detection, which is the preceding process in local feature extraction, sometimes fails because of excessive highlights or insufficient illumination. When the keypoint detection fails, the following feature description also fails.

For avoiding this problem, Nakamura et al. [10] proposed local feature extraction methods at grid points and randomly sampled points on images. Of course these methods do not require keypoint detection, and hence, these methods are not affected by keypoint detection failure. However, these methods produce redundant local features, which may increase the memory use consumes excessive computational power.

In contrast, we examine a method which extracts disappeared local features by fitting a face model to some extracting local features from an image. Most face models practically used are manually designed, e.g., Deformable Template [11], graphical models and Active Appearance Model [12]. However we mentioned before, bottom-up model construction is necessary for this task. As far as we surveyed, no works have been presented on bottom-up local feature based face model construction.

3. Proposed method

As we discussed, we will define an affinity measure between two image sets based on DD. Face images can be clustered just by iteratively merging two image sets having the maximum affinity. In this section, we first introduce the DD, define the affinity measure, propose clustering procedure, and describe EM-DD based common local feature extraction from clustered image.

3.1 Diverse Density

In the field of MIL [5], common local feature extraction has been regarded as an essential problem, which is formalized as an extremum search problem of a potential in the feature space. This potential is called Diverse Density (DD) [3, 4]. In the rest of this section, we will give a brief introduction of DD.

Bag $\mathcal{B}$: A set of instances. This corresponds to an image in our problem.

Label $+,-$: We assign positive labels to those bags in which common local features are to be found. Also, negative labels are assigned to those bags in which common local feature never exist. These are denoted by $\mathcal{B}_i^+, (\mathrm{i}=1,\dots,\mathrm{m})$ and $\mathcal{B}_i^-, (\mathrm{i}=1,\dots,\mathrm{n})$, respectively.

Instance $\boldsymbol{B}_{ij}^+$, $\boldsymbol{B}_{ij}^-$: An element belonging to a bag. This corresponds to a local feature vector. Positive and negative instances are denoted as $\boldsymbol{B}_{ij}^+ \in \mathcal{B}_i^+$ and $\boldsymbol{B}_{ij}^- \in \mathcal{B}_i^-$, respectively.

First, the following function represents a potential generated by instance $\boldsymbol{B}_{ij}$ at a point $\boldsymbol{x}$ in feature space as shown in Figure 3.

$$P(\boldsymbol{x}=\boldsymbol{t}_j|\mathcal{B}_i) = P(\boldsymbol{x}=\boldsymbol{t}_j|\boldsymbol{t}_j \in \mathcal{B}_i) = \exp\left(-\left\|\boldsymbol{B}_{ij}-\boldsymbol{x}\right\|^2\right). \quad (1)$$

The maximum and the minimum values of this potential are 1 and 0, respectively.

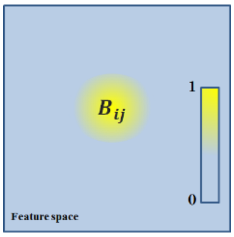


Figure 3: Potential generated by $\boldsymbol{B}_{ij}$

The following function represents the integrated potential $P(\boldsymbol{x}|\mathcal{B}_i^+)$ generated by instances in a positive bag $\mathcal{B}_i^+$.

$$P(\boldsymbol{x}|\mathcal{B}_i^+) = 1 - \prod_{\boldsymbol{t}_j \in \mathcal{B}_i^+} (1 - P(\boldsymbol{x} = \boldsymbol{t}_j|\mathcal{B}_i^+)). \tag{2}$$

Subtraction of an individual potential from 1 can be regarded as the similar meaning to negation and the product can be regarded as logical AND. Under this interpretation, Equation (2) can be regarded as integration by logical OR of the individual potentials in the bag by applying De Morgan's laws.

For negative bags, integrated potential from a negative bag $\mathcal{B}_i^-$ can be defined as follows.

$$P(\boldsymbol{x}|\mathcal{B}_i^-) = \prod_{\boldsymbol{t}_j \in \mathcal{B}_i^-} (1 - P(\boldsymbol{x} = \boldsymbol{t}_j|\mathcal{B}_i^-)). \tag{3}$$

Same as the interpretation of Equation (2), this integration can be regarded as logical NOR.

The potentials generated by positive and negative bags are further integrated by the product. The above-mentioned Diverse Density $DD(\boldsymbol{x})$ is defined as the product of integrated positive and negative potentials.

$$DD(\boldsymbol{x}) = \prod_i^m P(\boldsymbol{x}|\mathcal{B}_i^+) \prod_j^n P(\boldsymbol{x}|\mathcal{B}_j^-). \tag{4}$$

At a maximum of $DD(\boldsymbol{x})$ in the feature space, the point $\boldsymbol{x}$ can be regarded as a common local feature among the positive bags and does not contain similar features in all negative bags as shown in Figure 4.

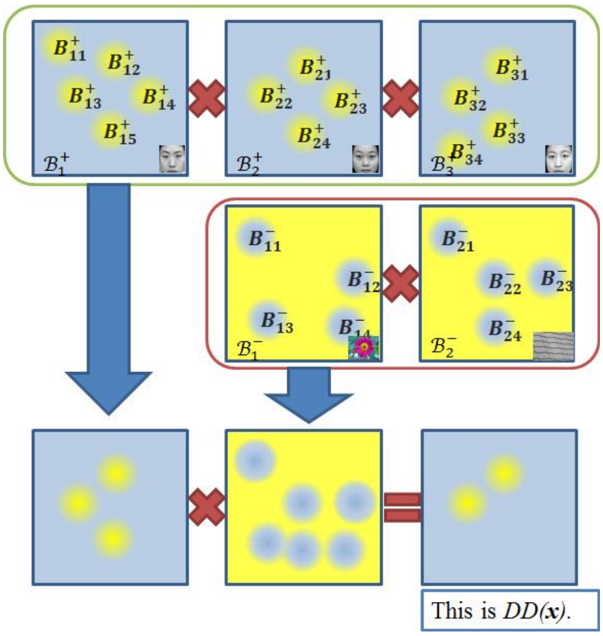


Figure 4: $DD(\boldsymbol{x})$ is defined as the product of positive and negative potentials.

3.2 Affinity between two image sets

In this section, we discuss how to define the affinity measure between two image sets. The measure should represent how many and/or strong common local features are preserved after merging two image sets. According to this principle, the measure $\mathcal{C}$ can easily be defined as Equation (5).

$$\mathcal{C}_{\mathbb{N}}(\mathbb{A}\cup\mathbb{B}) = \int_{\boldsymbol{x}\in\mathcal{F}} DD(\boldsymbol{x})d\boldsymbol{x}, \tag{5}$$

where $\mathbb{A}$ and $\mathbb{B}$ represent positive image sets, $\mathbb{N}$ negative image set, and $\mathcal{F}$ feature space. This affinity $\mathcal{C}_{\mathbb{N}}(\mathbb{A}\cup\mathbb{B})$ is obtained by integrating $DD(\boldsymbol{x})$ over feature space $\mathcal{F}$. However, this computation is practically impossible, because the feature space $\mathcal{F}$ is infinitely vast.

As the second candidate of affinity measure, the sum of the $DD(\boldsymbol{x})$ at all maxima $\mathcal{M}$ in $\mathcal{F}$ can be defined.

$$\mathcal{C}'_{\mathbb{N}}(\mathbb{A}\cup\mathbb{B}) = \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{M}}} DD(\boldsymbol{x}). \tag{6}$$

For this computation, we first have to compute $\mathcal{M}$, set of all maxima in $\mathcal{F}$. A maximum can be approximately be searched by EM-DD algorithm. However, the algorithm may miss some maxima and a single maximum can be found multiple times. Thus, correct $\mathcal{M}$ cannot be estimated practically.

Suppose that approximated $\mathcal{M}'$ is obtained by applying EM-DD starting from all positive features. As mentioned above, the number of $\mathcal{M}'$ is not reliable. Then an affinity measure can be defined as the expected value of $DD(\boldsymbol{x})$, which can be computed sample mean of $DD(\boldsymbol{x})$ by sampling points in $\mathcal{M}'$.

$$\mathcal{C}_{\mathbb{N}}^{\mathcal{M}'}(\mathbb{A}\cup\mathbb{B}) = \frac{1}{|\boldsymbol{\mathcal{M}}'|}\sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{M}}'} DD(\boldsymbol{x}). \tag{7}$$

Even in this approximation, estimating $\mathcal{M}'$, which is done by EM-DD starting from all positive points, is an expensive computation and is not feasible in most cases.

For reducing the computation, Equation (7) can be roughly approximated by Equation (8).

$$\mathcal{C}_{\mathbb{N}}^{\mathcal{S}_{\mathbb{B}}}(\mathbb{A}) = \frac{1}{|\mathcal{S}_{\mathbb{B}}|}\sum_{\boldsymbol{x}\in\mathcal{S}_{\mathbb{B}}} DD(\boldsymbol{x}), \tag{8}$$

where only the image set $\mathbb{A}$ is used as positive image set, and feature points $\mathcal{S}_{\mathbb{B}}$ belonging to image set $\mathbb{B}$ is used as the sampling point where $DD(\boldsymbol{x})$ is computed. Note that this computation does not require EM-DD.

Equation (8) is computationally inexpensive and feasible. The only problem is the asymmetric property as shown in Equation (9) and (10).

$$\mathcal{C}_{\mathbb{N}}^{\mathcal{S}_{\mathbb{B}}}(\mathbb{A}) \neq \mathcal{C}_{\mathbb{N}}^{\mathcal{S}_{\mathbb{A}}}(\mathbb{B}). \tag{9}$$

Especially, for $\mathcal{S}_{\mathbb{B}} \subset \mathcal{S}_{\mathbb{A}}$, the following inequality stands.

$$\mathcal{C}_{\mathbb{N}}^{\mathcal{S}_{\mathbb{B}}}(\mathbb{A}) > \mathcal{C}_{\mathbb{N}}^{\mathcal{S}_{\mathbb{A}}}(\mathbb{B}). \tag{10}$$

For guaranteeing the symmetric property, we employ the affinity measure defined by Equation (11) in this report.

$$\mathcal{C}_{\mathbb{N}}(\mathbb{A},\mathbb{B}) = \frac{1}{2}\left(\mathcal{C}_{\mathbb{N}}^{\mathcal{S}_{\mathbb{B}}}(\mathbb{A}) + \mathcal{C}_{\mathbb{N}}^{\mathcal{S}_{\mathbb{A}}}(\mathbb{B})\right). \tag{11}$$

3.3 Hierarchical clustering by using affinity

Affinity measure defined by Equation (11) can be utilized for hierarchical clustering by the following procedure.

Initialize: Form initial clusters, each of which consists of a single image.

Step1: Merge the cluster pair having maximum affinity measure among all cluster pairs.

Step2: If the number of cluster is greater than one, go to Step 1.

End

This is a greedy algorithm. Figure 5 shows a part of hierarchical clustering.

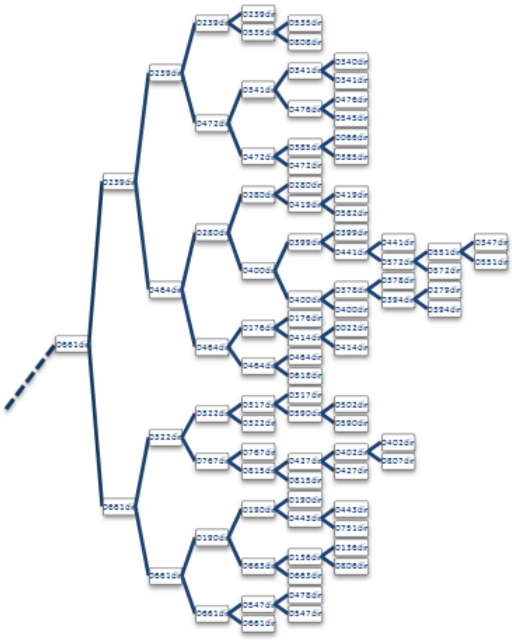


Figure 5: A part of hierarchical clustering: Leaf nodes represent face images. Internal nodes represent face-image clusters.

3.4 Face model construction by EM-DD

EM-DD [13] estimates a local maximum of DD in the feature space by hill climbing iterations, in which an accelerated approximation of DD like EM-algorithm is employed. EM-DD algorithm iteratively finds local maxima in the feature space starting from all positive instances. The DD is defined in Equation (4), but the computation using all positive and negative instances is cumbersome. For avoiding this, EM-DD approximates DD value only by using nearest instances each of which is selected from a bag. Since this selection process is similar with expectation process, and the hill climbing can be regarded as maximization process, this algorithm is called EM-DD.

The local maxima of DD in the feature space found by EM-DD correspond to common local features of the cluster. Based on these features, a face model is constructed from face images.

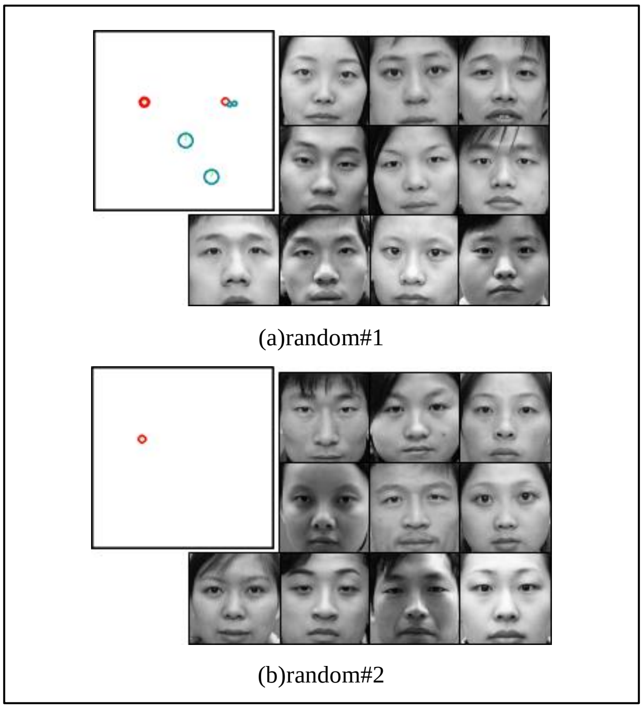
4. Experiments

We conducted experiments on face image clustering, common local feature extractions, and DD distribution comparisons. The face images used in these experiments are 1021 CAS-PEAL face images.

The local feature used in these experiments is 68D vector consisting of 64D SURF features, 2D keypoint location parameters, 1D orientation, and 1D scale. We employed integral SURF [14], mainly for the faster execution speed.

4.1 Common local features extracted from random and clustered images

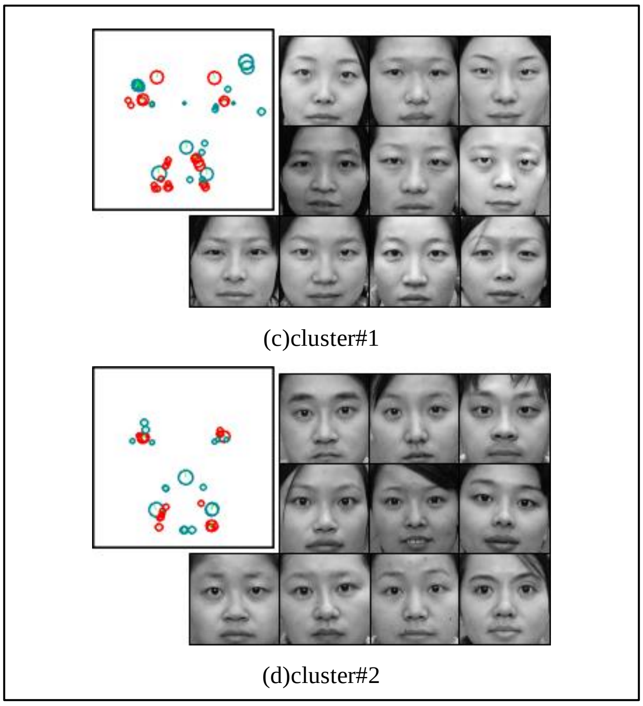
We constructed face models from clusters consisting of 10 face images. Figure 6 shows randomly selected face images and their corresponding common local features projected to x,y positions and scales. Figure 7 shows images clustered by using our method and their corresponding common local features denoted by circles. Each common local feature is extracted by thresholding the $DD(\boldsymbol{x})$ value by a threshold 0.0005.



(a)random#1

(b)random#2

Figure 6: Face models constructed from randomly selected 10 images.



(c)cluster#1

(d)cluster#2

Figure 7: Face models constructed by using our proposed method and image-sets for construction.

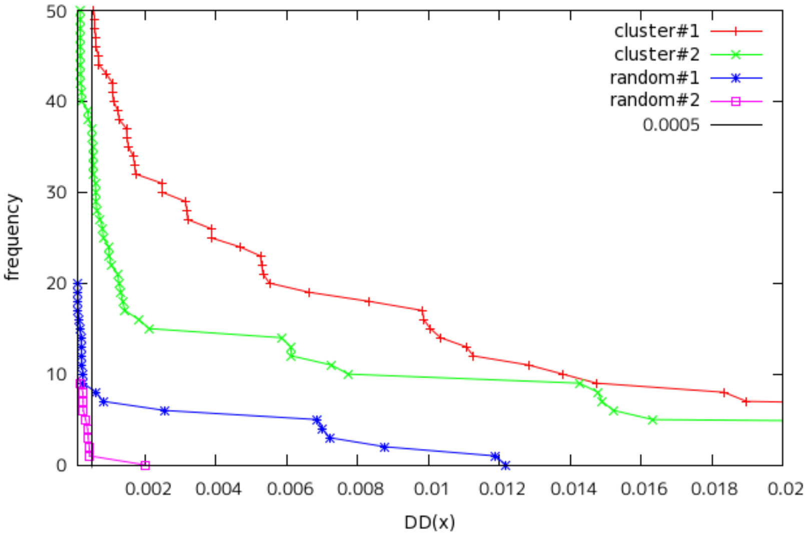


Figure 8: Inverse cumulative histogram of DD for random#1, random#2, cluster#1 and cluster#2:
Horizontal axis represents $DD(\boldsymbol{x})$, Vertical axis represents the frequency of features having smaller DD values than $DD(\boldsymbol{x})$.

From these figures, the numbers of extracted common local features of clustered images in Figure 7 are much bigger than that of randomly selected image sets shown in Figure 6. This result is obvious but surprising, because the common local features extracted by DD is easily be affected by an uncommon face image and our clustering exclude such uncommon images from the cluster.

These results are obtained under a certain threshold. For more quantitative evaluations, inverse cumulative histograms of DD for clusters in Figure 6 and 7 are shown Figure 8.

For incoherent image set, we can extract only few common local features. However, for coherent image clusters, we can extract so many local features from eyebrows, eyes, nose and mouth.

5. Conclusion

In this report, we proposed a method constructing a face model from actual face images. The key idea of this method is applying DD to clustered images, where the clustering is designed to preserve the common local features by a greedy algorithm. Through the experiments, we confirmed that our method can cluster the face images properly, and common local features are extracted by applying EM-DD to the clustered face images.

By using this hierarchical clustering result as a decision tree, we can roughly cluster an unknown face image, and some missing keypoints can be restored in the top-down manner. These should be done in the future works.

Reference

1) D.G. Lowe: Distinctive image features from scale-invariant keypoints, IJCV, Vol. 60, No. 2, pp. 91-110, 2004.
2) H. Bay, T. Tiytelaars, and L. J. Van Gool: SURF: Speeded Up Robust Features, In ECCV, pp. 404-417, 2006.
3) O. Maron and T. Lozano-Perez: *A Framework for Multiple-Instance Learning*, Advances in Neural Information Processing Systems 10, pp570-577, London, England, December 1997.
4) O. Maron and A. Ratan: Multiple-Instance Learning for Natural Scene Classification, Proceedings 15th International Conference on Machine Learning, pp341-349, Madison, Wisconsin, USA, July 1998.
5) T. G. Dietterich, R. H. Lathrop and T. Lozano-Perez: Solving the multiple-instance problem with axis-parallel rectangles, Artificial Intelligence, vol.89, no.1-2, pp31-71, January 1997.
6) J. Sivic and A. Zisserman: Video Google: A text retrieval approach to object matching in videos, In Proc. of ICCV, Vol.2, pp. 1470-1477, Oct 2003.
7) L. Fei-Fei and P. Perona: A Bayesian hierarchical model for learning natural scene categories, In Proc. of CVPR, Vol. 2, pp. 524-531, June 2005.
8) D. Nister and H. Stewenius: Scalable Recognition with a Vocabulary Tree, In Proc. of CVPR, June 2006, Vol.2, pp. 2161-2168, June 2006.
9) K. Kise, M. Iwamura, T. Nakai, K. Noguchi: Large-Scale Image Retrieval by Hashing of Local Features, DBSJ Journal, Vol. 8, No. 1, pp. 119-124, June, 2009. (in Japanese)
10) H. Nakamura, T. Harada, Y. Kuniyoshi: Dense sampling low-level statistics of local features, In Proc. of CIVR'09, Article No. 17, 2009.
11) A.L. Yuille: Deformable Templates for Face Recognition, Journal of Cognitive Neuro Science, Vol. 3, No. 1, pp. 59-70, 1991.
12) T.F. Cootes, G.J. Edwards, C.J. Taylor: Active appearance models, ECCV'98, Lecture Notes in /Computer Science. 1407. pp. 484, 1998.
13) Q. Zhang, S. A. Goldman: EM-DD: An Improved Multiple-Instance Learning Technique, Advances in Neural Information Processing System 14, pp1073-1080, Vancouver, British Columbia, Canada, December 2001.
14) Y. Yoshioka, T. Wada: Parallel Implementation of Image Keypoint Detection and Correspondence on FPGA, The 18th Symposium on Sensing via Image Information, noIS3-03, 2012. (in Japanese)