Efficient Facial Attribute Recognition with A Spatial Codebook

Yoshihisa Ijiri, Shihong Lao
Corporate R&D, Core Technology Center
OMRON Corporation
Kizugawa, Kyoto, Japan
{*yoshihisa_ijiri, shihong_lao*}*@omron.co.jp*

Tony X. Han
ECE Department
University of Missouri
Columbia, MO, USA
hantx@missouri.edu

Hiroshi Murase
Graduate School of Information Science
Nagoya University
Nagoya, Japan
murase@is.nagoya-u.ac.jp

***Abstract*—There is a large number of possible facial attributes such as hairstyle, with/without glasses, with/without mustache, etc. Considering large number of facial attributes and their combinations, it is difficult to build attributes classifiers for all possible combinations needed in various applications, especially at the designing stage. To tackle this important and challenging problem, we propose a novel efficient facial attributes recognition algorithm using a learned spatial codebook. The Maximum Entropy and Maximum Orthogonality (MEMO) criterion is followed to learn the spatial codebook. With a spatial codebook constructed at the designing stage, attribute classifiers can be trained on demand with a small number of exemplars with high accuracy on the testing data. Meanwhile, up to 600 times speedup is achieved in the on-demand training process, compared to current state-of-the-art method. The effectiveness of the proposed method is supported by convincing experimental results.**

***Keywords*-face; attribute; recognition; spatial codebook;**

I. Introduction

The number of surveillance cameras in operation is steadily growing. We have so many image data to be inspected, while human labors are far from meeting the demands. Therefore, computer aided surveillance technologies are highly desired. Since the major targets in surveillance are people, many face recognition algorithms have been proposed. However, current face recognition algorithms do not work well when face region is unclear or the resolution is low. For these scenarios, recognizing facial attributes such as hair style, hair colors, with/without glasses etc., is a more effective approach to search rough human identities.

However, facial attributes recognition, as illustrated in Fig.1, requires that attributes of interests can vary according to different scenarios. For example, sometimes we want to find old men with glasses, while young blond women may be of interest under different scenario. Considering the large number of possible attributes and their combinations, it is difficult for facial attributes recognition algorithm to deal with all the attributes at designing stage. Therefore learning new attributes on demand is a necessary function of attributes recognition algorithms. To learn the attributes on demand, problems are how many exemplars are needed and how long the training procedure takes.

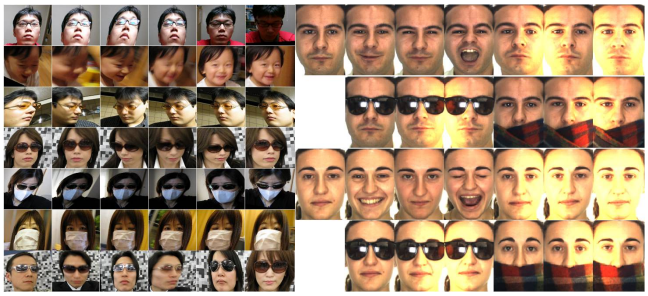

Figure 1. Examples of facial attributes from OMRON DB(left) and AR DB(right)

In this paper, to address the problems, we propose a novel facial attributes recognition framework which utilizes a "spatial codebook" learned at the designing stage. We notice that many facial attributes can be seen locally and expressed as combinations of relatively small number of local patterns, such as black, dark, or normal eye region, etc. If we can extract these local semantic patterns, diverse attributes can be decomposed into relatively simple patterns. Based on this observation, in the proposed framework, a set of representative local patterns, i.e. the spatial codebook, which composes facial attributes, is learned in advance[1]. Since all the computation for the spatial codebook to find semantic patterns is done in advance, when learning new attributes on demand, we only need to compute the simple distance based features for a few exemplars and train the attributes classifier. Therefore, in contrast to existing approaches which do feature selection and attributes classifier training at once, in our framework the on-demand training is efficient. A flowchart of the on-demand learning in comparison with current state-of-the art method is shown in Fig.2.

There have been several enlightening works on facial attributes recognition. Moghaddam[1], Zhuang[2] and Hosoi[3] proposed individual attribute recognition such as gender, age and ethnicity. Lyon[4] proposed simultaneous gender, ethnicity estimation. Although there are some other

[1]Local patterns which compose the spatial codebook are referred to spatial codewords subsequently. This process can be seen as effective feature selection in comparison with existing approaches.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.361

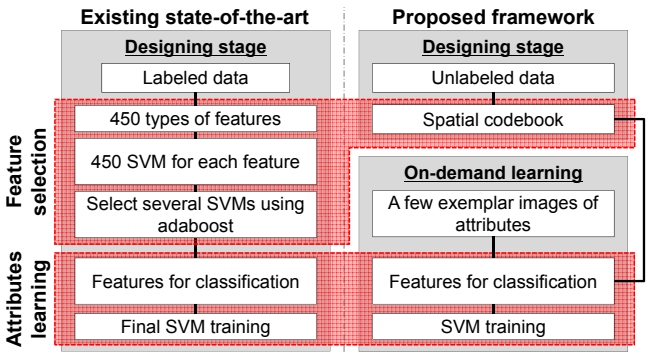

Figure 2. The proposed framework

similar works, common points in these early works are that they are specified to fixed attributes, and thus cannot recognize on-demand attributes.

Recently Kumar[5] proposed a state-of-the-art framework called Face Tracer to recognize diverse attributes. In the framework, 450 types of features are extracted at 10 local regions to build corresponding 450 Support Vector Machines (SVMs). Then several appropriate SVMs are selected by Adaboost[2] and the outputs of the selected SVMs are fused further by upper level SVM. Although extracting 450 types of redundant features and training corresponding SVMs are computationally expensive, all these procedures must be done when learning each attribute, as in Fig.2. Thus it is very difficult for their algorithm to learn new attributes instantly. Additionally in the Face Tracer, since pixel level features are usually with dimensions ranging from hundreds to thousands, large number of exemplars is required to prevent over-fitting. While in the proposed framework, the dimension of features corresponding to the number of the spatial codewords is the order of 10 to 100. Accordingly, the number of exemplars can be reduced by 10 times in our experiments to obtain similar accuracy of the Face Tracer. This fact means that the proposed framework can also be sped up, which make it easy to handle new attributes. These features are shown in experiments.

To summarize, our contribution is to propose a new framework that can learn diverse facial attributes on demand with a small number of exemplars.

II. Proposed Framework

As shown in Fig.2, using large number of unlabeled data $X = \{x_n; n = 1, \ldots, N\}$, the spatial codebook S, which is comprised of a set of spatial codewords $\{s_t; t = 1, \ldots, T\}$ at sub-region $\{l_t; t = 1, \ldots, T\}$, is learned in advance. When learning new attributes on demand, simple distances from training data to the spatial codewords are used as features $\{f_t; t = 1, \ldots, T\}$ to build SVMs $H = \{h_i; i = 1, \ldots, I\}$ corresponding to each attributes ω_i. In recognition step, the same feature extraction method is applied to input data to get an estimate $\hat{\omega}$ from SVMs.

[2]This part can be regarded as feature selection step.

Algorithm 1 Spatial codebook learning

Require: unlabeled training data X; number of spatial codewords T;

Ensure: a spatial codebook S by MEMO criterion

1: initialize by $S = \{\emptyset\}$
2: extract L number of sub-regions.
3: **for** each local sub-region $l = 1, \ldots, L$ **do**
4: do K-means in each sub-region l and obtain K number of cluster centers $\{C_{lk}; k = 1, \ldots, K\}$.
5: **end for**
6: compute membership functions $\phi_{lk}(x_i)$ against each training data $x_i \in X$.
7: **for** $t = 1, \ldots, T$ **do**
8: select a cluster C_{lk} based on MEMO, that is,
$$(\hat{l}_t, \hat{k}_t) = \arg\max_{\{l,k\}} \left\{ \frac{1}{||C_{lk}^T S||} + \alpha \left(-\sum_{x_i \in X} \mathrm{P}(\phi_{lk}) \log(\mathrm{P}(\phi_{lk})) \right) \right\}$$
$s_t = C_{\hat{l}_t \hat{k}_t}$, $S = S \cup s_t$, where $\alpha \geq 0$ is an appropriate balance parameter. When $t = 1$, use only second term (maximum entropy).
9: **end for**

Figure 3. An example of spatial codebook

A. Spatial codebook

We assume that human region is localized accurately by face or human body detector here. To represent diverse facial attributes, local patterns which comprise these attributes are extracted. The local patterns can be obtained by dividing input images into local sub-regions and applying K-means for each region. Let the number of sub-regions and clusters be L and K, respectively. We have KL clusters $\{C_{lk}\}(k = 1, \ldots, K; l = 1, \ldots, L)$ now. Using all obtained clusters, we have many redundant clusters. For an instance, if we have a cluster looks like sunglasses in a sub-region around left eye and another sunglasses-like cluster in a sub-region around right eye, data samples from the two clusters are very similar. Thus using all KL clusters are redundant and these redundant clusters should be removed to make the remaining set of clusters capable of expressing diverse attributes efficiently. For building this set, we select several clusters out of KL so that cluster memberships of the selected clusters become mutually orthogonal as much as

possible. On the other hand, clusters which include almost all data or almost no data is not efficient to represent attributes, since these clusters are generally corresponding to trends of the data or noises, rather than some specific attributes[6]. This property can be expressed by entropy of the membership function. Let a membership function to cluster C_{lk} for each data x_n be $\phi_{lk}(x_n)$. We define $\phi_{lk}(x_n) = 0$ in the case $x_n \notin C_{lk}$, otherwise $\phi_{lk}(x_n) = 1$, where $l = 1, \ldots, L$, $k = 1, \ldots, K$, $n = 1, \ldots, N$ are indices for local sub-regions, K number of clusters at each sub-region and unlabeled training data, respectively. By this function, clusters with maximum entropy in ϕ_{lk} and with maximum orthogonality against already selected clusters sequentially. Algorithm.1 shows detailed steps. In practice, during the selection it becomes difficult to select perfectly orthogonal ϕ_{lk} against already selected spatial codebook S, then we select a cluster by minimum inner product. Combination of Maximum Entropy and Maximum Orthogonality is called MEMO criterion for short. Fig.3 shows an example of resulting codebook. Rectangles surrounded by a bold rectangle in each face image shows locations of spatial codewords. We can see that these are corresponding to some specific attribute to some extent.

B. Extracting features with spatial codebook

Typical methods using codebooks do vector quantization based on the minimal distance between input patterns and codewords. In this case, for an instance, input patterns located in the middle of two codewords should be assigned to either of the codewords, thus the information the pattern is similar with both of the patterns is lost. The problem is called "hard assignment" problem[7]. To tackle this the problem, using distance between input patterns and codewords[7] to obtain higher accuracy than conventional vector quantization[8] is proposed. In this paper, we follow this scheme and extract features for classification f for each input image ξ as $f_t = \mathrm{dist}(s_t, \xi(\hat{l}_t))$, where $t = 1, \ldots, T$, $f \in \mathcal{R}^T$. Fig.4 shows how the feature extraction is done for each input image ξ, in which horizontal axis indicated index t for feature, and vertical axis indicates extracted feature f_t. Images shown below the horizontal axis are the spatial codewords giving shorter distances. The results show that the codewords with shorter distances are related to the similar patterns with input patterns.

C. Learning attributes on demand

When learning new attributes on demand, a small number of exemplars are used. Feature extraction is done by the way described in section.II-B to train classifier H_i for each added attributes ω_i. In this paper, we trained SVMs with "1 vs. All" scheme.

D. Recognizing attributes

During the recognition of the attributes, features are extracted in the same way as in learning step to get features for classification f. And then learned SVMs H_i are used to obtain an estimate as $\hat{\omega} = \arg\max_{\omega_i} H_i(f; \omega_i)$.

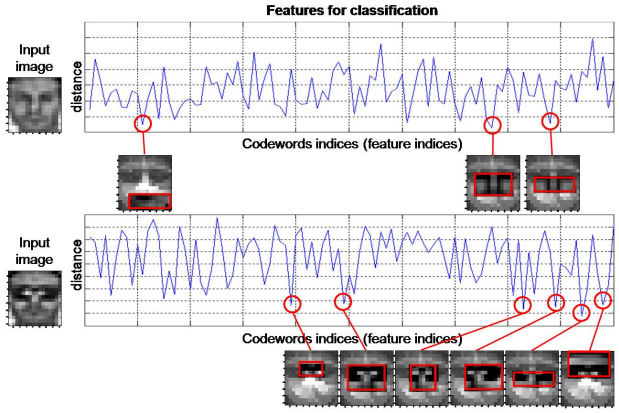

Figure 4. Extraction of features with spatial codebook

III. Experiments

To show the effectiveness of the proposed framework, experimental studies are shown in this chapter. To show training can be done with small number of exemplars, changing the number of exemplars, comparison with existing method is done. And then the training time is compared with current state-of-the-art method to show the efficiency of our framework. In the experimental study, two different datasets are used, AR Face dataset[9] and OMRON dataset. Though AR DB is originally created for face recognition, it consists of faces with sunglasses, scarf and extreme facial expressions, etc. In our experiments it is used for attributes recognition. In AR DB, faces are aligned manually, while in OMRON DB no ground-truth data is available then automatic face detection and facial parts localization are exploited. Furthermore, since OMRON DB is based on uncontrolled environment, the results can be regarded as performance of the real application. Table .I shows the detail of the experimental setup, where each value means how many samples are used. Facial images are rescaled to 64×75[pix] (28[pix] between eyes) using the perspective transform. All images are transformed to gray scale images. Sample images can be seen in Fig. 1. In AR DB, classification of faces with no special attribute, sunglasses, scarf, and scream was conducted, while in OMRON DB, normal faces, sunglasses, masks, and both sunglasses and masks were classified[3].

A. Accuracy against numbers of exemplars

By changing the number of exemplars for learning new attributes, the accuracy of the Face Tracer and the proposed framework is compared. When the number of exemplars is

[3]In implementation of the Face Tracer, we used only gray scale based features instead of gray scale, RGB, HSV, edge orientation and magnitude to do fair comparison. The number of feature types is reduced to 90 types from 450 types of original.

Table I
THE NUMBER OF TRAINING IMAGES

Dataset	Codebook construction	Testing
AR	2600	2600
OMRON	2032	9068

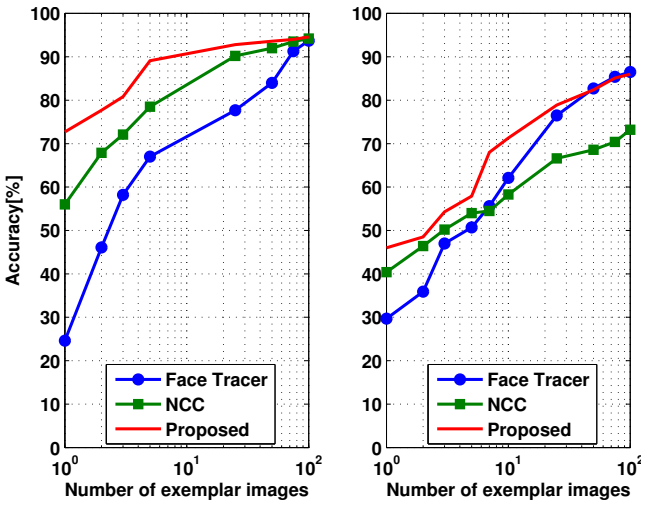

Figure 5. Accuracy vs. number of exemplars; AR DB(left) and OMRON DB(right)

small, normalized cross correlation (NCC) seems the most intuitive. Thus comparison with NCC is also conducted. The results are shown in Fig. 5 and Table II. Especially when the number of exemplars is small, the proposed method is performing nicely in both datasets. This fact indicates that on-demand learning can be done with smaller number of exemplars in the proposed method.

B. Run time for learning

In this experiment 100 exemplars are used for learning new attributes for both methods. [4] The result is shown in Tab. III. We can see that the proposed framework reduces the training time a lot. It is because only simple distances to learned spatial codewords in advance are used to obtain a final classifier, instead of doing redundant feature extraction and training of SVMs. This speedup especially important to attributes based human search system. For instance, if we have to wait for several hours at each time we give a query attribute, the system is of little value.

IV. CONCLUSION

We have presented a novel framework of using the spatial codebook for diverse facial attributes recognition. The new framework is effective especially when new attributes are to be learned on demand. Compared to traditional methods, the learning process on demand was sped up dramatically, and needed exemplars are reduced. All these aspects are supported by the experiments. Although the spatial codebook based pixel level representation was used in this paper, using better representation for handling illumination variations or other perturbations, the accuracy might be improved, which is a future work.

[4] A PC with Core2Duo 3GHz CPU, 2GByte memory and MATLAB were used.

Table II
ACCURACY VS. NUMBER OF EXEMPLARS

Dataset	# of samples	1	5	10	50
AR DB	FaceTracer	24.6	67.0	80.4	84.0
	NCC	56.0	78.5	81.9	92.0
	Proposed	72.7	89.1	91.0	93.6
OMRON	FaceTracer	29.7	50.7	62.1	82.7
	NCC	40.4	54.0	58.3	68.6
	Proposed	46.0	57.9	71.3	82.4

Table III
RUN TIME FOR LEARNING NEW ATTRIBUTES

Method	Run time [sec]
Face Tracer	6060 approximately
Proposed	11.5

REFERENCES

[1] B. Moghaddam and M.-H. Yang, "Learning gender with support faces," *IEEE Trans.on PAMI*, vol. 24, no. 5, pp. 707–711, 2002.

[2] X. Zhuang, X. Zhou, M. Johnson, and T. Huang, "Face age estimation using patch-based hidden markov model supervectors," in *Proc. of ICPR*, 2008, pp. 1–4.

[3] S. Hosoi, E. Takikawa, and M. Kawade, "Ethnicity estimation with facial images," in *Proc. of FGR*, 2004, p. 195.

[4] M. J. Lyons, J. Budynek, A. Plantey, and S. Akamatsu, "Classifying facial attributes using a 2-d gabor wavelet and discriminant analysis," *Proc. of FGR*, p. 202, 2000.

[5] N. Kumar, P. Belhumeur, and S. Nayar, "FaceTracer: A Search Engine for Large Collections of Images with Faces," in *Proc. of ECCV*, 2008, pp. 340–353.

[6] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. of ICCV*, 2005, pp. 604–610.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. of CVPR*, 2008, pp. 1–8.

[8] F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. of CVPR*, 2005, pp. 524–531.

[9] A. M. Martinez and R. Benavente, "The AR Face Database," *CVC Tech. Report*, p. 202, June 1998.