Face Photo Retrieval by Sketch Example

Hamed Kiani Galoogahi
School of Computing
National University of Singapore
Singapore, 117417
hkiani@comp.nus.edu.sg

Terence Sim
School of Computing
National University of Singapore
Singapore, 117417
tsim@comp.nus.edu.sg

ABSTRACT

Face photo-sketch matching has received great attention in recent years due to its vital role in law enforcement. The major challenge of matching face photo and sketch is difference of visual characteristics between face photo and sketch which is referred as modality gap. Earlier approaches have reduced the modality gap by synthesizing face photos and sketches in a same modality (photo or sketch). However, the effectiveness of these approaches is highly affected by synthesis results. That means a poor synthesis might degrade the performance of matching. Therefore, recent works have focused to directly match face photo and sketch of different modalities. However, the features used by these approaches are not robust against modality gap. In this paper, a modality-invariant face descriptor called Gabor Shape is proposed to retrieve face photos based on a probe sketch. Experiments on CUFS and CUFSF datasets show that the new descriptor outperforms the state-of-the-art approaches.

Categories and Subject Descriptors

I.4.7 [**IMAGE PROCESSING AND COMPUTER VISION**]: Feature Measurement – *Feature representation;* I.5.4 [**PATTERN RECOGNITION**]: Application – *Computer vision*

General Terms

Algorithms

Keywords

Photo Retrieval, Face Sketch, Gabor Filters, Radon Transform.

1. INTRODUCTION

Most of current face recognition systems focus on photo-based face identification; which are not practical when a probe face photo is not available [7]. For example, in many cases, the only available information is the recollection of eye-witness which can be used by forensic artist to draw a sketch of suspect's face. This sketch is very useful to automatically retrieve or narrow down face photos of potential suspects from police mug-shot databases.

The major challenge of sketch based photo retrieval is matching images of different modalities which referred as "modality gap" [3,16]. Basically, a face photo is captured by a digital camera, while a face sketch is drawn by an artist. Even for same human subject, the face photo and its sketch might be different. The face shape might be exaggerated by artist or facial texture might be lost or replaced by artistic rendering in face sketch. This problem will be more exacerbated for forensic investigations, when the eye-witness cannot exactly recollect the suspect's face.

To date, various works have addressed the problem of face photo-sketch matching; which can be categorized into two classes: intra-modality [7,10,11,14] and inter-modality approaches [3,4,16]. Intra-modality approaches synthesize pseudo photo (sketch) from input sketch (photo) for matching sketches and photos in a same modality (photo or sketch). Consequently, the performance of these methods is highly dependent on the effectiveness of image synthesis, which might be even harder than retrieval problem [16]. On the other hand, inter-modality approaches directly match face photos and sketches using discriminative features which are integrated with advanced classifiers [16,5]. However, most of these approaches use some features which are not originally designed to tackle with modality gap [16]. Therefore, a modality-invariant feature is eagerly needed to particularly deal with the presence of modality gap between face photo and sketch in photo-sketch matching systems.

The contributions of this paper can be summarized in two aspects: (1) We clearly explore the concept of modality gap and (2) A new modality invariant descriptor called Gabor Shape is proposed for face photo-sketch matching.

2. MOTIVATION

The modality gap is caused by the difference of visual cues which can be derived from face sketch and photo [3]. Generally, the visual cues of face image can be perceived from two different types of facial textures: fine and coarse textures. The fine texture (appearance) includes superficial and low contrast details of face skin such as flaws, moles, wrinkles as well as any shadow/reflection of lighting. Contrary, the coarse texture (shape) consists of the boundaries of main facial components with high contrast such as eyes, eyebrows, nose, lips, chin and ears which form the shape of face [3]. Thus, the amount of modality gap caused by coarse and fine textures can be explored separately.

The coarse texture is necessary for artists to draw the principles of face sketch. Moreover, it makes the facial components in face photo to be distinct from face skin. Therefore, the visual cues of coarse texture are firmly present in both face sketch and photo. That means the modality gap is not significantly affected by coarse texture [3]. In contrast, the fine texture of face photo might be lost or even replaced by artistic rendering in corresponding face sketch. Consequently, the visual cues derived from the fine texture of face sketches are much different from those of face photos. That causes a high amount of modality gap.

From the above discussion, it can be concluded that the modality gap is primarily caused by fine texture compared with coarse

MM'12, October 29-November 2, 2012, Nara, Japan.

texture. Therefore, coarse texture can be used to extract modality-invariant features from face photo and sketch. One solution to extract feature based on coarse texture is to use edge-preserving smoothing operators, e.g. Bilateral filters, to decompose images into coarse and fine layers [2]. Then, the coarse layer corresponded to coarse texture can be used for feature extraction. However, there is not a clear definition of layer's scales to separate fine and coarse layers precisely. Moreover, there are some range and domain parameters which have to be tuned precisely to control image decomposition [2]. Another possibility is to use both fine and coarse textures, but emphasizing coarse texture much more than fine texture in feature extraction stage. This is performed by the new descriptor Gabor Shape (GS) in two steps: First, the fine (coarse) texture is greatly attenuated (emphasized) using Gabor filters. Then the visual cues of face shape (coarse texture) are modeled by Radon transform as face descriptor. Indeed, by these two steps the fine texture is indirectly eliminated in feature extraction. The detail of Gabor Shape descriptor is presented by the following section.

3. PROPOSED METHOD

The overall framework of Gabor Shape is illustrated in Figure 1 (a). Given a face image, Gabor Shape is modeled by the following steps: (1) A bank of Gabor filters applied on the face image to compute the Gabor magnitudes; (2) Each Gabor magnitude is further divided into non-overlapping patches; (4) Each local patch is represented by a set of histograms which are obtained by Radon sampling; (4) The Radon histograms of all local patches are concatenated to form the Gabor Shape representation.

3.1 Gabor Filters

Gabor filters have been extensively used for facial feature extraction [15,17], due to its excellent capacities to capture salient visual properties such as spatial localization, orientation selectivity and spatial frequency [9,15]. The Gabor filters with orientation μ and scale ν are defined as [15]:

$$\psi_{\mu,\nu}(z) = \frac{\|k_{\mu,\nu}\|}{\sigma^2} e^{\left(-\frac{\|k_{\mu,\nu}\|^2 \|z\|^2}{2\sigma^2}\right)} \left[e^{ik_{\mu,\nu}z} - e^{\frac{\sigma^2}{2}}\right] \tag{1}$$

where $z = (x, y)$ denotes the pixel and $k_{\mu,\nu} = k_\nu e^{i\phi_\mu}$ is the wave vector with $k_\nu = k_{max}/\lambda^\nu$ and $\phi_\mu = \pi\mu/8$. k_{max} is the maximum frequency, and λ is the spacing factor between filters in frequency domain. The convolution of a face image $f(x, y)$ and Gabor filter $\psi_{\mu,\nu}(z)$ is calculated as:

$$G_{\psi f}(z, \mu, \nu) = f(x, y) * \psi_{\mu,\nu}(z) \tag{2}$$

where $*$ indicates convolution operator. The complex response $G_{\psi f}(z, \mu, \nu)$ can be represented as $G_{\psi f}(z, \mu, \nu) = M_{\mu,\nu}(z).\mathrm{e}^{\mathrm{i}\theta_{\mu\nu}(\mathrm{z})}$ where $M_{\mu,\nu}(z)$ and $\theta_{\mu,\nu}(z)$ denote the magnitude and phase, respectively. Due to highly sensitivity of phase to spatial variation, only its magnitude is used for feature extraction [15].

3.2 Radon Transform

Radon transform computes projections of image intensity along tracing lines. Each line is characterized by its distance to the origin of axes s and rotation angle from reference axes $\theta \in [0, \pi]$. The projection of a given image $f: (x, y) \to g$ in gray level along straight line $l(s, \theta)$ is computed by:

$$\mathcal{R}_{\theta,s}[f(x, y)] = \int_l f(x, y)dl \tag{3}$$

where all points on the line l satisfy the Equation (4):

$$x\,sin(\theta) - y\,cos(\theta) - s = 0 \tag{4}$$

therefore, the Equation (3) can be rewritten as:

$$\mathcal{R}_{\theta,s}[f(x, y)] = \iint f(x, y)\delta(x \sin\theta - y \cos\theta - s)dxdy \tag{5}$$

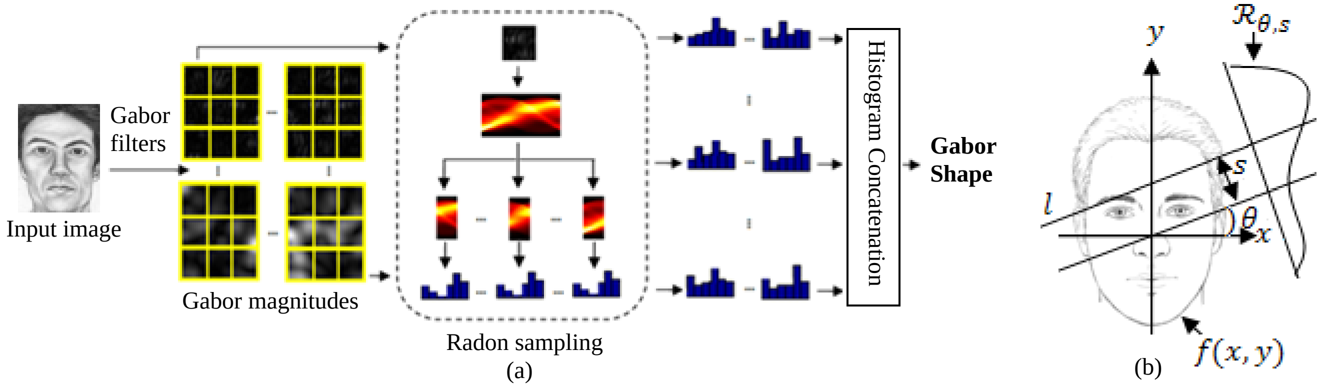
where $\delta(.)$ is the Dirac function. The Radon transform of the image is determined by a set of projections of the image along lines with different orientations and distances. The Radon transform is illustrated in Figure 1 (b).

3.3 Gabor Shape Representation

In order to capture the shape characteristics of Gabor magnitude $M_{\mu,\nu}(z)$, we locally encode the shape of $M_{\mu,\nu}(z)$ by Radon transform. First, each Gabor magnitude $M_{\mu,\nu}(z)$ is divided into $N(n_{hor} \times n_{ver})$ non-overlapping patches $\Omega = \{\omega^i_{\mu,\nu}(z) | \mu \in \{0, \ldots, 7\}, \nu \in \{0, \ldots, 4\}, 1 \leq i \leq N\}$, where $\omega^i_{\mu,\nu}(z)$ is the i^{th} local patch of magnitude $M_{\mu,\nu}(z)$. Then, each patch $\omega^i_{\mu,\nu}(z) \in \Omega$ is processed by Radon sampling as follows: (1) Each local patch $\omega^i_{\mu,\nu}(z)$ is transformed into Radon space by $\mathcal{R}_{\theta,s}[\omega^i_{\mu,\nu}(z)]$; (2) Each $\mathcal{R}_{\theta,s}[\omega^i_{\mu,\nu}(z)]$ of size $s \times \theta$ is divided into $n_{hor} \times n_{ver}$ non-overlapping local regions as Radon samples; (3) A histogram with b bins is calculated for each Radon sample to summarize the Radon values; (4) All histograms extracted from the Radon samples of all the Gabor magnitudes are concatenated into a single histogram as Gabor Shape of the given face image.

4. EXPERIMENTS

Two experiments are conducted on CUFSF [16] and CUFS [14] datasets to demonstrate the effectiveness of Gabor Shape for face photo-sketch matching. All 606 photos in CUFS are taken under a



(a)

(b)

Figure 1. (a) The framework of Gabor Shape face representation, (b) Radon transform of an image.

normal illumination condition. For each photo, there is a sketch without shape distortion which is drawn by an artist when viewing its corresponding photo. The CUFSF consists of 1194 photo/sketch pairs with lighting variations (for photos) and shape exaggeration (for sketches). All face images are translated, rotated and scaled such that the centers of eyes and mouth are at fixed positions. All images are then cropped to 160x128 pixels. Some examples of face photo/sketch pairs are shown in Figure 2.

In experiment 1, we compare Gabor Shape with popular face descriptors including LBP [1], SIFT [8] and LGBPHS [15] on CUFSF. The performance is evaluated by Receiving Operator Characteristic (ROC) curves as Verification Rate (VR) versus False Acceptance Rate (FAR). For all descriptors, the input image is divided into 7*5 non-overlapping local patches. The SIFT is obtained by concatenating 128-dimentional SIFT (4x4 spatial bins and 8 orientation bins over $[0-2\pi]$) of all the local patches [8]. The LBP descriptor is computed by concatenating LBP histograms of all the local regions. The LBP is parameterized by *r=8* (radius) and *p=8* (sampling points) [1]. For LGBPHS, we use 40 Gabor filters in 5 scales and 8 orientations. Each Gabor magnitude is encoded by LBP into local Gabor binary pattern (LGBP), which is further divided into local patches. The LGBPHS is formed by concatenation histogram sequences of LGBP local patches of all Gabor magnitudes [15].

The Gabor filters used for Gabor Shape is same as LGBPHS. A set of experiments with different configuration of parameters are conducted to explore the effect of varying free parameters on the performance of Gabor Shape, including the number of local patches for Gabor magnitude $n_{hor} \times n_{ver}$ (20×16, 10×8, 7×5, 5×4) and Radon sampling $m_{hor} \times m_{ver}$ (1×4, 2×4, 3×4, 1×6, 2×6, 3×6). The histogram bins b is selected as 8. The results show that the performance of Gabor Shape is not dominated by the setting of parameters. Eventually, these parameters are selected as 7×5 and 2×6 for $n_{hor} \times n_{ver}$ and $m_{hor} \times m_{ver}$, respectively. The dissimilarity between descriptors of photos and sketches is computed by Chi-square distance. The best match for a probe sketch is a photo with minimum dissimilarity.

The results of experiment 1 are illustrated in Figure 3. Obviously, Gabor Shape outperforms LBP, SIFT and LGBPHS. The worst performance is obtained by LBP (30.43%), due to the presence of large modality gap. Achieving higher performance by LGBPHS (45.10%) states that the modality gap is reduced by LGBPHS. Since, Gabor magnitudes which used by LGBPHS emphasize coarse texture which is not involved into modality gap. SIFT exploits both fine and coarse textures in feature extraction. Thus, its performance (37.41%) is less than our descriptor (60.48%).

The experiment 2 is designed to compare our method with the following state-of-the-art approaches on CUFSF and CUFS. The parameters of these approaches are tuned according to their papers. The training and testing set on the SUFSF dataset are generated by randomly selecting 500 and 694 subjects, respectively. On the CUFS dataset, we selected 306 subjects as training set; and the remaining 300 subjects form the testing set.

MRF+RS-LDA [14]: First, a MRF-based photo synthesis is trained to synthesize pseudo photos from input sketches. Then, a face recognition based on RS-LDA [13] is used to match pseudo photos against a gallery of face photos with known identities.

Kernel CSR [6]: A common discriminative subspace is learnt to directly match face photos and sketches in different modalities. The CSR model is separately trained for LBP and SIFT.

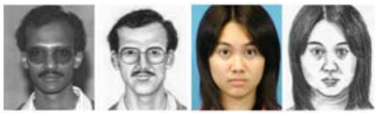


Figure 2. Examples of face photo/sketch pairs, (left) CUFSF and (right) CUFS dataset.

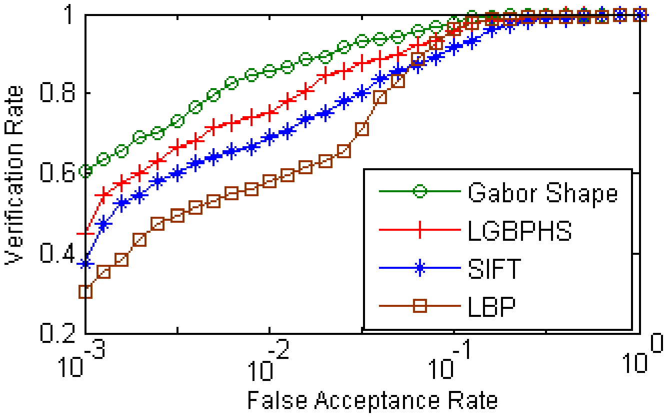


Figure 3. ROC curves of experiment 1: Comparing Gabor Shape with LGBPHS, SIFT and LBP on CUFSF dataset.

LFDA [5]: A discriminate projection is learnt by fusing different LBP and SIFT descriptors in a single feature vector. LBP is computed with four radii r = {1,3,5,7} and 8 sampling points. SIFT is computed from local patches with size 32 [5].

CITP [16]: CITP is characterized by five trees in CITP forest and 256 nodes for each tree. The pattern sampling is performed by a single ring with radius 2. A PCA-LDA classifier [12] with dimension of 600 is applied for feature reduction.

In this experiment, PCA-LDA [12] is applied to reduce the dimension of Gabor Shape. Given a projection matrix which is learnt by a training set, all Gabor Shape descriptors of testing images are first projected into the space with lower dimension. Then, the dissimilarity between projected features is computed by Chi-square distance. In order to choose the appropriate PCA-LDA dimension, we evaluate the new descriptor by different PCA-LDA dimensions {200, 400, …,999}, as shown in Figure 4. The results suggest that the VR at 0.1% FAR slightly changes from 800 to 999 (about 1%). Thus, we choose 800 as PCA-LDA dimension.

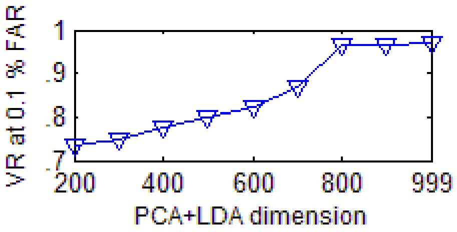


Figure 4. VR at 0.1% FAR vs. PCA-LDA dimension.

The results of experiment 2 on CUFSF are illustrated in Figure 5 and Table 1. Figure 5 shows that the best verification rate is achieved by Gabor Shape (96.32%) followed by CITP (93.95%). The lowest performance is obtained by MRF+RS-LDA (29.54%). Since, MRF-based photo synthesis is affected by the shape distortion. The verification rates of kernel CSR (for both LBP and SIFT) are much better than MRF+RS-LDA. That means the performance can be improved by learning appropriate common subspaces which are more robust against modality gap. Although the problem of modality gap is not directly addressed by LFDA, a promising result (90.78%) is achieved by fusing several features with different spatial partitions. Moreover, Gabor Shape obtains higher verification rate (96.32%) than CITP forest (93.95%). Since, a single CITP forest based on one sampling pattern is not able to capture all rich information of face images [16]. In order to boost the performance of CITP forest, a linear SVM is trained to fuse dissimilarities by different CITP forests of five different sampling patterns. The result in Table 1 shows that fusing five sampling patterns in CITP improves the verification rate at 0.1% FAR from 93.95% to 98.70%. Moreover, in order to fairly compare the fused CITP with Gabor Shape, we evaluate fused Gabor Shape. The fusion of Gabor Shape is performed by averaging the dissimilarities which are obtained from different Gabor Shape with different values of $n_{hor} \times n_{ver}$ (10×8, 7×5, 5×4). As shown in Table 1, the verification rate achieved by fused Gabor Shape (99.14%) is higher than fused CITP (98.70%).

Table 2 shows the Rank-1 recognition rates of experiment 2 on CUFS. According to Table 2, our method outperforms the other approaches. The recognition rate of MRF+RS-LDA verifies that it works well on CUFS with no large shape degradation.

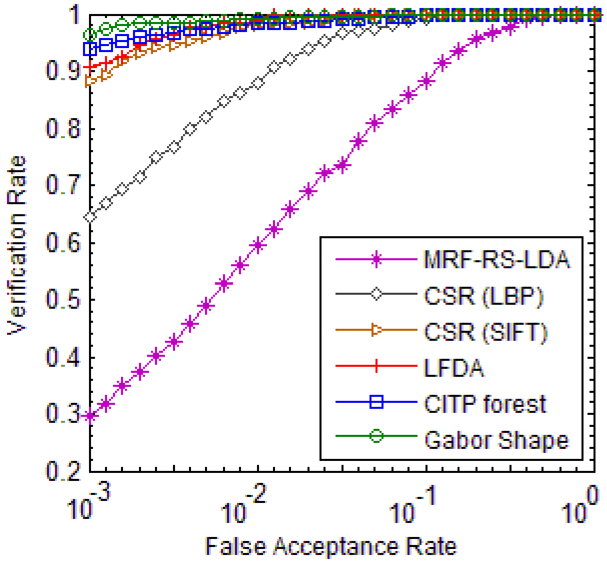


Figure 5. Results of experiment 2 on CUFSF: Comparing Gabor Shape with state-of-the-art approaches, VR vs. FAR.

Table 1. Comparing GS with CITP, VR at 0.1% FAR.

Single CITP	Fused CITP	Single GS	Fused GS
93.95%	98.70%	96.32 %	99.14%

Table 2. Rank-1 recognition rates on the CUFS dataset.

MRF+RS-LDA [14]	LFDA [5]	CITP [16]	GS
96.30%	99.47%	99.87%	99.91%

5. CONCLUSION

We proposed a new modality-invariant face descriptor for face photo retrieval by sketch example. Gabor Shape inspired by the fact that the modality gap between face photo and sketch is caused by fine texture not coarse texture. Therefore, the modality gap can be significantly reduced by features which are extracted from coarse texture (face shape). In Gabor Shape framework, first fine texture is attenuated by Gabor filters. Then, face shape which is represented by coarse texture is modeled by Radon transform. Experimental results on CUFS and CUFSF state that superior results are achieved by Gabor Shape compared to some state-of-the-art approaches. For future works, we will focus on face sketch recognition in presence of extreme shape degradation for real-world situations in which the eye-witness cannot properly recall the detail of suspect's face. Moreover, a face sketch dataset drawn based on recollection of eye-witness is required for future researches.

6. ACKNOWLEDGMENT

We acknowledge the generous support of NUS grant # R-252-000-451-112 for this work.

7. REFERENCES

[1] Ahonen, T., Hadid, A., and Pietikäinen, M. 2006. Face description with local binary patterns: Application to face recognition, *IEEE TPAMI*, 28, 2037-2041.

[2] Farbman, Z., Fattal, R., Lischinski, D., and Szeliski, R. 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. of Graphics*, 27, 3, 67:1-9.

[3] Kiani, H., and Sim, T. 2012. Inter-modality face sketch recognition. In *the Proc. of ICME'12*.

[4] Klare, B., and Jain, A. 2010. Sketch to photo matching: a feature-based approach. In *the proc. of SPIE Conference on Biometric Technology for Human Identification*.

[5] Klare, B., Li, Z., and Jain, A. 2009. Matching forensic sketches to mug shot photos. *IEEE TPAMI*, 33, 3, 639-646.

[6] Lei, Z., and Li, S. 2009. Coupled spectral regression for matching heterogeneous faces. In *the proc. of CVPR'09*, 1123-1128.

[7] Liu, Q., Tang, X., Jin, H., Lu, H., and Ma, S. 2005. A nonlinear approach for face sketch synthesis and recognition. In *the proc. of CVPR'05*, 1005-1010.

[8] Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2, 91-110.

[9] Shan, S.G., Gao, W., Chang, Y.Z., Cao, B., Yang, P. 2004. Review the strength of Gabor features for face recognition from the angle of its robustness to misalignment. In *the proc. of ICPR'04*, 338-341.

[10] Tang, X., and Wang, X. 2003. Face sketch synthesis and recognition. In *the proc. of ECCV'03*, 687-694.

[11] Tang, X., and Wang, X. 2004. Face sketch recognition. *IEEE Trans. Circuits and Systems for Video Tech.*, 14, 1, 50-57.

[12] Wang, X., and Tang, X. 2004. A unified framework for subspace face recognition. *IEEE TPAMI*, 26, 9, 1222–1228.

[13] Wang X. and Tang, X. 2006. Random sampling for subspace face recognition. *IJCV*, 70, 1, 91–104.

[14] Wang, X., and Tang, X. 2009. Face photo-sketch synthesis and recognition. *IEEE TPAMI*, 31, 11, 1955-1967.

[15] Zhang, W., Shan, S., Gao, W., Chen, X., and Zhang, H. 2005. Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A novel non-statistical model for face representation and recognition. In *the proc. of ICCV'05*, 786-791.

[16] Zhang, W., Wang, X., and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *the proc. of CVPR'11*, 513-520.

[17] Zhao, W.Y., Chellappa, R., Phillips, P.J., and Rosenfeld, A. 2003. Face recognition: A literature survey. *ACM Computing Survey*, 399-458.