Heterogeneous Face Image Matching Using Multi-scale Features

Sifei Liu[1], Dong Yi[1,2], Zhen Lei[1,2], Stan Z. Li[1,2*]
[1]Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
[2]China Research and Development Center for Internet of Thing
{sfliu,dyi,zlei,szli}@cbsr.ia.ac.cn

Abstract

Heterogeneous Face Recognition (HFR) refers recognition of face images captured in different modalities, e.g. Visual (VIS), near infrared (NIR) and thermal infrared (TIR). Although heterogeneous face images of a given person differ by pixel values, the identity of the face should be classified as the same. This paper focuses on NIR-VIS HFR. Light Source Invariant Features (LSIFs) are derived to extract the invariant parts between two types of face images. The derived LSIFs rely only on the variation patterns of the skin parameters so that the effects generated from light source can be largely reduced. A common feature extraction method is designed to capture LSIFs based on a group of differential-based band-pass image filters, and we show that the scale for filters is critical. Our results in CASIA HFB database validate the effectiveness of the model and our recognition approach.

1. Introduction

The concept of Heterogeneous Face Biometrics (HFBs) was initially proposed by *Li et al.*in [6, 7]. Recent developments have led to several proposals of HFBs, including matching between VIS and face sketch [13, 2, 4] and between VIS and NIR [16]. Among these works, using HFBs to solve lighting variation problem, specifically, matching between NIR and VIS, is increasingly studied in recent years. The NIR face recognition (FR) system [6] has been widely used and proved effective in varying illumination conditions. This system uses active near infrared sources to construct a light invariant recognition environment. Nevertheless, the visual image (VIS) is the most common image format in registration stage (such as photos from ID cards, web and surveillance videos), especially for the ID photo which provides good quality and is easy to get, and the registration can be continent and further standardized. Therefore, NIR-VIS HFR provides a way of both conquering light variations and staying convenient for users.

For heterogeneous NIR-VIS face matching, existing methods can be classified into three categories: (1) invariant feature-based [8, 15, 2], (2) common space-based [9, 16, 5], and (3) synthesis-based methods [13, 10, 14]. Similar to those in light variation problems, the invariant feature-based methods aim at finding features in common which are robust to light condition. While methods in [16, 5] treat spectral images as lying in the different subspaces, and a transformation is learned prior to matching. Comparatively, the invariant feature-based methods perform best among the three. However, other than modeling NIR-VIS pairs as taken in light variation conditions, no accurate illumination models between spectral images have been proposed.

Faces are distinct in appearances between NIR-VIS pairs. For a subject, we call the facial shape, skin and hair intrinsic properties. The extrinsic properties come from light sources, including light source spectra and distributions which differ a lot between NIR and VIS. In HFBs, the intrinsic properties are invariant to the same tester and independent of the extrinsic properties. However in heterogeneous face recognition, some intrinsic properties, such as skin absorption and scattering, are correlated to the light source spectra, so they are not completely invariant within the same identity. Thus, HFBs aims at encoding features that are invariant to light sources, including light source distribution and light source spectra.

In this paper, we propose a new feature set called the Light Source Invariant Features (**LSIFs**) that can (1) eliminate distinctions of heterogeneities that lie in low frequency of an image; (2) cut off noise of high frequency details that generated by different penetration extent of two light sources; and (3) retain useful information that relies on facial shape and variation patten of skin properties, which is invariant to light sources. This operator can be simply formulated by bandpass filters similar to the work of Liao et al. [8]. The optimal scale is decided by the extent of scattering and the structure of local facial regions, and we discover

[*] Corresponding author.

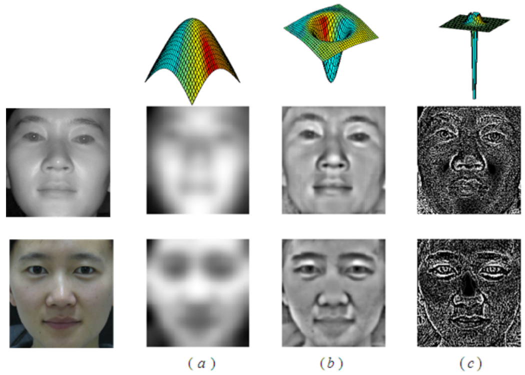


Figure 1. First row: image filters; Second row: NIR, Third row: VIS. (a)low frequency images by Gaussian filter (b)median frequency images by DoG filter (c)high frequency images by residue of Gaussian images

that the scale for differential operators is critical. We apply three local descriptors: HOG, GLOH and SIFT to further construct the feature space, and learn the most discriminative subspace via boosting and CSR [5] analysis.

2. Differential Facial Feature Operator

In this section, we introduce a Multi-DoG method to preprocess the heterogeneous facial images, and furthermore, apply local feature extraction and discriminative analysis to improve the classification.

We illustrate low, medium and high frequency of images from two modalities respectively in Fig. 1, with three different shape of kernels of Gaussian shown in the first row. The low frequency pairs, affected by both visual sources and Fresnel term, shows great distinctions in appearance. For the high frequency images, their appearances also vary a lot due to extent of “blurring” effect. We can see that the medium frequency pair is most consistent among the three and looks much similar than the original pair. As a result, the medium frequency image part generates patterns that are much more irrelevant to formulation of light sources. In this way an ideal preprocessing can be designed mostly by bandpass filters to cut off both high and low frequency of image components.

For NIR-VIS face matching task, we use Multi-Scale DoG (MSDoG) Face Representation to preprocess faces. A key issue in applying DoG filter is to find one or a group of optimal scale parameters that best fit for facial description. Rather than estimate the optimal scales through modeling both low frequency and high frequency images, we propose a Multi-Scale DoG(MSDoG) filters to generate a group of preprocessed images for each pair. In this phase, we can traverse the image scale space and identify potential useful features that are similar between two modalities in all sets. This is implemented by constructing a DoG pyramid as described in [11], and in this paper, we use the same MSDoG filter group to process both NIR and VIS images. For each photo I_i, the median frequency image group is represented as:

$$FI_i = \{fI_{s,r} | s \in [1, S], r \in [1, R]\} \tag{1}$$

where FI_i is the ith filtered image group composed of S scales of filters in R resolutions. As a result, altogether $S \times R$ filtered images are created to cover all possible fitting scales in order to capture effective invariant features.

Although in proper scales the preprocessed image group is similar between NIR and VIS in binary form, it still needs further process to obtain more robust feature. In this paper, we use three different feature descriptors to represent the heterogeneous face images: the HOG descriptor, the GLOH descriptor [12], and the SIFT descriptor [11]. HOG descriptors was successfully used for matching between face sketches and photographs [3, 4], and between NIR and VIS [2]. Both SIFT and GLOH are similar to HOG in feature formulation. For GLOH in our case, the PCA step has been removed as features will be concatenated and selected in a holistic way.

3. Discriminative Analysis

The proposed heterogeneous face classification framework uses training data to (1) select n “best” features, (2) uses discriminant analysis to further classify facial images across modalities. This unified framework is defined regardless of type of descriptors, and can be applied to new feature space in further research.

Given the whole set of multi-scale features, the descriptors would generate an over-complete representation with much redundant information contained. The fundamental learning problem here is to find a group of features or weak classifiers to construct a novel facial subspace representation. Note that unlike the work in[9] and [5] where features in different modalities are supposed to be in different subspace, we treat the heterogeneous features in the same subspaces after preprocess stage.

In this paper, we use Gentle Boost classifier, and the training stage are similar to the work in [6]. Rather than construct a strong classifier by weight each weak classifier, we retain the best N as selected features. There are two stages in training. We select for the first stage altogether 20,000 features, and for the second stage 5,000 features respectively for each descriptors.

We illustrate distribution of boosting selected features in Fig. 4. A gray-scale mesh grid shows for each facial region the selected amount, where a lighter block indicates

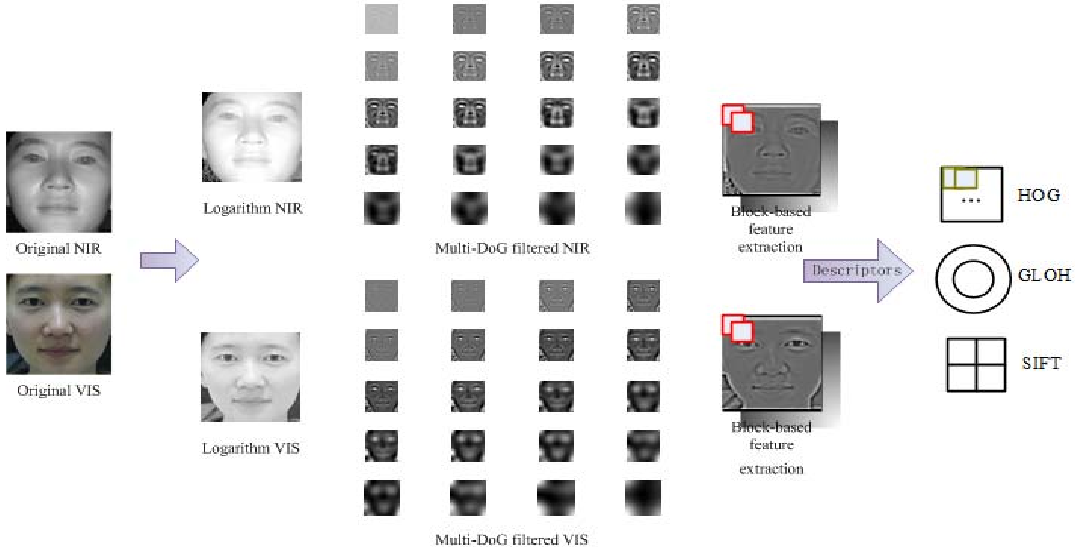


Figure 2. Multi-DoG: The heterogeneous faces are firstly logarithmic, and then a group of filtered DoG images are generated; Feature extractions are implemented after preprocessing, similar to that of homogeneous FR stages. The heterogeneous similarities existed in median frequency images generated by Multi-DoG, and can be selected through classification algorithm

more features selected in this region. Thus two conclusions can be drawn: (1) Mostly the regions of forehead, nose, and eyes contain more discriminative features for preprocessed heterogeneous data, and (2) All scales are useful, and their holistic selected feature distributions are differing in regions.

We also use linear Coupled spectral regression(LCSR) in [5] on selected feature space. Coupled spectral regression(CSR) framework is based on graph embedding framework and supposes that feature data occupy different positions in spaces. It calculates two projections for different modalities respectively to a common discriminant subspace. It is combined with regularization techniques, thus can achieve good generalization performance. Our experiment specially verified its performance.

In this work, although the boosting selected feature data is considered homogeneous, CSR can still achieve considerable generalization performances, better than Regularized LDA(RLDA) [1] in our experimental results.

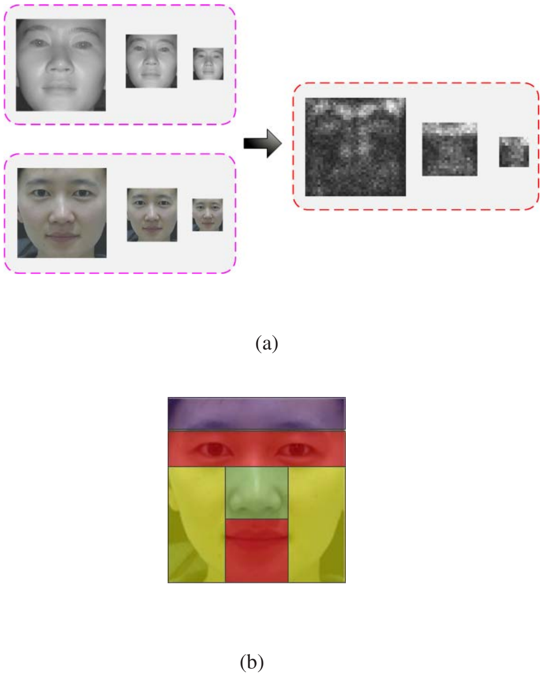
4. Score Level Fusion

Fusion of Score Level feature representation is an effective way to enhance the recognition performance. In our work we propose two score level fusion respectively on (1) fusion of different local descriptor representation, and (2) fusion of different facial component in score level.

Fusion of Local Descriptors Given a pair of NIR-VIS images and with three local descriptors D_1, D_2, D_3, we can obtain a sum of similarity scores $S(N,V) = S_{D_1}(N,V) + S_{D_2}(N,V) + S_{D_3}(N,V)$ that utilize all information for recognition. The score of $S_{D_i}(N,V)$ are normalized prior to summation.

Fusion of Facial Components We also extend our fusion strategy to facial component including EYE, EYEBROW, CHEEKS, NOSE, and MOUTH as in Fig. 4. Observing the boosting selected feature distributions, we can conclude that each part of the face contributes differently to the classification. The brow, eye and nose parts extract the most part of the selected features, which indicates that the classification can possibly be improved through giving more weight to them.

Fusion of facial parts is a simple and effective strategy to achieve this goal. Faces are firstly divided into five parts, then the stages of preprocessing and local feature extraction, feature selection and CSR discriminant analysis are performed the same as in holistic faces. Because the parts are not overlapped between each other, features are mostly the same with holistic operation except for a few locate along partition edges. In this work, we only tested SIFT descriptor to perform the components fusion because it is fast in extraction and can achieve good recognition performance.



(a)

(b)

Figure 3. (a) The most discriminative feature distributions of each scale over faces are illustrated on the right, where the left is the original images. A more lightening block indicates more features are selected from this region. (b) Five facial components corresponding to EYE, EYEBROW, CHECKS, NOSE, and MOUTH. There are no overlaps between each two parts.

5. Experiments

5.1. HFB Face Dataset

We evaluated the MSDoG preprocessing followed by boosting feature selection and CSR in the HFB face database [7]. This dataset has been tested for heterogeneous face recognition in [8, 5, 2, 4]. It consists of 202 subjects with 3,002 NIR probe images (captured in the near-infrared spectrum 780-1,100 nm) and 2,095 VIS gallery images captured in visible spectrum. We performed our training using the first 150 subjects with 7 images for each person, while 50 of the remaining subjects are selected as testing set.

5.2. Face Recognition for HFB Face Database

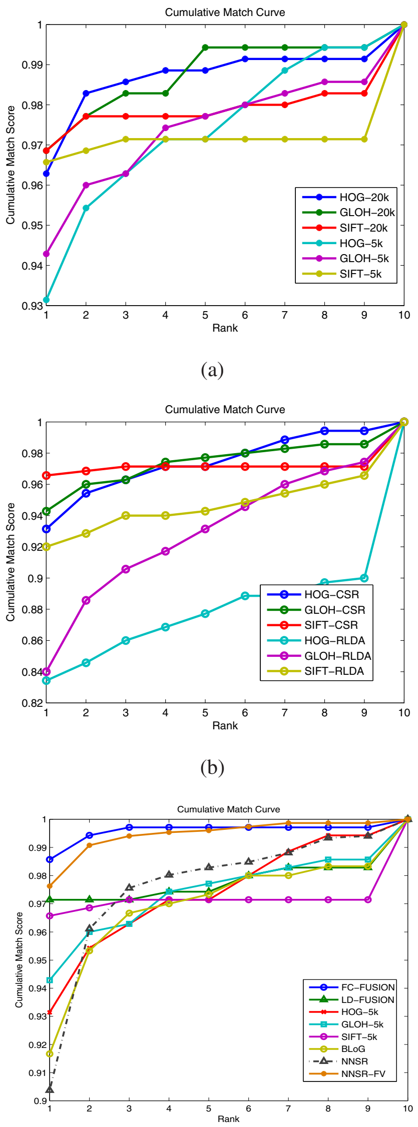
In face classification using MSDoG, three local descriptors has been applied in a dense sampling manner in preprocessed image group. For HOG, we use three rectangular patterns: 1:1, 1:2 and 2:1 in altogether 5 scales for minimal rectangular edge (range from 8 to 40 at step of 8 pixels) with 1/2 or 2/3 overlaps between neighboring blocks. For GLOH and SIFT, we use 6 pixels as the searching step in both vertical and lateral directions within all images of the group. The design of local descriptors is consistent with [12] and [11]. For image of size 120 × 120, there are totally 1,434,618, 1,728,832 and 1,088,000 features for HOG, GLOH and SIFT.

We use boosting to select the same number of features respectively for three local descriptors. For each one we use two stages of boosting to evaluate the performance. For the first stage, 20,000 features are selected, and for the second one, 5,000 features have been further trained and selected from the first stage. These features are utilized for the following CSR discriminant analysis. Considering the computational complexity, we only applied Linear CSR (LCSR) and has finally reduced the feature space dimension to 150. For CSR discriminant analysis we compare its performances on the second stage (5,000 selected features) with RLDA [1]. In Fig. 4, we illustrated the comparison performance between two stages and between RLDA and our proposed CSR.

In Fig. 4(a), from the results of two stages, we can see that all three descriptors can effectively encode heterogeneous images with MSDoG preprocessing. Among the three, SIFT performs best. Although the 20,000 features achieve better performance, the second stage with 5,000 features selected dose not significantly deteriorate and the SIFT with 5,000 features even get higher recognition rate than HOG with 20,000 features. The cumulative matching curves in Fig. 4(b) demonstrate that CSR achieves better discriminant performance than RLDA on the heterogeneous face dataset.

We also compared the proposed method with the work of Brendan Klare [2], where the best performance came from the fusion of NNSR and FaceVACS. From 4(c), all 3 descriptors outperform the NNSR and both SIFT and GLOH get a very close recognition rate to the best result, with a rank-1 accuracy of 96.86%. These results demonstrate the effectiveness of the proposed approach.

We also evaluate the heterogeneous matching performances for respectively Local-Descriptor fusion (LD-FUSION) and Face-Component fusion (FC-FUSION) strategies as described in section 4. For LD-FUSION, the similarity scores are computed after 5,000 feature selected and CSR discriminant analysis, where three sort of descriptors are equally weighted and summarized. For FC-FUSION, we use SIFT descriptor and the selection sets for each component are listed as table 1, the total selected feature number is 18,500. We can refer from the Rank-1 results for each component that brow contains the most useful matching information, while mouth performs worst due to its unstable coursed by expression and inaccurate local alignment. This performances are consistent with the boosting selected feature distributions as illustrated in Fig. 4. The combination of similarity scores of each component are weighted according to their ratio for selected feature numbers. The cumulative matching scores in Fig. 4(c) demonstrate the validation of our proposed fusion strategies, where



(a)

(b)

(c)

Figure 4. (a) CMC for stage1 with 20,000(20k) features selected, and stage2 with 5,000(5k) feature selected. (b) RLDA vs. CSR discriminant analysis. (c) Performances comparison for 3 local descriptors, two fusion strategies, BLoG method, and existed methods. The latest published best results from Brendan Klare [2] for HFB dataset in the same training and testing sets are also illustrated as (1)NNSR and (2) NNSR-FV, where our proposed FC-FUSION strategy outperforms the NNSR-FV, and all of our proposed strategies outperform NNSR.

the FC-FUSION strategy with Rank-1 accuracy of 98.57% even outperforms the score fusion of NNSR and FaceVACS which has up to now achieved the highest Rank-1 accuracy of 97.63%.

Moreover, we evaluated Yi's method [15] in HFB databases, denoted as BLoG. In this work, BLoG filter is used to preprocess NIR and VIS face images where the final features are represented by thresholding the LoG filtered images into binary codes. Thresholded binary images are more robust to noise than gray images. Therefore, the binary codes are directly construct the feature spaces and no further process of either local feature representations or training stage are needed, as we described in MSDoG for holistic faces. Instead, the binary features of corresponding patches are compared and distance computed using the Hamming distance. While the Hamming distance is suited for comparing two binary images, its very sensitive to occlusions, expression and mis-alignment. Therefore, a local search is performed and the minimum Hamming distance is finally found to measure the difference between two patches. These are illustrated in Figure5. All the minimum distances are fused by the sum rule to obtain the overall distance.

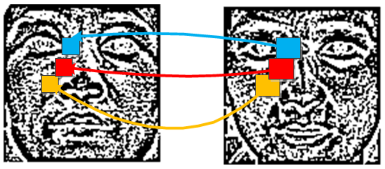


Figure 5. NIR-VIS face image matching strategy for BLOG. The 256×320 pixel area of a face is divided into 16×16 local patches.

	Size	Extracted	Selected	Rank 1
EYE	20×120	184,320	3,500	73.71%
BROW	20×120	184,320	5,000	**80.29%**
CHEEKS	$2 \times 80 \times 40$	506,880	5,000	52.86%
NOSE	35×40	113,664	2,200	48.86%
MOUTH	45×40	144,384	2,800	41.43%

Table 2. Experimental set and recognition results of each facial component in FC-FUSION strategy.

Because no training is needed, matching is performed in testing set with 50 subjects. This method is proved effective with a considerable Rank-1 accuracy of 91.67%, better than the training-based NNSR. Table 1 also provides a detailed results and comparison for all methods.

6. Conclusion

This paper focuses on the HFR problem in NIR-VIS face matching scenario. We propose MSDoG method that can capture the LSIFs for NIR and VIS facial images, and discriminative analysis are applied in extracted feature sets. The proposed method is prove through experiments in HFR database. We further improve the classification performances through proposing two fusion strategies, where the Face-Component fusion can achieve the best performance in HFR database.

Table 1. Experimental set and recognition results comparison of all strategies.

	FAR = 0.1%	FAR = 1%	Rank-1 Accuracy
Stage1-HOG	0.88	0.98	96.26%
Stage1-GLOH	0.92	0.98	96.86%
Stage1-SIFT	0.92	0.96	96.86%
Stage2-HOG	0.83	0.98	93.14%
Stage2-GLOH	0.94	0.96	94.29%
Stage2-SIFT	0.94	0.96	96.57%
LD Fusion	0.96	0.98	97.14%
FC Fusion	0.90	0.98	**98.51**%
BLoG	0.92	0.98	91.67%
S Liao's Approach[8]	0.68	0.87	Not Given
NNSR	0.79 ± 0.04	0.91 ± 0.02	90.38%
Cognitec	0.86 ± 0.02	0.94 ± 0.01	93.08%
NNSR+Cognitec	0.93 ± 0.01	0.97	97.63%

7. Acknowledgement

This work was supported by the Chinese National Natural Science Foundation Project #61070146, #61105023, #61103156, #61105037, National IoT R&D Project #2150510, and European Union FP7 Project #257289 (TABULA RASA http://www.tabularasa-euproject.org), and AuthenMetric R&D Funds.

References

[1] F. J. H. Regularized discriminant analysis. 1989. 3, 4

[2] B. Klare and A. K. Jain. Heterogeneous face recognition: Matching nir to visible light images. *Pattern Recognition, International Conference on*, 0:1513–1516, 2010. 1, 2, 4, 5

[3] B. Klare and A. K. Jain. Sketch-to-photo matching: a feature-based approach. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7667, Apr. 2010. 2

[4] B. F. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:639–646, 2011. 1, 2, 4

[5] Z. Lei and S. Li. Coupled spectral regression for matching heterogeneous faces. *CVPR 2009, IEEE Computer Society Conference on*, 0:1123–1128, 2009. 1, 2, 3, 4

[6] S. Li, R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):627–639, april 2007. 1, 2

[7] S. Li, Z. Lei, and M. Ao. The hfb face database for heterogeneous face biometrics research. In *CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 1–8, June 2009. 1, 4

[8] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li. Heterogeneous face recognition from local structures of normalized appearance. In *Advances in Biometrics*, volume 5558, pages 209–218. 2009. 1, 4, 6

[9] D. Lin and X. Tang. Inter-modality face recognition. In *Computer VisionCECCV 2006*, volume 3954, pages 13–26, 2006. 1, 2

[10] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 1005–1010 vol.1, june 2005. 1

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 2, 4

[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, oct. 2005. 2, 4

[13] X. Tang and X. Wang. Face sketch recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):50–57, jan. 2004. 1

[14] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):1955–1967, nov. 2009. 1

[15] D. Yi, S. Liao, Z. Lei, J. Sang, and S. Li. Partial face matching between near infrared and visual images in mbgc portal challenge. In *Advances in Biometrics*, volume 5558, pages 733–742. 2009. 1, 5

[16] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Li. Face matching between near infrared and visible light images. In *Advances in Biometrics*, volume 4642, pages 523–530. 2007. 1