Image Factorization and Feature Fusion for Enhancing Robot Vision in Human Face Recognition

Hui Yu, Zhaojie Ju and Honghai Liu, *Senior Member, IEEE*
School of Creative Technologies
University of Portsmouth
Portsmouth, PO1 2DJ, UK
{hui.yu; zhaojie.ju; honghai.liu}@port.ac.uk

***Abstract*—Illumination variation has been a challenging problem for face recognition in robot vision. To reduce the effect caused by illumination variation, a lot of studies have been explored. The Total Variation (TV) method is particular used to factorize images into a low frequency component and a high frequency one. However, the low frequency component still contains significant intrinsic features resulting in failure in face recognition in some cases. In this paper, we propose to further extract illumination invariant features from face images under uncontrolled varying lighting conditions. The Nonsampled Contourlet Transform (NSCT) method is employed to enhance the extraction of intrinsic feature. The combined factorization model is very effective in the experiment on the Yale database.**

Keywords—image factorization, robot vision, face recognition, total variation, contourlet transform, feature fusion

I. INTRODUCTION

Intelligent robot is considered as the partner of human beings rather than a machine. With the development of technologies, robots will increasingly undertake more tasks that only human beings are able to do nowadays. In the future, robots will be able to interact closely with humans in their daily lives in many applications such as health-care, cursing and entertainment etc. Robots can identify human by looking at their faces through their vision-based systems such as cameras. One of the critical tasks for mobile robots is to identify an individual person who it interacts with in the real world.

During last a few decades, face recognition has gained significant attention from researchers [1, 22, 23]. In spite of successes of those methods in some applications, the existing algorithms still face challenges from illumination variations caused by uncontrolled lighting conditions. Illumination variation is still one of the most challenging problems in face recognition, especially for appearance-based methods [24]. This is particular challenging for mobile robots, since it is not easy for robots to control lighting conditions of its surrounding environment by itself. To recognize a face, the query faces are compared with the known face database for "best matching". Usually face features are extracted to facilitate the matching procedure. Face recognition effect can be significantly reduced since face images of the same identity appear different under different illumination [7]. The straightforward solution to this problem is to normalize illumination variation to a canonical lighting condition.

To estimate illumination and reflectance from a single image is an ill-posed problem [9], so it is nontrivial to decouple reflectance component and illumination component from an image I. Land et al. [2] proposed a Retinex model for estimating these two components. In their method, the reflectance component R is regarded as the ratio of the input image I to the low-pass part L and illumination part is estimated as the smoothed version of image I. The reflectance component represents the intrinsic feature of the subject, which is invariant to illumination condition. Therefore, illumination variations mainly affect the low frequency component with low-frequency [6] rather than high frequency features such edges, line features etc. However, the low frequency features of a face can still be recognizable though with a much low recognition [11]. It suggested that both high frequency and large scale components could play parts in face recognition and synthesis, so the illumination part with large scale features still remains some valuable intrinsic features.

Thus, the key problem becomes how to effectively extract the illumination component from a single image but remove identification features as much as possible.

In this paper, we propose a method to tackle the illumination variation problem to enhance the possible of face recognition for robots. The proposed method combining Total Variation model [8] and Nonsampled Contourlet Transform [18], which intends to effectively extract intrinsic features from a given image along with removing noises and halo effect.

A variety of methods have been proposed with the intention to solve the illumination problem in face recognition. Illumination variations are caused mainly by 3D geometric face shapes when lighting come from different directions. Some 3D model assisted methods have been proposed for face recognition [19, 20]. Basri et al. [21] proposed that a set of all reflectance functions obtained from Lambertian objects under

This work was supported by EU seventh framework programme under grant agreement no 611391, Development of Robot-Enhanced Therapy for Children with Autism Spectrum Disorders (DREAM), and the projects under grant No. DMETKF2013001 by State Key Lab of Digital Manufacturing Equipment & Technology, China.

various lighting conditions can be used to approximate a 9D linear subspace. It shows that algorithms can be developed for face recognition based on this assumption. But this method requires a number of images of the object under different lighting conditions for the training purpose. Weiss [4] presented a method using maximum likelihood to estimate the reflectance component. Wang et al. [13] proposed a self-quotient image (SQI) method for illumination normalization by the division over a smoothed version of the input image itself. Though this method is effective for single image and easy to implementation, it suffers from noises and halo effects in the step regions.

Zhang et al. [5] proposed a multi-scale image decomposition method using the wavelet analysis in the logarithmic domain. Their method has the capability of edge-preserving in low frequency domain but cannot effectively decouple geometric information due to its weak ability in directionality. Xie et al. [14] proposed to use Nonsubsanpled Contourlet Transform in the logarithm domain to extract illumination invariant facial features from a single image for face recognition. Their method assumed that in the logarithm domain the low-pass subband of a face image along with the low frequency part of strong edges could be considered as the illumination component, whereas the weak edges and the high frequency part of strong edges could be regarded as the reflectance component. The Total Variation model is popularly used in image decomposition [3, 8, 17]. TV model is able to maintain the feature edges through penalizing small variations during the minimization procedure.

II. METHODOLOGY

According to the Lambertian theory, the intensity of a 2D face surface I can be described as

$$I(x,y) = R(x,y) \bullet L(x,y) \qquad (1)$$

where R represents the reflectance component determined by the surface material, so it can be regarded as illumination-invariant features. And L is the illumination component, so it represents the lighting condition on the surface. And the operator stands for pixel by pixel operation.

The above equation becomes an additive form if Logarithm transform is applied [3, 15]:

$$\log I(x,y) = \log R(x,y) + \log L(x,y) \qquad (2)$$

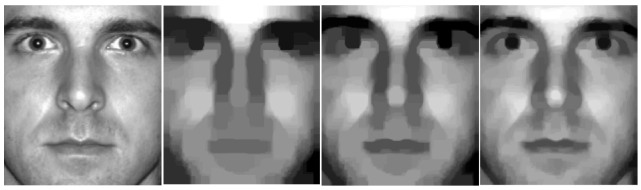
Once given the desired illumination $L^{'}$, the new face image $I^{'}$ can be expressed as:

$$\log I'(x,y) = \log R(x,y) + \log L'(x,y) \qquad (3)$$

Logarithm transform is well used in recent years, especially in image enhancement. By applying logarithm transform, the Lambertian equation can be converted from multiplicative form to an additive one, which makes it convenient to apply image processing algorithms. The logarithm transform does not affect the intrinsic structure of the image.

A. Total Variation Decomposition

The core of illumination normalization of a face image becomes finding an effective way for estimating these two components R and L. Chen et al. [8] developed a Logarithmic Total Variation (TV) model to decompose an image in to high frequency and low frequency components. The advantage of TV model is that it can preserve sharp feature edges of a face image due to the piecewise smooth regularization property of the TV norm. It factorizes an input image into low frequency part u and high frequency part v representing as illumination-dependent component and illumination–invariant component respectively. The illumination-dependent component u can be obtained by solving the following optimization problem:



(a) (b) (c) (d)

Fig. 1. low frequency part of TV decomposition under different λ values (a) raw face image, (b) $\lambda = 0.2$, (c) $\lambda = 0.3$,(d) $\lambda = 0.5$.

$$\min_{u} \int |\nabla u| + \lambda \| f - u \|_{L^1} \qquad (4)$$

Where f is the input image, λ is a scalar as a threshold on scale. And $\int |\nabla u|$ is the total variation (TV) of the unknown component u. The parameter selection for TV model is straightforward and remains as one of the main advantages of the TV model [8]. However, as a unique parameter for the equation, λ has a very high influence on the convergence and result of the above optimization. Since λ is in inverse proportion to the scale of the faces, low λ value results in larger face scale. Though through adjusting the value of λ, images can be decomposed into different scales, the u component still contains substantial information of the face features (Fig. 1). It usually contains enough facial profile features that can be recognized.

B. Fuse edges from Nonsampled Contourlet Transform

NSCT is developed by combining the nonsubsampled pyramid structure and the nonsubsampled directional filter bank structure. It can distinguish strong edges, weak edges and noises through factorizing signal into different frequency subbands. NSCT is a fully shift-invariant, multidirection and multi-scale transform, so it can capture geometrical structures of a face image. Through applying NSCT on the low frequency component, it is expected to further factorize intrinsic facial features and remained illumination.

Let the input signal in the j-th level be u_j. After applying high-pass filter f^1 and low-pass filter f^0, the nonsubsampled pyramid decomposes u_j into two subband: high-pass subband u_j^1 and low-pass subband u_j^0. This can be described as follows:

$$u_j^k = f^k \otimes u_j \quad k \in 0,1 \qquad (5)$$

Where $\otimes$ is a convolution operator, which can be formulated as:

$$u_j^k[m] = \sum_{i \in F(f^k)} f^k[i] u_j[m - i \bullet m'] \qquad (6)$$

$$k \in 0,1; m \in N \times N; m' = 2^{j-1} I_d$$

Where $F(f^k)$ is the compactly supported function and m' is a sampling matrix (I_d is the identity matrix). Nonsubsampled directional filter bank (NSDFB) is then applied to factorize the high-pass subband u_j^1 into several directional sbubands as follows:

$$e_{1,p} = f_p^{'} \otimes u_j^1, p \in 1,...,M^j \qquad (7)$$

Where p denotes the p-th direction and M^j is the number of directional subbands at level j. And low-pass subband u_j^0 becomes the input signal for the next-level decomposition.

The reconstruction of NSCT is an inverse procedure of decomposition, which can be formulated as the following equations:

$$\hat{u}_j = \hat{f}^0 \otimes \hat{u}_j^0 + \hat{f}^1 \otimes \hat{u}_j^1 \qquad (8)$$

$$\hat{u}_j^0 = \hat{u}_{j+1} \qquad (9)$$

$$\hat{u}_j^1 = \sum_{p=1}^{M^j} \hat{f}_p^{'} \otimes e_{j,p}, j \in 1,...J \qquad (10)$$

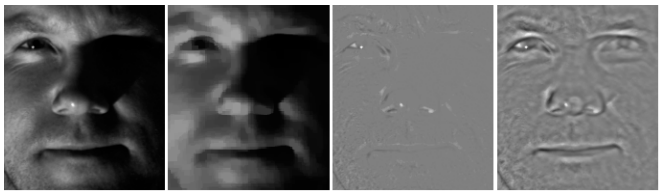
where $\hat{f}^0$ and $\hat{f}^1$ are filters for NSCT reconstruction corresponding to f^0 and f^1. And $\hat{f}_j^{'}$ is the corresponding filter of $f_j^{'}$ for direction p.

Since NSCT is a multiscale decomposition formulation, the procedure is conducted interactively. Specifically, by setting the low pass subband signal as the input of next level decomposition ($\hat{u}_{j+1} = \hat{u}_j^0$), the input signal can be reconstructed interactively.

In this paper, the Nonsampled Contourlet Transform (NSCT) method is applied to extract residual intrinsic features from the low frequency component. Specifically, the residual high frequency information, which mainly contains edge information of the face image, is further extracted from the u component from TV decomposition. It is then fused with the high frequency part v from TV model. Since direct fusing the intrinsic features from TV model with the residual high frequency part from NSCT model can be problematic, so we adopt a weighting parameter to control the second part .The fused facial features can be generated using the following equation:

$$I_v = v_{LTV} + \alpha * h_{NSCT} \qquad (11)$$

Where V_{LTV} is the v component from TV decomposition, h_{NSCT} is the residual high frequency features using NSCT model and α a weighting parameter. In the paper, we set $\alpha = 0.7$ for all experiments. Fig. 2 shows an example of TV decomposition and feature fusion results.



(a) (b) (c) (d)

Fig. 2. Example of face image factorization. (a) Raw face image from Yale Database B; (b) The low frequency component from the TV method, $\lambda = 0.6$; (c) The high frequency component from the TV method; (d) The high frequency part by the TV-NSCT method.

In the following, we will show how to use the image factorization to compensate face illumination. According to the reflectance model proposed by Chen et al. [8], an image I can be decomposed as follows:

$$I(x,y) = \frac{R}{R_l} \bullet (R_l L) = \rho \bullet S \qquad (12)$$

Where R_l denotes the albedos low frequency skin areas and background. Through solving the above equation, we can obtain S which contains extrinsic illumination and larger intrinsic structures. This larger intrinsic structure in S is defined compared with the smaller intrinsic structure ρ. The variation of image I under new illumination can be described as:

$$I' = \rho \bullet S' = R_l L' \qquad (13)$$

Applying logarithm transform to above equations, we can obtain the following form [14]:

$$\log S'(x,y) = \log S(x,y) + \varepsilon(x,y) \qquad (14)$$

where $\varepsilon(x,y)$ is the illumination difference between image I with original illumination and Image I' under new illumination in the logarithm domain. Thus, illumination compensation can be done in the low frequency component.

Since the objective is to compensate the illumination of a target image from the reference face image. During the illumination transfer, the frequency decomposition algorithm is applied to both face images. For the reference image, we want to keep the illumination but remove intrinsic facial features, whereas for the target image, we want to extract the intrinsic facial features. In this context, we assume (8) can still be valid when apply to the high frequency component.

III. EXPERIMENTS AND DISCUSSION

The proposed method has been tested on the Extended Yale Face Database B [16, 25], which provides different identities and lighting conditions. The extended Yale Face Database B consists of 38 human subjects, each with 64 illumination conditions. There are total 2414 images, since some images in the cropped imaged database labeled as “bad”. These”bad” images have not been used for training or testing. All face images for the recognition experiment have been cropped into

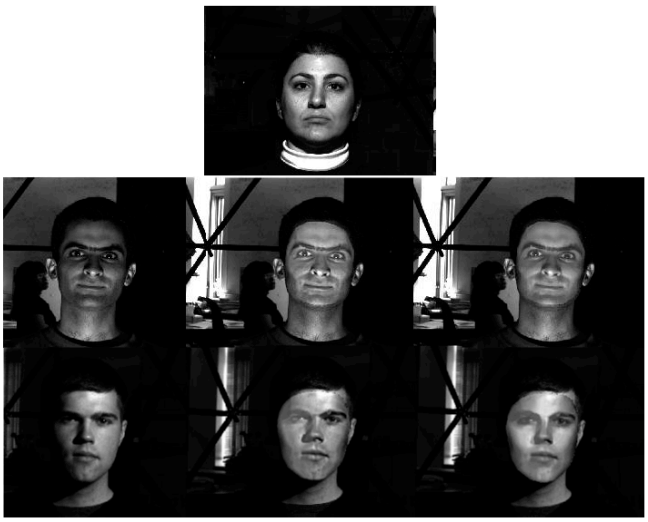


Fig. 3. Illumination compensation example for Extended Yale Face Database B. Top row: the reference face image. Both middle row and bottom row are target identities in which column one is the raw face image, column two compensated using TV-NSCT and column three compensated using TV.

192 x 168 pixels, which contain the majority of face region. The method was tested in two experiments. In the first experiment, we compare our method TV-NSCT with the TV method for face recognition on the Extended Yale Face database B. In the second experiment, we tried to normalize face illumination using the proposed method on the Extended Yale Face Database B without cropping the face region.

A. Experiment one

The method is able to compensate illumination of the target face from the source image. We have applied the illumination compensation algorithm to both grayscale and RGB images on two databases. This is rather visual compensation for face images with no recognition test involved. Fig. 3 gives illustration of illumination compensation on the Extended Yale Face Database B. We have manually labeled 76 landmarks on faces of Yale Face Database B. These landmarks are used as feature points to align the reference face and the target face for the purpose of accuracy. It shows that the TV-NSCT method can extract stronger facial features, eg. features around eyes and mouth areas. This is because the redundant feature information from the reference face image carried in the compensating illumination of the low frequency component. The result of TV shows uneven illumination changes on the face as shown on row 2 Fig. 3. And important features such eyes get blurred as well.

B. Experiment two

In this experiment, we compare our method with the TV method for face recognition under various illumination conditions. Since we concern the illumination influence, so only face images in front view with various illumination conditions were used in these databases. In recognition, the image pixel values of the invariant features were used to feed the SVM classifier. Fig. 4 shows examples of one identity with various lighting conditions the illumination invariant features. We compare our method with the TV method for recognition. In our experiments, $\lambda = 0.6$ for all TV factorization and three-level NSCT decomposition was applied.

The Extended Yale Database B are divided into 5 subsets (subset 1 to 5) in terms of the angles between the light source direction and the camera axis, but we only tested the front view face images in this paper. Since the cropped face images from the Extended Yale Database B have already been aligned, no further processing was conducted before applying the TV-NSCT method. One subset face images was used as the training set and the rest were used as the testing set. Fig. 5 demonstrates some examples of illumination invariant feature extraction from the Extended Yale B Database. The experiment showed that the TV-NSCT has a better performance than the TV method (see table 1). The average recognition rate for TV method is 95.87% and 99.36% for our method when using 8 random images per identity as training samples.

TABLE I. TABLE 1 COMPARISON OF RECOGNITION RATE FOR THE EXTENDED YALE

Method	Recognition Rate (%)			
	Subset 2	***Subset 3***	***Subset 4***	***Subset 5***
TV	100	98.22	95.51	91.39
TV-NSCT	100	100	98.19	99.27

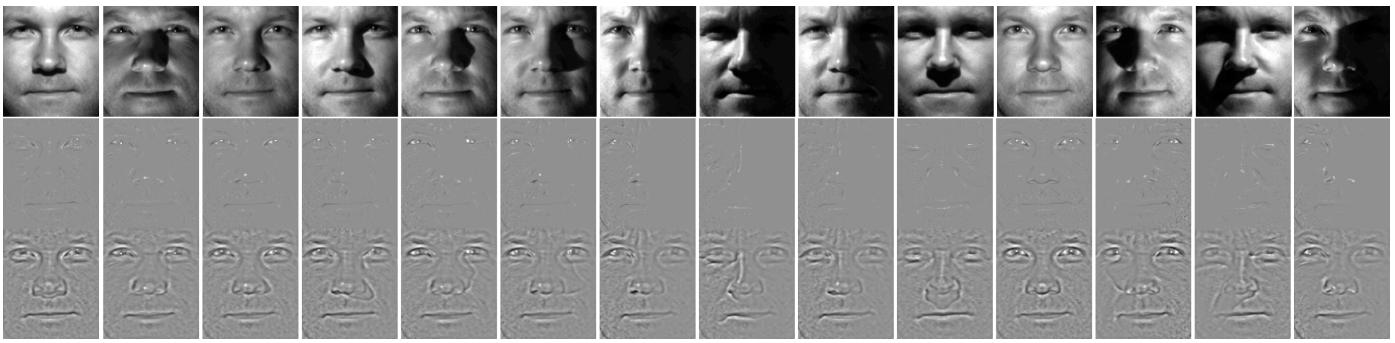


Fig. 4. The comparison of two methods. Top row: the raw images of a single subject from Yale Database B with various illumination conditions; Middle row: the small scale from the TV method with $\lambda = 0.6$; Bottom row: the high frequency using the TV-NSCT method with $\lambda = 0.6$.

Fig. 5 Examples of illumination invariant feature extraction from the Extended Yale B Database. Three blocks represent three identities. Each identity shows face images under various illumination conditions. For each block, top row is the raw face images; middle row is the small scale using TV method; bottom row is the result using TV-NSCT method.

IV. CONCLUSION AND FUTURE WORK

An illumination invariant feature extraction scheme has been proposed in this paper. NSCT was combined to further extract face features in multiscale level. This technique performed the factorization on both parts (large and small) of the TV decomposition. Useful information remained in the low frequency component from TV can be further extracted through discarding the redundant frequency part during the reconstruction of NSCT. Experiments have shown that the proposed method has achieved a better result than the TV method. An image-based illumination compensation approach has been presented using single face image as the reference.

In future, we will improve the combination weighting scheme, which can dynamically parameterize the weighting based on statistical analysis. We will apply the method to wild face images combining face detection and pose estimation algorithms etc. for automatic face recognition.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, Face Recognition: A Literature Survey, ACM Computing Surveys, Vol. 35, No. 4, pp. 399-458, 2003.

[2] E. H. Land and J. J. McCann. Lightness and Retinex Theory. Journal of the Optical Society of America (1917-1983), 61, Jan. 1971

[3] Q. Li, W. Yin, and Z. Deng. Image-Based Face Illumination Transferring Using Logarithmic Total Variation Models. Vis. Comput., 26(1):41–49, 2009

[4] Y. Weiss, Deriving intrinsic images from image sequences, in: Proceeding of IEEE International Conference on Computer Vision (ICCV), 2001

[5] T. Zhang, B. Fang, Y. Yuan, Y. Tang, Z. Shang, D. Li, F. Lang, Multiscale Facial Structure Representation For Face Recognition Under Varying Illumination, Pattern Recognition 42 (2009) 251–258.

[6] A. Shashua and T. Riklin-Raviv, "The quotient image: Class-based re-rendering and recognition with varying illuminations," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 2, pp. 129–139, Feb. 2001.

[7] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces,"IEEE Trans. Image Process., vol. 18, no. 11, pp. 2599–2606, Nov. 2009.

[8] T. Chen, W. Yin, X. Zhou, D. Comaniciu, T.S. Huang, Total variation models for variable lighting face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 1519–1524

[9] R. Rammamorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," inProc. ACM Conf. Siggraph, Los Angeles, CA, 2001, pp. 117–128.

[10] R. Rammamorthi, P. Hanrahan, A Signal-Processing Framework For Inverse Rendering, In: Proceedings of ACM SIGGRAPH, 2001.

[11] X. Xie, W. Zheng, J. Lai, P. Yuen, C. Y. Suen: Normalization of Face Illumination Based on Large-and High frequency Features. IEEE Transactions on Image Processing 20(7): 1807-1821 (2011)

[12] A.M. Martinez and R. Benavente. The AR Face Database. CVC Technical Report #24, June 1998

[13] H. Wang, S.Z. Li, and Y. Wang, "Generalized Quotient Image,"Proc. Int'l Conf. Computer Vision and Pattern Recognition,2004

[14] X. Xie., W.S. Zheng, J. Lai, P.C. Yuen, C.Y. Suen: Normalization of Face illumination Based on Large- and Small- Scale Features, IEEE Transactions on Image Processing, vol. 20, no. 7, pp. 1807-1821, 2011

[15] W. Chen, M. Er, S. Wu, Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain, IEEE Transactions on Systems, Man, and Cybernetics 36 (2) (2006).

[16] A. Georghiades, P. Belhumeur, and D. Kriegman. From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose, IEEE Trans. Pattern Anal. Mach. Intell. 23(6), 643–660, 2001.

[17] M. Nikolova, "Minimizers of Cost-Functions Involving Nonsmooth DataFidelity Terms"SIAM J. Numerical Analysis,vol. 40, no. 3, pp. 965-994, 2002.

[18] L. Cunha, J. Zhou, M.N. Do, The nonsubsampled contourlet transform: theory, design, and applications, IEEE Transactions on Image Processing 15 (2006) 3089–3101

[19] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2-4 September 2009, Genova, Italy

[20] G. Passalis, S. Zafeiriou ; G. Tzimiropoulos; M. Petrou; T. Theoharis; I.A. Kakadiaris, Bidirectional relighting for 3D-aided 2D face recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010

[21] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces,"IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 2, pp. 218–233, Feb. 2003.

[22] P. Pramod Kumar, Prahlad Vadakkepat, Ai Poh Loh, Hand Posture And Face Recognition Using A Fuzzy-Rough Approach International Journal of Humanoid Robotics Vol. 07, No. 03, pp. 331-356

[23] Y. Hu, A. S. Mian, R. Owens, "Face Recognition Using Sparse Approximated Nearest Points between Image Sets," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.34, no.10, pp.1992,2004, Oct. 2012

[24] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 low frequency results," inNational Institute of Standards and Technology, NISTIR.: , 2007, vol. 7408

[25] K. Lee; J. Ho; D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.27, no.5, pp.684,698, May 2005