An Open Source Framework for Standardized Comparisons of Face Recognition Algorithms

Manuel Günther, Roy Wallace, and Sébastien Marcel

Idiap Research Institute
Rue Marconi 19
CH - 1920 Martigny

Abstract. In this paper we introduce the facereclib, the first software library that allows to compare a variety of face recognition algorithms on most of the known facial image databases and that permits rapid prototyping of novel ideas and testing of meta-parameters of face recognition algorithms. The facereclib is built on the open source signal processing and machine learning library Bob. It uses well-specified face recognition protocols to ensure that results are comparable and reproducible. We show that the face recognition algorithms implemented in Bob as well as third party face recognition libraries can be used to run face recognition experiments within the framework of the facereclib. As a proof of concept, we execute four different state-of-the-art face recognition algorithms: local Gabor binary pattern histogram sequences (LGBPHS), Gabor graph comparisons with a Gabor phase based similarity measure, inter-session variability modeling (ISV) of DCT block features, and the linear discriminant analysis on two different color channels (LDA-IR) on two different databases: The Good, The Bad, and The Ugly, and the BANCA database, in all cases using their fixed protocols. The results show that there is not one face recognition algorithm that outperforms all others, but rather that the results are strongly dependent on the employed database.

1 Introduction

Over the last few decades, a great variety of face recognition algorithms have been proposed. To show their advantage over other existing algorithms, face recognition experiments have typically been executed on one or more publicly available facial image databases [1,2,3,4,5,6,7,8]. Unfortunately, often these databases are not accompanied by strict experimental protocols or the protocols that are provided are *biased*. A protocol for an image database defines which of the images within this database should be used for *training* the algorithms, which are for *enrolling* models and which images are finally used as *probes*. In a biased protocol, the identities used for training and for testing overlap, whereas in unbiased protocols training and testing identities are disjoint. For real-world scenarios, training and testing identities should be disjoint since it is impractical to retrain the algorithms each time a model of a new identity should be enrolled, and it is more realistic that imposters are unknown to the system.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 547–556, 2012.

Often, face recognition algorithms are tested only on a few of the available databases and only those results that are superior to some baseline results are published. Regrettably, published results are, thus, usually not comparable. Additionally, often the results of other researchers can not be reproduced since they do not publish all of the meta-parameters of their algorithms. Hence, face recognition surveys like [9,10,11,12,13] can only report the results of other researchers, so "it is really difficult to declare a winner algorithm" [9] since "different papers may use different parts of the database for their experiments" [13]. In an attempt to categorize the algorithms, [10] used a more advanced evaluation of the methods, but still they had to rely on the results published by the authors of the surveyed papers because they could not reproduce them themselves.

Some institutions already tried to provide an open source interface in which different algorithms can be tested, for example the CSU Face Identification Evaluation System [14]. Unfortunately, this library is not sufficient since:

1) the implemented algorithms are already outdated,
2) new algorithms must be implemented in their specific C environment, and
3) the main focus of the library is on the FERET image database [1].

Other comparisons of face recognition algorithms were done by the *face recognition vendor tests* (FRVT) [15,16,17] and similar tests held by the *National Institute for Standards and Technology* (NIST). Though these vendor tests already provide a fair comparison, unfortunately they are designed to compare commercial algorithms and, hence, the methodologies used by the participating vendors are usually kept secret. The results of such tests are, thus, largely useless for researchers who are interested in establishing what are the state-of-the-art face recognition techniques and expanding on the best performing methods.

The contribution of this paper is to present the *facereclib*, which is, to the best of our knowledge, the first tool to dependably compare face recognition algorithms that:

1) relies solely on open source software – all results are reproducible and there are no hidden tricks,
2) utilizes fixed protocols for most of the commonly used image databases – the generated results are, hence, comparable to previously and subsequently published results,
3) supports a broad variety of state-of-the-art face recognition algorithms,
4) allows the easy integration of already existing source code, and
5) permits rapid prototyping of novel ideas and testing meta-parameters of existing algorithms – an ideal playground for researchers.

The facereclib is a satellite package[1] of the recently released open source signal processing and machine learning toolbox *Bob* [18][2], which is written in C++

[1] Facereclib, the face recognition satellite package of Bob, will soon be available at `http://github.com/idiap/bob/wiki/Satellite-Packages`

[2] Bob is open sourced under a GPL v3 license. To download Bob, please visit `http://www.idiap.ch/software/bob`

and Python and itself contains an implementation of many face recognition algorithms and database protocols. Since the facereclib is also implemented in Python, it is easy to integrate other software. To show that capability, in this paper we incorporate one algorithm from the CSU Face Recognition Resources [19].

The facereclib is created for a fast design and execution of face recognition experiments. It includes Python scripts that take as arguments configuration files for:

1) the employed database, its protocol, and the location of the image files,
2) the parametrization of image alignment and preprocessing,
3) the type and the variation of the extracted features, and
4) the face recognition algorithm and its meta-parameters.

Additionally, a standardized evaluation of the results is provided. This makes it easy for researchers to follow the evaluation protocols since they do not have to implement them themselves, and we hope to encourage researchers to produce comparable results.

To illustrate the potential of the facereclib, we execute four representative face recognition experiments. We apply three state-of-the-art face recognition algorithms from Bob: *local Gabor binary pattern histogram sequences* [20] (LGBPHS), *Gabor graphs* [21] with a Gabor phase based similarity measure [22], and *inter-session variability modeling* (ISV) [23] on *discrete cosine transform* (DCT) block features, as well as one algorithm taken from the CSU Face Recognition Resources [19]: LDA-IR. These algorithms are tested on two popular and challenging facial image databases: *The Good, The Bad, and The Ugly* (GBU) [5] and the *BANCA* [24] database. All face recognition experiments are run using our facereclib. It assures that the recognition results are directly comparable since exactly the same processing chain is executed. Furthermore, all parameters of all steps of the processing chain are given in the configuration files and, importantly, the recognition results are reproducible.

The remainder of this paper is structured as follows: In Section 2 we give a short overview of the features and the algorithms that are used in this paper. Section 3 describes the image databases and the protocols that we consider. Section 4 presents experimental results, while Section 5 ends with a discussion of what we have achieved in this paper.

2 Algorithms

Face recognition algorithms can typically be described in three stages: *training*, *enrollment* and *deployment*. During training, face recognition algorithms adjust their parameters to fit a given set of training images. In enrollment, one or more images per identity are used to generate a *model* for each *client*. During deployment, an unseen *probe* image is compared to one or more of the models, and a *score* for each model/probe pair is computed. When the score exceeds a certain carefully selected threshold, the pair is accepted as a *client claim*, or it is rejected as an *imposter claim*.

The algorithms used in this paper cover a representative set of state-of-the-art approaches to semi-automatic face recognition. Since the aim of this paper is to compare face recognition algorithms rather than face detectors, hand-labeled eye positions were used to geometrically normalize the faces, throughout.

2.1 LDA-IR from CSU

Firstly, we consider the LDA-IR algorithm taken from [19], which extracts the *I layer* of YIQ color space and the *red channel* of RGB color space from 65×75 pixel images. After preprocessing, each of these images is projected into a PCA subspace, then an LDA subspace, both of which are always trained on the training set of the respective database. The Euclidean distance is used for comparison of feature vectors.

2.2 Algorithms from Bob

The algorithms below all work with 64×80 gray-level images, with eye positions at $(16, 16)$ and $(48, 16)$, and preprocessed with the Tan & Triggs algorithm [25].

Local Gabor Binary Pattern Histogram Sequences (LGBPHS) [20]. Using this algorithm, *local binary pattern* (LBP) histograms are calculated for non-overlapping 8×8 pixel blocks, after convolution with a set of 40 Gabor wavelets. The histograms of all blocks from all wavelets are concatenated into one long vector (188,800 dimensions) and these vectors are compared using the χ^2 measure. The Gabor wavelets are used in the common 8 orientations and 5 scales [21] while the size of the enveloping Gaussian was set to $\sigma = \sqrt{2}\pi$. For LBP extraction, we use uniform circular LBPs with 8 neighbors and in radius of 2 pixels (i. e. $\text{LBP}_{8,2}^{u2}$, see [26]).

Gabor Graphs with Gabor Phase Based Similarity [22]. This algorithm compares Gabor jets assembled in grid graphs. Both the magnitude and phase of the Gabor wavelet responses are used. The node positions of the grid graph correspond to the centers of the histogram blocks, leading to 80 Gabor jets per image. The Gabor wavelets are the same as those described for LGBPHS above. Gabor graphs are compared using the average similarity of corresponding Gabor jets, where Gabor jets are compared with the similarity measure S_{n+C} from [22].

Inter-session Variability Modeling (ISV) [23]. Local discrete cosine transform (DCT) features are sampled from overlapping 12×12 pixel blocks, resulting in 3657 feature vectors per image. Pixels are first normalized to zero mean and unit variance within each block. The 45 lowest frequency DCT coefficients form the feature vector for each block and these are normalized to zero mean and unit variance per image as in [27]. In the training stage of ISV, a *universal background model* (UBM) [28] is estimated, followed by a linear subspace (160 dimensions) that models the effects of within-class variability. Enrollment of a client involves adaptation of the UBM to a client-specific *Gaussian mixture model* (GMM).

To compare a probe image to a client's model, a likelihood ratio is calculated with respect to the UBM.

3 Databases, Protocols, and Evaluation Metrics

To estimate how database dependent the tested algorithms are, all algorithms are evaluated on two recent and challenging image databases. Face recognition protocols can be divided into *identification* and *verification* protocols. The two databases that we use are accompanied by face verification protocols.

3.1 The Good, The Bad, and The Ugly Database

The Good, The Bad, and The Ugly database (GBU) [5] includes high resolution images that were taken in uncontrolled illumination conditions, but all faces in the images are frontal. The GBU database defines training sets in four different sizes. Here we used the "x8" set to train the LDA-IR and the smaller "x2" for training the UBM and the ISV subspace[3]. The GBU database provides three different face verification protocols: the Good, the Bad, and the Ugly protocol. In each protocol, models are enrolled using a single image (there exist several models per client), and pairs of probe images and models are defined, which are used to compute the *false acceptance rate* (FAR) and the *correct acceptance rate* (CAR) curves. To evaluate the results of two algorithms, the *receiver operating characteristic* (ROC) curves are compared and if a single number is required, the CAR at FAR=0.1% is reported [17]. Please note that more accurate algorithms produce higher CAR values.

3.2 The BANCA Database

The second database is the BANCA database [24]. Here we use its P protocol. The BANCA database includes medium resolution images taken under controlled and uncontrolled illumination conditions. The poses of the faces are near-frontal. The protocols of BANCA define three sets: a training set, a development set, and a test set. The training set is used to train the LDA-IR and ISV algorithms. The development and test sets are split up into images that are used for model enrollment, and probe images. Probes are compared to the enrolled models in order to compute scores.

The scores of the development set are used to compute a score threshold θ_{dev}, which is applied to the test set in order to compute the final verification result. In this paper, the threshold is set at the *equal error rate* (EER), i. e., where the FAR and the FRR[4] curves of the development set intersect. On the test set, the *half total error rate* (HTER) is reported:

$$\mathrm{HTER}_{\mathrm{test}}(\theta_{\mathrm{dev}}) = \frac{\mathrm{FAR}_{\mathrm{test}}(\theta_{\mathrm{dev}}) + \mathrm{FRR}_{\mathrm{test}}(\theta_{\mathrm{dev}})}{2} \,. \tag{1}$$

[3] Neither the LGBPHS nor the Gabor graphs algorithm needs a training phase.

[4] The *false rejection rate* (FRR) can be computed as 100% - CAR.

The EER_{dev} and $\text{HTER}_{\text{eval}}$ measures are error measures. Hence, the better the algorithm performs, the lower the values are.

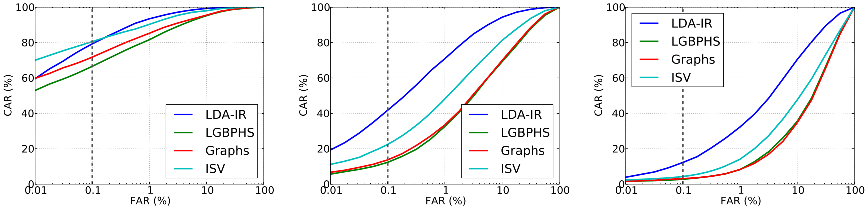
In contrast to the GBU protocols, BANCA provides several images per client to enroll the models. For enrollment using LGBPHS, the average histogram of the enrollment images is used, while for Gabor graphs and LDA-IR, all images are stored, and the maximum of the similarities of the probe image to all enrollment images is computed. For ISV, the features of all enrollment images are used to enroll the client specific ISV model.

4 Experiments

For a fair comparison of the algorithms in Section 2, we use the implementations in Bob for preprocessing, except for LDA-IR that defines its own preprocessing [19]. Further, facereclib and Bob were used throughout to compute the scores and generate the ROC curves and the EER/HTER results.

4.1 The Good, The Bad, and The Ugly

The ROC curves for the experiments on the GBU database are given in Figure 1, while the corresponding CARs at FAR=0.1% are reported in Table 1. Unsurprisingly, all algorithms work relatively well on the Good protocol, which is the most simple one. Noticeably, the ISV and LDA-IR algorithms, which made use of the training set, are better than the others, which did not. In contrast, on the Bad and Ugly protocols the algorithms perform much worse. Here, the



(a) Good (b) Bad (c) Ugly

Fig. 1. This figure shows ROC curves for the experiments on the GBU database, for (a) the Good, (b) the Bad, and (c) the Ugly protocol. The curves are displayed with a logarithmic FAR axis.

Table 1. This table details the resulting correct acceptance rates (CAR) at false acceptance rate (FAR) 0.1% on the GBU database

	LDA-IR	**LGBPHS**	**Gabor graphs**	**ISV**
Good	79.2%	66.6%	71.9%	80.5%
Bad	41.8%	12.3%	13.7%	22.5%
Ugly	12.3%	2.7%	3.1%	4.3%

LDA-IR algorithm works best, presumably since it is the only algorithm that uses a parametrization that is optimized for those protocols [19].

4.2 BANCA

The results on the BANCA database are reported in Table 2. Unlike on the GBU database, here the LDA-IR algorithm performs much worse than the other algorithms, which could be because the meta-parameters of LDA-IR are not optimized for this database, whereas, e. g., ISV was developed using BANCA in the initial work of [23]. Of the algorithms from Bob, ISV outperforms the LGBPHS and Gabor graphs slightly. Hence, in contrast to results on the GBU database, utilizing the training set of the BANCA database seems to give somewhat less of an advantage over the non-trained algorithms. It is also worth noting that ISV is the only one of the algorithms that is designed to make use of multiple enrollment images, which are used according to the BANCA protocol, in a principled way.

Table 2. This table presents the verification results of experiments performed on the BANCA database. It includes the equal error rates (EER) on the development set and the half total error rates (HTER) on the test set.

	LDA-IR	**LGBPHS**	**Gabor graphs**	**ISV**
EER	26.2%	13.2%	11.7%	10.0%
HTER	27.2%	16.1%	12.4%	10.9%

5 Conclusion

In this paper we have shown the capabilities of the Bob software library and its new satellite package facereclib to produce a fair comparison of open source face recognition systems. We used the facereclib to perform face verification experiments on the GBU and BANCA databases, using the unbiased protocols that are provided with them and implemented in Bob. It is important to note that interfaces to other image databases like FRGC [6], SCface [29], MOBIO [8], LFW [4], and AT&T [3] (to name only a few) including their fixed protocols are also available, and running experiments on these databases is as easy as changing one command line option.

In this paper, the facereclib was used to compute and compare results of three different face recognition algorithms implemented in Bob, as well as an algorithm implemented in another open source face recognition library that was integrated into the same experimental framework. The results of the algorithm comparison show that there is no one face recognition algorithm that outperforms all others, but rather that the results clearly depend on the image database and even on the protocol that is used. It also is beneficial to use meta-parameters that are optimized to a specific database. From the tested algorithms that are implemented in Bob, here the ISV algorithm performed the best overall, while both Gabor-based algorithms are approximately equal.

In this paper, we have not exhaustively shown the results of all of the face recognition algorithms that are implemented in Bob, for example we skipped the pure *eigenface* approach [30], the *Bayesian intrapersonal/extrapersonal classifier* [31,32], and *probabilistic LDA* (PLDA) [33]. Additionally, we have not tuned any parameter of any algorithm. Hence, the experiments performed in this paper surely provide a proof of concept and a good basis for observing initial trends, but they are not sufficient to judge the tested algorithms.

In future work, we will use facereclib to perform a deeper comparative analysis of algorithms that has never been done before and that is clearly needed. We will run the algorithms on more databases. We will also test different image preprocessing steps, different parameters, different feature comparison functions, and different combinations of features and algorithms. Furthermore, we will also test the impact of automatically detected faces, and the transferability of optimal parameters between image databases.

The results provided by this paper have successfully demonstrated the use of the proposed open source software framework facereclib to compare face recognition algorithms on standardized database protocols. Such a framework is critical to promote reproducible research and leads the way to deeper understanding of the state-of-the-art in this field into the future. For authors of other algorithms, Bob provides a platform to contribute new feature types, new face recognition algorithms, new database protocols, and other innovations to the face recognition community.

By the way: thanks to the facereclib and Bob, all experiments for this paper were designed, run, and evaluated in only three days.

Acknowledgements. The research leading to these results has received funding from the European Community's FP7 under grant agreements 238803 (BBfor2: bbfor2.net) and 284989 (BEAT: beat-eu.org). We want to thank the developers of Bob for their great work, and David S. Bolme for providing the source code of the CSU Face Recognition Resources.

References

1. Phillips, P., Rauss, P., Der, S.: FERET (face recognition technology) recognition algorithm development and test results. Technical report, Army Research Lab (1996)
2. Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., Zhao, D.: The CAS-PEAL large-scale Chinese face database and baseline evaluations. IEEE Transactions on Systems, Man, and Cybernetics 38, 149–161 (2008)
3. AT&T Laboratories Cambridge: AT&T database of faces (2004), `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`
4. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst (2007)

5. Phillips, P.J., Beveridge, J.R., Draper, B.A., Givens, G.H., O'Toole, A.J., Bolme, D.S., Dunlop, J.P., Lui, Y.M., Sahibzada, H., Weimer, S.: An introduction to the good, the bad, & the ugly face recognition challenge problem. In: Ninth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 346–353 (2011)
6. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Worek, W.: Preliminary face recognition grand challenge results. In: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp. 15–24 (2006)
7. Messer, K., Matas, J., Kittler, J., Luettin, J., Maître, G.: XM2VTSDB: the extended M2VTS database. In: Proceedings of the Second International Conference on Audio- and Video-Based Biometric Person Authentication (1999)
8. McCool, C., Marcel, S., Hadid, A., Pietikainen, M., Matejka, P., Cernocky, J., Poh, N., Kittler, J., Larcher, A., Levy, C., Matrouf, D., Bonastre, J.F., Tresadern, P., Cootes, T.: Bi-modal person recognition on a mobile phone: using mobile phone data. In: IEEE ICME Workshop on Hot Topics in Mobile Multimedia (2012)
9. Tan, X., Chen, S., Zhang, Z.: Face recognition from a single image per person: A survey. Pattern Recognition 39, 1725–1745 (2006)
10. Serrano, Á., de Diego, I.M., Conde, C., Cabello, E.: Recent advances in face biometrics with Gabor wavelets: A review. Pattern Recognition Letters 31, 372–381 (2010)
11. Huang, D., Member, S., Shan, C., Ardabilian, M., Wang, Y., Chen, L.: Local binary patterns and its application to facial image analysis: A survey. IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews 41, 765–781 (2011)
12. Jafri, R., Arabnia, H.R.: A survey of face recognition techniques. Journal of Information Processing Systems 5, 41–68 (2009)
13. Shen, L., Bai, L.: A review on Gabor wavelets for face recognition. Pattern Analysis and Applications 9, 273–292 (2006)
14. Beveridge, R., Bolme, D., Teixeira, M., Draper, B.: The CSU face identification evaluation system user's guide version 5.0. Technical report, Colorado State University (2003)
15. Blackburn, D., Bone, M., Phillips, P.: Face recognition vendor test 2000: evaluation report. Technical report, National Institute of Standards and Technology (2001)
16. Phillips, P., Grother, P., Micheals, R., Blackburn, D., Tabassi, E., Bone, M.: Face recognition vendor test 2002: evaluation report. Technical report, National Institute of Standards and Technology (2003)
17. Phillips, P., Scruggs, T., O'Toole, A., Flynn, P., Bowyer, K., Schott, C., Sharpe, M.: FRVT 2006 and ICE 2006 large-scale results. Technical report, National Institute of Standards and Technology (2007)
18. Anjos, A., Shafey, L.E., Wallace, R., Günther, M., McCool, C., Marcel, S.: Bob: a free signal processing and machine learning toolbox for researchers. In: 20th ACM Conference on Multimedia Systems. ACM Press (2012)
19. Beveridge, R., Bolme, D.S.: CSU Face Recognition Resources (2011), `http://www.cs.colostate.edu/facerec/algorithms/baselines2011.php`
20. Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition. In: IEEE International Conference on Computer Vision, vol. 1, pp. 786–791 (2005)
21. Wiskott, L., Fellous, J.M., Krüger, N., Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 775–779 (1997)

22. Günther, M., Haufe, D., Würtz, R.P.: Face Recognition with Disparity Corrected Gabor Phase Differences. In: Villa, A.E.P., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds.) ICANN 2012, Part I. LNCS, vol. 7552, pp. 411–418. Springer, Heidelberg (2012)
23. Wallace, R., McLaren, M., McCool, C., Marcel, S.: Inter-session variability modelling and joint factor analysis for face authentication. In: International Joint Conference on Biometrics (2011)
24. Bailly-Baillière, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., Thiran, J.P.: The BANCA Database and Evaluation Protocol. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 625–638. Springer, Heidelberg (2003)
25. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Transactions on Image Processing 19, 1635–1650 (2010)
26. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Recognition with Local Binary Patterns. In: Pajdla, T., Matas, J. (eds.) ECCV 2004, Part I. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
27. Wallace, R., McLaren, M., McCool, C., Marcel, S.: Cross-pollination of normalisation techniques from speaker to face authentication using Gaussian mixture models. IEEE Transactions on Information Forensics and Security (2012)
28. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41 (2000)
29. Grgic, M., Delac, K., Grgic, S.: SCface - surveillance cameras face database. Multimedia Tools and Applications 51, 863–879 (2011)
30. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
31. Moghaddam, B., Wahid, W., Pentland, A.: Beyond eigenfaces: Probabilistic matching for face recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 30–35 (1998)
32. Günther, M., Würtz, R.P.: Face detection and recognition using maximum likelihood classifiers on Gabor graphs. International Journal of Pattern Recognition and Artificial Intelligence 23, 433–461 (2009)
33. Prince, S.J.D.: Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of the International Conference on Computer Vision (2007)