Acknowledgment

This work was supported by the EC under contracts FP7-248984 GLOCAL and FP7-287911 LinkedTV.

Appendix A
Derivation of equations in Section II

A. Derivation of Eqs. (6) and (7)

The Gaussian mixture distribution concerning the i-th class in (5) can be derived in terms of latent variables [36], [37], as described in the following. Let $Z_i \in \mathbb{R}^{H_i}$ be a categorical latent random vector concerning the i-th class, whose parameter space $\mathcal{Z}_i$ is the standard base of $\mathbb{R}^{H_i}$, i.e., $\mathcal{Z}_i = \{\mathbf{e}_{i,1}, \dots, \mathbf{e}_{i,H_i}\}$, where only the j-th element of the unit vector $\mathbf{e}_{i,j}$ is equal to one and all other elements are equal to zero. Setting $p(Z_i = \mathbf{e}_{i,j}) = \pi_{i,j}$ and $p(\mathbf{x}|Z_i = \mathbf{e}_{i,j}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,j})$ the marginal and conditional densities, $p(\mathbf{z}_i)$ and $p(\mathbf{x}|\mathbf{z}_i)$, are expressed in terms of the mixing coefficients and mixture components respectively, $p(\mathbf{z}_i) = \prod_{j=1}^{H_i} \pi_{i,j}^{z_{i,j}}$, $p(\mathbf{x}|\mathbf{z}_i) = \prod_{j=1}^{H_i} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,j})^{z_{i,j}}$. Thus, using the product rule of probability we can express the i-th class-conditional joint density as

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}_i|\omega_i) &= p(\mathbf{z}_i|\omega_i) p(\mathbf{x}|\mathbf{z}_i, \omega_i) = p(\mathbf{z}_i) p(\mathbf{x}|\mathbf{z}_i) \\ &= \prod_{j=1}^{H_i} (\pi_{i,j} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,j}))^{z_{i,j}}, \end{aligned} \tag{52}$$

where we have used the fact that $\mathbf{x}$ is conditionally independent of ω_i given $\mathbf{z}_i$, and $\mathbf{z}_i$ is independent of ω_i. The i-th class-conditional marginal distribution of $\mathbf{x}$ can then be written as

$$p(\mathbf{x}|\omega_i) = \sum_{\mathbf{z}_i} p(\mathbf{x}, \mathbf{z}_i|\omega_i) = \sum_{j=1}^{H_i} \pi_{i,j} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,j}), \tag{53}$$

which is a Gaussian mixture equivalent to (5), and, using the Bayes' rule the posterior distribution is also derived

$$p(\mathbf{z}_i|\mathbf{x}, \omega_i) = \frac{\prod_{j=1}^{H_i} (\pi_{i,j} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,j}))^{z_{i,j}}}{\sum_{j=1}^{H_i} \pi_{i,j} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,j})}. \tag{54}$$

Therefore, under the i.i.d. assumption, the likelihood of the complete data set is expressed as (p.108, [27])

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) &= \prod_{i=1}^{C} \prod_{n=1}^{N_i} p(\mathbf{x}_i^n, \mathbf{z}_i^n|\omega_i) \\ &= \prod_{i=1}^{C} \prod_{n=1}^{N_i} \prod_{j=1}^{H_i} (\pi_{i,j} \mathcal{N}(\mathbf{x}_i^n|\boldsymbol{\mu}_{i,j}))^{z_{i,j}^n}. \end{aligned} \tag{55}$$

while the posterior distribution takes the form

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \propto \prod_{i=1}^{C} \prod_{n=1}^{N_i} \prod_{j=1}^{H_i} (\pi_{i,j} \mathcal{N}(\mathbf{x}_i^n|\boldsymbol{\mu}_{i,j}))^{z_{i,j}^n}, \tag{56}$$

where $\mathbf{Z} = \{\mathbf{Z}_1, \dots, \mathbf{Z}_C\}$ is the set of all categorical vectors. Observing that the posterior distribution is independent over $z_{i,j}^n$, the expectation of the categorical variables can be derived

$$\mathbb{E}[z_{i,j}^n] = \frac{\sum_{j=1}^{H_i} z_{i,j}^n (\pi_{i,j} \mathcal{N}(\mathbf{x}_i^n|\boldsymbol{\mu}_{i,j}))^{z_{i,j}^n}}{\sum_{j=1}^{H_i} \pi_{i,j} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,j})}, \tag{57}$$

and simplifying the above, we arrive to the definition of the responsibilities in (7).

Moreover, from (56) the log likelihood of the complete data set is retrieved

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \sum_{i=1}^{C} \sum_{n=1}^{N_i} \sum_{j=1}^{H_i} z_{i,j}^n (\ln \pi_{i,j} + \ln \mathcal{N}(\mathbf{x}_i^n|\boldsymbol{\mu}_{i,j})). \tag{58}$$

Applying the expectation operator to the above expression and substituting $\mathbb{E}[z_{i,j,n}]$ from (7) the expectation of the complete data log-likelihood is expressed as

$$\begin{aligned} &\mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] \\ &= \sum_{i=1}^{C} \sum_{n=1}^{N_i} \sum_{j=1}^{H_i} h_{i,j}^n (\ln \pi_{i,j} + \ln \mathcal{N}(\mathbf{x}_{i,n}|\boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma})) \\ &= \sum_{i=1}^{C} \sum_{j=1}^{H_i} \tilde{N}_{i,j} \ln \pi_{i,j} - \frac{NF}{2} \ln(2\pi) + \frac{N}{2} \ln |\boldsymbol{\Sigma}^{-1}| \\ &- \frac{1}{2} \sum_{i=1}^{C} \sum_{n=1}^{N_i} \sum_{j=1}^{H_i} h_{i,j,n} (\mathbf{x}_{i,n} - \boldsymbol{\mu}_{i,j})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{i,n} - \boldsymbol{\mu}_{i,j}). \end{aligned} \tag{59}$$

Using the identity $(\mathbf{x}_i^n - \boldsymbol{\mu}_{i,j})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i^n - \boldsymbol{\mu}_{i,j}) = (\mathbf{x}_i^n - \bar{\mathbf{x}}_i^n)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i^n - \bar{\mathbf{x}}_{i,j}) + (\bar{\mathbf{x}}_{i,j} - \boldsymbol{\mu}_{i,j})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_{i,j} - \boldsymbol{\mu}_{i,j}) + 2(\mathbf{x}_i^n - \bar{\mathbf{x}}_{i,j})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_{i,j} - \boldsymbol{\mu}_{i,j})$ along with the fact that $\sum_{n=1}^{N_i} (\mathbf{x}_i^n - \bar{\mathbf{x}}_{i,j})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_{i,j} - \boldsymbol{\mu}_{i,j}) = 0$, and multiplying both sides by two, we arrive to (6).

B. Derivation of Eq. (18)

The constraint that the mixing coefficients should sum to one can be incorporated in (17) using C lagrange multipliers $\eta_i, i = 1, \dots, C$. Therefore, we need to find the stationary point of

$$\begin{aligned} &\sum_{i=1}^{C} \sum_{n=1}^{N_i} \sum_{j=1}^{H_i} h_{i,j}^n (\ln \pi_{i,j} + \ln \mathcal{N}(\mathbf{x}_i^n|\boldsymbol{\mu}_{i,j})) \\ &\qquad + \sum_{i=1}^{C} \eta_i (\sum_{j=1}^{H_i} \pi_{i,j} - 1) \end{aligned} \tag{60}$$

with respect to $\pi_{i,j}$ and η_i. Optimizing over $\pi_{i,j}$ we arrive to $\tilde{N}_{i,j}/\pi_{i,j} + \eta_i = 0$. If we multiply both sides with $\pi_{i,j}$ and sum over all subclasses of the i-th class we get $\eta_i = -N_i$. Eliminating η_i we obtain (18).

References

[1] Y. Ephraim and H. L. V. Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[2] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[3] R. Duda, P. Hart, and D. Stork, *Pattern Classification, (2nd ed.)*. New York, USA: John Wiley & Sons, Inc., 2001.

[4] S. Ji and J. Ye, “Generalized linear discriminant analysis: A unified framework and efficient model selection,” *IEEE Trans. Neural Netw.*, vol. 19, no. 10, pp. 1768–1782, Oct. 2008.

[5] C. B. Moler and G. W. Stewart, “An algorithm for generalized matrix eigenvalue problems,” *SIAM Journal on Numerical Analysis*, vol. 10, no. 2, pp. 241–256, Apr. 1973.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[7] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Oct. 2006.
[8] C. S. Dhir and S.-Y. Lee, "Discriminant independent component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 6, pp. 845–857, Jun. 2011.
[9] K. Muller, S. Mika, G. Ratsch, S. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–202, Mar. 2001.
[10] L. Wang, K. L. Chan, P. Xue, and L. Zhou, "A kernel-induced space selection approach to model selection in KLDA," *IEEE Trans. Neural Netw.*, vol. 19, no. 12, pp. 2116–2131, Dec. 2008.
[11] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 755–761, Apr. 2009.
[12] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 117–126, Jan. 2003.
[13] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. and Learning Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.
[14] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1119–1132, Jul. 2011.
[15] B.-C. Kuo and K.-Y. Chang, "Feature extractions for small sample size classification problem," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 756–764, Mar. 2007.
[16] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 155–176, Jul. 1996.
[17] M. Zhu and A. Martinez, "Subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1274–1286, Aug. 2006.
[18] S.-W. Kim and R. P. W. Duin, "On using a pre-clustering technique to optimize LDA-based classifiers for appearance-based face recognition," in *Proc. 12th Iberoamerican Congress on Pattern Recognition*, Vina del Mar-Valparaiso, Chile, Nov. 2007, pp. 466–476.
[19] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Face detection using multimodal density models," *Computer Vision and Image Understanding*, vol. 84, no. 2, pp. 264–284, Oct. 2001.
[20] A. Pnevmatikakis and L. Polymenakos, "Subclass linear discriminant analysis for video-based face recognition," *Journal of Visual Communication and Image Representation*, vol. 20, no. 8, pp. 543–551, Nov. 2009.
[21] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, Nov. 2011.
[22] D. Wu and K. L. Boyer, "Resilient subclass discriminant analysis," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV 2009)*, Kyoto, Japan, Sep./Oct. 2009, pp. 389–396.
[23] F. Oveisi, "Subclass discriminant analysis using dynamic cluster formation for EEG-based brain-computer interface," in *Proc. IEEE/EMBS 4th Int. Conf. on Neural Engineering*, Antalya, Turkey, May 2009, pp. 303–306.
[24] S.-W. Kim, "A pre-clustering technique for optimizing subclass discriminant analysis," *Pattern Recogn. Lett.*, vol. 31, no. 6, pp. 462–468, Apr. 2010.
[25] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "High-level event detection in video exploiting discriminant concepts," in *Proc. 9th International Workshop on Content-Based Multimedia Indexing (CBMI 2011)*, Madrid, Spain, Jun. 2011, pp. 85–90.
[26] ——, "Mixture subclass discriminant analysis," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 319–332, May 2011.
[27] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate analysis*. Academic Press, 1979.
[28] N. A. Campbell, "Canonical variate analysis - A general model formulation," *Australian & New Zealand Journal of Statistics*, vol. 26, no. 1, pp. 86–96, 1984.
[29] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.
[30] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 623–627, Jun. 2000.
[31] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 762–766, Jul. 2001.
[32] B. Chen, L. Yuan, H. Liu, and Z. Bao, "Kernel subclass discriminant analysis," *Neurocomputing*, vol. 71, no. 1–3, pp. 455–458, Dec. 2007.
[33] D. You, O. C. Hamsici, and A. M. Martinez, "Kernel optimization in discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 631–638, Mar. 2011.
[34] V. Vapnik, *Statistical learning theory*. New York: Willey, 1998.
[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.
[36] G. Mclachlan and D. Peel, *Finite Mixture Models*. New Work: Wiley-Interscience, 2000.
[37] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
[38] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York, USA: Springer, 1996.
[39] N. Vlassis and A. Likas, "A kurtosis-based dynamic approach to gaussian mixture modeling," *IEEE Trans. Syst., Man, Cybern. A*, vol. 29, no. 4, pp. 393–399, Jul. 1999.
[40] L. Wang and J. Ma, "A kurtosis and skewness based criterion for model selection on gaussian mixture," in *2nd Int. Conf. on BioMedical Engineering and Informatics*, Tianjin, China, Oct. 2009, pp. 1–5.
[41] I. T. Jolliffe, *Principal Component Analysis*. New York, USA: Springer, Oct. 2002.
[42] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml
[43] "Gunnar raetsch's benchmark datasets," http://theoval.cmp.uea.ac.uk/~gcc/matlab/default.html#benchmarks, accessed 2012-05-01.
[44] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, Madison, WI, USA, Jun. 2003, pp. II–409–15.
[45] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "Automatic event-based indexing of multimedia content using a joint content-event model," in *ACM Multimedia 2010, (EiMM10)*, Firenze, Italy, Oct. 2010.
[46] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," in *Face Recognition: From Theory to Applications, Computer and Systems Sciences*, H. Wechsler et al., Ed. NATO ASI Series F, 1998, vol. 163, pp. 446–456.
[47] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota FL, USA, Dec. 1994, pp. 138–142.
[48] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.
[49] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
[50] D. Cai, X. He, Y. Hu, J. Han, and T. S. Huang, "Learning a spatially smooth subspace for face recognition," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, Minneapolis, Minnesota, USA, Jun. 2007, pp. 138–142.
[51] "Four face databases in matlab format," http://http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html, accessed 2012-05-01.
[52] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for adaboost," *Mach. Learn.*, vol. 42, no. 3, pp. 287–320, Mar. 2001.
[53] C. E. Thomaz, D. F. Gillies, and R. Q. Feitosa, "A new covariance estimate for bayesian classifiers in biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 214–223, Feb. 2004.
[54] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947.
[55] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.
[56] Y. Aksu, D. J. Miller, G. Kesidis, and Q. X. Yang, "Margin-maximizing feature elimination methods for linear and nonlinear kernel-based discriminant functions," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 701–717, May 2010.
[57] F. Song, D. Mei, and H. Li, "Feature selection based on linear discriminant analysis," in *IEEE Int. Conf. Intell. Syst. Design and Eng. Appl.*, vol. 1, Changsha, China, Oct. 2010, pp. 746–749.
[58] S. Huh and D. Lee, "Linear discriminant analysis for signatures," *IEEE Trans. Neural Netw.*, vol. 21, no. 12, pp. 1990–1996, Dec. 2010.