

Maximum-Likelihood Dimensionality Reduction in Gaussian Mixture Models with an Application to Object Classification

Massimo Piccardi, Hatice Gunes, Ahmed Fawzi Otoom

*Faculty of Engineering and Information Technology, University of Technology, Sydney (UTS),
Sydney, Australia*

E-mail: {massimo, haticeg, afaotoom}@it.uts.edu.au

Abstract

Accurate classification of objects of interest for video surveillance is difficult due to occlusions, deformations and variable views/illumination. The adopted feature sets tend to overcome these issues by including many and complementary features; however, their large dimensionality poses an intrinsic challenge to the classification task. In this paper, we present a novel technique providing maximum-likelihood dimensionality reduction in Gaussian mixture models for classification. The technique, called hereafter mixture of maximum-likelihood normalized projections (mixture of ML-NP), was used in this work to classify a 44-dimensional data set into 4 classes (bag, trolley, single person, group of people). The accuracy achieved on an independent test set is 98% vs. 80% of the runner-up (MultiBoost/AdaBoost).

1. Introduction

Classification of objects of interest, in particular unattended stationary objects, is very important for video surveillance to assess potential security threats [1-5]. Classification is challenging due to occlusions, deformations of non-rigid objects (such as stationary humans), variable views and illumination, and the high intra-class variability of certain classes (such as pieces of luggage). Approaches based on simple feature sets such as form factors or basic size measurements are deemed to fail in any realistic scenarios. The solution that is most often explored by video surveillance researchers is that of choosing a feature set containing many, partially redundant features of different nature such as global and local shape descriptors, texture measurements, perspective-compensated sizes, and more (e.g. [4]). Consequently, feature sets tend to be large, in the order of tens or hundreds of features, and

pose an inherent dimensionality challenge to the classification step.

A viable approach to mitigate this curse of dimensionality is the use of dimensionality reduction techniques such as the popular Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [6]. PCA is an orthogonal projection of the data onto a lower dimensionality linear space where the variance of the projected data is maximized. LDA exploits class information, maximizing the ratio between between-class scatter and within-class scatter. Non-linear methods include Kernel PCA (KPCA) [7] and Nonlinear Independent Component Analysis (NL-ICA) [8]. Other sophisticated methods have also been proposed including mixtures of local PCA [9], Probabilistic PCA [6] and mixtures of Probabilistic PCA [10].

In a typical maximum-likelihood classification approach, a training set is used to learn probability density functions (pdf) adequately describing the class-conditional likelihoods. Often, a class cannot be well represented by a single, compact mode and the need arises for a finite mixture model that can approximate more complex distributions [11]. One of the most well-known finite mixture models is the Gaussian mixture model (GMM), where a number of Gaussian components are combined together to provide a multimodal density. A widely applied method for estimating the GMM parameters is the Expectation-Maximization algorithm (EM) [11]. GMM is implemented widely for classification problems in various applications (e.g. [9,10,11]).

According to the results presented in [9] and [10], feature reduction and GMM may be combined to achieve more accurate classification. [9] proposes to combine a set of local PCAs, each describing one mode of the desired density in a compressed space. [10] proposes the use of a mixture of Probabilistic Principal Component Analyzers and derives efficient variants of the EM algorithm for their learning. The

main idea is that Probabilistic PCA, unlike basic PCA, offers an explicit density model for which likelihood can be formally maximized. Probabilistic PCA is based on the assumption that a noise component of spherical Gaussian density exists in the mapping from the original to the compressed space. If such a noise component tends to zero, the framework returns the conventional PCA, together with an undesirable singular solution for the maximum likelihood.

Differently from the aforementioned approach [10], in this paper we propose a new method for maximum-likelihood dimensionality reduction within GMM, called mixture of maximum-likelihood normalized projections (mixture of ML-NP) hereafter. Our model is explicitly maintained in the compressed space and assumes a deterministic transformation between the original and the compressed spaces. In order to prevent the singular solution, a normalization constraint is imposed on the transformation matrix at each iteration of the learning algorithm. The re-estimation formula for the transformation matrix we originally derive guarantees maximum likelihood like for the other model's parameters. Differently from other local PCA approaches, our method does not require hard clustering and retains the elegance and accuracy of the conventional EM.

2. Mixture of maximum-likelihood normalized projections

In this section, we describe our method, the mixture of maximum-likelihood normalized projections, for dimensionality reduction within GMM. Our goal is to project the data at each iteration of the EM algorithm onto linearly reduced dimensions in a way that maximizes the likelihood of the model given the data. After performing the projection, the remaining GMM parameters are learned in the compressed space. In the following, we derive the Ω_l matrix providing the transformation for each l -th component, $l = 1..M$, in the GMM. We recall that a GMM is defined as:

$$p(z) = \sum_{l=1}^M \alpha_l G(z | \mu_l, \Sigma_l) \quad (1)$$

where the $p(z)$ density is provided by the mixture of M normally-distributed components of parameters: weights, α_l ; means, μ_l ; and covariance matrices, Σ_l ; $l = 1..M$.

We then consider a set of i.i.d. observations, $\{z_i\}_{i=1..N}$, in a high-dimensional space with P dimensions. For each z_i , we pose $x_{li} = \Omega_l z_i$, where Ω_l is a $D \times P$ real matrix with $D \ll P$ achieving the desired dimensionality reduction. The projected data, $\{x_{li}\}_{i=1..N}$, differ for each component. Our problem is

then that of finding values for Ω_l maximizing the likelihood. To this aim, we write Ω_l as $P, D \times 1$ column vectors, $\Omega_l = [\omega_{lj}]_{j=1..P}$, and consider the auxiliary function:

$$\begin{aligned} Q(\theta | \theta^g) &= \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | x_{li}, \theta^g) + \\ &+ \sum_{l=1}^M \sum_{i=1}^N \log(G(x_{li} | \mu_l, \Sigma_l)) p(l | x_{li}, \theta^g) \end{aligned} \quad (2)$$

with θ and θ^g representing, respectively, the new and old model's parameters. In the EM framework, (2) represents the expectation of the joint log-likelihood of the data and their latent variables and increasing it guarantees to increase the data likelihood. We ignore the first term as it does not depend on x_{li} and re-write the second term as:

$$\begin{aligned} &\sum_{l=1}^M \sum_{i=1}^N \log(G(\Omega_l z_i | \mu_l, \Sigma_l)) p(l | x_{li}, \theta^g) = \\ &= \sum_{l=1}^M \sum_{i=1}^N \left(-\frac{1}{2} \log(|\Sigma_l|) - \frac{1}{2} (\omega_{l1} z_{i1} + \dots + \omega_{lP} z_{iP} - \mu_l)^T \right. \\ &\quad \left. \times \Sigma_l^{-1} (\omega_{l1} z_{i1} + \dots + \omega_{lP} z_{iP} - \mu_l) \right) \times p(l | x_{li}, \theta^g) \end{aligned} \quad (3)$$

We differentiate (3) with respect to each ω_{lj} and equate to zero to obtain the solution for the ω_{lj} . In view of that, we can ignore all multiplicative constants and terms that do not depend on ω_{lj} and write:

$$\sum_{i=1}^N \left((\omega_{l1} z_{i1} + \dots + \omega_{lP} z_{iP} - \mu_l)^T \right) p(l | x_{li}, \theta^g) \quad (4)$$

since ω_{lj} does not appear in $M - 1$ terms of the external sum. We then differentiate (4) in ω_{lj} and obtain:

$$\frac{\partial(4)}{\partial \omega_{lj}} = \sum_{i=1}^N (2 \Sigma_l^{-1} (\omega_{l1} z_{i1} + \dots + \omega_{lP} z_{iP} - \mu_l)) z_{ij} p(l | x_{li}, \theta^g) \quad (5)$$

Let us now equate (5) to zero and manipulate it to, for instance, extract ω_{l1} :

$$\begin{aligned} &\sum_{i=1}^N \omega_{l1} z_{i1}^2 p(l | x_{li}, \theta^g) + \\ &+ \sum_{i=1}^N (\omega_{l2} z_{i2} + \dots + \omega_{lP} z_{iP} - \mu_l) z_{i1} p(l | x_{li}, \theta^g) = 0 \\ &\rightarrow \omega_{l1} \sum_{i=1}^N z_{i1}^2 p(l | x_{li}, \theta^g) = \\ &= \sum_{i=1}^N (-\omega_{l2} z_{i2} - \dots - \omega_{lP} z_{iP} + \mu_l) z_{i1} p(l | x_{li}, \theta^g) \end{aligned}$$

$$\rightarrow \omega_l = \frac{\sum_{i=1}^N (-\omega_{l2} z_{i2} - \dots - \omega_{lp} z_{ip} + \mu_l) z_{i1} p(l | x_{i1}, \Theta^g)}{\sum_{i=1}^N z_{i1}^2 p(l | x_{i1}, \Theta^g)} \quad (6)$$

Eq. (6) gives the desired re-estimation formula for ω_l ; similarly, we derive the re-estimation formulas for the remaining column vectors. As the following step, we normalize the updated value of Ω_l by its Frobenius norm and use it for compressing the data in the current iteration. The re-estimation formulas for the other GMM parameters are as in the conventional EM [11]:

$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^N p(l | x_{i1}, \Theta^g) \quad (7)$$

$$\mu_l^{new} = \frac{\sum_{i=1}^N x_{i1} p(l | x_{i1}, \Theta^g)}{\sum_{i=1}^N p(l | x_{i1}, \Theta^g)} \quad (8)$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^N (x_{i1} - \mu_l^{new})(x_{i1} - \mu_l^{new})^T p(l | x_{i1}, \Theta^g)}{\sum_{i=1}^N p(l | x_{i1}, \Theta^g)} \quad (9)$$

From a computational point of view, we can compare the computational complexity of the model's evaluation (1) for a GMM in the original space and for our mixture of ML-NP. Given a generic datum, z , the GMM evaluation is $O(MP^2)$ in the case of a full covariance matrix, while the evaluation of the mixture of ML-NP requires $O(MDP)$ for compression and $O(MD^2)$ for the actual evaluation, only linear overall in the higher dimensionality, P . Therefore, the cost for generating the reduced-dimensionality data is more than compensated for by the faster computation of the Gaussian exponent.

3. Experiments and Analysis

Experiments were conducted on a video surveillance data set in order to evaluate the performance of the proposed method against that of state-of-the-art classifiers. For this purpose, we clipped 600 images of trolleys, bags, single persons, and groups of people from video footage acquired at a number of airports world-wide. Objects of interest in these images appear under different viewpoints, illumination conditions and scale, and cluttered with background and other objects. The feature set is described in detail in [12]; in brief, it consists of statistics of various local features such as line segments, circles, corners, and global shape descriptors such as fitted ellipses and bounding boxes. We then divided the images into two data sets: a training set (400 images) and a test set (200 images), with equal

number of images for each class so as to have equal class priors. The performance of the various classifiers is evaluated here in terms of classification accuracy (or detection rate, or recall) for each class and overall. The accuracy is simply calculated as the percentage of the number of objects correctly detected against the total number of objects. Given that the number of test cases is known and is the same for all classifiers, the overall precision (or false alarm rate) is complementary to the overall recall.

Firstly, we trained and tested with a number of well-known classifiers such as the Bayesian-based classifier BayesNet (BN), C4.5 Decision Trees, Support Vector Machine (SVM) with Sequential Minimal Optimization, AdaBoost (AB.M1), and MultiBoost/AdaBoost (i.e., MB, a variant of AdaBoost combining wagging and boosting) [13].

Secondly, we used a GMM trained on the original data i.e. without dimensionality reduction. To this aim, we trained a separate GMM for each class with 100 training samples by using the standard Expectation-Maximization algorithm (EM) with soft clustering. The number of components, M , was set manually to two for each mixture. We then classified the test samples for each class using the estimated parameters of the model and assigning the sample to the class providing the highest log-likelihood.

Thirdly, we reduced the dimensions of the training and testing data (i.e. down from $P = 44$ features to $D = 3$ features) by learning the PCA sub-space from the overall training set. We then trained a GMM in the three-dimensional sub-space (GMM-PCA) and used it for classification.

Eventually, we applied the method proposed in this paper (mixture of ML-NP) for classification. Again, we used 100 training and 50 testing samples for each class, $D = 3$ as number of reduced dimensions and $M = 2$ as number of components per class. The classification results of classification are reported in Table 1.

Table 1 shows that the highest accuracy is achieved by the proposed approach at 98.0% overall. The gap with respect to the runner-up is huge, with 18.5% accuracy improvement, and is, in a way, impressive and beyond our expectations. We wish to point out that experiments with the other classifiers were carried out in the most genuine way by setting any tunable parameters to achieve the highest performance. The performance for the GMM with the original data (65.5%) is worse than for the other classifiers, proving that the high data dimensionality and the consequent model's parameterization challenge its accuracy. However, reducing the dimensionality by a simple technique such as PCA that is not informed by an explicit density model even leads to decreased

accuracy (52.5%). Although the reported results are based on parameters $D = 3$ and $M = 2$, we have experimented with other values, confirming the accuracy behaviors.

Table 1. Accuracy/recall (%) for each class across the tested classifiers.

| classifier | trolley | person | bag | group | overall |
|---------------|---------|--------|-----|-------|-------------|
| C4.5 | 62 | 82 | 92 | 56 | 73.0 |
| BN | 68 | 78 | 92 | 50 | 72.0 |
| SVM | 72 | 84 | 88 | 56 | 75.0 |
| AB.M1 | 74 | 74 | 96 | 70 | 78.5 |
| MB | 80 | 86 | 96 | 56 | 79.5 |
| GMM | 46 | 72 | 80 | 64 | 65.5 |
| GMM-PCA | 52 | 40 | 92 | 26 | 52.5 |
| Mixture ML-NP | 92 | 100 | 100 | 100 | 98.0 |

4. Conclusion

In this paper, we have described a novel maximum-likelihood dimensionality reduction approach for GMM and its application to object classification in video surveillance. Dimensionality reduction is provided by a linear transformation learned through an EM algorithm in a similar way to that of the other GMM parameters. By interlacing dimensionality reduction and parameter estimation at the iteration level, the proposed method proves capable of providing a combined optimal solution. Singular solutions are prevented by normalizing the estimated linear transformation by its Frobenius norm after every update. When trained and tested on a classification task of video surveillance objects, the proposed method achieved an accuracy of 98.0% against the 79.5% of the runner-up (MultiBoost/AdaBoost) with an improvement of above 18%.

The classifier presented in this paper is meant to become an integral part of a video surveillance system capable of classifying unattended stationary objects for security of crowded areas such as railway stations and airport terminals. However, thanks to the general nature of its design, it promises accurate classification of high-dimensional data also in other domains.

5. Acknowledgments

This research is partially supported by the Australian Research Council and iOmniscient Pty Ltd under the ARC Linkage Project Grant Scheme 2006 - LP0668325.

6. References

- [1] M. D. Beynon, D. J. van Hook, M. Seibert, A., Peacock, D., Dungeon, "Detecting abandoned packages in a multi-camera video surveillance system", *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance*, Florida, USA, 221-228, 2003.
- [2] L.M. Brown, "View independent vehicle/person classification", *Proc. of ACM WS on Video Surveillance & Sensor Networks*, New York, USA, 114-123, 2004.
- [3] M. Tsuchiya, H. Fujiyoshi, "Evaluating feature importance for object classification in visual surveillance", *Proc. of IEEE Int. Conf. on Pattern Recognition*, 978-981, Hong Kong, China, 2006.
- [4] S. Ferrando, G. Gera, C. Regazzoni, "Classification of unattended and stolen objects in video-surveillance system", *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Sydney, Australia, 21:6 pp, 2006.
- [5] S. Lu, J. Zhang., D. D. Feng, "Detecting unattended packages through human activity recognition and object association", *Pattern Recognition*, 40(8):2173-2184, 2007.
- [6] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
- [7] B. Scholkopf, A. Smola, K. R. Muller, "Non linear component analysis as a kernel eigenvalue problem", *Neural Computation*, 10: 1299-1319, 1998.
- [8] A. Hyvarinen and P. Pajunen, "Nonlinear independent component analysis: existence and uniqueness results", *Neural Networks*, 12(3): 429-439, 1999.
- [9] H.-C. Kim, D. Kim, S. Y. Bang, "A numeral character recognition using the PCA mixture model", *Pattern Recognition Letters*, 23(1-3):103-111, 2002.
- [10] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, 11(2): 443-482, 1999.
- [11] P. Paalanen, J.-K. Kamarainen, J. Ilonen, H. Kälviäinen, "Feature representation and discrimination based on Gaussian mixture model probability densities - practices and algorithms", *Pattern Recognition*, 39(7): 1346-1358, 2006.
- [12] A. F. Otoom, H. Gunes, and M. Piccardi, "Towards automatic abandoned object classification in visual surveillance systems", *Proc. of Asia-Pacific WS on Visual Information Processing*, Tainan, Taiwan, 143-149, 2007.
- [13] G. I. Webb, "MultiBoosting: a technique for combining boosting and wagging", *Machine Learning*, 40(2):159-196, 2000.