LEARNING PROBABILISTIC DISTRIBUTION MODEL FOR MULTI-VIEW FACE DETECTION

Lie Gu, Stan Z Li, Hong-Jiang Zhang

Microsoft Research China, Beijing, China, 100080
Email: {liegu, szli, hjzhang}@microsoft.com

ABSTRACT

Modeling subspaces of a distribution of interest in high dimensional spaces is a challenging problem in pattern analysis. In this paper, we present a novel framework for pose invariant face detection through multi-view face distribution modeling. The approach is aimed to learn a set of low-dimensional subspaces from an originally nonlinear distribution by using the mixtures of probabilistic PCA [16]. From the experiments, we found the learned PPCA models are of low dimensionality and exhibit high local linearity, and consequently offer an efficient representation for visual recognition. The model is then used to extract features and select "representative" negative training samples. Multi-view face detection is performed in the derived feature space by classifying each face into one of the view classes or into the nonface class, by using a multi-class SVM array classifier. The classification results from each view are fused together and yields the final classification results. The experimental results demonstrate the performance superiority of our proposed framework while performing multi-view face detection.

1. INTRODUCTION

Face images are commonly represented as high dimensional data points in some feature space. Despite the high dimensionality, the common spatial structure of faces implies that the data may lie in a lower-dimensional manifold. In recent years, a growing research interest is focused on statistically identify and parameterize the low-dimensional manifold of such data with an objective to provide a meaningful representation while performing visual recognition tasks. However, face appearance in nature scenes varies drastically with changes in viewpoint, illumination conditions, and facial shapes. Such variations cause the face distribution to be highly nonlinear and complex in the original image space [1] [15].

To address this problem, a natural treatment is to divide face images into several subsets according to view angles and model each view subspace respectively [2]. Such view-based scheme is preferred because it is avoided to explicitly establish 3D model from images or appearances of the objects from possible views, which often tends to be a more complicate problem. Furthermore, the learning of distribution is simplified to be performed only within a fixed view class. In this paper, we propose a novel framework which follows the view-based scheme to learn nonlinearly distributed view subspaces via the PPCA mixtures model, and apply the learned representation to face detection.

Learning a subspace representation usually involves deriving a set of basis components from training data. As a powerful technique for data reduction, principal component analysis has been applied to face analysis [3, 4]. A subspace spanned by a set of eigen-pictures is generated and used to encode face images. This method leads to a significant improvement in the performance of both face recognition and detection because it provides a compact description of face appearance and automatically identifies the degrees-of-freedom of the underlying statistical variability. However, as a global linear model, PCA assumes that the data distribution is single Gaussian, which is not the case in many real-world applications. To obtain a better description of face variations, we may consider one of the following two methods: 1) Approximate the face distribution with a mixture of local linear subspace model [5] [6] [7] [8]. 2) Nonlinear subspace analysis algorithms, such as principle curves [9], splines [10] and kernel principal analysis [11] may be used for analyzing face manifolds.

The basic assumption of linear subspace mixtures is that the intrinsic structure of highly complex data can be captured by a set of local linear model. Sung and Poggio [5] modeled the distribution of frontal face and nonface with twelve Gaussian clusters (six for each) where the first 75 eigenvectors of each cluster span one subspace. Two distances, DIFS and DFFS computed from training patterns and clusters are used to train a MLP classifier. Moghaddam and Pentland [6] proposed a density estimation technique derived from PCA algorithm. The target density is divided into two uncorrelated parts: the densities in principal subspace and its orthogonal complement. The two parts were multiplied together to give a complete evalua-

tion of the likelihood. The derived density is subsequently used for view-based face recognition and frontal face detection. The authors argue that the introduction of density in orthogonal supplementary subspace can remove a great number of false alarms. Brendan and Huang [7] applied a mixture of factor analysis model to face recognition and demonstrated it outperforms traditional PCA methods.

In [21], a nonlinear kernel machine based approach is presented for learning such nonlinear mappings. The aim is to provide an effective view-based representation for multi-view face detection and pose estimation. Assuming that the view is partitioned into a number of distinct ranges, one nonlinear view-subspace is learned for each (range of) view from a set of example face images of that view (range), by using KPCA. Projections of the data onto the view-subspaces are then computed as view-based nonlinear features. Multi-view face detection and pose estimation are performed by classifying a face into one of the facial views or into the nonface class, by using a multi-class kernel support vector classifier (KSVC).

In this paper, we address the problem of multi-view face detection from three aspects. First, the principle manifolds of target object (multi-view faces, in our case) are learnt under a maximum likelihood framework by fitting training examples into a mixture of probabilistic subspaces. To be specific, a collection of view specific PPCA mixture models are trained to capture the "intrinsic" dimensionality of face manifolds, in contrast to Stan Z Li et al. 's work [21] where nonlinear KPCA was used to achieve this. Given an input sample, its posterior probability density can be robustly evaluated under a group of constrained Gaussian clusters; Second, a set of non-positive training patterns, which is generally hard to be directly modeled, are collected in the light of the robust density estimation offered by the probabilistic PCA model. An efficient process to select a "representative" non-positive training set is introduced and demonstrated to be more efficient than common bootstrapping methods; finally, the projections onto each subspace and the local reconstruction errors are combined as feature vectors to train a SVM array classifier. The outputs of each SVM are fused together in the last stage to give the final decision.

The rest of the paper is organized as follows: Section 2 introduces the formulation of PPCA and PPCA mixture model. We present the details of our proposed framework in section 3. Section 4 provides the experimental results and the conclusions are drawn in section 5.

2. MODELS FOR DISTRIBUTION ANALYSIS

Probabilistic PCA is a density estimation technique which is well grounded in the theory of factor analysis and latent variable models. The major advantage of PPCA is that it offers a robust likelihood measure and simultaneously provides an efficient computation procedure derived from a Gaussian latent variable model.

2.1 Probabilistic Formulation of PCA

Tipping and Bishop developed PPCA model [16] [17] by reformulating PCA as a maximum likelihood solution of a specific form of latent variable model. The d-dimensional observed variable t is related to the q-dimensional latent variable x by a linear mapping as $t = Wx + \mu + \varepsilon$, where W is a $d \times q$ weight matrix and ε is an isotropic Gaussian noise. We can compute the probability distribution over t for a given x as

$$p(t \mid x) = (2\pi\sigma^2)^{-d/2} \exp\{-\frac{1}{2\sigma^2}\|t - Wx - \mu\|^2\}$$

If assuming x is independently distributed Gaussian with unit variance $p(x) \sim N(0, I_q)$, we can obtain the likelihood of t by integrating over x,

$$p(t) = (2\pi)^{-d/2} \mid C \mid^{-1/2} \exp\{-\frac{1}{2}(t-\mu)^T C^{-1}(t-\mu)\}$$

which is also a normal distribution with covariance matrix

$$C = o^2 I + WW^T$$

The problem here is to find the optimal weight matrix W to maximize the log-likelihood of observed data.

With the parameterized probability model of the data, an EM based maximum likelihood approach can be naturally employed to solve this problem. Considering a "complete" data set consisted of both t and x, we derive the *complete-data log-likelihood* as $p(t,x) = p(x \mid t)p(t)$. The objective of maximizing data likelihood can be achieved by an iterative two-step procedure of maximizing the expectation of $p(t,x)$: a) First calculate the expectation of $p(t,x)$ with respect to the posterior distribution $p(x \mid t, W, o^2)$ over x; b) Find the new parameters W and σ^2 which maximize the $p(t,x)$.

The log-likelihood (2) of observed data is maximized while the weight matrix W and the noise variance σ^2 take the form of

$$W_{ML} = U_q(\Lambda_q - \sigma^2 I)^{1/2} R \qquad \sigma^2_{ML} = \frac{1}{d-q}\sum_{j=q+1}^{d} \lambda_j$$

Where the column vectors of matrix U_q are first q eigenvectors of the data covariance matrix, with corresponding eigenvalues in the diagonal matrix Λ_q, R is an arbitrary orthogonal rotation matrix. Given a new data point t_{new}, its posterior distribution $p(x_{new} \mid t_{new})$ can be easily obtained by

using Bayes' rule and its projection in latent variable space can be computed as the mean of the distribution.

2.2 Mixtures of Probabilistic PCA Model

The probabilistic formulation of PCA offers a graceful extension to model complex data structures with a mixture of local PPCA models. Again, all parameters of the model, i.e. the mean μ_i and the variance Σ_i of each local modal, the noise variance σ^2 can be determined within an EM framework while the likelihood is maximized. For a given data point t, its likelihood is given by a weighted sum of M probability density $p(t \mid i)$ associated with corresponding local model i:

$$p(t) = \sum_{i=1}^{M} \pi_i \, p(t \mid i) ,$$

where π_i is the mixing proportion of component i, with $\sum_{i=1}^{M} \pi_i = 1$.

3. DISTRIBUTION MODELLING AND FACE DETECTION

Assume that a set of view-labeled training face are provided for learning; see fig. 1 for some examples. All left in-depth rotated face are mirrored to right side so that each view angle is between 0° and 90°. The pose is quantized into 10 discrete values:

$$\theta_i : i = 0..9 \text{, where } \theta_0 = 0^\circ, \theta_1 = 10^\circ, \cdots, \theta_9 = 90^\circ$$

and grouped into three view classes: frontal, right profile and half right profile. All nonfaces are put into one class, and this produces $L+1$ classes (where $L = 3$). Our goal is to differentiate face patterns from nonfaces and simultaneously decide which classes they belong to.

The learning process for face detection carries out in three stages: it first learns local linear subspaces with low dimensionality from original nonlinear distribution; A large amount of nonface patterns with low likelihoods are strained out while the left face-like patterns are kept for training; Finally one multi-class SVM is trained for each view and fused with other SVMs to perform classification based on the projection coefficients and local reconstruction errors.

3.1 Learning Local Structures of Multi-view Face Distribution

The first stage aims to learn low-dimensional subspace from multi-view face images. In our system, this is achieved by fitting the training faces into a set of view based PPCA mixture models. Face images are first divided into subsets according to view angles. The mixture of PPCA model is then applied to learn the structures within each view subset.

The choice of the latent-space dimensionality q and the number of Gaussian components M is crucial for identifying PPCA mixture model. To find the optimal values of q and M, we empirically tune the parameters and evaluate the exactness of the model by using the density model to classify the testing faces and nonface images. Section 4.2 accounts for how we set these parameters in detail. We evaluate the density of each testing image under the model. If it exceeds some threshold, the image will be classified as a face.

A meaningful characteristic of the derived model we observe in the experiments is that, while a fairly low subspace dimensionality is adopted (for example, $q = 7$ for profile view) for modeling, the classification results based on PPCA mixtures model can still be better than an unconstrained Gaussian mixtures model with high-dimensional subspaces, as illustrated in figure 3, section 4.2. It may partly come from the fact that in the conventional mixtures model, the partition of clusters is commonly determined by an iterative process which involves calculating a hard Euclidean distance or a likelihood measure with a full covariance matrix. It may give rise to biased local clustering results while the training data is insufficient to recover all parameters of the full covariance matrix. We can argue that with a set of low-dimensional sub-models and less free parameters, the PPCA mixtures model can actually approximate the data manifolds more faithfully. This characteristic of high local linearity can consequently offer an efficient representation for the following classification process.

3.2 Choosing Face-like Nonface Samples

While a great deal of images can served as nonface training patterns, selecting a representative set of nonfaces remains a tricky issue in most of face detection systems. To cope with this problem, a "boot-strapping" method [5] [12] [13] [14] is often used to iteratively update nonface set when starting with a relatively smaller data set. "Boot-strapping" based nonfaces selection scheme can lead to a significant rising in training time since each iteration will inevitably involves training a new classifier.

Instead of collecting nonface samples in training phase, we seek to an alterative approach based on density measure offered by PPCA, to obtain a nonface training set that can be viewed as "representative" before the training is started. Let x be a new cropped nonface pattern. To which extend it is similar with faces in view subset i can be evaluated by its likelihood under the corresponding PPCA mixtures

model. The following is an outline of our probability based nonface selection process:

1) Randomly select a set of nonface (over 8000) samples from a group of images.
2) Compute the likelihood of all nonface samples under the PPCA mixtures model; choose the first fifteen percent of samples with maximal probability likelihood as an initial training set. The minimal likelihood value τ_i in the initial set is recorded as a threshold τ_i.
3) Continue to crop image patches from an extended image set where there is no face contained and add those whose likelihoods exceed the threshold τ_i to the training set until enough nonface patterns are collected.
4) Repeat first three steps for every view subset.
5) Add some randomly selected new nonface samples into the training set.

By this strategy we gather a great number of nonface patterns that visually like faces or geometrically locate near the face clusters. Some of nonfaces with high likelihood under the PPCA mixtures model of frontal faces are showed in figure 2. In theory the selected nonfaces can be more similar to faces if we set a higher threshold; but longer time will be spent to accumulate enough samples. It is also worth to note that the different "representative" nonface patterns are obtained in different views. We expect these patterns to improve the classification performance since they can clearly characterize the boundary between faces and nonfaces in each view class.

3.3 Training SVM Array

We use a set of view-based multi-class SVM classifiers [18] to identify face patterns from nonfaces. One SVM is trained for each view to perform $L+1$-class classification based on the features in the corresponding PPCA mixture subspace. The one-against-the rest method [19] [20] is used for solving the multi-class problem. SVM is used because it can lead to high generalization ability while only few parameters need to be tuned. More importantly, SVM is a boundary sensitive classifier; for example, most of supports vectors from the nonfaces set are quite similar to faces [13]. Note that we select a great number of faces-like nonfaces which locate near the face clusters; SVM can obtain enough evidences to learn a boundary between faces and nonfaces.

In a specific view, the projections onto each Gaussian subspace

$$\vec{p} = (\sigma^2 I + W^T W)^{-1} W^T (t - \mu)$$

and the local reconstruction errors

$$e = ((t-\mu)^2 - \vec{p}^2)^{1/2}$$

under the corresponding view are combined together and arranged in a row as feature vectors. We use such local linear features instead of distance based metrics like DIFS in face classification for two reasons: First, in our experiments, distance based metrics achieve good performance in detecting frontal face, but while applied to profile or half profile face, the detection rate drop down. This can be partly interpreted by the fact that non-frontal face patterns generally include more artifacts like backgrounds or hairs in the appearance. The estimated distance as a similarity measurement can be insufficient while applied to non-frontal faces. Second, the feature vectors composed by projection coefficients and reconstruction errors contain richer information while still remain a reasonable size due to the low dimensionality of local PPCA subspaces.

Reconstruction error is one of the important features used in training the classifier. While PPCA mixture model tries to maximize the log-likelihood of training data in its iterative optimization, it doesn't explicitly grant to minimize reconstruction errors. For some data points, their reconstruction errors may be high even in the nearest PPCA cluster. We can expect ignoring this information may introduce performance drop in face detection.

4. EXPERIMENTAL RESULTS

The purpose here is to demonstrate the PPCA mixtures modeling approach to learning low dimensional manifolds and evaluate the performance of our proposed framework in multi-view face detection.

4.1 Data Preparation

More than 6,000 face samples are collected by cropping from various sources (mostly from video). The view is in the range of $[0^\circ, 90^o]$ with 0^o representing the side view and 90^o representing the frontal view. A total number of about 25,000 multi-view face images are generating from the 6,000 samples by artificially shifting or rotation in the image panel. Face samples are grouped into three view classes (frontal, half profile and profile). Face samples labeled with $90^\circ, 80^\circ, 70^\circ$ are grouped as frontal faces, those with $60^\circ, 50^\circ, 40^\circ$ are grouped as half profile faces and the faces in $30^\circ, 20^\circ, 10^\circ, 0^\circ$ are grouped as profile faces. The faces from all three views are divided into three data sets for the different usages. Table 1 shows the partition and the sizes of the three data sets. Set 1 is used for learning 3 PPCA mixture models and training SVMs. Set 2 is used in parameters selection of the mixture model. Set 3 is used for testing. A set of $10,000 \times 3$ nonface samples (10,000 for each view class) are collected for training in the way illustrated in section 3.2, (see figure 2 for some

Table 1. Composition of three data sets

View	Set 1	Set 2	Set 1
Frontal	3000	500	4500
Half Profile	3000	500	4500
Profile	4000	500	6000
Total Faces	10000	1500	15000
Nonfaces	10000x3	2000	10000

Figure 1: Multi-view face examples

Figure 2. (Up) Typical nonface samples collected by density measure comparison under the frontal PPCA mixtures model. The images have been preprocessed. (Bottom) Randomly selected nonface samples.

examples) where additional 10,000 nonfaces are used for testing.

Each windowed sub-image is normalized into 20×20 pixels, approximately aligned and preprocessed by illumination correction and histogram equalization in a way similar to what was done in most of existing systems [12] [14].

4.2 Face Distribution Modeling

During the training phase, only face samples of one view subset are presented to the corresponding mixture model. Three view-specific PPCA mixtures, corresponding to frontal, profile and half profile, are trained with a same component number $M = 9$. Each component shares a common dimension of $q = 7$. To illustrate how well the model can capture the face distribution, we use the density model to perform 2-class classification in set 2. The testing sets used in each PPCA mixture model include faces samples in the corresponding view class and all other nonfaces samples.

To explore how the model changes with different parameters, we evaluate the density of face images of Set 2 under a set of PPCA mixture models. The component number is fixed with $M = 9$ while the dimensionality of components decreases from 49 to 7. We note the fact that the classification accuracy remains almost unchanged. This manifests within a specified view, PPCA mixture model is capable of capturing the local structures of high complex distribution with fairly low dimensionality.

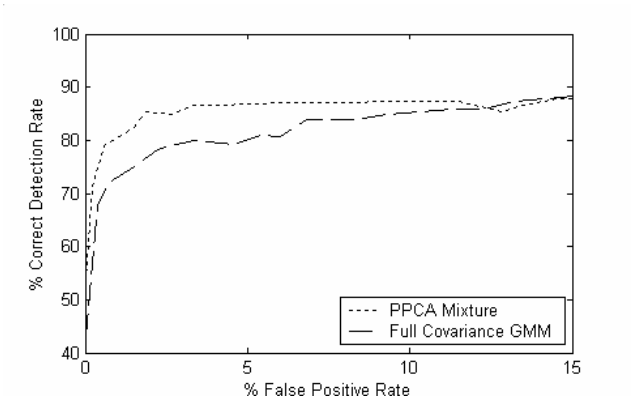

Figure 3: ROC curves for comparing a PPCA mixture model and a full covariance GMM. 3000 profile faces are used to train both of the models. Testing image is classified as face while its density exceeds some threshold.

Additional experiments have done for comparing PPCA mixture with traditional linear subspace mixture model, i.e. unconstrained GMM, in view-based face distribution modeling. Both PPCA mixtures and a full covariance GMM with a relatively optimal choice of parameters ($M = 7$ and $q = 75$) are used to model profile face manifolds and identify profile faces from nonface. This leads to an ROC curve, as shown in figure 3. The result highlights the improvement of the performance while changing from an unconstrained Gaussian mixture model to a PPCA mixture model.

4.3 Face Detection by SVM Array Classification

For the SVM training, an RBF kernel $K(x, y) = e^{-\|x-y\|^2/\sigma^2}$ is selected, with $\sigma = 9$. The parameters are empirically selected and largely dependent on the experiments. All images in Set 1 are projected to the Gaussian components of three derived PPCA mixtures. Their projection coefficients and reconstruction errors are used for training. Images in Set 3 are used for testing. The classification results are demonstrated in classification matrices (c-matrices); see figure 4. The entry (i, j) of the c-matrix gives the number of examples whose ground truth class label is i (in row) and which are classified into class j (in column). The first $L = 3$ rows and 3 columns correspond to the 3 view

C-matrix for frontal view

4301	68	1	84
108	4149	121	44
27	206	5700	221
64	77	178	9960

C-matrix for half-profile view

4199	164	10	189
180	4127	127	64
18	159	5694	101
103	50	169	9655

C-matrix for profile view

4234	94	2	102
192	4053	170	59
9	243	5615	187
65	110	213	9661

Figure 4. Classification statistics as demonstrated by c-matrices

classes (frontal, half-profile, profile) of the ground truth and classification result, respectively, whereas the last row and column correspond to the nonface class.

From these c-matrices, the corresponding missing and false alarm rates for face detection can be calculated, as shown in **Table 2**; the classification consequences from all views are fused together in a simple voting scheme. Results of the fusion show in the bottom of the table.

Table 2. Face Detection Error Rates

View Classes	Missing (%)	False A. (%)
Frontal	2.13	3.49
Half-Profile	2.15	3.54
Profile	2.59	3.48
3 SVM's Fused	2.13	3.28

Finally we apply the system to real images by exhaustively scanning for face-like patterns at all possible positions. Multiple scales are handled by examining sub-window images taken from the scaled image. Testing images collected from VCD movies (in 352×288 pixels) are used for the evaluation. Figure 5 shows some examples.

CONCLUSION

We have presented a view-based framework for learning the low-dimensional subspace representation of multi-view faces and for pose invariant face detection. The main part of the work is to model the distribution of multi-view faces using PPCA mixture models. Faces are identified from

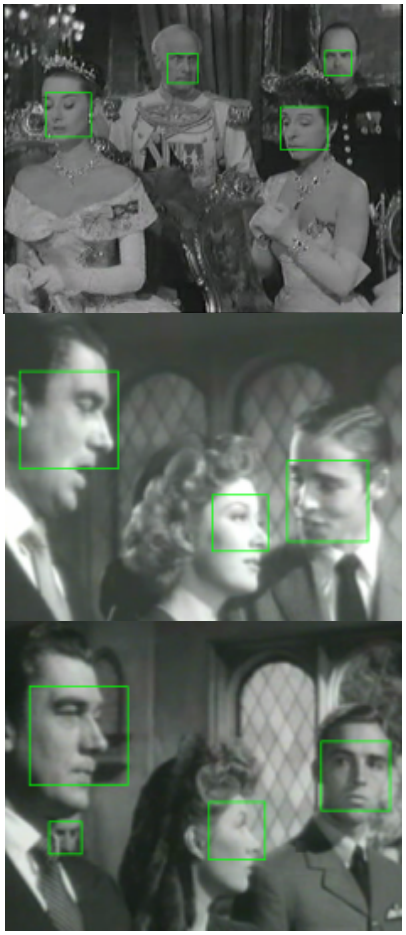

Figure 5: pose invariant face detection results in some real images.

nonface by a set of SVMs. Given this framework demonstrates good performance in multi-view face detection, we stress that the underlying architecture is fairly general and can be applied to other appearance based object detection tasks.

REFERENCES

[1] M.Bichsel and A.P. Pentland. Human face recognition and the face image set's topology. CVGIP: Image Understanding, 59:254--261, 1994.

[2] A.P. Pentland, B.Moghaddam, and T.Starner. View-based and modular eigenspaces for face recognition. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 84--91, 1994.

[3] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 586--591, Hawaii, June 1991.

[4] M. Kirby and L. Sirovich, Application of the Karhunen-Loeve procedure for the characterization of human faces, IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 1, pp. 103-108, January 1990.
[5] K.K. Sung and T.Poggio. Example-based learning for view-based human face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1):39--51, 1998.
[6] B.Moghaddam and A.Pentland. Probabilistic visual learning for object representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 7:696--710, July 1997.
[7] Brendan J. Frey, Antonio Colmenarez, Thomas S. Huang. Mixtures of Local Linear Subspaces for Face Recognition. In CVPR, 1998
[8] M. H. Yang, N. Ahuja, and D Kriegman. Mixtures of linear subspaces for face detection, In Proceeding of the Fourth International Conference on Automatic Face and Gesture Recognition, pages 70-76, 2000
[9] T.Hastie and W.Stuetzle. Principal curves. Journal of the American Statistical Association, 84(406):502--516, 1989.
[10] H.Murase and S.K. Nayar. Visual learning and recognition of 3-D objects from appearance. International Journal of Computer Vision, 14:5--24, 1995.
[11] B.Scholkopf, A.Smola, and K.R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10:1299--1319, 1998. Technical Report No. 44, 1996, Max Planck Institut fur biologische Kybernetik, Tubingen.
[12] H.A. Rowley, S.Baluja, and T.Kanade. Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1):23--28, 1998.
[13] E.Osuna, R.Freund, and F.Girosi. Training support vector machines: An application to face detection. In CVPR, pages 130--136, 1997.
[14] H.Schneiderman and T.Kanade. A statistical method for 3d object detection applied to faces and cars. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2000.
[15] S.Gong, S.McKenna, and J.Collins. An investigation into face pose distribution. In Proc. IEEE International Conference on Face and Gesture Recognition, Vermont, 1996.
[16] M.E. Tipping and C.M. Bishop. Mixtures of probabilistic principal component analyzers. Neural Computation, 11(2):443-482, 1999.
[17] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B, 21(3):611-622, 1999.
[18] V.N. Vapnik. Statistical learning theory. John Wiley & Sons, New York, 1998.
[19] B.Scholkopf, C.Burges, and V.Vapnik. Extracting support data for a given task. In U.M. Fayyad and R.Uthurusamy, editors, Proceedings, First International Conference on Knowledge Discovery & Data Mining, Menlo Park, 1995. AAAI Press.
[20] J.Weston and C.Watkins. Multi-class support vector machine. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, Egham, UK, 1998.
[21] S.Z. Li, Q.D. Fu, L. Gu, B. Scholkopf, Y.M. Cheng, H.J. Zhang. ``Kernel Machine Based Learning for Multi-View Face Detection and Pose Estimation''. In Proceedings of 8th IEEE International Conference on Computer Vision. Vancouver, Canada. July 9-12, 2001.