On the use of Independent Tasks for Face Recognition

Àgata Lapedriza
Computer Vision Center (CVC)
Universitat Autònoma de Barcelona (UAB)
Edifici O, Campus UAB, Barcelona (Spain)
agata@cvc.uab.es

David Masip
Universitat Oberta de Catalunya
Rambla del Poblenou 156, Barcelona (Spain)
dmasipr@uoc.edu

Jordi Vitrià
Computer Vision Center (CVC)
Universiat de Barcelona
Gran Via de les Corts Catalanes 585, Barcelona (Spain)
jordi@cvc.uab.es

Abstract

We present a method for learning discriminative linear feature extraction using independent tasks. More concretely, given a target classification task, we consider a complementary classification task that is independent of the target one. For example, in face classification field, subject recognition can be a target task while facial expression classification can be a complementary task. Then, we use labels of the complementary task in order to obtain a more robust feature extraction, being the new feature space less sensitive to the complementary classification. To learn the proposed feature extraction we use the mutual information measure between the projected data and both labels from the target and the complementary tasks. In our experiments, this framework has been applied to a face recognition problem, in order to inhibit this classification task from environmental artifacts, and to mitigate the effects of the small sample size problem. Our classification experiments show an improved feature extraction process using the proposed method.

1. Introduction

The task of a classification method is to automatically learn a correspondence between a training data set (numeric or symbolic) and a predefined set of discrete labels. The goal is the optimal identification of new unseen samples using a classifier that is learned by minimizing some defined empirical loss function. The empirical nature of this optimization problem has been usually approached by making use of the training data set and, moreover, incorporating any available prior knowledge as an additional term in the loss function.

During the past decades different classification algorithms have been developed and successfully applied in many situations. However, one of the main drawbacks of visual data classification is the small sample size problem, what makes difficult the generalization capability of any classifier. For instance, in face recognition field, the identity of a set of subjects is modelled using a training set that is, in theory, a sufficient sample of the full data set [14]. Nevertheless, training face sets often suffer from lack of data [11], and this fact is even worse if we consider the full range of possible appearance variations for a face: non uniform illumination, facial expression effects or even partial occlusions. In that case, when the training data set is small, the robustness of the learned classifier will be poor as well as the capability of successful recognition in non-controlled environments. This phenomenon has been mitigated with the use of dimensionality reduction techniques to extract relevant and discriminant information [5, 3].

In real world applications related to classification problems, we often have at our disposal additional information about the problem domain that is usually neglected in the classic pattern recognition field. This information might be independent from the original classification task, but it can be helpful if it is complementarily used in the learning process. For example, in the case of face recognition mentioned above, faces can be labelled according to the subject that appears in the image or according to the conditions of acquisition. In that case, this second categorization can be used as complementary information during the training process of subject recognition task.

In this paper we propose a new use of systematic appear-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

ance variations to learn a more robust feature extraction for subject recognition. The idea is to consider these effects as a new classification problem (for example, to classify the dominant illumination in a face image, or to classify the facial expression) that can be seen as an independent task. In this way, this extra face classification task can be used to inhibit the confusion caused by these artifacts, that can be consequence of non-controlled environment's conditions. More concretely, the extracted features keep the information from the original data that is useful to perform the subject recognition task and, moreover, this new feature space does not preserve the information that allows the classification according the systematic appearance variations.

The outline of the paper is as follows: section 2 includes a brief review of related work, while section 3 introduces formally the problem and the notation. In section 4 we describe in detail the proposed feature extraction method. After that, section 5 shows the performed performed and discuss the obtained results, and, finally, section 5 concludes this work.

2. Related Work

A topic that is closely related to the proposed feature extraction method is Multi-Task Learning (MTL). The MTL approach is a new classifying methodology based on jointly training multiple related tasks taking advantage of its relationship to improve the learning process of the individual classification tasks. The previous works on MTL show interesting improvements at two different levels: the accuracies of the methods increase (parallel transfer) when problems are jointly trained, and the number of samples needed to train reliable classifiers decreases (sequential transfer). The advantages of MTL have been experimentally validated in the first works of Thrun [10] and Baxter [6]. Two different approaches to MTL can be identified in the recent literature: (i) a functional approach, where the tasks share the hypothesis space [2] and (ii) a structural approach, where the representation of the data is supposed to share a common generative process, that can be used in the hypothesis selection [4].

The MTL paradigm applied to feature extraction can be a useful tool to focus the projection vectors on the general recognition task, discarding the intra-variations due to illumination or other artifacts. In this context, the proposed method can be seen as a functional MTL approach to learn a new faces feature space that keeps subjects identity information being as less sensitive as possible to illumination changes and partial occlusions.

3. Problem Statement

Let be $X = \{x_1, ..., x_i, ..., x_N\}$ a data set in $\mathbb{R}^D$. Consider a target classification task T_T, and the corresponding labels $C_T(X)$, $\{c_1, ..., c_N\}$, according to T_T, where each $c_i \in \{1, ..., N_T\}$.

Let be T' another classification task for the elements in X and suppose that we have the labels of the elements in X according this task, $C'(X)$, $\{c'_1, ..., c'_N\}$, where each $c'_i \in \{1, ..., N'_T\}$.

Definition: T_T and T' are *independent tasks* if C_T and C' are independent random variables. That is, if

$$P((C_T(x) = c) \cap (C'(x) = c')) = \quad (1)$$
$$= P(C_T(x) = c)P(C'(x) = c') \quad (2)$$

for all $x \in X$, $c \in \{1, ..., N_T\}$ and $c' \in \{1, ..., N'_T\}$. A task T' that is independent of the target class will be denoted as T_I. On the other hand, we denote as P_T the data partition according to T_T in classes $\{c_1, ..., c_N\}$, while the independent partition of the data according T_I will be denoted by P_I.

Independent tasks examples can be found in real problems of computer vision. For instance, a set of manuscript symbols can be partitioned according to which symbol appears in the image or according to the person who drew it. In that case, these two tasks are independent if we assume that the probability of writing a concrete symbol is the same for all of the authors. On the other hand, considering a set of face images having some kind of expression (smile, anger, scream or neutral) we can divide the set according to the subject that is in the image or according to the expression. Then, supposing that the expression do not depend on the subject, we have also two independent classification tasks.

In several real situations, a task T_T should be learned from a reduced set of training samples, where the variability according an independent task T_I is not represented. For instance, in face classification field, we can consider the target task of subject recognition. In most of these cases, we will have just a few number of training samples per class. Moreover, in real situations, these images will be captured in controlled environments, appearing poor local changes in the illumination and no representations of the subject with partial occlusions. However, the goal will be in general to recognize this person in any uncontrolled condition, although our training set X do not allow us to model the independent partition P_I of the faces set according the conditions of the image acquisition (for instance highlight in a particular faces part, or partial occlusion of a face fragment).

Suppose, in this framework, that we have another faces set X', where the partition P_I is well represented, but $X \bigcap X' = C(X) \bigcap C(X') = \varnothing$. In that case, we can learn from X' and the labels C' to be poorly sensitive to the partition P_I, finding a feature space where this variability is as irrelevant as possible.

In this paper we propose a linear feature extraction method based on mutual information that uses both (X', C') and (X, C) to learn an appropriate feature space for the task T_T of recognizing the subjects in X. We will consider the case where all the elements in X belong to a same class $c' \in \{1, .., N_I\}$. The idea is to use (X, C) for learning the target task properly and, moreover, to use (X', C') to find a feature space where the variabilities appearing in this second set are not represented.

4. Linear Feature Extraction for Independent Tasks

Given the framework presented above, we propose to find a linear projection $W : \mathbb{R}^D \rightarrow \mathbb{R}^d$, $d < D$, such that: (i) the new feature space has high mutual information between the projected data $Y = WX$ and the class labels C and (ii) at a time, it has as low information as possible between Y and C'. That is, we seek the linear projection

$$W = \arg\max_A J_\lambda(A) \tag{3}$$

$$J_\lambda(A) := I(AX, C) - \lambda I(AX, C') \tag{4}$$

where λ is a positive weight. However, given that C' can not be modelled from X (we are supposing all the elements of X in the same class $c' \in \{1, .., N_I\}$), we will use $Y' = WX'$ to approximate this value.

Given that Shannon's mutual information between data and labels is difficult to estimate, we use in this work the Quadratic Mutual Information proposed by Torkkola [13]. The main idea is to use Renyi quadratic entropy instead of Shannon's definition of entropy. In that case, the Renyi quadratic entropy can be estimated as a sum of local interactions if the density functions of the variables are estimated using Parzen window method. [9]. That is, the probability distribution function of Y will be represented by

$$p(y) = \frac{1}{N} \sum_{i=1}^{N} G(y - y_i, \sigma I) \tag{5}$$

where I is the $d \times d$ identity matrix and G is a d-dimensional Gaussian kernel,

$$G(y, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\Sigma)^{\frac{1}{2}}} \exp(-\frac{1}{2} y^T \Sigma^{-1} y) \tag{6}$$

for a covariance matrix Σ.

Moreover, to take benefit from the kernel properties in the mutual information estimation, Torkkola uses as divergence measure a functional proposed by Kapur [7], instead of the Kullbach-Leibler divergence. Thus, after some manipulations, the quadratic mutual information between the continuous valued Y and discrete C is expressed as

$$I(Y, C) = V_{IN} + V_{ALL} - 2V_{BTW} \tag{7}$$

computing each term from the data as follows

$$V_{IN} = \frac{1}{N^2} \sum_{c=1}^{N_T} \sum_{j=1}^{N_c} \sum_{k=1}^{N_c} G(y_{cj} - y_{ck}, 2\sigma^2 I) \tag{8}$$

$$V_{ALL} = \frac{1}{N^2} (\sum_{c=1}^{N_T} (\frac{N_c}{N})^2) \sum_{j=1}^{N} \sum_{k=1}^{N} G(y_j - y_k, 2\sigma^2 I) \tag{9}$$

$$V_{BTW} = \frac{1}{N^2} \sum_{c=1}^{N_T} \frac{N_c}{N} \sum_{j=1}^{N_c} \sum_{k=1}^{N} G(y_{cj} - y_k, 2\sigma^2 I) \tag{10}$$

where a sample is denoted by one index y_i, $i \in \{1, .., N\}$ if the class is irrelevant, and it is denoted by two indexes y_{cj} when its class is relevant. In this second case $c \in \{1, .., N_T\}$ represents the class index and $j \in \{1, .., N_c\}$ represents the within-class index, being N_c the number of elements in class c.

On the other hand, given that

$$\frac{\partial}{\partial y_i} G(y_i - y_j) = G(y_i - y_j, 2\sigma^2 I) \frac{(y_j - yi)}{2\sigma^2} \tag{11}$$

the corresponding derivatives $\frac{\partial V_{IN}}{\partial y_{ci}}$, $\frac{\partial V_{ALL}}{\partial y_{ci}}$ and $\frac{\partial V_{BTW}}{\partial y_{ci}}$ are respectively

$$\frac{\sum_{j=1}^{N_c} G(y_{cj} - y_{ci}, 2\sigma^2 I)(y_{cj} - y_{ci})}{N^2 \sigma^2} \tag{12}$$

$$\frac{(\sum_{r=1}^{N_T} (\frac{N_r}{N})^2) \sum_{j=1}^{N} G(y_j - y_{ci}, 2\sigma^2 I)(y_j - y_i)}{N^2 \sigma^2} \tag{13}$$

$$\frac{\sum_{r=1}^{N_T} \frac{Nr + N_c}{2N} \sum_{j=1}^{N_r} G(y_{rj} - y_{ci}, 2\sigma^2 I)(y_{rj} - y_{ci})}{N^2 \sigma^2} \tag{14}$$

what allows the use of gradient ascent techniques to optimize the criterion of the expression 4.

We use in this work a two-sample stochastic gradient ascent to learn the projection W, given that a classical gradient ascent procedure is not computationally feasible for large sets of high dimensional data. It is specially appropriated the use of two-sample subsets to approximate J_λ and their derivatives because of the expressions' simplification.

The actualization of this gradient ascent is performed by

$$W_{t+1} = W_t + \xi \frac{\partial J_\lambda}{\partial W} \tag{15}$$

where $\partial J_\lambda / \partial W$ is approximated at each iteration using a predetermined number Q of sample pairs. Thus, let be

$Z = \{X, X'\}$ and $S = \{Y, Y'\}$. Given $\tilde{Z} = \{z_1, z_2\} \subset Z$ and the respective $\tilde{S} = \{s_i = W z_i\}$,

$$\frac{\partial J_\lambda}{\partial W} = \sum_{i=1}^{2} \frac{\partial J_\lambda}{\partial s_i} \frac{\partial s_i}{\partial W} = \sum_{i=1}^{2} \frac{\partial J_\lambda}{\partial s_i} z_i^T \tag{16}$$

and

$$\frac{\partial J_\lambda}{\partial s_i} = \frac{\partial I(\tilde{S}, C)}{\partial s_i} - \lambda \frac{I(\tilde{S}, C')}{\partial s_i} \tag{17}$$

From equations 10, we have

$$I(\tilde{S}, C) = \frac{1}{4}(G(0, 2\sigma^2 I) - G(s_1 - s_2, 2\sigma^2 I)) \tag{18}$$

if $c(s_1) \neq c(s_2)$ and $I(\tilde{S}, C) = 0$ otherwise, and

$$\frac{\partial I(\tilde{S}, C)}{\partial s_1} = -\frac{\partial I(\tilde{S}, C)}{\partial s_2} = -\frac{1}{8\sigma^2} G(s_1 - s_2, 2\sigma^2 I)(s_2 - s_1)) \tag{19}$$

if $c(s_1) \neq c(s_2)$, and $\frac{\partial I(\tilde{S}, C)}{\partial s_1} = \frac{\partial I(\tilde{S}, C)}{\partial s_2} = 0$ otherwise.

The same expressions can be applied to compute $I(\tilde{S}, C')$ and the partial derivatives. For more details of these derivations see [12]. The stochastic gradient ascent to optimize J_λ is shown in table 1.

Notice that essential points in this procedure are: *(i)* to decide whether two elements $x \in X$ and $x' \in X'$ are or not in the same class according partitions P_I, and *(ii)* the use of the labels $C(X')$ in the algorithm. Both points should be considered and appropriately approached depending on the tasks and the data sets.

5. Experiments

We use in our experiments the publicly available ARFace database [8] and the subset of FRGC [1] database composed by the images acquired in non-controlled environments.

ARFace database is composed by face images from 126 different subjects (70 men and 56 women). The database has from each person 2 sets of images, acquired in two different sessions, with the structure detailed in figure 2. On the other hand, the mentioned FRGC database subset is composed by 1886 images from 275 subjects, having from 2 to 16 images per person. Figure 3 shows some examples of these images.

We have aligned all the images according to the eyes and we have used in our experiments just the internal part of the face, in a resolution of 36×33 pixels.

In all the experiments we consider subject recognition as a target task T_T and image type classification as an independent task T_I. (The independence assumption between these two tasks has been discussed in section 3). We have in all the cases a set X and a set X'. The set X is always split

- Initialize W_0 (randomly, by PCA or by LDA, using (X, C))
- Set $Y = W_0 X$ and $Y' = W_0 X'$
- $\sigma = \sigma(Y)$ (for example, maximum distance between samples / 2)
- while $\sigma < \sigma_f$ (for example, mean distance between samples)
  - repeat
    - draw Q sample pairs $\{\tilde{S}_1, .., \tilde{S}_Q\}$ from $S = \{Y, Y'\}$ at random
    - approximate J_λ as the mean of all $J_\lambda(\tilde{S}_q)$, $q = 1, .., Q$, $\tilde{J}_\lambda$
    - approximate the gradient $\partial(J_\lambda)/\partial W$ as the mean of $\partial(J_\lambda(\tilde{S}_q)/\partial W$, $\partial(\tilde{J}_\lambda)/\partial W$
    - actualize $W_{t+1} = W_t + \xi \partial(\tilde{J}_\lambda)/\partial W$
    - project the data in the new space, $Y = W_{t+1} X$ and $Y' = W_{t+1} X'$
  - until $\tilde{J}_\lambda$ does not decrease
- end while

Figure 1. Algorithm pseudocode

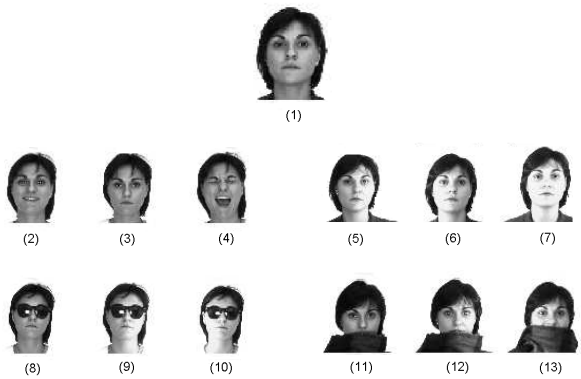

Figure 2. One sample from each of the image types in AR Face Database. The image types are the following: (1) neutral expression, (2) smile, (3) anger, (4) scream, (5) left light on, (6) right light on, (7) all side lights on, (8) wearing sun glasses, (9) wearing sun glasses and left light on, (10) wearing sun glasses and right light on, (11) wearing scarf, (12) wearing scarf and left light on, (13) wearing scarf and right light on.

in two subsets: X_{train} and X_{test}. Then, we learn the linear feature extraction matrix W using X_{train} and compute the classification accuracy of X_{test} using the Nearest Neighbor Classifier (NN) in the new feature space.

In all the test we show the obtained classification accura-

Figure 3. Examples of images included on the FRGC database acquired in uncontrolled scenes.

cies using the following features: *(i)* original feature space (NN), *(ii)* feature extraction using PCA ($PCA{+}NN$) *(iii)* feature extraction using FLD ($FLD + NN$), *(iv)* the maximization of $I(WX, C)$ ($J_0 + NN$), and *(v)* the proposed J_λ, for $\lambda > 0$, ($J_\lambda + NN$).

5.1. Subject Recognition with the ARFace database

Here we test the proposed method making a subject recognition experiment using the ARFace database. In this experiment we will use a reduced training set composed of two (expression and illumination) neutral faces per person corresponding to type 1 ARFace database images. The target task will be their recognition under the other imaging conditions (ARFace types 2 to 13). The complementary task will be based on the classification of the different imaging artifacts (corresponding to ARFace types 1 to 13) for a non overlapping set of subjects.

The total number of subjects have been split in 5 subsets of 17 persons. We perform the experiments according the following protocol: for all the possible subsets combinations,

- X is composed of 3 of these subsets (51 subjects), while the other 2 belong to X' (34 subjects).
- X_{train} include just images of type 1 (neutral expressions) and the rest are included in X_{test}.

In this case, elements in X_{train} are labelled according to both T_T and T_I. On the other hand, given that the target task is to recognize the 51 subjects that compose X_{train}, we decided to label X' as $C(X') = 52$. Then, when $I(WZ, C)$ is computed, we take into account just the pairs of elements in X_{train} belonging to a different class, or the pairs composed of one element in X_{train} and the other in X'. This is a way of ignoring the discriminant information related to the subjects in X', that we actually do not need to model. However, notice that all the samples in $Z = \{X_{train}, X'\}$ are used to estimate $I(WZ, C')$.

Table 5.1 includes the mean accuracies and the confidence intervals obtained in this experiment.

Table 1. Subject recognition using the ARFace database (51 subjects), using 2 neutral frontal images per subject in the training set and testing with images having expressions, high local changes in the illumination and partial occlusions.

Method	Accuracy
NN	$32.43\% \pm 3.84\%$
$PCA + NN$	$31.45\% \pm 3.58\%$
$FLD + NN$	$31.94\% \pm 3.31\%$
$J_0 + NN$	$40.66\% \pm 3.62\%$
$J_1 + NN$	$51.94\% \pm 2.62\%$

Table 2. Subject recognition using the FRGC database (166 subjects), where images are acquired in uncontrolled environments.

Method	Accuracy
NN	$44.01\% \pm 3.90\%$
$PCA + NN$	$41.37\% \pm 3.84\%$
$FLD + NN$	$46.28\% \pm 4.01\%$
$J_0 + NN$	$67.67\% \pm 4.50\%$
$J_{0.5} + NN$	$78.93\% \pm 3.09\%$

5.2. Subject Recognition with the FRGC database using the ARFace as a Complementary Set

In this experiment, the set X is composed by images from the 166 subjects having more than 4 samples. The training set is composed of subject faces acquired in non controlled environments. The target task is their recognition in the same kind of scenarios. The complementary task in this experiment is based on face images belonging to a different database, and its objective is the classification of the different imaging artifacts (corresponding to ARFace types 1 to 7). In this case we expect that the complementary task can inhibit the feature extraction task from systematic imaging conditions, even when using a different database.

We perform 10 experiments rounds according the following protocol:

- X_{train} is composed by 50% the images per subject randomly selected and the rest of images are included in the test set, X_{test}.
- X' is composed by all the images in the ARFace database belonging to image types from 1 to 7.

Here, elements in X are labelled just according to T_T. Given that they are not labelled according to T_I, we have supposed that none of them cannot be univocally identified with any concrete image types (from 1 to 7) of T_I . For this reason we label them as $C'(X_{train}) = 8$. Moreover, as in the experiment presented above we use $C(X') = 167$.

The obtained results and the confidence intervals are shown in table 5.2.

5.3. Discussion

We can see that in both experiments the best accuracy is obtained by the proposed feature extraction method.

In the first experiment the training set is composed of just 2 neutral frontal images per subject. In that case, we obtain the same results using NN in the original space, $PCA + NN$ and $FLD + NN$. This indicates that the principal variance of the data set is represented by the variability of the different subjects. For this reason, PCA is able to keep the relevant information of the original space, while FLD is not able to improve PCA, given that there is no information about the within class variability, and mean class samples are poorly estimated.

On the other hand, the criteria J_0 that maximizes the mutual information between the extracted features and the target labels improves the other feature extraction systems. Moreover, the proposed method J_1 is able to learn a feature space that is less sensitive to the appearance variation. This is because it can extrapolate the information about the data variability from the complementary set, although the real within class variability of the target subjects is not actually represented.

In the second experiment the training set is composed of 50% of the images per subject, which are acquired in non-controlled environments. In this situation, we have some within class variability represented in the training set. For this reason we can see that $FLD + NN$ improves both $PCA + NN$ and NN in the original space. Nevertheless, once again the criterion J_0 that maximizes the mutual information between the new features and the target labels improve $FLD+NN$. On the other hand, the proposed feature extraction method $J_{0.5}$ outperforms again the criterion J_0. In that case, the extra information is obtained from a complementary set belonging to a different database.

6. Conclusions and Future Work

In this paper we propose a linear feature extraction algorithm based on considering independent complementary classification tasks. The mutual information statistic is used for this purpose. Given a principal classification problem, we seek for the linear feature extraction that maximizes the mutual information between the data and the target labels, simultaneously minimizing the mutual information with the complementary task. The framework can be applied to many real world problems, such as handwritten letter recognition, speech, audio or automatic text classification. In this paper, the method has been applied to the face recognition field, in order to: (i) mitigate the effects of the small sample size problem, being the method able to extract from the complementary tasks useful subspace information for the main classification problem, and (ii) inhibit the feature extraction task from known environmental artifacts that can be incorporated as prior knowledge.

We plan as a future work the study of alternative criteria J_λ to add the complementary tasks. More sophisticated expressions could be used, at the expense of increasing the optimization problems complexity. Also, the addition of sparsity priors could benefit the isolation of features focused only on the main classification problem. The optimal initialization of the matrix W and possible non linear extensions of the method are also subjects of further development.

Acknowledgements

This work was partially supported by MEC grant TIC2006-15308-C02-01 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

[1] The 2005 ieee workshop on face recognition grand challenge experiments. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, page .45, Washington, DC, USA, 2005. IEEE Computer Society. 4

[2] J. Baxter. A model of inductive bias learning. *Journal of Machine Learning Research*, 12:149–198, 2000. 2

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, Jul 1997. 1

[4] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. 2

[5] R. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, 7:179–188, 1936. 1

[6] J.Baxter. Learning internal representations. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, pages 311–320, 1995. 2

[7] J. Kapur and H. K. Kesavan. *Entropy optimization principles with applications*. Academic Press, San Diego, London, 1992. 3

[8] A. Martinez and R. Benavente. The AR Face database. Technical Report 24, Computer Vision Center, june 1998. 4

[9] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962. 3

[10] S.Thrun and L.Pratt. *Learning to Learn*. Kluwer Academic, 1997. 2

[11] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. *Pattern Recogn.*, 39(9):1725–1745, 2006. 1

[12] K. Torkkola. On feature extraction by mutual information maximization. In *ICASSP*, 2002. 4

[13] K. Torkkola. Feature extraction by non parametric mutual information maximization. *J. Mach. Learn. Res.*, 3:1415–1438, 2003. 3

[14] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003. 1