FACE VERIFICATION BY SHARING KNOWLEDGE FROM DIFFERENT SUBJECTS

David Masip[1], Àgata Lapedriza[2] and Jordi Vitrià[2]

[1] *Department of Applied Mathematics and Analysis (MAiA), University of Barcelona (UB)*
Edifici Històric Gran Via de les Corts Catalanes 585, Barcelona, Spain

[2]*Computer Vision Center, Department of Computer Science*
Universitat Autònoma de Barcelona, Edifici O, Bellaterra 08193, Spain

Keywords: Face verification, Computer Vision, Logistic Regression Model, Multi-task Learning.

Abstract: In face verification problems the number of training samples from each class is usually reduced, making difficult the estimation of the classifier parameters. In this paper we propose a new method for face verification where we simultaneously train different face verification tasks, sharing the model parameter space. We use a multi-task extended logistic regression classifier to perform the classification. Our approach allows to share information from different classification tasks (transfer knowledge), mitigating the effects of the reduced sample size problem. Our experiments performed using the publicly available AR Face Database, show lower error rates when multiple tasks are jointly trained sharing information, which confirms the theoretical approximations in the related literature.

1 INTRODUCTION

Face verification can be defined as a binary classification problem where we receive as input a high-dimensional data vector $\mathbf{x} \in \mathbb{R}^D$ and a claimed identity for the individual. Then, the goal is to verify the correctness of the identity using a model of the person to be verified.

Different facts make that Face Verification is still nowadays an unsolved problem. For example, the dimensionality of the face data is large, making the estimation of the classifier parameters more difficult. On the other hand, the shortage of samples from a single person is frequent, what produces classifiers that are not robust enough.

Most of the face verification algorithms found in the literature are focused on the classification in high dimensional spaces problem. The procedure in that cases is usually divided in two steps. First a feature extraction step is performed to reduce the data complexity and second a classifier in the new reduced space is trained. Thus, the problem of the high dimensional data vectors classification has been addressed with two complementary methodologies: feature extraction techniques and classification algorithms.

One of the most spread unsupervised feature extraction techniques is PCA, where the goal is to find the linear projection that minimizes the mean squared error criterion, obtaining a decorrelated representation of the data. Recently, more sophisticated techniques have appeared, imposing extra restrictions on the extracted feature space, such as sparsity (Lee and Seung, 2000) or independence (Hyvarinen, 1999), which have been successfully applied to face classification in presence of occlusions and strong changes in the illumination. On the other hand, supervised feature extraction techniques use the data label in the dimensionality reduction process, finding the subspace that maximizes some separability criterion on the training data. Linear Discriminant Analysis (Fisher, 1936) is the most popular supervised linear technique (LDA), which seems to outperform the PCA "eigenfaces" approach (Moghaddam et al., 1998) (Belhumeur et al., 1997). Nevertheless, LDA uses the class-scatter matrices of the training data as a separability measure, being blind beyond second order statistics. Recent methods such as Non Parametric Discriminant Analysis (NDA) (Fukunaga and Mantock, 1983) or Boosted Discriminant Projections (Masip and Vitrià, 2006; Masip et al., 2005) have been shown to outperform the classic approach.

In this paper we focuss our attention in the short-

Masip D., Lapedriza À. and Vitrià J. (2007).
FACE VERIFICATION BY SHARING KNOWLEDGE FROM DIFFERENT SUBJECTS.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 286-289

age of training samples in the Face Verification field. We propose to use a face verification scheme where multiple face verification tasks are simultaneously learned.

The idea of sharing knowledge by training related classification tasks was proposed by Caruana (Caruana, 1997). He introduced the term Multi-task learning to describe a technique that learns a neural network classifier from a set of related tasks, improving thus the generalization error. This behavior is justified by the fact that the bias learned in a multiple related tasks environment is likely to be less specific than in a single task problem. It has been shown that using a multiple related tasks learning scheme the number of samples needed decreases with the number of tasks, achieving also better generalization results (Thrun and Pratt, 1997).

The multi-task learning paradigm has been recently applied to different classifiers, such as SVM (Evgeniou et al., 2005), Adaboost (Torralba et al., 2004), or probabilistic frameworks (Ando and Zhang, 2005). Nevertheless, up to our knowledge it has not been still applied to face classification tasks.

There no exists in the literature a formal definition of "related tasks". Here we consider each verification problem as a task and suppose that these kind of tasks are related one to each other.

The paper is organized as follows: in the next section we present our method, that is based on applying the sharing knowledge paradigm to multiple related logistic regression classifiers, section 3 describes the experiments performed on a publicly available data set and section 4 concludes this work.

2 SHARED LOGISTIC REGRESSION MODEL FOR CLASSIFICATION

A binary classification task is the problem of assigning to a sample $\mathbf{x} \in \mathbb{R}^D$ its corresponding label $L \in \{-1,1\}$. A common procedure to solve such a task is to assume that the data follows a specific statistical model and fix the parameters of the model from a set of known samples that is called the *training data set*.

Classic logistic regression model is a statistical approach for solving a binary classification task and it assigns to a sample $\mathbf{x} \in \mathbb{R}^D$ a label $L \in \{-1,1\}$ as

$$L = \arg\max_{c \in \{-1,1\}} P(L = c|\mathbf{x}) \tag{1}$$

where

$$P(L = 1|\mathbf{x}) = \frac{1}{1 + e^{-\beta \mathbf{x}^\mathbf{T}}} \tag{2}$$

and $\beta \in \mathbb{R}^D$ is the parameters vector that is usually estimated by the maximum likelihood criterion. Notice that $P(L = -1|x_i^n, \beta) = 1 - P(1|x_i^n, \beta)$ given that P is a probability distribution.

Our proposal is to extend this logistic regression approach to a multi-task learning framework, where we have multiple related binary tasks $T_1, \ldots, T_M$.

Let be $Z_i = \{((x_{i1}, L_{i1}), \ldots, (x_{iN_i}, L_{iN_i}))\}_{i=1,..,M}$ training data for each one of the M tasks and $\mathbf{Z} = \{Z_i\}_{i=1,..,M}$ the entire training samples set. Suppose that we model each task using the logistic regression explained above, what yields a parameters matrix B

$$B = \begin{pmatrix} \beta_1^1 & \ldots & \beta_1^M \\ \vdots & \vdots & \vdots \\ \beta_D^1 & \ldots & \beta_D^M \end{pmatrix}$$

where each $\beta^i = (\beta_1^i, \ldots, \beta_D^i)$ is the parameter vector for the i-th task. Thus, to assign the label L to an input $x \in \mathbb{R}^D$ according the i-th task, we should follow the criterion

$$L = \arg\max_{c \in \{-1,1\}} P_{T_i}(L = c|\mathbf{x}) = \frac{1}{1 + e^{-\beta^i \mathbf{x}^\mathbf{T}}} \tag{3}$$

Given that situation, normally the negated log-likelihood $N(\mathbf{Z}, \mathbf{B})$ is used to estimate the parameters adding a regularization term, usually $\frac{1}{\sigma^2}\|B\|_2$, to avoid a complex probability distribution on the parameters set.

Nevertheless, our goal is to enforce the different tasks to share some information given that we are assuming that they are related. For this aim, we hierarchically impose a prior distribution on each row of the matrix $\mathbf{B}$.

Let us define the mean vector $\bar{\beta} = (\bar{\beta}_1, \ldots, \bar{\beta}_D)$ as:

$$\bar{\beta_\mathbf{j}} = \frac{\sum_{i=1}^M \beta_j^i}{M} \tag{4}$$

and impose a gaussian centered prior to the mean vector $\bar{\beta}$. We want to enforce each row of the matrix $\mathbf{B}$ to be gaussian distributed with mean $\bar{\beta}_j$. The resulting optimization function is then $G(\mathbf{B}) = N(\mathbf{Z}, \mathbf{B}) + R(\mathbf{B})$ where

$$R(\mathbf{B}) = \frac{1}{\sigma_1^2}\|\bar{\beta}\|_2 + \frac{1}{\sigma_2^2}\sum_{i=1}^M \|\beta_j^i - \bar{\beta}\|_2 \tag{5}$$

and (σ_1^2, σ_2^2) are the corresponding variances of the imposed priors.

In this work we optimize the criterion $G(\mathbf{B})$ using the gradient descent algorithm given that this loss

function is differentiable and we can compute all the partial derivatives

$$\frac{\partial G(\mathbf{B})}{\partial \beta_k^{(s)}} = \frac{\partial N(\mathbf{Z},\mathbf{B})}{\partial \beta_k^{(s)}} + \frac{\partial R(\mathbf{B})}{\partial \beta_k^{(s)}} \quad (6)$$

The first term $N(\mathbf{Z},\mathbf{B})$ only depends on the parameter matrix $\mathbf{B}$ and can be directly obtained differentiating the negated log-likelihood estimator of the tasks set $\mathbf{Z}$.

The second term $R(\mathbf{B})$ depends on $\mathbf{B}$ and $\bar{\beta}$, and can be rewritten as follows

$$R(\mathbf{B}) = \sum_{\mathbf{j=1}}^{\mathbf{D}} [\frac{\bar{\beta}_\mathbf{j}^2}{\sigma_\mathbf{1}^2} + \frac{\mathbf{1}}{\sigma_\mathbf{2}^2} \sum_{\mathbf{i=1}}^{\mathbf{M}} (\beta_\mathbf{j}^\mathbf{i} - \bar{\beta}_\mathbf{j})^\mathbf{2}] \quad (7)$$

Given that the second term of the sum depends also on β, we need to express it as a function of $\mathbf{B}$. Nevertheless, notice that we we can obtain an expression of each β_j depending only $(\mathbf{B})$ since we want to minimize $R(\mathbf{B})$. Thus, we have

$$\bar{\beta}_j = \arg\min_b (\frac{b^2}{\sigma_1^2} + \frac{1}{\sigma_2^2} \sum_{i=1}^{M} (\beta_j^i - b)^2) \quad (8)$$

and this expression has a global minimum at

$$\bar{\beta}_j(\mathbf{B}) = \frac{\sigma_1^2 \Sigma_{i=1}^M \beta_j^i}{\sigma_2^2 + M\sigma_1^2} \quad (9)$$

Given that the $\frac{\partial \bar{\beta}_j(B)}{\partial \beta_k^{(s)}} = 0$ if $j \neq k$, we obtain:

$$\frac{\partial \bar{\beta}_j(B)}{\partial \beta_j^{(s)}} = \frac{\sigma_1^2}{\sigma_2^2 + M\sigma_1^2} \quad (10)$$

Then, we get a final expression for the derivatives of the term $R(\mathbf{B})$ substituting here

$$\frac{\partial R(B)}{\partial \beta_k^{(s)}} = \frac{2\bar{\beta}_k}{\sigma_1^2} \frac{\partial \bar{\beta}_k}{\partial \beta_k^{(s)}} + \frac{2}{\sigma_2^2} \sum_{i=1}^{M} [(\beta_k^{(i)} - \bar{\beta}_k) \frac{\partial \bar{\beta}_k}{\partial \beta_k^{(s)}}] \quad (11)$$

the expressions in equations 9 and 10.

3 EXPERIMENTS

To test this proposed shared logistic regression model in face classification field we have performed different subject verification experiments using images from the public AR Face Database (Martinez and Benavente, 1998).

To perform the experiments we have used only the internal part of the face images and this fragments have been resized to be 16×16 pixels. All the images used in the training and test step are aligned by the center pixel of each eye. Some examples of these images are shown in figure 1.

Figure 1: The corresponding processed training and test images: resized fragments of the original images including the internal part of the face.

The experiments have been repeated 10 times, and the subjects identifiers, the training images and the test images have been always randomly selected. Given that the shared models are specially appropriated when the training set has small size, to train this verifications we have used only 2 positive samples (images from the subject we want to verify) and 4 negative samples (images from other subjects). To test the system we have used 20 positive images and 40 negative samples in each case.

We have considered different task groups that have from 1 to 10 different tasks, using the classical single-task logistic regression model and the shared presented model to train all the tasks of each group at a time. The results are shown in table 1, considering that a correct verification is the correct classification of a subject in its corresponding positive or negative label.

The subject classification results obtained show that when more than 4 tasks are simultaneously trained sharing information, the general accuracies increase. The performance of our proposal progressively increases as new tasks are added to the system. However, when there are a few tasks to train our shared logistic regression model, the optimization method fails in obtaining the proper model parameters.

4 CONCLUSIONS

In this paper we introduce a shared logistic regression classifier applied to a face verification problem. We show that the theoretic benefits of the inductive transfer knowledge in the machine learning process stated in the recent literature can be practically applied in a probabilistic modelling of a real life problem. The results obtained in the face verification application, show that considering the training of the different subject classification tasks sharing the model information yields improved accuracies. Moreover,

Table 1: Mean accuracies and 95% confidence intervals of the logistic regression method trained separately (first row) and following our shared logistic approach (second row). When more than 4 verification tasks are simultaneously trained, the error rates of the shared approach become lower.

	1	2	3	4	5
Logistic	68.1 ± 8.2	65.5 ± 6.4	69.5 ± 5.3	68.2 ± 4.2	69.8 ± 3.9
Shared Logistic		59.4 ± 4.2	64.2 ± 5.4	67.9 ± 5.2	71.3 ± 5.2
	6	7	8	9	10
Logistic	68.2 ± 3.6	68.6 ± 3.3	70.4 ± 3.3	69.8 ± 3.0	70.4 ± 2.9
Shared Logistic	72.8 ± 4.2	76.4 ± 3.1	78.2 ± 2.8	82.5 ± 2.4	84.6 ± 2.3

the improvement is more significant when the number of jointly trained verification tasks increases, being a 15% higher in the case of 10 simultaneous verifications.

The probabilistic modelling presented in this paper suggests new lines of future research. In our first formulation, the sharing knowledge property is imposed by constraining the parameter space of the classifiers along the multiple tasks. Other approaches could be followed, such as a more complex modelling based on a hidden model that generates the parameter space.

Moreover, the addition of extra related tasks from different domains could be studied. For example, a gender or ethnicity recognition problem. The enlargement of the task pool should benefit the amount of shared information between the related tasks, mitigating the effects of the small sample size problem in face verification.

ACKNOWLEDGEMENTS

This work is supported by MEC grant TIN2006-15308-C02-01, Ministerio de Ciencia y Tecnologia, Spain.

REFERENCES

Ando, R. and Zhang, T. (2005). A framework for learning predicitve structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

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

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

Evgeniou, T., Micchelli, C., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, 7:179–188.

Fukunaga, K. and Mantock, J. (1983). Nonparametric discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(6):671–678.

Hyvarinen, A. (1999). The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Process. Lett.*, 10(1):1–5.

Lee, D. and Seung, S. (2000). Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562.

Martinez, A. and Benavente, R. (1998). The AR Face database. Technical Report 24, Computer Vision Center.

Masip, D., Kuncheva, L. I., and Vitria, J. (2005). An ensemble-based method for linear feature extraction for two-class problems. *Pattern Analysis and Applications*, 8:227–237.

Masip, D. and Vitrià, J. (2006). Boosted discriminant projections for nearest neighbor classification. *Pattern Recognition*, 39(2):164–170.

Moghaddam, B., Wahid, W., and Pentland, A. (1998). Beyond eigenfaces: Probabilistic matching for face recognition. In *Proc. of Int'l Conf. on Automatic Face and Gesture Recognition (FG'98)*, pages 30–35, Nara, Japan.

Thrun, S. and Pratt, L. (1997). *Learning to Learn*. Kluwer Academic.

Torralba, A., Murphy, K., and Freeman, W. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.