

TRAFFIC SIGN CLASSIFICATION USING ERROR CORRECTING TECHNIQUES

Sergio Escalera, Petia Radeva

*Computer Vision Center, Dept. Computer Science, UAB, 08193 Bellaterra, Spain
sescalera@cvc.uab.es, petia@cvc.uab.es*

Oriol Pujol

*Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via 585, 08007, Barcelona, Spain
oriol@cvc.uab.es*

Keywords: Traffic Sign Classification, Error Correcting Output Codes, Ensemble of Dichotomies, Multiclass Classification.

Abstract: Traffic sign classification is a challenging problem in Computer Vision due to the high variability of sign appearance in uncontrolled environments. Lack of visibility, illumination changes, and partial occlusions are just a few problems. In this paper, we introduce a classification technique for traffic signs recognition by means of Error Correcting Output Codes. Recently, new proposals of coding and decoding strategies for the Error Correcting Output Codes framework have been shown to be very effective in front of multiclass problems. We review the state-of-the-art ECOC strategies and combinations of problem-dependent coding designs and decoding techniques. We apply these approaches to the Mobile Mapping problem. We detect the sign regions by means of Adaboost. The Adaboost in an attentional cascade with the extended set of Haar-like features estimated on the integral shows great performance at the detection step. Then, a spatial normalization using the Hough transform and the fast radial symmetry is done. The model fitting improves the final classification performance by normalizing the sign content. Finally, we classify a wide set of traffic signs types, obtaining high success in adverse conditions.

1 INTRODUCTION

Traffic sign classification in uncontrolled environments is a hard task in Computer Vision due to the high variability of symbol appearance caused by illumination changes, lack of visibility, or occlusions. In the last years, several approaches to deal with the problem have been proposed. Usually, traffic sign recognition strategies are divided into two main groups: color-based and grey scale-based. Grey scale-based approaches focus on object geometry, whereas color-based techniques allow to prevent false positives detection. Traffic sign recognition is studied for several purposes, like autonomous driving under certain simplified conditions or for assisted driving (Handmann et al., 1998). We focus on the goal of mobile mapping (Casacuberta et al., 2004), as a technique used to compile cartographic information from a mobile system. One of the main difficulties that makes this problem hard is the great number of classes and the high resemblance among signs in poor

resolution images. In order to deal with these hindrances, robust multiclass classifiers must be considered.

Error Correcting Output Codes were born as an alternative for handling multiclass problems using binary classifiers (Dietterich and Bakiri, 1995). It is well-known that ECOC, when applied to multiclass learning problems, can improve the generalization performance (Windeatt and Ghaderi, 2003)(Allwein et al., 2002). One of the reasons for this improvement is its property to decompose the original problem into a set of complementary two-class problems -coded in the ECOC matrix- that allows sharing of classifiers across the original classes.

Recently, there has been a renewed interest in the design of Error Correcting Output Codes. The common pre-designed coding strategies (one-versus-one and one-versus-all) have been improved with problem-dependent designs (Pujol et al., 2006)(Escalera et al., 2006b). On the other hand, new studies on the decoding step (Escalera et al., 2006a) have

shown that the performance of the ECOC classification can be improved considering carefully the decoding strategy applied. The new approaches take into account that when one use a third symbol (zero) in the ECOC matrix, that means that a particular class is not considered by a classifier. In those cases, the behavior of the decoding strategies should be adapted to the influence of the zero symbol (Escalera et al., 2006a).

In this paper, we deal with the problem of traffic sign classification. We use the information obtained from a Mobile Mapping System (Casacuberta et al., 2004) to analyze the road video sequences. We use Adaboost with the Haar-like features estimated over the integral image (Viola and Jones, 2002) to detect regions with high probability of containing a traffic sign. After applying a spatial normalization and model fitting, we classify the candidate signs in their different categories. We compare the recently proposed coding and decoding strategies in the framework of Error Correcting Output Codes, showing the improvement of these last techniques when problem-dependent ECOC designs are combined with proper decoding strategies. The proposed ECOC designs robustly classify several types of signs with high variability.

The paper is organized as follows: section 2 overview the Error Correcting Output Codes and the state-of-art on coding and decoding strategies. Section 3 explains the system for traffic signs classification. Section 4 shows experimental results, and section 5 concludes the paper.

2 ERROR CORRECTING OUTPUT CODES

The basis of the ECOC framework is to create a codeword for each of the N_c classes. Arranging the codewords as rows of a matrix, we define a "coding matrix" M , where $M \in \{-1, 0, 1\}^{N_c \times n}$ in the ternary case, being n the code length. From the point of view of learning, M is constructed by considering n binary problems (dichotomies), each corresponding to a matrix column. Joining classes in sets, each dichotomy defines a partition of classes (coded by +1, -1, according to their class set membership, or 0 if the class is not considered by the dichotomy). In fig.1 we show an example of a ternary matrix M . The matrix is coded using 7 dichotomies $\{h_1, \dots, h_7\}$ for a four multiclass problem (c_1, c_2, c_3 , and c_4). The white regions are coded by 1 (considered as positive for its respective dichotomy, h_i), the dark regions by -1 (considered as negative), and the grey regions correspond to the

zero symbol (not considered classes for the current dichotomy). For example, the first classifier is trained to discriminate c_2 versus c_1, c_3 and c_4 , the second one classifies c_3 versus c_1 , and so on. Applying the n trained binary classifiers, a code is obtained for each data point in the test set. This code is compared to the base codewords of each class defined in the matrix M , and the data point is assigned to the class with the "closest" codeword (Allwein et al., 2002)(Windeatt and Ghaderi, 2003). In the case of the figure, a new test input x is evaluated by all the classifiers and the systems assigns the label c_1 with minor Euclidean decoding distance $d(x, y^i) = \sqrt{\sum_{j=1}^n (x_j - y_j^i)^2}$, where y is a class codeword, and n is the total number of binary classifiers.

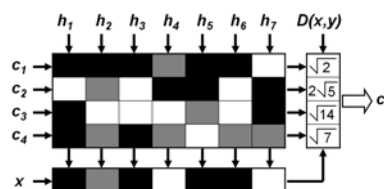


Figure 1: ECOC design and input test classification.

3 TRAFFIC SIGN CLASSIFICATION SYSTEM

We focus on the goal of mobile mapping to compile cartographic information from a mobile system. In particular, we use the video sequences obtained from the Mobile Mapping System of (Casacuberta et al., 2004). In this system, the position and orientation of the different traffic signs are measured in movement with the car video cameras. The system has a stereo pair of calibrated cameras, which are synchronized with a GPS/INS system. Therefore, the result of the acquisition step is a set of stereo-pairs of images with their position and orientation information.

The traffic sign recognition system used is divided in three main steps: object detection, model fitting, and classification. Each of these steps must be robust enough to minimize the propagation of errors in the system.

The detection process is based on the face detector presented by Viola and Jones in (Viola and Jones, 2002). In particular, we use the Discrete version of Adaboost with decision stumps (Friedman et al., 1998). The weak classifiers are trained using the attentional cascade based on the extended set of Haar-like features (that is, including the rotated ones) estimated on the integral image (Viola and Jones, 2002).



Figure 2: Detected traffic signs.

As a result of the detection process, we obtain results as in fig. 2.

Given an image where the Adaboost learning algorithm detected a road sign, a region of interest (ROI) that contains a sign is determined (circular or triangular). However, since we have missing information about sign scale and position, before the recognition process we apply a spatial normalization to improve final recognition. In particular, the Hough transform (Morse, 2000) and fast radial symmetry (Loy and Zelinsky, 2003) are applied in order to fit the model since they offer great robustness against noise.

The fast radial symmetry is calculated over a set of one or more ranges, depending on the scale of the features one is trying to detect. The value of the transform at a range indicates the contribution to radial symmetry of the gradients at a distance n away from each point. At each range n , we examine the gradient g at each point p , from which a corresponding positively-affected pixel $p_{+ve}(p)$ and negatively-affected pixel $p_{-ve}(p)$ are determined and accumulated in the orientation projection image O_n : $P_{\pm ve}(p) = p \pm \text{round}(\frac{g(p)}{\|g(p)\|}n)$, $O_n(P_{\pm ve}(p)) = O_n(P_{\pm ve}(p)) + 1$. Now, to locate the center of radial symmetry, we search for the position (x, y) of maximal value in the accumulated orientations matrix $O^T = \sum_{i=1}^n O_n$. Locating that maximum we determine the radius length. This procedure allows to obtain robust results for circular traffic signs fitting.

The Hough transform has been shown to allow the detection of straight lines in a robust way. We apply this procedure in order to look for the three representative lines of the triangular sign and calculate their intersections to transform the image. Nevertheless, we need to consider additional restrictions to obtain the three representative border lines of a triangular traffic sign. Each line has associated a position in relation to the others. Once we have the three detected lines we calculate their intersection. Given the parameters a and b that define the equation $y = a \times x + b$ for each of the three lines, the intersection point (X, Y) for each pair of lines is defined as follows:

$$X_t = (b_2^i - b_1^i) / (a_1^i - a_2^i), \quad Y_t = a_1^i X_t + b_1^i \quad | \quad t, i \in [1, \dots, 3] \quad (1)$$

To assure that the lines are the expected ones, we complement the procedure searching for a corner at a circular region at each intersection surroundings:

$$S = \{(x_i, y_i) \mid \exists p < ((x - x_i)^2 + (y - y_i)^2 - r^2)\} \mid i \in [1, \dots, 3] \quad (2)$$

where S is the set of valid intersection points, and p corresponds to a corner point to be located in a neighborhood of the intersection point.

Once the sign model is fitted using the commented methods, the next procedure is the spatial normalization of the shape before classification. The steps are: transform the image to make the recognition invariant to small affine deformations re-scaling to the signs database size (32×32 pixels), filter with Weickert anisotropic filter, and mask the image to exclude background at the classification step. To prevent the effects of illumination changes, the histogram equalization improves image contrast and obtains a uniform histogram.

From 10 analyzed DVD video sequences, we have obtained the classes in fig. 3. The classes are divided in three main groups: speed, circular, and triangular, with a total of 27 different classes to recognize. The speed signs are treated as a special case due to their similarity and difficulty to discriminate in adverse conditions. The three attentional cascades (one for each group) have been trained using a total of 1500 positive samples divided into the tree different groups.

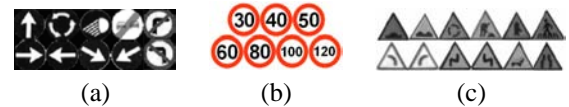


Figure 3: (a) Speed classes. (b) Circular classes. (c) Triangular classes.

Applying the three attentional cascades at Mobile Mapping System video sequences, the detected and normalized regions are classified, depending of the type of the detected sign, using different classification strategies combining the presented coding and coding strategies of Error Correcting Output Codes.

4 RESULTS

Applying the attentional cascades over a test set of 15000 road frames obtained from the Mobile Mapping System, we have detected 1200 regions that contain traffic signs. The detected regions are divided in the three groups of fig. 3. After applying the model fitting and the spatial normalization, all pixels are treated as a 1024 feature vector. The strategies used to validate the classification are Discrete Adaboost with decision stumps, and linear SVM with the regularization parameter C set to 1. These two classifiers generate the set of binary problems to embed in the set of ECOC configurations: one-versus-one, one-versus-all, dense-random, DECOC (Pujol et al.,

Table 1: Multiclass SVM results.

| Problem | Accuracy |
|------------|------------|
| Speed | 76.96±0.84 |
| Circular | 97.02±0.77 |
| Triangular | 95.74±0.99 |

2006), and ECOC-ONE (Escalera et al., 2006b). Each of the ECOC strategies are evaluated using different decoding strategies: Euclidean distance, Laplacian (Escalera et al., 2006a) and Pessimistic β -Density decoding (Escalera et al., 2006a). For the dense random case, where we have selected n binary classifiers for a fair comparison with one-versus-all and DECOC designs in terms of a similar number of binary problems. The classification tests are performed using stratified ten-fold cross-validation with 95% of the confidence interval.

We have generated three types of experiments, each one for each of the three different traffic signs groups. The mean rankings for each classification strategy using the results of the three presented experiments. The ranking is shown in fig. 4. One can observe that the best position is obtained by the ECOC-ONE strategy, followed by one-versus-one, DECOC, one-versus-all, and finally dense random strategy. It is important to note that for each ECOC designs, the Laplacian, and β -Density increase the classification accuracy of Euclidean decoding for all the cases.

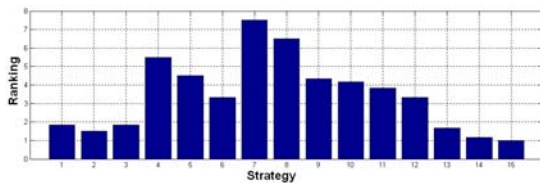


Figure 4: Ranking position for each classification strategy. From left to right: (1)-one-versus-one Euclidean, (2)-one-versus-one Laplacian, (3)-one-versus-one β -Density, (4)-one-versus-all Euclidean, (5)-one-versus-all Laplacian, (6)-one-versus-all β -Density, (7)-dense random Euclidean, (8)-dense random Laplacian, (9)-dense random β -Density, (10)-decoc Euclidean, (11)-decoc Laplacian, (12)-decoc β -Density, (13)-ecoc-one Euclidean, (14)-ecoc-one Laplacian, (15)-ecoc-one β -Density.

To show the robustness of the presented classification framework, we compare the results obtained with the ECOC methods with a built-in multiclass SVM. The results are shown in table 1. One can observe that the linear multiclass SVM obtains inferior results to the ones obtained by one-versus-one and ECOC-ONE strategies.

5 CONCLUSIONS

In this paper, we presented a classification scheme that obtains a very high performance for the problem of traffic sign classification. The system has three main stages: traffic sign detection, model fitting and spatial normalization, and sign categorization. The multiclass classification techniques are evaluated on real video sequences obtained from a Mobile Mapping System. We compared the state-of-the-art and recently proposed designs for Error Correcting Output Codes, and we combined them with robust decoding strategies, showing high robustness and better performance than traditional ECOC designs and the state-of-the-art multiclassifiers. In particular, the Laplacian and β -Density decoding strategies when applied to the coding designs improve the system performance. The presented traffic sign recognition system obtains robust classification results in front of high variability of the objects appearance.

ACKNOWLEDGEMENTS

This work was supported in part by the projects, FIS-G03/1085, FIS-PI031488, MI-1509/2005, and TIN2006-15308-C02-01.

REFERENCES

- Allwein, E., Schapire, R., and Singer, Y. (2002). Reducing multiclass to binary: A unifying approach for margin classifiers. In *Journal of Machine Learning Research*, volume 1, pages 113–141.
- Casacuberta, J., Miranda, J., Pla, M., Sanchez, S., A.Serra, and J.Talaya (2004). On the accuracy and performance of the geomobil system. In *International Society for Photogrammetry and Remote Sensing*.
- Dietterich, T. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. In *Journal of Artificial Intelligence Research*, volume 2, pages 263–286.
- Escalera, S., Pujol, O., and Radeva, P. (2006a). Decoding of ternary error correcting output codes. In *CIARP, Lecture notes in Computer Science*.
- Escalera, S., Pujol, O., and Radeva, P. (2006b). Ecoc-one: A novel coding and decoding strategy. In *International Conference on Pattern Recognition*.
- Friedman, J., Hastie, T., and Tibshirani, R. (1998). Additive logistic regression: a statistical view of boosting. In *Technical Report*.
- Handmann, U., Kalinke, T., Tzomakas, C., Werner, M., and von Seelen, W. (1998). An image processing system for driver assistance. In *IEEE International Conf. on Intelligent Vehicles*, pages 481–486.

- Loy, G. and Zelinsky, A. (2003). Fast radial symmetry for detecting points of interest. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 25.
- Morse, B. (2000). Segmentation (edge based, hough transform). In *Technical report*.
- Pujol, O., Radeva, P., and Vitrià, J. (2006). Discriminant ecoc: a heuristic method for application dependent design of error correcting output codes. In *T. Pattern Analysis and Machine Intelligence*, volume 28, pages 1007–1012.
- Viola, P. and Jones, M. (2002). Robust real-time object detection. In *International Journal of Computer Vision*.
- Windeatt, T. and Ghaderi, R. (2003). Coding and decoding for multiclass learning problems. In *Information Fusion*, volume 1, pages 11–21.

FACE VERIFICATION BY SHARING KNOWLEDGE FROM DIFFERENT SUBJECTS

David Masip¹, Àgata Lapedriza² and Jordi Vitrià²

¹ *Department of Applied Mathematics and Analysis (MAiA), University of Barcelona (UB)
Edifici Històric Gran Via de les Corts Catalanes 585, Barcelona, Spain
davidm@maia.ub.es*

² *Computer Vision Center, Department of Computer Science
Universitat Autònoma de Barcelona, Edifici O, Bellaterra 08193, Spain
{agata,jordi}@cvc.uab.es*

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 ex-

traction 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-

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 (Torralla 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

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

where

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

and $\beta \in \mathbb{R}^D$ is the parameters vector that is usually estimated by the maximum likelihood criterion. Notice that $P(L = -1 | \mathbf{x}, \beta) = 1 - P(L = 1 | \mathbf{x}, \beta)$ 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, \dots, T_M .

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

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

where each $\beta^i = (\beta_1^i, \dots, \beta_D^i)$ 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

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

Given that situation, normally the negated log-likelihood $N(\mathbf{Z}, \mathbf{B})$ is used to estimate the parameters adding a regularization term, usually $\frac{1}{\sigma^2} \|\mathbf{B}\|_2$, 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 $\bar{\beta} = (\bar{\beta}_1, \dots, \bar{\beta}_D)$ as:

$$\bar{\beta}_j = \frac{\sum_{i=1}^M \beta_j^i}{M} \quad (4)$$

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

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

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

$$R(\mathbf{B}) = \sum_{j=1}^D \left[\frac{\bar{\beta}_j^2}{\sigma_1^2} + \frac{1}{\sigma_2^2} \sum_{i=1}^M (\beta_j^i - \bar{\beta}_j)^2 \right] \quad (7)$$

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

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

and this expression has a global minimum at

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

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

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

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

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

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

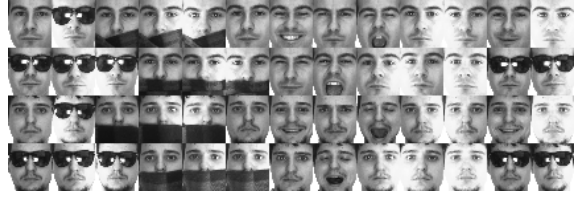


Figure 1: The corresponding processed training and test images: resized fragments of the original images including the internal part of the face.

the center pixel of each eye. Some examples of these images are shown in figure 1.

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 Tecnología, Spain.

REFERENCES

- Ando, R. and Zhang, T. (2005). A framework for learning predictive 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.
- Torrallba, 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*.