A CO-TRAINING APPROACH TO AUTOMATIC FACE RECOGNITION

Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay

Multimedia Communication Department, EURECOM
2229 Route des Cretes , BP 193, F-06560 Sophia-Antipolis Cedex, France
email: {zhaox, evans, dugelay}@eurecom.fr

ABSTRACT

Semi-supervised face recognition using both labelled and unlabelled data has received considerable interest in recent years. Co-training is one of the most well-known semi-supervised learning methods, but its application in face recognition almost remains unexplored because its assumption of conditional independence can be rarely satisfied between two facial features. However, even if two facial features are not completely independent, their different characteristics produce a so-called "classification margin" between two classifiers based on them, and hence there is the possibility of mutual training. In this paper, we report a semi-supervised face recognition algorithm which applies co-training on two classifiers based on Linear Discriminant Analysis (LDA) and Local Binary Patterns (LBP) features respectively.Experimental results show not only that the proposed co-training algorithm significantly improves the recognition accuracy over supervised methods using only labelled training data, but also demonstrates the superiority of co-training over self-training methods which only use one facial feature.

1. INTRODUCTION

Automatic Face Recognition (AFR) systems aim to recognise or verify the identity of a person from a digital image or a video source. The standard approach involves the learning of a face model (or template) for each client using appropriate features extracted from sufficient training images, and its subsequent comparison to test images according to some distance metric. However, many practical AFR scenarios are characterized by weakly trained models involving only a small number of labelled training data. When these face models are used to identify test images which inevitably contain inter-session variations in illumination, occlusion, pose and expression, performance can be unacceptable when these variations are not reflected in the models. Meanwhile, in some applications, a large pool of unlabelled auxiliary data can often be obtained easily since its collection does not entail costly manual labelling. For example, in access control applications, labelled training images acquired in the enrollment step are generally obtained in a single session and often limited in quantity. In this case the training data is rarely representative of variations in appearance due to ageing or illumination for example. But, during the operation of the system and without time constraints, huge amounts of unlabelled face images can be acquired. When collected over sufficient duration they should be more representative of intersession variation and may thus be used to enhance templates or models.

Semi-supervised learning (SSL) refers to a general class of machine learning techniques that make use of both labelled and unlabelled data for training, typically a small amount of labelled data and a larger amount of unlabelled data [1]. Existing SSL approaches include: self-training and co-training [2], semi-supervised SVM [3], graph-based semi-supervised learning [4], etc. A number of attempts to develop SSL approaches to face recognition have also been reported previously. Roli and Marcialis [5] proposed one of the first whereby a PCA-based classifier is initially weakly trained with a small number of manually labelled examples before it is used to classify unlabelled auxiliary data to augment the training set. In related work, also applied to PCA-based classifiers, Roli [6] proposed a variation in which 3 independent classifiers where used. In this work unlabelled auxiliary data are added to augment the labelled dataset only if more than two classifiers agree on the classification result. Both approaches, use a projected PCA sub-space as the feature space. The approach, however, is not robust to lighting and pose variations which is hence an inherent limitation. Linear Discriminant Analysis (LDA) is one of the most popular discriminative linear projection techniques for feature extraction, and is a powerful tool for face recognition when sufficient and representative training examples are available [7]. A semi-supervised face recognition algorithm based on LDA self-training is proposed in [9]. Results show that it can outperform PCA-based methods by a large margin.

Co-training is a well-known SSL algorithm which was proposed by Blum and Michell [2] in 1998. The basic idea is that features exhibit some redundancy and can thus be separated into two feature subsets where each of them is sufficient for correct classification. First two classifiers can be trained weakly using a small number of labelled examples and two different feature sets respectively. Each classifier is then used to classify the unlabelled data. The most positive examples are then used to train the other classifier. The process is iterative and is repeated several times. The key property is that some examples which are mis-classified by one classifier are confidently and correctly labelled by the other. These examples are thus highly informative in training the first classifier. In the orginal approach of co-training [2], Blum and Mitchell claimed that co-training can be applied only if the two features are conditionally independent, but later work of Goldman and Zhou [10] has shown that this independence assumption can be released to some extent.

Many different features, with different levels of conditional independence, have been successfully applied in AFR problems. Well-known examples include the sub-space projections of the original face image vector, for example, Principle Component Analysis (PCA) [11] and Linear Discriminant Analysis (LDA)[7], or features which aim to capture local image structures such as Local Binary Pattern (LBP) [12]. AFR classifiers which exploit different features may have different characteristics, for example, different levels of robustness to lighting and pose variations. A face image mis-classified by classifier A could be correctly classified by classifier B, and vise versa. This so-called "classification margin" between different features implies the potential improvement that co-training can bring over other semi-supervised learning methods based on single features (even if such features do not entirely satisfy the conditional independence assumption).

In this paper, we propose a new co-training face recognition approach based on LDA and LBP features. LDA is a supervised dimension reduction technique that has been successfully applied in AFR problems; satisfactory performance is typically obtained when large and representative labelled training samples are available. LBP is one of the most successful unsupervised feature extraction techniques for AFR and is also dependent on sufficient training data. Due to the distinctive nature of the two feature extraction techniques, we can safely assume a certain classification margin between the two classifiers. The hypothesis under investigation in this paper is that they are complementary and thus that there is potential for them to be successfully harnessed within a co-training scenario. There is obvious utility where labelled data is scarce.

The remainder of this paper is organized as follows. The LDA and LBP feature extraction and classification techniques are described in Section 2. The LDA-LBP co-training algorithm is described in Section 3. Experiments and results are detailed in Section 4 before our conclusions are presented in Section 5.

2. LDA AND LBP FACIAL FEATURE EXTRACTION AND RECOGNITION

In this section, the basic LDA and LBP feature extraction and classification methods will be briefly presented.

2.1 Baseline LDA face recognition

Linear subspace analysis has been used for AFR over many years and is now a well-known simple, efficient and proven approach. LDA is a supervised algorithm which, according to an optimised projection W_{opt}, projects data vectors x_i into a new space where the ratio between the inter-class (or between, S_B) and intra-class (or within, S_W) scatter is maximized. S_W and S_B are determined according to:

$$S_W = \sum_{j=1}^{c}\sum_{i=1}^{l_j}(x_i^j - \mu_j)(x_i^j - \mu_j)^T, \qquad (1)$$

and:

$$S_B = \sum_{j=1}^{c} l_j(\mu_j - \mu)(\mu_j - \mu)^T, \qquad (2)$$

where x_i^j is the i^{th} sample, μ_j is the mean, and where c is the number of classes, and l_j is the number of samples, all in class j, and where the global mean, subsuming all classes, is denoted by μ. We further define the total scatter according to:

$$S_T = \sum_{i=1}^{l}(x_i - \mu)(x_i - \mu)^T, \qquad (3)$$

where l is the total number of samples such that $S_T = S_B + S_w$. W_{opt} is then obtained according to the objective function:

$$W_{opt} = arg\,max_W \frac{W^T S_B W}{W^T S_T W} = [\mathbf{w_1}, \cdots, \mathbf{w_m}], \qquad (4)$$

where $[\mathbf{w_1}, \cdots, \mathbf{w_m}]$ are the eigenvectors of S_B and S_T which correspond to the m largest generalized eigenvalues according to:

$$S_B \mathbf{w_i} = \lambda_i S_T \mathbf{w_i},\ i = 1, \cdots, m, \qquad (5)$$

where λ_i is the ith largest eigenvalue. Note that there are at most $c-1$ non-zero generalized eigenvalues. Since S_W is often singular it is common to first apply principal component analysis (PCA) to reduce the original image vector to a g-dimensional vector, where $l > g > c-1$, before to LDA is used to obtain $(c-1)$-dimensional vectors.

Described above is the well-known Fisherface [7] algorithm. It gives satisfactory performance but tends to require a relatively high number of labelled training samples to learn reliable projections. When the quantity of training data is low, the S_w can be unreliable and result in poor performance [7].

2.2 Baseline LBP face recognition

The Local Binary Pattern (LBP) operator was introduced by Ojiala *et al.* [12] as a method of texture analysis. For each pixel, the operator considers a 3 × 3 neighborhood and thresholds the neighboring pixels with the center pixel value. Formally, the LBP operator takes the form:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n, \qquad (6)$$

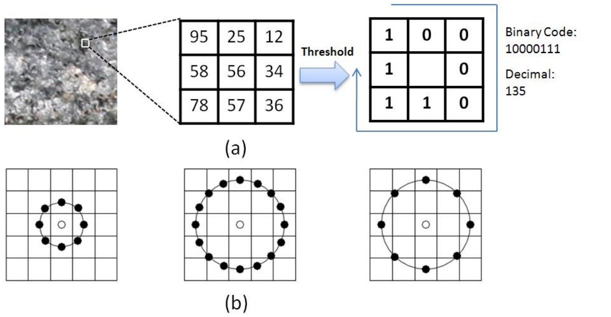

Figure 1: (a) basic LBP operator, (b) the circular (8,1), (16,2) and (8,2) neighborhood

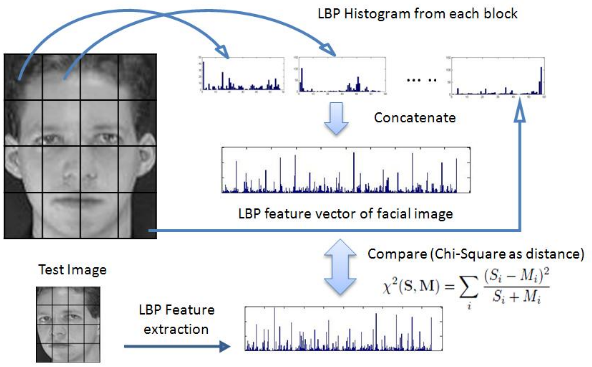
LBP Histogram from each block
Concatenate
LBP feature vector of facial image
Compare (Chi-Square as distance)
$\chi^2(S,M)=\sum_i \frac{(S_i-M_i)^2}{S_i+M_i}$
Test Image
LBP Feature extraction

Figure 2: LBP face recognition

where i_c is the intensity value of the center pixel and i_n is the intensity value of the 8 neighboring pixels, and where the index n of the summation corresponds to the 8 binary number. $s(u)$ is 1 if $u \geq 0$ and 0 otherwise. The result is considered as an 8-bit binary number and is assigned to the center pixels. As a result, each pixel of the image has an LBP value between 0 and 255. After that, a 256-bin histogram of these LBP values of the whole picture is then calculated, and is used as a feature vector. The LBP encoding process is illustrated in Figure 1(a). The LBP concept was later extended in two ways [13]. First, in order to deal with textures at different scales, the LBP operator was extended to use neighborhoods of different sizes. The local neighborhood is defined as a set of sampling points evenly spaced on a circle, and binary interpolation is applied when the sample point does not fall in the center of a pixel. The notation (P, R) implies P sampling points on a circle of radius R. See Figure 1(b) for an example. The second extension defined the so-called *uniform patterns*: an LBP is "uniform" if it contains at most one 0-1 and one 1-0 transition when viewed as a circular bit string. For example, the LBP code in Figure 1(a) is uniform. It is noticed that only 57 of the 256 8-bit patterns are uniform, but they account for 90% of all observed image neighbourhoods[13]. In the computation of LBP histogram, uniform patterns are used so that the histogram has a separate bin for every uniform pattern and all non-uniform patterns are assigned to a single bin. In this way, the number of bins are significantly reduced without losing too much information.

The application of LBP in AFR problems was first introduced by Ahonen *et al.* in [14]. The LBP face recognition process is illustrated in Figure 2. The facial image is divided into local regions and texture descriptors are extracted from each region independently. The descriptors are then concatenated into a single long vector to form a global description of the face. In the recognition step, the LBP features of the test image is extracted and compared to the LBP features of training images, and a Chi-square distance metric is often applied. Distance between two vectors x and ξ is defined as:

$$\chi^2(\mathbf{x},\xi) = \sum_i \frac{(x_i-\xi_i)^2}{x_i+\xi_i}. \tag{7}$$

3. LDA-LBP CO-TRAINING ALGORITHM

In this section, we describe how the LDA and LBP face recognition systems are combined in a co-training framework.

The LDA self-training methods proposed in [9] shows that, provided an auxiliary unlabelled dataset, the performance of an LDA face recognition system can be enhanced by iteratively adding most positive samples into the labelled training set. However, a theoretical deficiency of self-training methods is that, a data point which could be confidently labelled by a classifier contains little new information with which to retrain the same classifier. Its robustness to different variations not seen in the training set remains unchanged. The co-training algorithm, on the other hand, generally uses two classifiers based on two conditionally independent features. In our case, the assumption of conditional independence does not apply since they are different views of the same image. There is, however, a so called "classification margin" between two classifiers since they use different features, that is a data point which could be correctly classified by one classifier might be misclassified by the other, and hence it should be informative for training the second classifier.

We propose to apply an LDA based face recognizer and a LBP based face recognizer in a co-training scheme. The input to the system is a labelled dataset $\mathbf{D_l}$ and a larger unlabelled auxiliary dataset $\mathbf{D_u}$. First the LBP features of the $\mathbf{D_l}$ are extracted, and for each class, a template is calculated by averaging all the LBP features of the image in the same class. The LDA algorithm is also applied on $\mathbf{D_l}$, so the original image vectors can be projected into the $(c-1)$-dimensional LDA feature space. A template is calculated using the projected mean of images in each class. The LBP and LDA features of the unlabelled images in $\mathbf{D_u}$ are then extracted, and for each feature, the unlabeled samples are

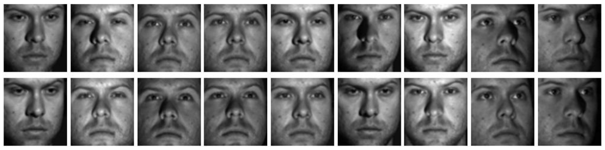

Figure 3: Sample images of Yale face database in subset 1 and 2

assigned the label of its nearest template. Chi-square and euclidean distance metrics are used for LBP and LDA systems respectively. Then, for each feature and for each class, the single example which is nearest to the corresponding template is removed from $\mathbf{D_u}$ and added in $\mathbf{D_l}$. The enlarged $\mathbf{D_l}$ is then used to relearn the LDA projection and LBP templates. This process is repeated iteratively until $\mathbf{D_u}$ is empty. A less conservative strategy can also be used whereby, upon each iteration, more than one automatically labelled example is added to the training data for each class. This results in a less computational demanding algorithm but one which does not capitalise on all the additional training data when each individual sample is selected. Improved computational efficiency thus comes at the cost of reduced performance.

4. EXPERIMENTAL RESULTS

The goal of our experiments is to evaluate the capability of the LDA-LBP co-training algorithm to used unlabelled image data and hence to improve the recognition performance over the supervised algorithms which only use labelled data. The co-training algorithm is trained with a small amount of labelled training data and a large quantity of automatically labelled face images which include variations such as pose, illumination and expression. To this end, we conducted experiments with Yale University face database B [15]. This database contains 5760 single light source images of 10 subjects (persons). Each subject has 9 poses and each pose has 64 different illumination conditions. The size of each image is 640×480 and, for computational efficiency, the images are resized to 64×48. We note that even such aggressive down sampling has only a small impact on performance. The images are divided into five subsets according to the light-source angle θ: Subset 1 ($\theta < 15°$ from optical axis), Subset 2 ($20° < \theta < 25°$), Subset 3 ($35° < \theta < 50°$), Subset 4 ($60° < \theta < 77°$), and Subset 5 (others). In our experiment, in order to avoid extreme lighting conditions, only images from Subset 1 and 2 are used, which include 1710 images in total (171 for pictures each subject). Figure 3 shows sample images used in our experiments and serves to illustrate the range of pose, illumination and expression.

Out of the 171 images of each subject, 85 images were randomly selected to be training images, while the others constituted a test image set. After that, l images of the training set were randomly selected as labelled images, while the others are used as unlabelled images. For the LBP recognizer, we used a 4×4 division of the face images and apply an (8, 2) uniform LBP operator. Results reported below are obtained from 10 repetitions of each experiment.

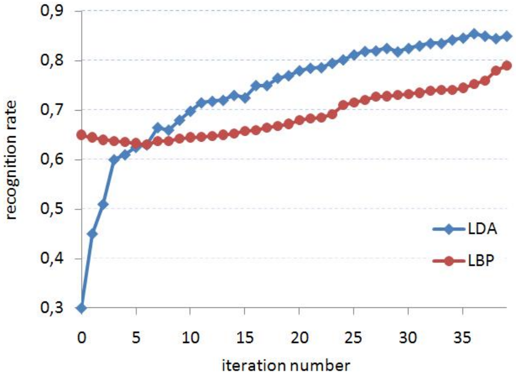

Figure 4: Average accuracy on test set as function of iteration number of co-training, initialized with 2 labelled examples per subject

Results for independent LBP and LDA classifiers are illustrated in Figure 4. For this experiment, 2 labelled samples per subject are used to initialize the algorithm. Performance is shown as a function of the number of co-training iterations. In every iteration, 20 images are added to the labelled training set. The profiles show that the recognition accuracy of both LDA and LBP classifiers increases when a greater number of images is incorporated into the training set with co-training: LBP classification accuracy increases from 65% to 79% while LDA classification accuracy increases from 30% to 86%.

In order to demonstrate the advantage of co-training, which is based on the "classification margin" of different classifiers over the self-training methods, we compared its performance with self-training methods. Besides the LDA self-training algorithm proposed in [9], we also implemented an LBP self-training algorithm: a nearest template classifier based on LBP features is built based on the means of LBP features of a few labelled images, and the classifier is then used to label the unlabelled dataset and the most positive examples are added to the labelled training set, and the process is repeated.

We conducted experiments with different numbers of initial labelled images per subject. The baseline performances are achieved with the supervised LBP and LDA algorithms respectively, trained with l (l=2 to 6) labelled examples per subject. Then, provided with the auxiliary set of unlabelled data, LBP self-training, LDA self-training, and co-training are applied respectively and the performance is summarzied in Table 1. In each column of the table, we observe that: (1) with different number of labelled data, self-training always improves the performance of LDA classifier while it does not improve the LBP classifier when the labelled dataset is sufficiently large ($l \geq 5$); (2) co-training always improves the per-

	$l = 2$	$l = 3$	$l = 4$	$l = 5$	$l = 6$
Baseline LBP	63%	70%	72%	78%	80%
Baseline LDA	30%	57%	67%	78%	84%
LBP self-training	72%	73%	75%	76%	78%
LDA self-training	78%	88%	92%	93%	94%
LBP co-training	80%	82%	84%	85%	85%
LDA co-training	86%	91%	93%	95%	95%

Table 1: Comparison of performances with different number of initial labelled examples per subject

formance of both classifiers over their supervised version; (3) co-training provides a more significant gain in performance to both LBP and LDA classifiers than self-training methods. From the observations we conclude that co-training is more stable, and makes more efficient use of the unlabelled data.

5. CONCLUSION

This paper presents a new semi-supervised face recognition algorithm based on LDA-LBP co-training. Two different classifiers based on two different feature sets are first weakly trained with a few labelled examples and are used to label an auxiliary unlabelled dataset. The most positive samples identified by one classifier are used to retrain the other classifier. The process is iterative and can use any quantity of unlaballed auxiliary data. The LBP-LDA co-training algorithm performs better than the self-training of each single classifier and demonstrates the potential of exploring the "classification margin" between classifiers in a semi-supervised face recognition scenario. In the case presented in this paper the two features do not satisfy the assumption of conditional independence but satisfactory results are still obtained. Better results might be obtained when the feature sets are conditionally independent.

6. ACKNOWLEDGEMENT

This work was partially supported by the TABULA RASA project funded under the 7th Framework Programme of the European Union (EU), grant agreement number 257289, and by a Futur et Ruptures award from Institut TELECOM.

REFERENCES

[1] X. Zhu, "Semi-supervised Learning Literature Survey, " Technical report, Univ. Wisconsin, Madison, USA, Jan. 2006.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann, pp. 92-100, 1998

[3] K. P. Bennett and A. Demiriz, "Semi-Supervised Support Vector Machines," *Advances in Neural Information Processing Systems*, 11, pp. 368-374, 1999

[4] X. J. Zhu, "Semi-supervised learning with graphs," ISBN:0-542-19059-1, Order Number:AAI3179046, Year of Publication: 2005

[5] F. Roli, G.L. Marcialis, "Semi-supervised PCA-based face recognition using self training," in *Proc. Joint IAPR Int. Work. on Structural and Syntactical Pattern Recognition and Statistical Techniques in Pattern Recognition*, S+SSPR06, Springer LNCS 4109: 560-568, 2006.

[6] F. Roli. "Semi-supervised multiple classifier systems: Background and research directions," In *Proc. of the 6th International Workshop on Multiple Classifier Systems*, pp. 1-11, 2005.

[7] A.M. Martinez and A.C. Kak, "PCA versus LDA, " in *IEEE Transaction on PAMI*, vol. 23(2), PP. 228-233, Feb. 2001.

[8] D. Cai, X. He, and J. Han. "Semi-supervised discriminant analysis," In *Proc. ICCV*, 2007

[9] X. Zhao, N. Evans and J. Dugelay. "Semi-supervised Face Recognition with LDA Self-training," *2011 International Conference in Image Processing*, Brussels, Belguim, 2011

[10] S. Goldman, Y. Zhou. "Enhancing supervised learning with unlabeled data," in *Proc. 17th International Conference on Machine Learning*, San Francisco, USA, 2000

[11] M. Turk and A. Pentland. "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, 3(1):71-86, 1991.

[12] T. Ojala, M. Harwood. "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, 29, 1996

[13] T. Ojala, M. Pietikainen, T. Maenpaa. "Multiresolution gray-scale and rotation invarianat texture classification with local binary patterns," *IEEE Transaction on PAMI*, 24(7), pp. 971-987, 2002

[14] T. Ahonen, A. Hadid, M. Pietikainen. "Face recognition with local binary patterns," in *Proc. ECCV 2004*, LNCS, vol. 3021, pp. 469-481. Springer, Heidelberg, 2004.

[15] Georghiades, A.S. and Belhumeur, P.N. and Kriegman, D.J., "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," in *IEEE Transaction on PAMI*, Vol 26(6), pp. 643-660, 2001.