FEATURE SELECTION VIA SIMULTANEOUS SPARSE APPROXIMATION FOR PERSON SPECIFIC FACE VERIFICATION

Yixiong Liang, Lei Wang, Shenghui Liao, Beiji Zou

School of Information Science and Engineering, Central South University,
Changsha, Hunan 410083, China
{yxliang, wanglei, lsh, bjzou}@mail.csu.edu.cn

ABSTRACT

There is an increasing use of some imperceivable and redundant local features for face recognition. While only a relatively small fraction of them is relevant to the final recognition task, the feature selection is a crucial and necessary step to select the most discriminant ones to obtain a compact face representation. In this paper, we investigate the sparsity-enforced regularization-based feature selection methods and propose a multi-task feature selection method for building person specific models for face verification. We assume that the person specific models share a common subset of features and novelly reformulated the common subset selection problem as a simultaneous sparse approximation problem. The effectiveness of the proposed methods is verified with the challenging LFW face databases.

Index Terms— Person specific face verification, feature selection, multi-task learning, simultaneous sparse approximation

1. INTRODUCTION

Although face recognition has achieved significant progress under controlled conditions in the past decades, it is still a very challenging problem in the uncontrolled environment such as the web where pose, lighting, expression, age, occlusion and makeup variations are more complicated. As local areas are often more descriptive and more appropriate for dealing with those variations, there is an increasing use of some imperceivable local features for face verification. Those local descriptors are generally extracted by performing some transformation (both linear or nonlinear) on the local region only or followed by some explicitly spatial pooling means such as the spatial histogramming scheme [1]. These initial representation is often redundant or over-completed, whereas only a relatively small fraction of them is relevant to the recognition task. Thus feature selection is a crucial and necessary step to select the most discriminant ones from the local features to obtain a compact face representation, which can not only improve performance but also decrease the computational burden.

Adaboost-based method is the most popular and impressive feature selection methods in face recognition Scenario [2, 3, 4, 5]. It applies the simple weak classifier, which only consists in a threshold on the value of a single feature, many times on differently weighted version of data and therefore obtaining a sequence of weak classifiers corresponding to the selected features. One possible problem of these methods is very time consuming because of the need of training and evaluating a different classifier for each feature. An alternative is the sparsity-enforced regularization techniques [6] which is the state-of-the-art feature selection tool in bioinformatics and recently has been successful applied in face detection and verification [7]. The main advantages of the regularization approach are its effectiveness even in the high dimensionality small sample size cases coupled with the support of well-grounded theory [6].

The concern of this work is mainly about how to build person specific models for both feature selection and face verification in unconstrained environments. In this case, although the face verification can be seen as a binary classification problem (accept or reject), it is in fact several binary classification problems (one for each client model) and thus its essence is by nature a multiple binary classification problem. Most existing approaches train a generic model for all individuals [2, 3, 4], which may fail to capture the variations among different individuals and therefore are suboptimal. Other approaches build person specific models for different individuals separately [7] and often lead to overfitting due to the small sample size of each individual. To combat over the overfitting problem, Wang et al. [5] adopted multi-task learning to improve the generalization performance of the Adaboost-based methods.

In this paper, we investigate the multi-task generalization of regularized methods and propose a multi-task feature selection method for person specific face verification. We assume that different person specific models share a common subset of relevant features and novelly reformulate the common subset selection problem as a simultaneous sparse approximation problem. The classification can be done by simple linear regression such as ridge regression. The experiment results on

the LFW face database [8] demonstrate the advantages and effectiveness of the proposed methods.

2. NOTATION AND SETUP

Suppose that there are L individuals to be verified. Given a training image set of size N, among them N_l images correspond to the subject l, while the remaining images are of other subjects excluding the known L subjects. From each image we can obtain a d-dimensional feature vector $\mathbf{f}$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{N \times d}$ be the data matrix with each row an input feature vector, and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L] \in \mathbb{R}^{N \times L}$ be the corresponding indicator response matrix where $\mathbf{y}_l$ is a N-dimensional vector with its ith entry equal to 1 if the ith samples come from the subject l and else equal to 0. Therefore $\mathbf{Y}$ is a matrix of $0's$ and $1's$ with each row having at most a single 1. We write $\mathbf{c}_l$ for the l-th column of the matrix $\mathbf{C}$ and $\mathbf{c}^l$ for the l-th row.

3. SINGLE-TASK FEATURE SELECTION

We restrict ourselves to the case of regression models that are linear in the components of feature. For class l, this linear relationship can be characterized in matrix notation

$$\mathbf{y}_l = \mathbf{X}\mathbf{c}_l + b_l \mathbf{1}, (l = 1, \ldots, L) \qquad (1)$$

where $\mathbf{c}_l$ is a d-dimensional coefficient vector and b_l is the *bias* in the model of class l respectively, while $\mathbf{1}$ being a vector with its entries equal to 1. The square error loss function

$$\text{Err}(\mathbf{c}_l, b_l) = \|\mathbf{y}_l - \mathbf{X}\mathbf{c}_l - b_l \mathbf{1}\|_2^2 \qquad (2)$$

is adopted to fit the above linear model to the given training set. Minimizing the square error loss function directly yields a unique solution known as the least squares solution, which is typically non-sparse and thus do not provide the feature selection in the sense. A natural generalization for feature selection is l_0 regularization

$$\min_{\mathbf{c}_l, b_l} \text{Err}(\mathbf{c}_l, b_l) + \lambda \|\mathbf{c}_l\|_0, \qquad (3)$$

where $\| \cdot \|_0$ is the l_0 *quasi-norm* counting the nonzero entries of a vector and λ quantifies how much improvement in the approximation error is necessary before we admit an additional term into the approximation. It is a classic combinatorial sparse approximation problem which is a NP-hard in general. A lot of numeric methods has been proposed to solve the above combinatorial sparse approximation problem and two most common approaches are greedy methods and convex relaxation methods. Greedy techniques such as OMP abandon exhaustive search but iteratively construct a sparse approximate one step at a time by selecting the columns maximally reduces the residual and use it to update the current approximation. Convex relaxation methods replace the combinatorial

sparse approximation problems with a related convex version that can be solved more efficiently. As the l_1 *norm* provides a natural convex relaxation of the l_0 *quasi-norm*, the basis pursuit (BP) method solves the sparse approximation problem by introducing an l_1 *norm* in place of the l_0 *quasi-norm*

$$\min_{\mathbf{c}_l, b_l} \text{Err}(\mathbf{c}_l, b_l) + \lambda \|\mathbf{c}_l\|_1, \qquad (4)$$

which is an unconstrained convex function and can be solved by some standard mathematical programming softwares. Similarly, the parameter λ negotiates a compromise between approximation error and sparsity. It is also known as LASSO [6].

Provided the regularization coefficient λ is same across different individuals, then solving each of these problems independently is equivalent to solving the global problem obtained by summing the objectives:

$$\min_{\mathbf{C}, \mathbf{b}} \sum_{l=1}^{L} \frac{1}{N_l} \text{Err}(\mathbf{c}_l, b_l) + \lambda \sum_{l=1}^{L} \|\mathbf{c}_l\|_0, \qquad (5)$$

Where $\mathbf{C}$ is the coefficient matrix with $\mathbf{c}_l$ in columns and $\mathbf{b} = [b_1, \ldots, b_L]^T$ is the bias vector. Similarly, the corresponding l_1 *norm* relaxation objective is

$$\min_{\mathbf{C}, \mathbf{b}} \sum_{l=1}^{L} \frac{1}{N_l} \text{Err}(\mathbf{c}_l, b_l) + \lambda \sum_{l=1}^{L} \|\mathbf{c}_l\|_1. \qquad (6)$$

4. MULTI-TASK FEATURE SELECTION

In this section, we will describe our proposed multi-task feature selection in details. As mentioned before, we assume each face shares common subset of the redundant and imperceivable local features. It's reasonable because each face shares a common structure, i.e. face is composed of eyebrow, eye, nose, mouth, etc. In the regularized feature selection frame, sharing a small subset of features means that the coefficient matrix $\mathbf{C}$ has many rows which are identically equal to zero and the corresponding features will not be used for all tasks. Thus the global common feature selection can be formulated as searching minimum number of nonzero rows of $\mathbf{C}$ while balancing the error loss function

$$\min_{\mathbf{C}, \mathbf{b}} \sum_{l=1}^{L} \frac{1}{N_l} \text{Err}(\mathbf{c}_l, b_l) + \lambda \|\mathbf{C}\|_{row-l_0}, \qquad (7)$$

where $\| \cdot \|_{row-l_0}$ is the row-l_0 *quasi-norm* which denotes the number of nonzero rows and is given by

$$\|\mathbf{C}\|_{row-l_0} = |\bigcup_{l=1}^{L} \text{supp}(\mathbf{c}_l)|, \qquad (8)$$

where $\text{supp}(\cdot)$ denotes the support of a vector. When the matrix $\mathbf{C}$ is a column vector, the row-support degenerates to the

support of the vector and the row-l_0 *quasi-norm* degenerates to the usual l_0 *quasi-norm*. If we regard $\mathbf{X}$ as a dictionary and $\mathbf{y}_l(l = 1, \cdots, L)$ as a serious of signals to be approximated, the problem (7) is indeed a simultaneous sparse approximation problem.

It is immediately clear that the combinatorial optimization problem (7) is at least as hard as combinatorial optimization problem (3) and thus it is a more complicated NP-hard problem in general. Some greedy pursuit algorithms such as simultaneous orthogonal matching pursuit (SOMP) [9] are proposed to solve this combinatorial optimization problem. Another approach to simultaneous sparse approximation is to replace the row-l_0 *quasi-norm* by a closely related convex function [10]. There are many different ways to relax the row-l_0 *quasi-norm* and one may define an entire family of relaxations of the following form

$$\|\mathbf{C}\|_{p,q} = \sum_{i=1}^{d}(\|\mathbf{c}^l\|_q)^{p/q} = \sum_{i=1}^{d}[\sum_{j=1}^{L}|c_{ij}|^q]^{p/q}. \qquad (9)$$

This relaxation can be done by first applying the l_q *norm* to the rows of $\mathbf{C}$ and then applying the l_p *norm* or *quasi-norm* to the resulting vector of l_p *norm*. On the one hand, we want to obtain row-sparse of $\mathbf{C}$. On the other hand, we want the selected feature to contribute to as many individuals as possible. This requires most rows of $\mathbf{C}$ should be zero but the nonzero rows should have many nonzero entries. Therefore we have $p \le 1$ and $q > 1$. The rational behind this is that minimizing the $l_p(p \le 1)$ *norm* promotes sparsity whereas minimizing the $l_q(q > 1)$ *norm* promotes non-sparsity. If set $p = 1$ and $q = 2$, our method is equivalent to the multi-task feature selection frame proposed in [11]. In our implementation, we set $q = \infty$ since the l_∞ *norm* can provide better non-sparsity than l_2 *norm*. Replacing the row-l_0 *quasi-norm* in the combinatorial optimization problem (7) by its relaxation (9) with $p = 1$ leads to the following convex program

$$\min_{\mathbf{C},\mathbf{b}} \sum_{l=1}^{L} \frac{1}{N_l} \mathrm{Err}(\mathbf{c}_l, b_l) + \lambda\|\mathbf{C}\|_{p,q}, \qquad (10)$$

which can be solved by some standard mathematical programming software [10].

Recalled that the above feature selection frame can be used for classification directly in that it fits linear regression models to the class indicator variables. One can also consider its usage as a pure feature selection tool and explore some other common classifiers for classification. Moreover, the proposed method is not specific for face verification but to any other classification or regression problem, providing that the tasks share the same training data.
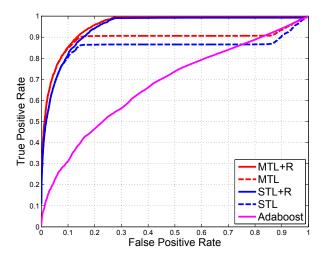
5. EXPERIMENTAL RESULTS

We carry out some experiments on the LFW face database [8]. The LFW face database contains $13,233$ labeled face images collected from news sites in the Internet. These images belong to $5,749$ different individuals and have high variations in position, pose, lighting, background, camera and quality, which make it appropriate to evaluate face verification methods in realistic and unconstrained environments. As there is not available protocol along with the database for person specific face verification, we select 158 people with at least 10 images in the database as the known people, i.e. $L = 158$. For each known people, we choose the former 5 images for training and the remaining for testing. We also select 210 people with only one image in the database as the background person (or unknown person) for training. Hence we have a training set of size $1,000$ corresponding to 368 people and a testing set of size $3,534$ from the known 158 people. Note that our training set is overwhelmingly *unbalanced* (5 positive samples and 995 negative samples with their ratio be close to $1 : 200$ for each person).

In our experiments, each image is rotated and scaled so that centers of the eyes are placed on specific pixels and then was cropped to 64×64 pixels. We choose Gabor feature as the initial representation due to its peculiar ability to model the spatial summation properties of the receptive fields of the so called "bar cells" in the primary visual cortex. We use 40 Gabor filters with five scales $\{0, \cdots, 4\}$ and eight orientations $\{0, \cdots, 7\}$ which are common in face recognition area to obtain the Gabor feature. The dimension d of the resulting feature is then $64 \times 64 \times 40 = 163,840$.

We apply both single-task and multi-task feature selection approach to select the most informative 300 features from the original $163,840$-dimensional Gabor features. From a runtime point of view, OMP and SOMP are adopted to solve the single-task and multi-task feature selection problems, respectively. The outputs of OMP and SOMP include both the indexes of the selected features and the corresponding weights and therefore can be used for verification directly. We also utilize the ridge regression method to determine the weights of the selected features. The corresponding verification methods are denoted as "STL", "MTL" and "STL+R" and "MTL+R". In addition, we adopt the Adaboost-based method as the baseline for feature selection and verification.

Those methods all can verify the training set exactly, but perform very differently on the testing set. We adopted the average ROC curves and the average area under ROC curves (AUC) to evaluate their performance across different persons. The comparative performance is shown in Fig. 1 and Table 1. The Adaboost-based method may suffer from the unbalance of the training set and performs much worse than the regularization-based methods. The proposed multi-task feature selection methods ("MTL" and "MTL+R") perform better than the corresponding single-task feature selection methods ("STL" and "STL+R"). Another observation is that the ridge regression-based verification does marginally improve the performance compared with directly using the feature selection frame for verification. This can be attributed to the fact

Fig. 1. Average ROC curves of verifying images of 158 known people only using 300 Gabor features

Table 1. The average true positive rates (TPR) using different methods when the false positive rate (FPR) is fixed at 0.1 and the average AUC

Methods	TPR(std. dev.)	AUC(std. dev.)
STL	0.8046($\pm$ 0.1600)	0.8506($\pm$ 0.1255)
MTL	0.8465($\pm$ 0.1458)	0.8901($\pm$ 0.0969)
STL+R	0.8185($\pm$ 0.1636)	0.9444($\pm$ 0.1458)
MTL+R	**0.8525($\pm$ 0.1480)**	**0.9586($\pm$ 0.0288)**
Adaboost	0.3112($\pm$ 0.1708)	0.6811($\pm$ 0.1066)

that the sparsity-enforcement in the feature selection frame may underestimate the resulting coefficients and hereby obtain the worse performance.

6. CONCLUSIONS

We have proposed a multi-task learning method for building of personal specific models both for feature selection and face verification. The person specific models are jointly learned by sharing the training data and then the multi-task feature selection problem can be reformulated as a simultaneous sparse approximation problem which can be solved by some greedy algorithms such as SOMP or some related convex relaxation methods. The experimental results show that the proposed multi-task feature selection method can overcome the potential overfitting issues due to the lack of training data and the adoption of ridge regression for verification can marginally improve the performance.

7. ACKNOWLEDGEMENT

8. REFERENCES

[1] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *TPAMI*, vol. 33, pp. 43–57, 2011.

[2] M. Jones and P. Viola, "Face recognition using boosted local features," in *ICCV*. IEEE, 2003.

[3] G.C. Zhang, X.S. Huang, S.Z. Li, Y.S.Wang, and X.H. Wu, "Boosting local binary pattern (lbp)-based face recognition," in *Proc. Advances in Biometric Person Authentication*, 2004, pp. 179–186.

[4] P. Yang, S.G. Shan, W. Gao, S.Z. Li, and D. Zhang, "Face recognition using ada-boosted gabor features," in *AFGR*. IEEE, 2004, pp. 356–361.

[5] X.G. Wang, C. Zhang, and Z.Y. Zhang, "Boosted multi-task learning for face verification with applications to web images and video search," in *CVPR*. IEEE, 2009, pp. 142–149.

[6] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction (2nd edition)," *Springer*, 2009.

[7] A. Destrero, C.De Mol, F. Odone, and A. Verri, "A regularized framework for feature selection in face detection and authentication," *Int. J. Comput. Vis.*, vol. 83, pp. 164–177, 2009.

[8] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *University of Massachusetts, Amherst, Technical Report 07-49*, 2007.

[9] J.A. Tropp, A.C. Gilbert, and M.J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, 2006.

[10] J.A. Tropp, A.C. Gilbert, and M.J. Strauss, "Algorithms for simultaneous sparse approximation. part ii: Convex relaxation," *Signal Processing*, vol. 86, pp. 589–602, 2006.

[11] G. Obozinski, B. Taskar, and M.I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Journal of Statistics and Computing*, pp. 1–22, 2009.