

# TRANSFORM-INVARIANT DICTIONARY LEARNING FOR FACE RECOGNITION

Shu Zhang, Man Zhang, Ran He and Zhenan Sun

Center for Research on Intelligent Perception and Computing,  
National Laboratory of Pattern Recognition, Institute of Automation,  
Chinese Academy of Sciences, Beijing 100190, China  
{shu.zhang, zhangman, rhe, znsun}@nlpr.ia.ac.cn

## ABSTRACT

Dictionary learning has important applications in face recognition. However, large transformation variations of face images pose a grand challenge to conventional dictionary learning methods. A large portion of misleading dictionary atoms are usually learned to represent transformation factors, which will cause ambiguity in face recognition. To address this problem, this paper proposes a general framework for transform-invariant basis matrix learning. Specifically, we present a transform-invariant dictionary learning method which explicitly incorporates an appearance consistent error term to the original objective function in dictionary learning. The unified objective function is effectively optimized in an alternating iterative way. An ensemble of aligned images and a discriminative transform-invariant dictionary for sparse coding can be obtained by solving the formulated objective function. Experimental results on two public face databases demonstrate our algorithm's superiority compared with two state-of-the-art dictionary learning methods and the recently proposed transform-invariant PCA method.

**Index Terms**— Transform-invariant, dictionary learning, sparse coding, face recognition

## 1. INTRODUCTION

Sparse coding learns to represent a  $n$ -dimensional signal  $y$  with a sparse linear combination of dictionary atoms  $d_i \in R^n$ , where  $i = 1, 2, \dots, N$  and  $N$  is the size of the dictionary. Practical applications of sparse coding in computer vision including image denoising [1] and restoration [2] prove it to be a powerful model. Recent advances in dictionary learning (DL) reveal that the dictionary learned directly from observed data [3] performs better than a predefined one. The application of DL on classification tasks has also attracted great attention from the computer vision community recently.

The main purpose of the conventional DL framework is signal representation rather than signal recognition. However,



**Fig. 1.** Dictionary atoms learned by LC-KSVD (left) and our TIDL method (right) with the largest 16 sparse coding coefficients on the LFW subsets. The LC-KSVD dictionary exhibits many transformation-related information, whereas our dictionary atoms are approximately eye-aligned.

the discriminative power can be obtained with the incorporation of some sorts of classification error term on the original reconstructive objective function. The learning of discriminative dictionaries can be classified into two categories based on how the classification phase is implemented. The first category [4][5] learns a sub-dictionary for each class respectively and concatenates them into a single dictionary. As in the classification stage, they follow the procedure of SRC [6] and reconstruction error is employed to indicate the classification results. Effective as they are, all these methods involve class specific dictionary learning. When dealing with thousands of classes, the training and subsequent sparse coding stage can be prohibitively slow.

Another category of methods involves training a single compact dictionary and then adopting the sparse coding coefficients for classification. Most methods belonging to this category learn a dictionary and a classifier at the same time with the incorporation of a classification error term, as in [7], the cost function of logistic regression, to the original DL objective function. Although such a technique is sophisticated, the resultant objective function is highly complex and hard

This work is funded by the National Natural Science Foundation of China (Grant No. 61135002, 61273272) and the Instrument Developing Project of the Chinese Academy of Sciences (Grant No. YZ201266).

to solve. The methods [8] and [9] incorporate a multi-class linear classification error term and a label consistent term respectively to enforce the discriminative power of the sparse coding coefficients, and elegantly solve the unified objective function with K-SVD [3] algorithm.

The development of discriminative DL [8][9] indicates it to be a promising trend for face recognition [10][11]. The DL framework is robust to illumination, expressions and even disguise and occlusion on face recognition, whereas it is severely affected by image transformations. Transformation in training images lead to blurred dictionary atoms accounting for transformations rather than discriminative structure of face images. Those atoms can't provide useful discriminative information but rather cause ambiguity for the classification task. This issue has caught attention of many researchers in the computer vision community [12][13][14]. Chen et al. [15] proposed to learn dictionaries in the radon transform domain to enforce the rotational invariant property of learned dictionaries. Deng et al. [16] derived a Transform-Invariant PCA (TIPCA) approach for automatic face alignment and representation. Improved experimental results prove the significance of employing transform invariant basis (principal components in PCA or dictionary atoms in DL) for face recognition task. However, the TIPCA framework can't guarantee recognition rate as it is an unsupervised learning process.

In this paper, we present a transform-invariant dictionary learning (TIDL) method for face recognition. The proposed method aims to simultaneously learn a transform-invariant discriminative dictionary and align the training ensembles. In virtue of the representative power of DL, the alignment of face images and the construction of face basis matrix are integrated into a unified objective function. Fig.1 demonstrates a comparison of dictionary atoms learned by LC-KSVD [9] and our method respectively. It is obvious that our dictionary atoms represent intrinsic structure of human faces that are approximately eye-aligned, whereas the LC-KSVD dictionary contains much transformation related information. Compared to previous work, this paper's contributions concentrate on the following three aspects: (1) we formulate a general framework for learning of transform-invariant basis matrix, and further point out that TIPCA is a special case of our general framework; (2) we explicitly incorporate an appearance consistent error term for learning of a transform-invariant dictionary for face recognition, and develop an alternating optimization approach to efficiently optimize the objective function; (3) we conduct experiments on two public face databases with various transform variations, the result of which validate our proposed work's ability in recognizing misaligned faces with dictionary learning framework.

The rest of the paper is organized as follows: Section 2 proposes a transform-invariant matrix basis learning framework and gives a detailed derivation of the TIDL method, Section 3 evaluates our proposed method on two public databases, Section 4 concludes this paper with possible future re-

search directions.

## 2. PROPOSED METHOD

For a set of  $N$  misaligned face images, we aim to construct a transform-invariant dictionary for accurate face recognition. Conventional methods construct dictionary directly from the misaligned image ensembles, thus learn many transform-related dictionary atoms. On the other hand, we align the image ensembles while learning the dictionary. To this end, we determine a warp  $\tau_i$  for each face image  $I_i$  in the face ensembles to get the aligned image  $I_i^o = I_i \circ \tau_i$ .

### 2.1. General Framework for Transform-Invariant learning

In this section, we propose a general framework for simultaneously learning a transform-invariant basis matrix  $U$  and aligning the training images. First we stack each column of an image  $I_i$  as a  $n$ -dimensional column vector  $y_i \in R^n$ . Suppose  $Y$  is a set of  $N$  images, i.e.  $Y = [y_1 \dots y_N] \in R^{n \times N}$ , and we want to represent  $Y$  as  $Y = UV$ , where  $U = [u_1 \dots u_K] \in R^{n \times K}$  is the basis matrix, with each column being a basis vector; and  $V = [v_1 \dots v_N] \in R^{K \times N}$  represent the coding parameters. We can formulate the the learning of  $U$  and  $V$  as:

$$\langle U, V \rangle = \arg \min_{U, V} \|Y - UV\|_F^2 \quad (1)$$

This is a general representative model for  $n$ -dimensional signals and the basis matrix  $U$  can either be a dictionary in DL or principal components in PCA. By minimizing the reconstruction error  $\|Y - UV\|_F^2$  subject to certain constraints, e.g. sparse constraint for  $V$  in DL, we can obtain a basis matrix  $U$ . To address the problem of aligning misaligned images or ensembles within the representative model, we explicitly incorporate the warp implementation  $Y \circ \tau$ ,  $\tau = [\tau_1 \dots \tau_N]$  for each training image in the original model, thus we can get an unified objective function which seeks to find a transform-invariant basis matrix and warp parameters at the same time:

$$\langle U, V, \tau \rangle = \arg \min_{U, V, \tau} \|Y \circ \tau - UV\|_F^2 \quad (2)$$

where  $Y \circ \tau$  denotes the image domain transform for all the images with a set of transformations  $[\tau_1 \dots \tau_N]$ . After the transformation, we will get a set of aligned images  $[y_1 \circ \tau_1 \dots y_N \circ \tau_N]$ . It is easy to derive that the TIPCA and TIDL algorithms are both special cases of this framework. Simply substituting the basis matrix  $U$  with principal components and imposing orthogonal constraint for the basis matrix in equation (2), we will get the TIPCA method introduced in [16]. In next section, we will give a detailed derivation of the TIDL method.

### 2.2. Transform-Invariant Dictionary Learning

As discussed above, by substituting the  $U, V$  with dictionary  $D$  and sparse coding coefficients  $X$  respectively, and impos-

ing sparsity constraint on each column of  $X$ , we can intuitively extend the above framework to DL, which is formulated as follows:

$$\langle D, X, \tau \rangle = \arg \min_{D, X, \tau} \|Y \circ \tau - DX\|_F^2 \text{ s.t. } \forall i, \|x_i\|_0 \leq T \quad (3)$$

where  $T$  is the sparsity constraint factor. By iteratively solving  $D, X$  and  $\tau$ , a transform-invariant dictionary can be obtained. In TIPCA [16], for fast convergence of its objective function, they introduce a scheme of using low-to-high dimensional eigenspace for alignment. However, in DL, there's no explicit indication for the rank of the dictionary atoms as in PCA. In order to ensure convergence, we propose to impose an appearance consistent error term  $\|Y \circ \tau - \mu(Y)\|_F^2$  on equation (3), where  $\mu(Y) \in R^{n \times N}$  and each column of  $\mu(Y)$  is the mean of the current aligned image ensembles (also called average face), denoted as  $\mu(y)$ . These two terms together will force each image  $y_i$  to align towards  $(\sum_{j=1}^K d_j x_j - \mu(y))/2$ , which is the mean of its sparse reconstruction  $\sum_{j=1}^K d_j x_j$  and the average face  $\mu(y)$ . After some iterations, the faces will be aligned to the same appearance model.

To obtain a dictionary with discriminative power for face recognition, we further combine previously proposed label consistent term [9] and classification error term [8] with the above objective function. A complete version of our proposed transform-invariant discriminative dictionary learning framework can be formulated as follows:

$$\begin{aligned} \langle D, W, A, X, \tau \rangle &= \arg \min_{D, W, A, X, \tau} \|Y \circ \tau - DX\|_F^2 \\ &+ \|Y \circ \tau - \mu(Y)\|_F^2 + \alpha \|Q - AX\|_F^2 + \beta \|H - WX\|_F^2 \\ &\text{s.t. } \forall i, \|x_i\|_0 \leq T \end{aligned} \quad (4)$$

where  $\|Q - AX\|_F^2$  is the label consistent term [9] to enforce discriminability of sparse coding coefficients,  $Q = [q_1 \dots q_N] \in R^{K \times N}$ , and  $q_i = [q_i^1 \dots q_i^K]^t = [0 \dots 1, 1, \dots 0]^t \in R^K$ . The non-zero value of  $q_i$  only occurs when signal  $y_i$  and the dictionary item  $d_k$  share the same label.  $A$  is a linear transformation matrix. This term enforces an explicit correspondence between dictionary atoms and class labels. The term  $\|H - WX\|_F^2$  represents the classification error,  $W$  are the parameters for a linear classifier  $H = WX$ , the non-zero position of each column in  $H$  is adopted to determine the class label of a certain signal  $y_i$ , e.g.  $h_i = [0, 0, \dots 1, \dots, 0, 0]^t \in R^m$ , where  $m$  denotes the number of classes.  $\alpha$  and  $\beta$  are scalars controlling the relative contribution of the corresponding terms.

The above formulation can be effectively solved by iterating over a two-step alternating optimization procedure, i.e. we iteratively optimize over  $\{D, W, A, X\}$  and  $\tau$  with the others fixed. To begin with, the same initialization setup is adopted as in [9] to learn a dictionary  $D$  for each class using

K-SVD and  $\{W, A, X\}$  can also be computed accordingly. Next, we would iterate over the following two steps.

**Step 1:** fix  $\{D, W, A, X\}$ , update warp parameter  $\tau$  (we use similarity transformation for  $\tau$  in our implementation). In this step, equation (4) is reduced to the following form:

$$\begin{aligned} \langle \tau \rangle &= \arg \min_{\tau} \|Y \circ \tau - (DX + \mu(Y))/2\|_F^2 \\ &= \arg \min_{[\tau_1 \dots \tau_N]} \sum_{i=1}^N (y_i \circ \tau_i - (\sum_{j=1}^K d_j x_j - \mu(y))/2) \end{aligned} \quad (5)$$

This equation can be obtained with the first two terms in (4) written together. This formulation can be seen as a typical image alignment [17] problem which solves for a  $\tau_i$  to register each image  $y_i$  to its template  $(\sum_{j=1}^K d_j x_j - \mu(y))/2$ . Following the convention of inverse compositional algorithm introduced in [17], this problem can be solved very efficiently.

**Step 2:** fix  $\tau$ , update  $\{D, W, A, X\}$  with the following optimization problem:

$$\begin{aligned} \langle D, W, A, X \rangle &= \arg \min_{D, W, A, X} \\ &\|Y' - DX\|_F^2 + \alpha \|Q - AX\|_F^2 + \beta \|H - WX\|_F^2 \\ &\text{s.t. } \forall i, \|x_i\|_0 \leq T \end{aligned} \quad (6)$$

where  $Y'$  is the aligned image from last step, calculated using  $Y \circ \tau$  with  $Y$  and  $\tau$  from the last step. This optimization problem can be effectively solved by LC-KSVD algorithm proposed in [9]. After iterations over the above two steps until the objective function stops to decrease, finally we can get a transform-invariant dictionary  $D$ , classifier parameters  $W$  and the aligned image ensembles. As for test stage, the probe images are first aligned using equation (5) with the learned  $D$  and aligned training ensembles  $Y_{align}$ , and then sparse codes of the aligned images are used as the input for classifier  $H = WX$  to yield the classification results.

### 3. EXPERIMENTS AND RESULTS

We evaluate our approach on two public face databases: the Extended YaleB database [18] and the LFW database [19]. To obtain image ensembles with transform variations, frontal faces in Extended YaleB database are perturbed with similarity transform using Matlab codes provided in [14]. For comparison purpose, we also report recognition results using D-KSVD [8], LC-KSVD [9], TIPCA [16] and PCA [20].

#### 3.1. Results on the Extended YaleB Database

The Extended YaleB database contains face images of 28 individuals with various poses and illumination conditions [18]. In our experiment, we only use a subset of the database, which contains 49 frontal face images (images with extreme illumination conditions are excluded) for each person, of which 30 images are used for training and the rest are used for testing.



**Fig. 2.** Face images used in our experiment.

	Extended YaleB	lfw
Methods	Acc±Std(%)	Acc(%)
D-KSVD[8]	90.51±0.98	60.00
LC-KSVD[9]	90.15±1.51	62.11
TIPCA[16]	80.06±1.72	43.16
PCA[20]	76.84±1.75	33.68
<b>TIDL</b>	<b>94.17±0.81</b>	<b>80.00</b>

**Table 1.** Recognition results on the Extended YaleB and LFW database.

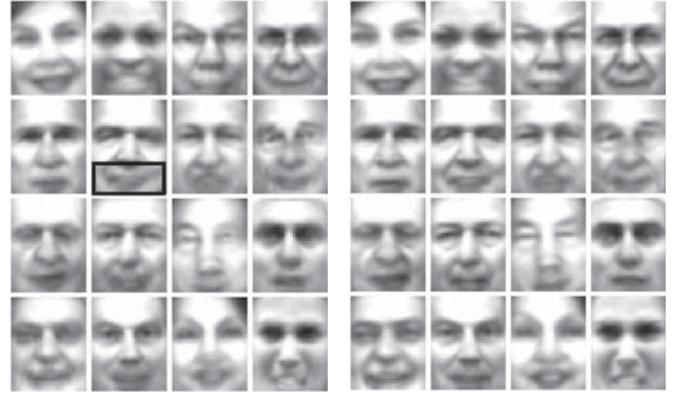
To accommodate the frontal face images to our experiments, we firstly eye-aligned each image to  $(15, 42)$  and  $(65, 42)$  in a  $100 \times 80$  crop with RASL toolbox [14]. Eye-aligned images are perturbed with similarity transformation to get image ensembles with transform variations. To further facilitate the search of optimal warp parameter with the inverse compositional algorithm, we crop the obtained image to  $80 \times 64$ , and resize them to  $40 \times 32$  for our experiment. As for face features, we stack each column of the face image to a single column vector as the input feature. We conduct the experiment 10 times, each time with 20 percent of images perturbed. Fig.2 illustrates some examples of our image ensembles.

The mean and standard derivation of the recognition results are reported in the second column Table 1, where the dictionary size is 308 (11 atoms per person) for DL based methods, the sparsity constraint factor  $T$  is set to 25, scalar  $\alpha$  and  $\beta$  are set to 2 and 3 respectively and the number of principal components used for recognition is 100 for PCA based methods. Histogram equalization is implemented in advance on the image ensembles for all five methods.

As the results indicate, both our TIDL and TIPCA outperform their counterparts, which validate the effectiveness of our proposed general framework and appearance consistent error term. By advocating learning a transform-invariant basis matrix for representation and classification, ambiguity caused by transform-related bases is excluded, thus recognition rate is improved accordingly. Another fact needed to point out is that as supervised learning methods, all 3 DL based methods outperform the unsupervised PCA based methods in face recognition task, which justify the importance of learning a discriminative dictionary.

### 3.2. Results on the LFW Database

The LFW database is designed for studying the problem of unconstrained face recognition [19]. Face images in this database exhibit various poses, illumination, expression variations, so this database is supposed to be very difficult for appearance based recognition methods like DL and PCA. We



**Fig. 3.** Average faces of 16 persons before and after TIDL learning, the left are average faces of the original training ensembles, the right are that of TIDL aligned ones. More detail information of the average face is revealed using our method, e.g. the mouth marked in a black rectangle

evaluate 5 methods on a subset of the LFW database provided by [14], with images from 19 persons, each have 30 images. Since the original images already have very large transform variations, there's no need to perturb them, but other setups remain the same with the last experiment on Extended YaleB. Among all the images, 25 images from each person are used for training, the other 5 are used for testing.

Recognition results are reported in the third column of Table 1. For DL based methods, the dictionary size is 190 (10 atoms per person), the sparsity constraint factor  $T$  is set to 15, scalar  $\alpha$  and  $\beta$  are set to 3 and 4 respectively; for PCA based methods, the number of principal components used for recognition is 100. Although the recognition results are not as satisfying as the last database, they demonstrate the same trend as analyzed before. A significant improvement in detail of the average faces after employing TIDL can be observed in Fig.3. This, along with Fig.1 which shows some learned transform-invariant dictionary atoms, validate the effectiveness of our algorithm's learning of transform-invariant dictionaries.

## 4. CONCLUSIONS

In this paper, we propose a general framework for transform-invariant basis matrix learning to address the problem of recognizing face images with large transformation variations. Specifically we derive a discriminative TIDL method with the introduction of an appearance consistent error term to enforce alignment while learning the dictionary. With its two main features, transform-invariance and discriminability, our method outperforms two state-of-the-art discriminative DL methods and the recently proposed TIPCA method on two public databases. Future work on this subject might include developing more robust alignment term and extending this framework to object classification.

## 5. REFERENCES

- [1] Michael Elad and Michal Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [2] Julien Mairal, Michael Elad, and Guillermo Sapiro, “S-sparse representation for color image restoration,” *Image Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 53–69, 2008.
- [3] Michal Aharon, Michael Elad, and Alfred M. Bruckstein, “k-svd: An algorithm for designing overcomplete dictionaries for sparse representation,” *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [4] Meng Yang, Lei Zhang, Jian Yang, and David Zhang, “Metaface learning for sparse representation based face recognition,” in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1601–1604.
- [5] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro, “Classification and clustering via dictionary learning with structured incoherence and shared features,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3501–3508.
- [6] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma, “Robust face recognition via sparse representation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.
- [7] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman, “Supervised dictionary learning,” *arXiv preprint arXiv:0809.3083*, 2008.
- [8] Qiang Zhang and Baoxin Li, “Discriminative k-svd for dictionary learning in face recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2691–2698.
- [9] Zhuolin Jiang, Zhe Lin, and Larry S Davis, “Learning a discriminative dictionary for sparse coding via label consistent k-svd,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1697–1704.
- [10] Zhen Lei, Rufeng Chu, Ran He, Shengcai Liao, and Stan Z Li, “Face recognition by discriminant analysis with gabor tensor representation,” in *Advances in Biometrics*, pp. 87–95. Springer, 2007.
- [11] Ran He, Wei-Shi Zheng, Bao-Gang Hu, and Xiang-Wei Kong, “Two-stage nonnegative sparse representation for large-scale face recognition,” *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 1, pp. 35–46, 2013.
- [12] Qi Li, Zhenan Sun, Ran He, and Tieniu Tan, “Joint alignment and clustering via low-rank representation,” in *Pattern Recognition (ACPR), 2013 2nd IAPR Asian Conference on*. IEEE, 2013, pp. 591–595.
- [13] Ran He, Zhenan Sun, Tieniu Tan, and Wei-Shi Zheng, “Recovery of corrupted low-rank matrices via half-quadratic based nonconvex minimization,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2889–2896.
- [14] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma, “Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 763–770.
- [15] Yi-Chen Chen, C.S. Sastry, V.M. Patel, P.J. Phillips, and R. Chellappa, “Rotation invariant simultaneous clustering and dictionary learning,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 1053–1056.
- [16] W. Deng, J. Hu, J. Lu, and J. Guo, “Transform-invariant pca: A unified approach to fully automatic face alignment, representation, and recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.
- [17] Simon Baker and Iain Matthews, “Lucas-kanade 20 years on: A unifying framework,” *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.
- [18] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.
- [19] Gary B Huang, Marwan Mattar, Tamara Berg, Eric Learned-Miller, et al., “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” in *Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition*, 2008.
- [20] Matthew A Turk and Alex P Pentland, “Face recognition using eigenfaces,” in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR ’91., IEEE Computer Society Conference on*. IEEE, 1991, pp. 586–591.