Simultaneous Image Transformation and Sparse Representation Recovery

Junzhou Huang
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854, USA
jzhuang@cs.rutgers.edu

Xiaolei Huang
Lehigh University
19 Memorial Drive West
Bethlehem, PA 18015, USA
xih206@lehigh.edu

Dimitris Metaxas
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854, USA
dnm@cs.rutgers.edu

Abstract

Sparse representation in compressive sensing is gaining increasing attention due to its success in various applications. As we demonstrate in this paper, however, image sparse representation is sensitive to image plane transformations such that existing approaches can not reconstruct the sparse representation of a geometrically transformed image. We introduce a simple technique for obtaining transformation-invariant image sparse representation. It is rooted in two observations: 1) if the aligned model images of an object span a linear subspace, their transformed versions with respect to some group of transformations can still span a linear subspace in a higher dimension; 2) if a target (or test) image, aligned with the model images, lives in the above subspace, its pre-alignment versions would get closer to the subspace after applying estimated transformations with more and more accurate parameters. These observations motivate us to project a potentially unaligned target image to random projection manifolds defined by the model images and the transformation model. Each projection is then separated into the aligned projection target and a residue due to misalignment. The desired aligned projection target is then iteratively optimized by gradually diminishing the residue. In this framework, we can simultaneously recover the sparse representation of a target image and the image plane transformation between the target and the model images. We have applied the proposed methodology to two applications: face recognition, and dynamic texture registration. The improved performance over previous methods that we obtain demonstrates the effectiveness of the proposed approach.

1. Introduction

The sparse representation theory has shown that sparse signals can be exactly reconstructed from a small number of linear measurements [4, 5, 7, 8, 17]. It leads to the problem: given the linear measurements $y \in \mathbb{R}^m$ of a sparse signal $x \in \mathbb{R}^n$, $y = Ax$, how to reconstruct the sparse signal x from its linear measurements y ? Obviously, this problem can be formulated with l^0 minimization:

$$x_0 = argmin\|x\|_0 \quad while \quad y = Ax \tag{1}$$

where $\|\cdot\|_0$ denotes the l^0-norm that counts the number of nonzero entries in a vector. This problem is NP-hard. In the general case, no known procedure can correctly find the sparsest solution more efficiently than exhausting all subsets of the entries for x. Recent developments in compressive sensing [6] show that if the original signal x is sparse enough, the above l^0-minimization problem is equivalent to the following l^1-minimization problem:

$$x_0 = argmin\|x\|_1 \quad while \quad \|y - Ax\|_2 < \epsilon \tag{2}$$

where ε denotes the noise level. If the solution is really sparse and has k nonzero entries, it can be efficiently solved by the homotopy algorithms in $O(k^3 + n)$ time [8]. If the signal has k nonzero entries, $m = O(k * log(n/m))$ linear measurements are sufficient to reconstruct the original signal exactly with high probability [4, 6]. Moreover, if the signals are not exactly k sparse but can be represented by k of active elements as well as contaminated with noise, sparse representation theory in compressive sensing can also handle this case with random projection analysis.

A novel and comprehensive approach for face recognition is recently proposed based on the sparse representation theory. [21]. The assumptions are that the training images of a single object span a subspace and that a target test image can be sparsely represented by the entire set of training images. Therefore, the face recognition problem is treated as searching for a sparse representation of a given test image. This treatment dexterously casts recognition as a globally sparse representation problem, and the sparse presentation theory in compressive sensing can then be utilized to efficiently solve it. Experimental results [21] showed that the sparse representation recognition approach achieved favorable performance compared with other state-of-the-art

methods under various conditions. It further demonstrates the power of sparse representation via l^1 minimization. One limitation of the method, however, is that it only handles cases in which all training images and the test image are well aligned and have the same pose. While the training images can be easily aligned off-line, aligning each test image to model images is a difficult task in practical applications. It is thus desirable to develop a transformation-invariant image sparse representation to overcome the difficulty.

In this paper, we propose a new algorithm to make sparse representation invariant to image-plane transformations. The proposed approach aims to simultaneously recover the image plane transformation and sparse representation when a test image is not aligned with the model images. It is inspired by two simple observations: 1) if the model images span a linear subspace, their transformed versions w.r.t. some group of small transformations still can span a linear subspace in a higher dimension; 2) if a transformed version of the test image, which is aligned with the model images, lives in the above subspace, the test image after applying more and more accurately estimated transformations will get gradually closer to the subspace. When the transformation between the test image and model images is small, the first observation motives us to convert a nonlinear model representation to a linear one by increasing the dimensionality of the model representation. However, this scheme is no longer effective in the presence of large transformations. To resolve this problem, we turn to the second observation and recent developments in random projection manifolds [2, 13], by iteratively projecting the unaligned test image to random-projection manifolds of an extended linear model. The projections can then be separated into the aligned projection target and some residue due to misalignment. The more accurate the estimated transformation parameters are, the closer the transformed version of the test image should be to the linear sparse representation. Under this framework, we can simultaneously recover the sparse representation of the target image based on the model images and the image plane transformation between the target image and the model images.

The remainder of the paper is organized as follows. Section 2 introduces the related work and the interested problem. Problem formulation and solution are detailed in section 3. Section 4 presents the experimental results when applying the proposed method to face recognition and dynamic texture registration, respectively. We conclude this paper in section 5.

2. Related Work

2.1. Sparse solution by l^1 minimization

As introduced above, sparse solutions can be obtained by performing l^1 minimization instead of l^0 minimization [6]. Thus, efficient l^1 minimization becomes the core of the sparse representation problem. Recently, several efficient algorithms are developed and they only require matrix-vector products operations [8, 15, 10].

The l_1-magic package implements the algorithms introduced in [5, 8]. The l^1 minimization is recasted as a second-order cone program and then a primal log-barrier approach is applied. In the process, only multiplications by A (eqn. 2) and its transpose are required. One l^1 regularization algorithm is developed for a large-scale least squares reconstruction problem [15]. A specialized interior-point method is employed to solve this problem, which uses preconditioned conjugate gradient method to approximately solve linear systems in a truncated-Newton framework. Each search step requires only multiplications by A and its transpose. The proposed $l1$-ls package is reported to outperform all previous implementations for l^1 minimization, including the l_1-magic package. Gradient Projection for Sparse Reconstruction (GPSR) is another interior point approach, which considers the sparse presentation reconstruction problem as a bound-constrained quadratic program. In order to accelerate convergence, a variant of Barzilai-Borwein steps is optionally applied when the projected gradient steps are used. Their experiments show the GPSR appears to be faster than state-of-the-art algorithms,including $l1$-ls, especially in large-scale settings. Moreover, it does not require application-specific tuning. Considering these advantages, the l^1 minimization in our algorithm is based on their *GPSR*-4.0 package.

2.2. Randomfaces

A pioneering attempt was conducted to use the sparse representation theory for face recognition and the proposed "Randomfaces" algorithm obtained very good face recognition results [21]. We briefly review their method here.

First, each image with size $w \times h$ is stacked as a vector $I_{i,n_i} \in \mathbb{R}^m$, where i is the subject number and n_i is the image number of each subject. The whole training image model can be represented as follows:

$$A = [I_{1,1}, I_{1,2}, ..., I_{1,n_1}, ..., I_{k,n_k}] \in \mathbb{R}^{m \times n} \tag{3}$$

Here, k is the total number of the subjects and n is the total number of training images. Based on the assumption that the vectors of each subject span a subspace for this subject, the new test image $y \in \mathbb{R}^m$ of subject i can be represented as a linear combination of the training images of subject i:

$$y = \sum_{j=1}^{n_i} \alpha_{i,j} I_{i,j} \tag{4}$$

where $\alpha_{i,j}, j = 1, ..., n_i$ are weights. Then the test image y of subject i can be sparsely represented in terms of all

training images:

$$y = Ax_0 \in \mathbb{R}^m \tag{5}$$

where $x_0 = [0, ..., 0, \alpha_{i,1}, ..., \alpha_{i,n_i}, 0, ..., 0] \in \mathbb{R}^n$ is a coefficient vector whose entries are zero except those associated with subject i. The sparse representation is obtained if subject number k is reasonably large. The only problem is that the dimension of the data is very high. Motivated by the theoretical results in [4, 6], random projection is used to reduce the data dimension:

$$\widetilde{y} = Ry = RAx_0 = \widetilde{A}x_0 \in \mathbb{R}^d \tag{6}$$

where $R \in \mathbb{R}^{d \times m}$ with $d \ll m$ is a random projection matrix. Until now, the face recognition is dexterously formulated as the linear sparse representation problem:

$$x_0 = argmin\|x\|_0 \quad while \quad \widetilde{y} = \widetilde{A}x \tag{7}$$

As introduced above, this problem is equivalent to the l^1 minimization problem in equation 2, which can be efficiently solved. The remaining problem is to identify the test image y by encoding x_0 after l^1 minimization:

$$identity(y) = argmin_i E[r_i], \quad E[r_i] = \frac{1}{l}\sum_{t=1}^{l} r_i^t \tag{8}$$

where r_i is the residual and $r_i(y) = \|\widetilde{y} - \widetilde{A}\delta_i(x)\|_2$. In the new vector $\delta_i(x)$, the entries in x associated with subject i keep unchanged and others are set as zeros.

Figure 1(c) shows several sparse representation examples by the *Randomfaces* algorithm [21]. We implemented the algorithm according to their paper as there is not public code available. The images are also from the Extended Yale B database [12]. This database consists of 2,414 frontal-face images of 38 individuals. The image size 192×168. All images are aligned and normalized. We randomly select half of the images for training (32 images per subject), and the other half for testing. In Figure 1, different rows represent different subjects. In each row, Column (a) shows one of training images of the subject. Column (b) shows one test image of this subject, whose sparse reconstruction based on sparse solution and the model images by Randomfaces algorithm is shown in column (c). Very good reconstruction results were obtained using their approach, which demonstrates the effectiveness of sparse representation.

However, the current Randomfaces algorithm can not handle the case where the test images are not aligned with the training images. In an experiment, we introduced small translations (15 pixels in both horizontal and vertical directions) to a set of test images so that they are not aligned with the model training images. Then all the images are cropped to the size 177×153. Several sparse representation results by the Randomfaces algorithm on the unaligned test images are shown in Figure 2(c). One can clearly see that there

(a) (b) (c) (d)

Figure 1. Sparse representation results on aligned images. One of training images (a), the test image (b), the linear sparse representation result [21] (c), and the sparse representation result using the proposed method.

are ghost effects due to the misalignment between the test images and the model images, which shows that the current linear sparse image representation approach depends on correct alignments and lacks of invariance to images plane transformations.

3. Transform-invariant Sparse Representation

In this section, we describe a sparse representation invariant to image-plane transformations.

3.1. Problem formulation

When there exist image plane transformations between the test images and the model images, the problems (1) and (2) can be reformulated as follows:

$$(x_0, \beta_0) = argmin\|x\|_1, \quad T(y, \beta_0) = Ax \tag{9}$$

where β_0 is the parameter of the image transformation between the test image y and model images A. Here, A is assigned according to equation 3 and $T(y, \beta_0)$ represents the transformed version of image y with parameter β_0. In this problem, given the model A and the unaligned image y, we attempt to simultaneously recover the sparse solution x_0 and the image plane transformation β_0. It is a typical Chicken-and-Egg problem. If we know the exact image plane transformation, the sparse solution can be easily obtained by l^1 minimization in problem (2) just as done in a previous approach [21]. If we know the exact sparse solution, we can obtain the sparse representation according to the sparse solution, and then the image plane transformation can be easily estimated by classical motion estimation

methods [3]. However, we know neither the image plane transformation nor the sparse solution. We therefore face a highly ill-posed problem.

(a) (b) (c) (d)

Figure 2. Sparse representation results given unaligned test images. Training images (a), test images (b), results by Randomfaces [21] (c), results by the proposed approach.

3.2. Algorithm

Our task is to simultaneously recover the image plane transformation and the sparse representation while the model images are aligned but the test image is not aligned to the model images. For convenience, we use face recognition as an example to introduce our algorithm below.

We consider the translation transformation first. Let $I(\mathbf{x})$ be an image where $\mathbf{x} = (x_1, x_2)$. $I(\mathbf{x}+\beta)$ represents its translated version with parameter $\beta = (a_1, a_2)$. When the transformation parameter β is small, we have:

$$T(I,\beta) = I(\mathbf{x}+\beta) \approx I(\mathbf{x}) + a_1 I_{x_1} + a_2 I_{x_2} \tag{10}$$

where I_{x_1} is $\frac{\partial I}{\partial x_1}$ and I_{x_2} is $\frac{\partial I}{\partial x_2}$.

Similarly, the affine transformed version of I can be represented as:

$$\begin{aligned} T(I,\beta) \approx I(\mathbf{x}) + a_1 I_{x_1} + a_2 x_1 I_{x_1} + a_3 x_2 I_{x_1} \\ + a_4 I_{x_2} + a_5 x_1 I_{x_2} + a_6 x_2 I_{x_2} \end{aligned} \tag{11}$$

where $\beta = (a_1, a_2, a_3, a_4, a_5, a_6)$. Now, let us consider the problem in equation 9.

In the case of translation transformation, let $\beta_0 = (a_1, a_2)$ represent the transformation between the model images and the test image y of subject i. Then, the translated version of the test image y', with parameter β_0, can be represented as:

$$y' = T(y, -\beta_0) = \sum_{j=1}^{n_i} \alpha_{i,j} I_{i,j} \tag{12}$$

Considering that translating y toward y' is equivalent to translating the aligned model images toward y, we can rewrite the above equation:

$$\begin{aligned} y = T(y', \beta_0) = \sum_{j=1}^{n_i} \alpha_{i,j} T(I_{i,j}, \beta_0) \\ = \sum_{j=1}^{n_i} [\alpha_{i,j} I_{i,j} + \alpha_{i,j} a_1 I_{i,j,x_1} + \alpha_{i,j} a_2 I_{i,j,x_2}] \end{aligned} \tag{13}$$

where I_{i,j,x_1} is $\frac{\partial I_{i,j}}{\partial x_1}$ and I_{i,j,x_2} is $\frac{\partial I_{i,j}}{\partial x_2}$. The equations form a linear system:

$$\begin{aligned} y = Bx, B = [A_0, A_1, A_2], x = [z_0, z_1, z_2]^T \\ z_0 = [0, ..., 0, \alpha_{i,1}, ..., \alpha_{i,n_i}, 0, ..., 0] \\ z_1 = [0, ..., 0, a_1\alpha_{i,1}, ..., a_1\alpha_{i,n_i}, 0, ..., 0] \\ z_2 = [0, ..., 0, a_2\alpha_{i,1}, ..., a_2\alpha_{i,n_i}, 0, ..., 0] \\ A_0 = [I_{1,1}, I_{1,2}, ..., I_{1,n_1}, ..., I_{k,n_k}] \\ A_1 = [I_{1,1,x_1}, I_{1,2,x_1}, ..., I_{1,n_1,x_1}, ..., I_{k,n_k,x_1}] \\ A_2 = [I_{1,1,x_2}, I_{1,2,x_2}, ..., I_{1,n_1,x_2}, ..., I_{k,n_k,x_2}] \end{aligned} \tag{14}$$

where we now obtain the linear image model $B \in \mathbb{R}^{m \times 3n}$ for translation transformations [1], as similarly done in [18]. In this way, the unaligned image y of subject i can be sparsely represented in terms of all training images and their derivatives. The random projection is used to reduce the data dimensionality:

$$\tilde{y} = Ry = RBx = \tilde{A}x \in \mathbb{R}^d \tag{15}$$

where $R \in \mathbb{R}^{d \times m}$ with $d \ll m$ is a random projection matrix. l^1 minimization instead of l^0 minimization is performed to derive the sparse solution:

$$x_0 = argmin\|x\|_1 \quad while \quad \|\tilde{y} - \tilde{A}x\|_2 < \epsilon \tag{16}$$

With the computed sparse solution $x_0 = [z_0, z_1, z_2]^T$, the random projection $\tilde{y}$ can be separated into the aligned projection target $\tilde{y'_{est}}$ and the residue $\tilde{y''}$:

$$Ry'_{est} = \tilde{y'_{est}} = \tilde{A}_0 z_0 = RA_0 z_0 \tag{17}$$

where y'_{est} is the estimation of the aligned version y' of the test image y. The recent developments [2, 13] in random projection manifolds provide the following scheme for estimating the aligned target y' from equation 17.

Lemma 1 *Let $\mathcal{M}$ be a compact k dimensional manifold in $\mathbb{R}^m$ having volume V and condition number $1/\tau$. Fix $0 < \varepsilon < 1$ and $0 < \rho < 1$. Let R be a random orthoprojector from $\mathbb{R}^m$ to$\mathbb{R}^d$ and*

$$d \geqslant \mathcal{O}(\frac{k * log(mV\tau^{-1}) log(\rho^{-1})}{\varepsilon^2}) \tag{18}$$

[1] Similarly, we can obtain $B \in \mathbb{R}^{m \times 7n}$ for affine transformations.

suppose $d < m$, then, with probability $1 - \rho$, the following statement holds: for every pair of points $x, y \in \mathcal{M}$, and $i \in \{1, 2\}$,

$$(1-\varepsilon)\sqrt{\frac{d}{m}} \leq \frac{\|Rx - Ry\|_i}{\|x - y\|_i} \leq (1+\varepsilon)\sqrt{\frac{d}{m}} \tag{19}$$

A fundamental connection between this Lemma and the sparse representation theory has been identified in compressive sensing [1, 2]. It states that, when the projections of two points in a random projection manifold are close, then the two original points are also close, only if these two points live in the same compact manifold. According to this, we can get:

$$\frac{\|\tilde{y'_{est}} - \tilde{A}_0 z_0\|_2}{\|y'_{est} - A_0 z_0\|_2} \approx \sqrt{\frac{d}{m}} \tag{20}$$

Then, $y'_{est} \approx A_0 z_0$ can be obtained from equation 17 and 20. Since y'_{est} is the estimation of the aligned version y' of the test image, we can optimize the translations between y and y'_{est} by a model based approach [3]:

$$\triangle\beta = argmin_\beta \|T(y, \beta) - y'_{est}\|_2 \tag{21}$$

With the estimated transformation parameters, the test image y is warped towards y'_{est}. Then, the warped image is projected again onto the manifolds defined by the model matrix B and the random projection matrix; this process repeats until the residue is gradually reduced to a certain level. The complete procedure is summarized in algorithm 1.

It is worth noting that, the above process can also be implemented in a coarse-to-fine framework, where the procedure is applied at each level of the pyramid.

3.3. Simultaneous Face Alignment and Recognition

The proposed approach can be useful in face detection followed by identification, where the target image obtained by the detection module is possibly not aligned with the model images although all the model images are already aligned. In this scenario, Algorithm 1 can be directly used for simultaneous face alignment and recognition. Moreover, alignment and recognition can interact in a loop to improve each other's performance. Better alignment leads to more accurate sparse solution, which in turn makes possible better recognition performance. On the other hand, more accurate sparse solution allows to perform better alignment.

Figure 1(d) shows several sparse representation examples by the proposed algorithm on aligned test images. For the first and second subjects, there are almost no differences between the sparse representation results by the proposed approach and the Randomfaces algorithm. For the third subject, there exists slight rotation between the test image and the model images. Our result using the translation model is not perfect, but it is still better than that by Randomfaces, which produces severe ghost effects. This further confirms our conclusion: simultaneous transformation and sparse representation recovery is very important. We also tested our approach using test images that are not aligned with model images (15-pixels shift in both horizontal and vertical directions). Figure 2(d) shows several examples. The results are very promising and we were able to obtain both the sparse representation and the translation motion; this demonstrates that the proposed approach can generate transformation invariant sparse representation.

Algorithm 1. *Transform-invariant Sparse Representation (TSR)*

Input: The training image matrix A_0 from k subjects, a test image $y \in \mathbb{R}^m$ and iteration number s.
Build the model matrix $B = [A_0, A_1, A_2] \in \mathbb{R}^{m \times 3n}$ (For affine model, $B = [A_0, A_1, A_2, A_3, A_4, A_5, A_6] \in \mathbb{R}^{m \times 7n}$)
Generate l random projections $R^1, ..., R^l \in \mathbb{R}^{d \times m}$.
for all $p = 1, ..., l$ **do**
  $\beta = \mathbf{0}$
  for all $q = 1, ..., s$ **do**
    Compute $y'_{est} = T(y, -\beta)$
    Compute $\tilde{y} = R^p y$ and $\tilde{A} = R^p B$, normalize $\tilde{y}$ and columns of $\tilde{A}$
    Perform l^1 minimization:

$$x_0 = argmin\|x\|_1 \quad while \quad \|\tilde{y} - \tilde{A}x\|_2 < \epsilon$$

    Compute $y'_{est} = A_0 z_0$, here $z_0 = x_0(1:n)$.
    Compute $\triangle\beta = argmin_\beta \|T(y, -\beta) - y'_{est}\|_2$.
    Compute $\beta = \beta + \triangle\beta$ until $\triangle\beta$ small enough.
  end for
  Compute $r_i^p = \|\tilde{y} - \tilde{A}\delta_i(x_0)\|$ for $i = 1, ..., k$
  Compute $\beta^p = \beta$
end for
Compute $identity(y) = argmin_i E[r_i]$
Compute $transform(y) = E[\beta]$
Output: $identity(y)$ and $transform(y)$.

3.4. Online Dynamic Texture Registration

Online video registration is required by many video analysis applications when a video sequence is captured by a moving camera. Traditional methods generally make the brightness constancy assumption [3]: $I(x_1, x_2, t) = I(x_1, x_2, t-1)$, where (x_1, x_2) denotes the spatial coordinates and t represents the time frame. However, this assumption is often violated in dynamic scenes.

Fitzgibbon [11] proposed to perform dynamic scene registration by minimizing the entropy function of an auto regressive process, which results in a difficult non-linear optimization problem. Dynamic Texture Constancy Constraint

(DTCC) is introduced in [20] to solve this problem, instead of the brightness constancy. In [14], another solution is proposed by jointly optimizing over registration parameters, the average image, and the dynamic texture model according to certain prior models. These three methods involve complex optimization and do not suit well the needs of online video registration. One online video registration method proposed in [16] attempts to solve two independent subproblems: 1) the extrapolation of the preceding frames using block based video synthesis techniques; 2) the alignment of a new image frame to best fit the above extrapolation [3].

In this paper, we propose a new online dynamic texture registration approach, based on the sparse representation constancy assumption instead of the traditional brightness constancy assumption. The sparse representation constancy assumption states that, given a new frame, its aligned version should be represented as a linear combination of as few preceding image frames as possible. As we know, a dynamic scene is called a dynamic texture when it is captured by a static camera and its temporal evolution exhibits certain stationarity [9]. Thus, our assumption is reasonable for dynamic-texture image sequences. Our experimental results in the next section also confirm the validity of this assumption. As a matter of fact, the traditional brightness constancy assumption seeks that the aligned version of the current image frame can be best represented by a single preceding frame, while the proposed sparse representation constancy assumption seeks that the aligned version of the current image frame can be best represented by all preceding image frames via l_1 minimization. Thus, the former can be thought as a special case of the latter.

Suppose a video sequence consists of frames $I_1, ..., I_n \in \mathbb{R}^m$. Without loss of generality, we can assume that the first k frames have already been aligned to the k^{th} frame. Let $A_0 = [I_1, ..., I_k] \in \mathbb{R}^{m \times k}$. Considering the translation model, our task is to estimate the translation motion between the $(k+1)^{th}$ frame and the preceding frames:

$$(x_0, \beta_{\mathbf{0}}) = argmin\|x\|_1, \quad T(y, \beta_{\mathbf{0}}) = A_0 x \qquad (22)$$

where β is the motion parameter. Obviously, this problem is equivalent to the problem in equation 9 and can be efficiently solved by Algorithm 1. After recovering the motion β between the $(k+1)^{th}$ frame and preceding frames, we can warp all preceding frames toward the $(k+1)^{th}$ frame according to the estimated motion parameter β. The same procedure can be applied to aligning with the $(k+2)^{th}$ frame, and so on.

For long video sequences, it is impractical to build a model matrix $A_0 = [I_1, ..., I_{t-1}] \in \mathbb{R}^{m \times (t-1)}$, where t denotes the current frame number. In order to cope with this case, we can set a time window width parameter τ. We then build the model matrix, $A_0 = [I_{t-\tau}, ..., I_{t-1}] \in \mathbb{R}^{m \times (t-\tau)}$, for the t^{th} frame, which can avoid the memory requirement blast for a long video sequence. The complete algorithm for online dynamic texture registration is summarized below.

Algorithm 2. *TSR Based Online Dynamic Texture Registration*

1: **Input:** The video sequence $I_1, ..., I_n$, the number k which means $1^{st} \sim k^{th}$ have been aligned to k^{th} frame, the time window width $\tau \leq k$
2: **for all** $t = k+1, ..., n$ **do**
3: Set $A_0 = [I_{t-\tau}, ..., I_{t-1}]$
4: Set $y = I_t$ and iteration number s.
5: Perform Algorithm 1, $\beta_t = TSR(B, y, s)$
6: Warp $I_1, ..., I_{t-1}$ toward I_t according to β_t
7: **end for**
8: **Output:** The registered $I_1, ..., I_n$ and $\beta_{k+1}, ..., \beta_n$.

4. Experiments

The proposed transformation invariant sparse representation is applied to face recognition and online dynamic texture registration respectively.

4.1. Face Recognition

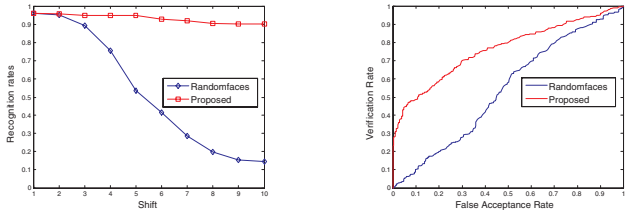
In this section, we validate respectively the identification and verification performance of Algorithm 1 for face recognition using a public face database, namely, the Extended Yale B database [12]. This database consists of 2,414 frontal-face images of 38 individuals. The image size is 192×168. We randomly selected 20 subjects, half of whose images are used for training (32 images per subject), and the other half for testing. There are a total of 640 images from the 20 subjects for training. In the identification experiment, there are 640 images for testing. In the verification experiment, there are 1198 test images, half of which are true outliers.

All images are aligned and normalized in the Extended Yale B database. To evaluate the identification performance of the proposed approach, we generated shifted test images according to different shift values. For example, if the shift value is 7, each test image is shifted with random parameters between 0 and 7 pixels in both horizontal and vertical directions. The training images are kept unchanged. For fair comparison, the implementations of the Randomfaces algorithm and the proposed algorithm use the same parameters as those introduced in [21] (random projection matrix number is $l = 5$, error distortion $\varepsilon = 0.05$, and the reduced dimension d is 504). Figure 3(a) shows the recognition performance of the proposed algorithm and the Randomfaces algorithm, as a function of shift values. One can see that the proposed algorithm outperforms the Randomfaces algorithm. When the shift value is smaller than 2 pixels, our results are slightly better than Randomfaces. When

Affine	Group 1	Group 2	Group 3
Eigenfaces [19]	75.16%	49.38%	38.75%
Randomfaces [21]	75.94%	49.69%	38.28%
Proposed	91.56%	89.22%	81.25%

Table 1. Identification rates under affine transformations

the shift value exceeds 2 pixels, the Randomfaces recognition performance is much degraded, which further demonstrates that the previous sparse image representation is sensitive to image plane transformations. In comparison, the proposed transform-invariant sparse representation achieves better and more stable recognition performance.



(a) Identification (b) Verification

Figure 3. Identification results and ROC curves

The verification performance of the proposed algorithm is evaluated using 1198 shifted test images with the shift value 7. Among the 1198 images, 558 images are true outliers. Since the role of verification is to reject test images that are difficult to classify, we also use the Sparse Concentration Index (SCI) [21] as an indication of confidence:

$$SCI(x_0) = \frac{k * max_i \|\delta_i(x_0)\|_1 / \|x_0\|_1 - 1}{k - 1} < \sigma \quad (23)$$

where k is the subject number, $\sigma \in [0, 1]$ is a preselected threshold, and $\delta_i(x_0)$ represents the entries in x_0 associated with subject i keep unchanged and others are set as zeros. We plot the Receiver Operating Characteristic (ROC) curve according to different σ values in Figure 3(b). As expected, the proposed algorithm outperforms Randomfaces.

We also tested the proposed algorithm's invariance and robustness to the affine transformation model. We scaled up the 640 test images by a factor of 1.02 and then rotated them by $2°$, $4°$ and $6°$, respectively. This gave us 3 groups of test images (640 in each). Table 1 tabulates the comparison of the identification results, which shows our algorithm has better invariance to affine transformations.

4.2. Dynamic Scene Registration

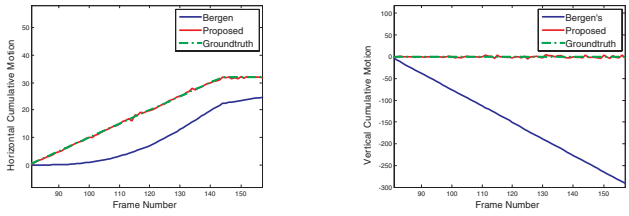
The first set of experiments uses the Escalator sequence (shown in Figure 4). It includes 157 image frames. We resized each image frame to 120×160 pixels. In order to evaluate the proposed approach, we generated 3 new video sequences by transforming each image frame with a known

Figure 4. The Escalator Sequence [22]

Generated Sequence	1st	2nd	3rd
Bergen's [3]	22.55%	21.86%	23.25%
Proposed	1.39%	1.48%	1.36%

Table 2. FEF of horizontal cumulative motion

motion and record the motion as ground truth. For comparison, we implemented the classic model-based motion estimation method [3], which we call Bergen's method. Since the proposed method is an online registration method and assumes that the beginning frames have been aligned, we only compared the motion estimations from the 81st to the last frame in this experiment. The comparison results on one generated sequence are shown in Figure 5. The proposed method almost performs perfect motion estimation, while Bergen's result is not as good. It is easy to interpret these trends in these results. Bergen's method is based on the assumption of Brightness Constancy, thus it considers that the local/nonrigid motion in dynamic textures is also caused by camera motion. On the other hand, our approach is based on the more accurate sparse representation constancy assumption and seeks the optimal estimation in terms of all preceding image frames via l_1 minimization.



(a) Horizontal direction (b) Vertical direction

Figure 5. The cumulative motion estimation

For quantitative evaluation, the false estimation fraction (FEF) is used to indicate the difference between the ground-truth M_{True} and the estimated motion M_{Est}: $FEF = |M_{Est} - M_{True}|/|M_{True}|$. Table 2 records the FEF of registration results on the three generated image sequences by different implementations. Our algorithm gives very good motion estimation results.

The moving flower bed sequence shown in Figure 6 has been used as a registration example by [11, 20, 16, 14]. The whole sequence includes 554 image frames. The camera motion in this sequence is a horizontal translation. The ground truth of the cumulative horizontal motion in the whole sequence is 110 pixels based on manual motion labelling of one red flower. While quantitative motion estimation results were not reported in [11], the FEF of the

cumulative motion is 29.41% by Vidal's approach on a subsequence with 250 frames [20]. The FEF of the cumulative motion on this sequence is reported as 1.7% in [16] and 4.98% in [14]. For quantitative comparison with these methods, we also tested the proposed algorithm on this sequence. The cumulative motion along the horizontal direction is estimated as 107.7 pixels by our approach, thus a 2.09% FEF of cumulative motion; this performance is close to the reported result by the extrapolation based registration approach [16]. Based on the efficient l^1 minimization, our algorithm takes less than 5 seconds to register each image frame on a $1.5GHz$ laptop PC in MATLAB environment, which is faster than extrapolation based registration using block-based video synthesis.

Figure 6. A sequence of moving flower bed [11, 16, 20, 14].

4.3. Discussions

All of the above experimental results have validated the proposed transformation invariant sparse representation algorithm.

1. The sparse image representation, successfully extended to be invariant to a desired group of image-plane transformations, is easily applied to image analysis related problems.
2. The sparse representation constancy assumption, proposed in place of brightness constancy assumption for motion estimation, has been validated and improves performance.
3. Compared to previous algorithms, the proposed algorithm can handle less constrained cases and promises better performance at the cost of more memory usage (3 and 7 times more for translation and affine transformations, respectively).

5. Conclusions

In this paper, we extend the sparse representation to be invariant to a desired group of image-plane transformations of an ensemble of unaligned images. By coupling the recently emerged theories on compressive sensing and random projection manifold, the proposed approach can efficiently recover not only sparse representation of a target image but also the image plane transformation between the target and the model images. Experiments on face databases and real video sequences demonstrate the performance of our method and show marked improvement over previous approaches.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 2007. To appear. 5

[2] R. Baraniuk and M. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 2007. 2, 4, 5

[3] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings of ECCV*, pages 237–252, 1992. 4, 5, 6, 7

[4] E. Candes. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, 2006. 1, 3

[5] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2006. 1, 2

[6] E. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006. 1, 2, 3

[7] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006. 1

[8] D. Donoho and Y. Tsaig. Fast solution of l1-norm minimization problems when the solution may be sparse, 2006. Preprint. 1, 2

[9] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003. 6

[10] M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal on Selected Topics in Signal Processing*, 2007. Accepted. 2

[11] A. Fitzgibbon. Stochastic rigidity: Image registration for nowhere-static scenes. In *Proceedings of ICCV*, 2001. 5, 7, 8

[12] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001. 3, 6

[13] C. Hegde, M. Wakin, and R. Baraniuk. Random projections for manifold learning. In *Proceedings of NIPS*. Appear. 2, 4

[14] J. Huang, X. Huang, and D. Metaxas. Optimization and learning for registration of moving dynamic textures. In *Proceedings of ICCV*, 2007. 6, 7, 8

[15] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for large-scale l1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, 2007. Accepted. 2

[16] A. Rav-Acha, Y. Pritch, and S. Peleg. Online registration of dynamic scenes using video extrapolation. In *Workshop on Dynamical Vision at ICCV*, pages 151–164, 2005. 6, 7, 8

[17] Y. Sharon, J. Wright, and Y. Ma. Computation and relaxation of conditions for equivalence between l1 and l0 minimization, 2007. Submitted to IEEE Transactions on Information Theory. 1

[18] A. Shashua, A. Levin, and S. Avidan. Manifold pursuit: a new approach to appearance based recognition. In *Proceedings of ICPR*, 2002. 4

[19] M. Turk and A. Pentland. Eigenfaces for recognition. In *Proceedings of CVPR*, 1991. 7

[20] R. Vidal and A. Ravichandran. Optical flow estimation and segmentation of multiple moving dynamic textures. In *Proceedings of CVPR*, pages 516–521, 2005. 6, 7, 8

[21] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007. Accepted. 1, 2, 3, 4, 6, 7

[22] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In *Proceedings of ICCV*, 2003. 7