Hierarchical Sparse Representation for Robust Image Registration

Yeqing Li, *Member, IEEE*, Chen Chen, *Student Member, IEEE*,
Fei Yang, *Member, IEEE*, and Junzhou Huang, *Member, IEEE*

Abstract—Similarity measure is an essential component in image registration. In this article, we propose a novel similarity measure for registration of two or more images. The proposed method is motivated by the fact that optimally registered images can be sparsified hierarchically in the gradient domain and frequency domain with the separation of sparse errors. One of the key advantages of the proposed similarity measure is its robustness in dealing with severe intensity distortions, which widely exist on medical images, remotely sensed images and natural photos due to differences of acquisition modalities or illumination conditions. Two efficient algorithms are proposed to solve the batch image registration and pair registration problems in a unified framework. We have validated our method on extensive and challenging data sets. The experimental results demonstrate the robustness, accuracy and efficiency of our method over nine traditional and state-of-the-art algorithms on synthetic images and a wide range of real-world applications.

Index Terms—Image registration, hierarchical sparse representation, sparse learning

1 Introduction

Image registration is a fundamental task in image processing and computer vision [1], [2], [3]. It aims to align two or more images into the same coordinate system, and then these images can be processed or compared. Accuracy and robustness are two of the most important metrics to evaluate a registration method. It has been shown that a mean geometric distortion of only 0.3 pixel will result in a noticeable effect on the pixel-to-pixel image fusion process [4]. Robustness is defined as the ability to get close to the accurate results on different trials under diverse conditions. Based on the feature used in registration, existing methods can be classified into feature-based registration (e.g., [5], [6], [7]) and pixel-based registration ([8], [9], [10], [11]). Feature-based methods rely on the landmarks extracted from the images. However, extracting reliable features is still an open problem and an active topic of research [3]. In this article, we approach image registration by directly using their pixel values. In addition, we successfully registered the images from a variety of applications in subpixel-level accuracy, as precisely as possible.

One key component for image registration is the energy function to measure (dis)similarity. The optimized similarity should lead to the correct spatial alignment. However, finding a reliable similarity measure is quite challenging due to the unpredicted variations of the input images. In many real-world applications, the images to be registered may be acquired at different times and locations, under various illumination conditions and occlusions, or by different acquisition modalities. As a result, the intensity fields of the images may vary significantly. For instance, slow-varying intensity bias fields often exist in brain magnetic resonance images [12]; the remotely sensed images may even have inverse contrast for the same land objects, as multiple sensors have different sensitivities to wavelength spectrum [13]. Unfortunately, many existing pixel-based similarity measures are not robust to these intensity variations, e.g., the widely used sum-of-squared-difference (SSD) [2].

Recently, the sparsity-inducing similarity measures have been repeatedly successful in overcoming such registration difficulties [14], [15], [16], [17]. In RASL (Robust Alignment by Sparse and Low-rank decomposition) [15], the images are vectorized to form a data matrix. The transformations are estimated to seek a low rank and sparse representation of the aligned images. Two online alignment methods, ORIA [16] (online robust image alignment) and t-GRASTA [17] (transformed Grassmannian robust adaptive subspace tracking algorithm) are proposed to improve the scalability of RASL. The sparse and low-rank approaches (RASL and the following works) have demonstrated some promising results. Similar ideas have been applied to many applications such as face synthesis [18], face recovery in video restoration [19], background and foreground separation [20], and object detection [21]. All of these methods (RASL, t-GTASTA, etc.) assume that the large errors among the images are sparse (caused by shadows and partial occlusions) and separable. However, as we will show later, many real-world images contain severe spatially-varying intensity distortions. These intensity variations are not sparse and, therefore, are difficult to be separated

- *Y. Li, C. Chen, and J. Huang are with the University of Texas at Arlington and Tencent AI Lab, respectively, Arlington, TX 76019. E-mail: yeqing.li@mavs.uta.edu, chenchen.cn87@gmail.com, jzhuang@uta.edu.*
- *F. Yang is with Facebook Inc., Menlo Park, CA 94025. E-mail: feiyang@cs.rutgers.edu.*

Manuscript received 8 May 2016; revised 4 Aug. 2017; accepted 16 Aug. 2017. Date of publication 31 Aug. 2017; date of current version 13 Aug. 2018. (Corresponding author: Junzhou Huang.)
Recommended for acceptance by P. V. Gehler.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2748125

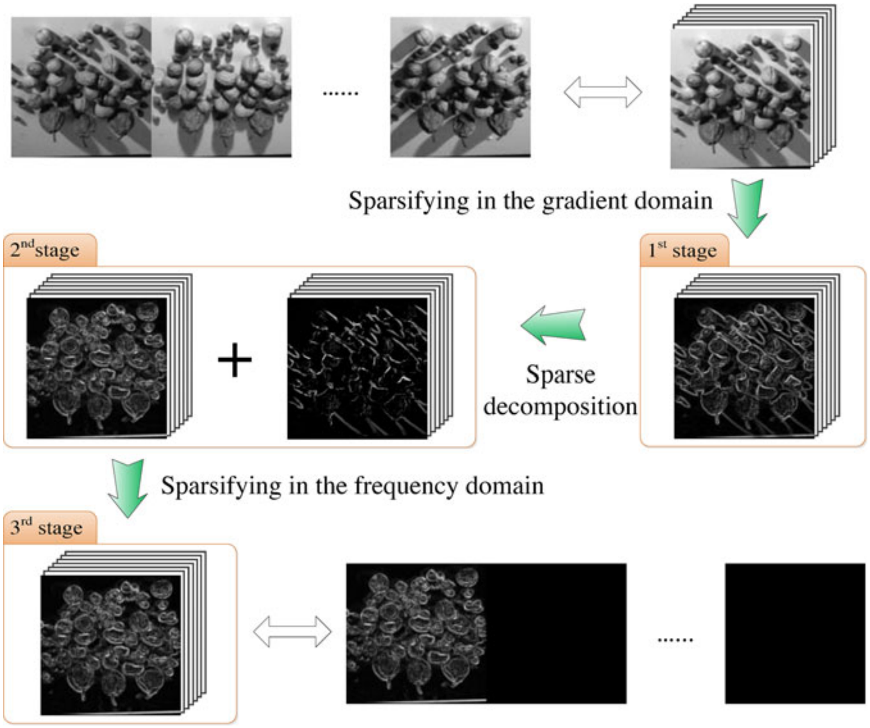

......
Sparsifying in the gradient domain
1st stage
2nd stage
Sparse decomposition
+
Sparsifying in the frequency domain
3rd stage
......

Fig. 1. Hierarchical sparse representation of the optimally registered images. First we sparsify the image tensor into the gradient tensor (1st stage). The sparse error tensor is then separated out in the 2nd stage. The gradient tensor with repetitive patterns are sparsified in the frequency domain. Finally we obtain an extremely sparse frequency tensor (composed of Fourier coefficients) in the 3rd stage.

by these methods. As a result, the above measures may fail to find the correct alignment and thus are less robust in these challenging tasks.

The residual complexity (RC) [14] is one of the best measures for registering two images corrupted by severe intensity distortion [22], which uses the discrete cosine transform (DCT) to sparsify the residual of two images. RC is primary used in medical image registration [3] such as MRIs. In this scenario, the images are not only under simple global deformations like affine transformations but may also have local deformations. This makes the problem challenging. Furthermore, those images may also be corrupted by slow-varying intensity bias fields [12], heterogeneous illumination and reflectance artifacts [23]. Such distortions will violate the assumptions of many existing approaches and will jeopardize the registration performance. It has been demonstrated that RC is able to handle those effects better than previous counterparts in pair-image registration. However, for a batch of images, RC has to register them pair-by-pair, and the solution may be sub-optimal. In addition, DCT and inverse DCT are required in each iteration, which slows down the overall speed of registration. Finally, although RC is robust to intensity distortions, the ability of RC to handle partial occlusions is unknown.

In this article, we proposed a novel similarity measure for robust and accurate image registration based on the hierarchical sparse representation (HSR) of the natural images. Unlike previous works that vectorize each image [15], [16], [17], we arranged the input images into a 3D tensor to keep their spatial structure. With this arrangement, the optimally registered image tensor can be sparsified into a sparse frequency tensor and a sparse error tensor (Fig. 1). Severe intensity distortions and partial occlusions will be sparsified and separated out in the first and second stages, while any misalignment will increase the sparseness of the frequency tensor (third stage). We propose a novel similarity measure based on such hierarchical sparse representation of the natural images. Compared with the low rank similarity measure which requires a batch of input images, the proposed similarity measure still works even when there are only two input images. An efficient algorithm based on the Augmented Lagrange Multiplier (ALM) method is proposed for the batch mode, while the gradient descent method with backtracking is presented to solve the pair registration problem. Both algorithms have very low computational complexity in each iteration. We compare the proposed method with nine traditional and state-of-the-art algorithms on a wide range of natural image data sets, including medical images, remotely sensed images and photos. Extensive results demonstrate that our method is more robust to different types of intensity variations and always achieves higher sub-pixel accuracy over all the tested methods.

Some of the preliminary results were initially published in a CVPR 2015 paper [24]. This journal article has considerable changes compared with the conference paper, which mainly includes:

(1) Theoretical convergence analysis. New discussion of the convergence property of the proposed method has been added, which uses the theory of strong uniqueness.
(2) Extension to non-rigid registration. Non-rigid transformation is another challenging class of registration problem. Our journal version has extended the conference one and proposed a new model to handle this problem.

(3) Larger data set and broader application. Two case studies are added to the preliminary data set, including face images and non-rigid transformed brain images. Experiments on the enlarged data set are more convincing. Besides, these experimental results better demonstrate the flexibility and effectiveness of our approach.

(4) Clearer explanation. The proposed HSR and related techniques are further explained with more details. Especially, elaborate illustrations and mathematical derivation are added to better demonstrate our approach.

Organization. The remainder of this article is organized as follows: In Section 2, we reviewed the basic idea of sparse and low-rank model for image registration. In Section 3, we introduce the hierarchical sparse representation approaches for image registration, where we first discuss models for both the batch and pair image registration and then proposed the algorithms for solving the hierarchical sparse representation problems. We provide experimental results in Section 5 to demonstrate the efficiency and effectiveness of our methods. Then in Section 6 we extend our approach on non-rigid registration problems and give experimental results. Section 8 provides concluding remarks and proposed potential extensions to our approaches.

2 Related Work

2.1 Image Registration

The measurement of similarities is the key component for image registration problems. Suppose we have a batch of gray-scale images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N \in \mathbb{R}^{w\times h}$ to be registered, where N denotes the total number of images. First, we consider the simplest case where all input images are perturbed from a set of transformations $\tau = \{\tau_1, \tau_2, \ldots, \tau_N\}$.

One of the most popular ways to estimate the transformations is to minimize the rank of the data matrix [15]: Measurement is based on matrix rank and the registration problem can be formulated as

$$\min_{\mathbf{A},\tau} \ \mathrm{rank}(\mathbf{A}), \ s.t. \ \mathbf{D}\circ\tau = \mathbf{A}, \tag{1}$$

where $\mathbf{D}\circ\tau \doteq [\mathrm{vec}(\mathbf{I}_1) \ \mathrm{vec}(\mathbf{I}_2) \ \ldots \ \mathrm{vec}(\mathbf{I}_N)] \in \mathbb{R}^{M\times N}$; vec: $\mathbb{R}^{w\times h} \longmapsto \mathbb{R}^M$ denotes the vectorization operation; $\mathrm{vec}(\mathbf{I}_t^0)\circ\tau_t$ denotes image $\mathbf{I}_t$ warped by τ_t for $t = 1, 2, \ldots, N$. It is assumed that the rank of $\mathbf{A}$ is low (ideally, it is one), since that each column of $\mathbf{A}$ is the same after the transformations are correctly estimated. This extreme case happens when all the images are identical after they are aligned.

However, the low-rank assumption usually fails in practice because of apparent differences in images even when they are aligned. This may be due to some partial occlusions or pixel corruptions. In order to compensate this situation, a sparse error term $\mathbf{E}$ is usually introduced. By replacing the non-convex norm with the convex surrogate and using the Lagrangian method, the object function can be written as

$$\min_{\mathbf{A},\mathbf{E},\tau} \ \|\sigma(\mathbf{A})\|_1 + \lambda\|\mathbf{E}\|_1, \ s.t. \ \|\mathbf{D}\circ\tau - \mathbf{A} + \mathbf{E}\|_F \le \epsilon, \tag{2}$$

where the hyper-parameter $\epsilon > 0$ controls the tolerance of the noise level; thus $\lambda > 0$ is the hyper-parameter used to balance the rank constraint and sparse constraint, while $\sigma(\mathbf{A})$ is an operator that computes all the eigenvalues of $\mathbf{A}$. Here, $\mathrm{svd}(\mathbf{A}) = \mathbf{U}\Sigma\mathbf{V}^T$ is the singular value decomposition of $\mathbf{A}$, where $\mathbf{U}$, $\mathbf{V}$ are orthogonal matrices and Σ is a diagonal matrix that contains all the eigenvalues of $\mathbf{A}$ on its diagonal. Therefore, we have $\sigma(\mathbf{A}) = \mathrm{diag}(\Sigma) = \mathrm{diag}(\mathbf{U}^T\mathbf{A}\mathbf{V})$. The $\|\sigma(\mathbf{A})\|_1$ is also known as nuclear norm $\|\mathbf{A}\|_*$.

Equation (2) can be effectively solved by an off-the-shelf optimization technique such as the augmented Lagrange multiplier Methods [15], [25].

Here, we did not choose the low rank model since it treats each image as a 1D signal without considering the spatial prior information of natural images. Instead, the spatial information will be utilized in our model.

2.2 Total Variation

Total Variation (TV) has been widely used in the image process and in computer vision literature, especially as a sparse regularization. This model was proposed by Rudin-Osher and Fatemi (ROF) in [26], whose main goal was placing proper constraint on edges and removing noise in a given image. Let $\mathbf{X} \in \mathbb{R}^{w\times h}$ be a 2D matrix. The basic form of TV (semi)-norm can be written as $\|\mathbf{X}\|_{TV} = \|\Delta(\mathbf{X})\|_\ell$, where $\Delta(\cdot)$ is the differential operation and $\ell \in \{1, 2\}$. In image processing problems, a discrete version of TV is usually used. There are two commonly used discrete TVs. One is an ℓ_2-norm based isotropic TV

$$\begin{aligned}\|\mathbf{X}\|_{TV} = &\sum_{i=1}^{w-1}\sum_{j=1}^{h-1}\sqrt{(\mathbf{X}_{i,j}-\mathbf{X}_{i+1,j})^2 + (\mathbf{X}_{i,j}-\mathbf{X}_{i,j+1})^2} \\ &+ \sum_{i=1}^{w-1}|\mathbf{X}_{i,h}-\mathbf{X}_{i+1,h}| + \sum_{j=1}^{h-1}|\mathbf{X}_{w,j}-\mathbf{X}_{w,j+1}|\end{aligned} \tag{3}$$

The other one is an $\ell 1$-norm based anisotropic TV

$$\begin{aligned}||\mathbf{X}||_{TV} = &\sum_{i=1}^{w-1}\sum_{j=1}^{h-1}\{|\mathbf{X}_{i,j}-\mathbf{X}_{i+1,j}| + |\mathbf{X}_{i,j}-\mathbf{X}_{i,j+1}|\} \\ &+ \sum_{i=1}^{w-1}|\mathbf{X}_{i,h}-\mathbf{X}_{i+1,h}| + \sum_{j=1}^{h-1}|\mathbf{X}_{w,j}-\mathbf{X}_{w,j+1}|\end{aligned} \tag{4}$$

In both cases, we assume the borders satisfy the following reflexive condition [27]

$$\forall j \in 1, \ldots, h, \mathbf{X}_{w+1,j} - \mathbf{X}_{w,j} = 0, \tag{5}$$

$$\forall i \in 1, \ldots, w, \mathbf{X}_{i,h+1} - \mathbf{X}_{i,h} = 0. \tag{6}$$

In this article, we mainly use the anisotropic TV. Although, TV has proved to be useful in many tasks, e.g., denoising, deblurring and reconstruction [27], [28], [29]. However, solving a TV-regularized problem is non-trivial. This is due to the fact that TV operation is nontrivial. Another difficulty is the scale of the these problems. This problem has attracted a lot of attention in the literature. Many algorithms have been applied to efficiently solve TV-based problems, such as FISTA-TV [27], alternating direction method of multipliers (ADMM) [30], primal-dual approach [31], [32]. The benefits of TV have also motivated several other variances of TV such as non-local TV [28], directional total variation [33], and higher-order total variation [34], [35].

Natural images are often piece-wise smooth and they have sparse gradients. Minimizing the TV norm is equivalent to sparsifying the images in the gradient domain. In this article, we will apply key idea of TV in our hierarchical sparse representation pipeline.

3 Image Registration via Hierarchical Sparse Representation

In this section, we use bold letters to denote multi-dimensional data. For example, $\mathbf{x}$ denotes a vector, $\mathbf{X}$ denotes a matrix and $\mathcal{X}$ is a 3D or third-order tensor. $\mathcal{X}_{(i,j,t)}$ denotes the entry in the ith row, jth column and tth slice. $\mathcal{X}_{(:,:,t)}$ denotes the whole tth slice, which is, therefore, a matrix. The ℓ_1 norm is the summation of absolute values of all entries, which applies to vector, matrix and tensor.

3.1 Batch Mode

We introduce our hierarchical sparsity architecture in the inverse order for easy understanding. Suppose the input images are arranged into a 3D tensor $\mathcal{D} \in \mathbb{R}^{w\times h\times N}$

$$\mathcal{D}_{(:,:,t)} = \mathbf{I}_t, \quad t = 1, 2, \ldots, N. \tag{7}$$

Again, let us consider the simplest case wherein all the input images are identical and perturbed from a set of transformations $\tau = \{\tau_1, \tau_2, \ldots, \tau_N\}$ (which can be affine, non-rigid, etc.). After removing the transformation perturbations, the slices show repetitive patterns. We have seen how the matrix of the vectorized similar/same images will demonstrate low rank properties. Besides, such periodic signals are also extremely sparse in the frequency domain. Here, one can view the discrete Fourier transform as the counterpart to the spectral decomposition in Equation (2) with a known basis, i.e., the Fourier matrix. Ideally the Fourier coefficients from the second slice to the last slice should all be zeros. We can minimize the ℓ_1 norm of the Fourier coefficients to seek the optimal transformations

$$\min_{\mathcal{A},\tau} \ \|\mathcal{F}_N\mathcal{A}\|_1, \ s.t. \ \mathcal{D}\circ\tau = \mathcal{A}, \tag{8}$$

where $\mathcal{F}_N$ denotes the Fourier transform in the third dimension.

To understand the motivation of the Fourier transform (FT) here, one can think of an extreme example of performing the Fourier transform on a constant vector (i.e., all the elements in this vector are the same). In this case, only one element of the Fourier coefficients will be non-zero, which is a very sparse signal. This is due to the nature of FT, which represent a signal in the frequency domain. So the more the signal appears to repeat itself, the sparser the coefficients will be.

The above model may perform poorly on practical cases due to the corruptions and partial occlusions in the images. Similar to the findings in [15], we assume the noise is negligible in magnitude as compared to the error caused by occlusions. Let $\mathcal{E}$ be the error tensor. We can separate it from the image tensor if it is sparse enough and use the ℓ_1 norm to induce sparseness

$$\min_{\mathcal{A},\mathcal{E},\tau} \ \|\mathcal{F}_N\mathcal{A}\|_1, \ s.t. \ \mathcal{D}\circ\tau = \mathcal{A}+\mathcal{E}, \ \|\mathcal{E}\|_0 \le k, \tag{9}$$

where $\|\mathcal{E}\|_0$ counts the number non-zero entries in $\mathcal{E}$ and k is a constant to constrain the sparseness. The nonconvexity and nonsmoothness renders the ℓ_0-norm impractical for real-world applications. Therefore, we use ℓ_1 norm to encourage sparsity instead of the ℓ_0 norm

$$\min_{\mathcal{A},\mathcal{E},\tau} \ \|\mathcal{F}_N\mathcal{A}\|_1 + \lambda\|\mathcal{E}\|_1, \ s.t. \ \mathcal{D}\circ\tau = \mathcal{A}+\mathcal{E}, \tag{10}$$

where $\lambda > 0$ is a regularization parameter.

The above approach requires that the error $\mathcal{E}$ is sparse. However, in many real-world applications, the images are corrupted with spatially-varying intensity distortions. Existing methods such as RASL [15] and t-GRASTA [17] may fail to separate these non-sparse errors. The last stage of our method comes from the intuition that the locations of the image gradients (edges) should be roughly keep the same, even under severe intensity distortions. Therefore, we register the images in the gradient domain

$$\min_{\mathcal{A},\mathcal{E},\tau} \ \|\mathcal{F}_N\mathcal{A}\|_1 + \lambda\|\mathcal{E}\|_1, \ s.t. \ \nabla\mathcal{D}\circ\tau = \mathcal{A}+\mathcal{E}, \tag{11}$$

where $\nabla\mathcal{D} = \sqrt{(\nabla_x\mathcal{D})^2 + (\nabla_y\mathcal{D})^2}$ denotes the gradient tensor along the two spatial directions. This is based on a mild assumption that the intensity distortion fields of natural images often change smoothly.

With this rationale, the input images can be sparsely represented in a three-stage architecture, which is shown in Fig. 1. We call it hierarchical sparse representation of images. Compared with existing popular low rank representations [15], our modeling has two major advantages. First, the low-rank representation treats each image as a 1D signal, while our modeling exploits the spatial prior information (piece-wise smoothness) of natural images. Second, when the number of input images is not sufficient to form a low rank matrix, our method is still effective. Next, we will demonstrate how our method can register only two input images.

3.2 Pair Mode

In this section, we turn to the registration problem of two images, which is a special case of multiple image registration. In the pair registration case, we usually have one image as the reference image and the goal is to register the other image (usually called the source image) to the reference image. For registering a pair of images, our model can be simplified and accelerated. After two-point discrete Fourier transform (DFT) on the registered images, the first entry is the sum and the second entry is the difference. The difference term is much sparser than the sum term after the two images have been registered. To better understand the situation, suppose the gradient of two identical images, if we sum them together, we will get the same number of non-zero elements as the gradient of one image; on the other hand, if we subtract one from the other, all the values should be canceled out and we get a zero image, which is much sparser than the summed result.

This property enables us to discard the sum term to seek a sparser representation and only optimize on the difference term. Let $\mathbf{I}_1$ be the reference image, and $\mathbf{I}_2$ be the source image to be registered. The problem (11) can be simplified to

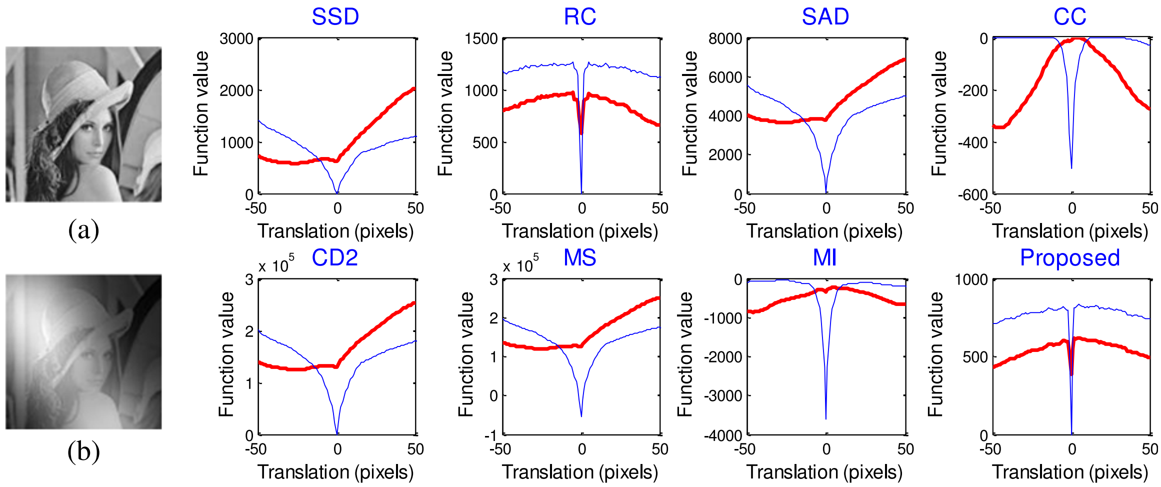


Fig. 2. A toy registration example with respect to horizontal translation using different similarity measures (SSD [2], RC [14], SAD [2], CC [36], CD2 [9], MS [37], MI [10] and the proposed pair mode). (a) The Lena image (128×128). (b) A toy Lena image under a severe intensity distortion. Blue curves show registration between (a) and (a); red curves show registration between (b) and (a).

$$\begin{aligned}\min_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{E},\tau} &\ \|\mathbf{A}_1-\mathbf{A}_2\|_1+\lambda\|\mathbf{E}\|_1,\\ s.t.\ &\nabla\mathbf{I}_1=\mathbf{A}_1, \nabla\mathbf{I}_2\circ\tau=\mathbf{A}_2+\mathbf{E}.\end{aligned} \tag{12}$$

Both ℓ_1 norms in (12) imply the same property, i.e., sparseness of the residual image $\mathbf{E}$. Therefore, we can further simplify the above energy function

$$\min_{\tau}\ \|\nabla\mathbf{I}_1-\nabla\mathbf{I}_2\circ\tau\|_1. \tag{13}$$

It's interesting that (13) is equivalent to minimizing the total variation of the residual image. TV has been successfully utilized in many image reconstructions [38], [39] and non-rigid registration [40] problems.

We compare the proposed similarity measure with SSD [2], RC [14], sum-of-absolute value (SAD) [2], correlation coefficient (CC) [36], CD2 [9], MS [37] and mutual information (MI) [10] on a toy example. The Lena image is registered with respect to the horizontal translations. The blue curves in Fig. 2 show the responses of different measures, all of which can find the optimal alignment at the zero translation. After adding intensity distortions and rescaling, the appearance of the source image shown in Fig. 2b is not consistent with that of the original Lena image. The results denoted by the red curves show that only RC and the proposed pair mode can handle this intensity distortion, while other methods fail.

4 Algorithms

4.1 Batch Mode

Problem (11) is difficult to solve directly due to the non-linearity of the transformations τ. We used the local first order Taylor approximation for each image

$$\nabla\mathbf{I}_t\circ(\tau_t+\triangle\tau_t)\approx\nabla\mathbf{I}_t\circ\tau_t+\mathcal{J}_t\otimes\triangle\tau_t \tag{14}$$

for $t=1,2,\dots,N$, where $\mathcal{J}_t=\frac{\partial}{\partial\zeta}(\nabla\mathbf{I}_t\circ\zeta)|_{\zeta=\tau_t}\in\mathbb{R}^{w\times h\times p}$ when τ_t is defined by p parameters. The *Tensor-Vector Product* of the last term is defined by:

Definition 1 (Tensor-Vector Product). *The product of a tensor $\mathcal{A}\in\mathbb{R}^{n_1\times n_2\times n_3}$ and a vector $\mathbf{b}\in\mathbb{R}^{n_3}$ is a matrix $\mathbf{C}\in\mathbb{R}^{n_1\times n_2}$. It is given by $\mathbf{C}=\mathcal{A}\otimes\mathbf{b}$, where $\mathbf{C}_{(i,j)}=\sum_{t=1}^{n_3}\mathcal{A}_{(i,j,t)}\mathbf{b}_{(t)}$, for $i=1,2,\dots,n_1$ and $j=1,2,\dots,n_2$.*

Based on this, the batch mode (11) can be rewritten as

$$\begin{aligned}\min_{\mathcal{A},\mathcal{E},\triangle\tau} &\ \|\mathcal{F}_N\mathcal{A}\|_1+\lambda\|\mathcal{E}\|_1,\\ s.t.\ &\nabla\mathcal{D}\circ\tau+\mathcal{J}\otimes\triangle\tau=\mathcal{A}+\mathcal{E},\end{aligned} \tag{15}$$

This constrained problem can be solved by the augmented Lagrange multiplier algorithm [15], [25]. The augmented Lagrangian problem is to iteratively update $\mathcal{A},\mathcal{E},\triangle\tau$ and $\mathcal{Y}$ by

$$\begin{aligned}(\mathcal{A}^{k+1},\mathcal{E}^{k+1},\triangle\tau^{k+1}) &=\arg\min_{\mathcal{A},\mathcal{E},\triangle\tau}\mathcal{L}(\mathcal{A},\mathcal{E},\triangle\tau,\mathcal{Y}),\\ \mathcal{Y}^{k+1} &=\mathcal{Y}^k+\mu^k h(\mathcal{A}^k,\mathcal{E}^k,\triangle\tau^k),\end{aligned} \tag{16}$$

where k is the iteration counter and

$$\begin{aligned}\mathcal{L}(\mathcal{A},\mathcal{E},\triangle\tau,\mathcal{Y}) = &< \mathcal{Y},h(\mathcal{A},\mathcal{E},\triangle\tau) > +\|\mathcal{F}_N\mathcal{A}\|_1\\ &+\lambda\|\mathcal{E}\|_1+\frac{\mu}{2}\|h(\mathcal{A},\mathcal{E},\triangle\tau)\|_F^2,\end{aligned} \tag{17}$$

where the inner product of two tensors is the sum of all the element-wise products and

$$h(\mathcal{A},\mathcal{E},\triangle\tau)=\nabla\mathcal{D}\circ\tau+\mathcal{J}\otimes\triangle\tau-\mathcal{A}-\mathcal{E}. \tag{18}$$

A common strategy to solve (16) is to minimize the function against one unknown at one time. Each subproblem has a closed form solution

$$\begin{aligned}\mathcal{A}^{k+1} &=\mathcal{F}_N^{-1}\Big[\mathcal{T}_{1/\mu^k}(\mathcal{F}_N(\nabla\mathcal{D}\circ\tau+\mathcal{J}\otimes\triangle\tau+\frac{1}{\mu^k}\mathcal{Y}^k\\ &\quad-\mathcal{E}^k))\Big]\\ \mathcal{E}^{k+1} &=\mathcal{T}_{\lambda/\mu^k}\left(\nabla\mathcal{D}\circ\tau+\mathcal{J}\otimes\triangle\tau+\frac{1}{\mu^k}\mathcal{Y}^k-\mathcal{A}^{k+1}\right)\\ \triangle\tau_t^{k+1} &=\mathcal{J}_t^{\dagger}\otimes\Bigg(\mathcal{A}^{k+1}_{(:,:,t)}+\mathcal{E}^{k+1}_{(:,:,t)}-\nabla\mathcal{D}_{(:,:,t)}\circ\tau\\ &\quad-\frac{1}{\mu^k}\mathcal{Y}^k_{(:,:,t)}\Bigg),\quad \text{for } t=1,2,\dots,N\end{aligned} \tag{19}$$

where the $\mathcal{T}_\alpha()$ denotes the soft thresholding operation with threshold value α. In the third equation of (19), we use the Tensor-Matrix Product and Tensor Transpose defined as follows:

Definition 2 (Tensor-Matrix Product). *The product of a tensor* $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ *and a matrix* $\mathbf{B} \in \mathbb{R}^{n_2 \times n_3}$ *is a vector* $\mathbf{c} \in \mathbb{R}^{n_1}$. *It is given by* $\mathbf{c} = \mathcal{A} \otimes \mathbf{B}$, *where* $\mathbf{c}_{(i)} = \sum_{j=1}^{n_2} \sum_{t=1}^{n_3} \mathcal{A}_{(i,j,t)} \mathbf{B}_{(j,t)}$, *for* $i = 1, 2, \ldots, n_1$.

Definition 3 (Tensor Transpose). *The transpose of a tensor* $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ *is the tensor* $\mathcal{A}^T \in \mathbb{R}^{n_3 \times n_1 \times n_2}$.

The registration algorithm for the batch mode is summarized in Algorithm 1. Let $M = w \times h$ be the number of pixels of each image. We set $\lambda = 1/\sqrt{M}$ and $\mu_k = 1.25^k \mu_0$ in the experiments, where $\mu_0 = 1.25/\|\nabla D\|_2$. For the inner loop, applying the fast Fourier transform (FFT) costs $\mathcal{O}(N \log N)$ for each pixel. All the other steps cost $\mathcal{O}(MN)$. Therefore, the total computation complexity of our method is $\mathcal{O}(MN \log N + MN)$, which is faster than $\mathcal{O}(N^2 M)$ when applying SVD decomposition in RASL (if $M \gg N$).

Algorithm 1. Image Registration via HSR - Batch Mode

```
Input: Images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N$, initial transformations
$\tau_1, \tau_2, \ldots, \tau_N$, regularization parameter $\lambda$.
repeat
  1)  Compute $\mathcal{J}_t = \frac{\partial}{\partial \zeta}(\nabla \mathbf{I}_t \circ \zeta)|_{\zeta = \tau_t}$, $t = 1, 2, \ldots, N$;
  2)  Warp and normalize the gradient images:
        $\nabla \mathcal{D} \circ \tau = \left[ \frac{\nabla \mathbf{I}_1 \circ \tau_1}{\|\nabla \mathbf{I}_1 \circ \tau_1\|_F}; \ldots; \frac{\nabla \mathbf{I}_N \circ \tau_N}{\|\nabla \mathbf{I}_N \circ \tau_N\|_F} \right]$;
  3)  Use (19) to iteratively solve the minimization problem
      of ALM:
        $\mathcal{A}^*, \mathcal{E}^*, \triangle\tau^* = \arg\min \mathcal{L}(\mathcal{A}, \mathcal{E}, \triangle\tau, \mathcal{Y})$;
  4)  Update transformations: $\tau = \tau + \triangle\tau^*$;
until Stop criteria
```

4.2 Pair Mode

Similar to that in the batch mode, we have

$$\nabla \mathbf{I}_2 \circ (\tau + \triangle\tau) \approx \nabla \mathbf{I}_2 \circ \tau + \mathcal{J} \otimes \triangle\tau \tag{20}$$

where $\mathcal{J} \in \mathbb{R}^{w \times h \times p}$ denotes the Jacobian. Thus, the pair mode (13) is to minimize the energy function with respect to $\triangle\tau$

$$E(\triangle\tau) = \|\nabla \mathbf{I}_1 - \nabla \mathbf{I}_2 \circ \tau - \mathcal{J} \otimes \triangle\tau\|_1 \tag{21}$$

The ℓ_1 norm in (21) is not smooth. We can have a tight approximation for the absolute value: $|x| = \sqrt{x^2 + \epsilon}$, where ϵ is a small constant (e.g., 10^{-10}). Let $\mathbf{r} = \nabla \mathbf{I}_1 - \nabla \mathbf{I}_2 \circ \tau - \mathcal{J} \otimes \triangle\tau$, and we can obtain the gradient of the energy function by the chain rule

$$\nabla E(\triangle\tau) = \mathcal{J}^T \otimes \frac{\mathbf{r}}{\sqrt{\mathbf{r} \circ \mathbf{r} + \epsilon}} \tag{22}$$

where $\circ$ denotes the Hadamard product. Note that the division in (22) is element-wise.

Gradient descent with backtracking is used to minimize the energy function (21), which is summarized in Algorithm 2. We set the initial step size $\mu^0 = 1$ and $\eta = 0.8$. The computational complexity of each iteration is $\mathcal{O}(M)$, which is much faster than $\mathcal{O}(M \log M)$ in RC when fast cosine transform (FCT) is applied [14]. Hierarchal estimation is used for both rigid and non-rigid registration [41]. The function value is calculated on the overlapped area of two images. Similar to the batch mode, we used the normalized images to rule out the trivial solutions. We used a coarse-to-fine hierarchical registration architecture for both the batch mode and pair mode.

Algorithm 2. Image Registration via HSR - Pair Mode

```
input: $\mathbf{I}_1, \mathbf{I}_2, \eta < 1, \tau, \mu^0$.
repeat
  1)  Warp and normalize $\mathbf{I}_2$ with $\tau$;
  2)  $\mu = \mu^0$;
  3)  Compute $\triangle\tau = -\mu \nabla E(\mathbf{0})$;
  4)  If $E(\triangle\tau) > E(\mathbf{0})$,
          set $\mu = \eta\mu$ and go back to 3);
  5)  Update transformation: $\tau = \tau + \triangle\tau$;
until Stop criteria
```

4.3 Convergence and Optimality

The alternative optimization approach is a common practice for solving difficult problems. Specifically, the original problem is decomposed to a sequence of easier subproblems and each subproblem is solved alternatively. The properties of this approach have been extensively studied and presented in the optimization literature. For example, the batch mode of HSR is similar to the RASL [15] and can be viewed as a Gauss-Newton method for optimizing a composited objective of a nonsmooth function and a smooth, nonlinear mapping. There are many existing studies on the convergence properties of these problems, and it still an active research topic [42]. Here, we discuss the convergence of our approach using the previous results of Jittorntrum and Osborne [43] and Cromme [44].

Given a problem of minimizing a composited function of a norm $\|\cdot\|_\diamond : \mathbb{R}^n \mapsto \mathbb{R}$

$$\min_{x \in \mathbb{R}^p} \|f(x)\|_\diamond, \tag{23}$$

where $f : \mathbb{R} \mapsto \mathbb{R}^n$ is a C^2 mapping. For the following iterative algorithm

$$\delta_k = \underset{\delta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| f(x_k) + \frac{\partial f}{\partial x}(x_k)\delta \right\|_\diamond, \tag{24}$$

$$x_{k+1} = x_k + \delta_k, \tag{25}$$

the authors of [43] and [44] demonstrated that if $x^* \in \mathbb{R}^p$ is a *strictly unique* optima to problem (23), then the sequence of (24) and (25) will have quadratic convergence to x^*. The concept of *strictly unique* is that

$$\begin{aligned} &\exists a > 0, \forall \delta \in \mathbb{R}^p, \\ &\left\| f(x^*) + \frac{\partial f}{\partial x}(x^*)\delta \right\|_\diamond \geq \|f(x^*)\|_\diamond + \alpha\|\delta\|. \end{aligned} \tag{26}$$

In order to apply the results of [43] and [44] to our HSR method, we need to demonstrate the connection of our problem to (23). For the case of the batch mode, we need to determine if the following equation is indeed a norm $\|\cdot\|_\diamond$.

$$\|\mathcal{M}\|_\diamond \equiv \min_{\mathcal{M} = \mathcal{A} + \mathcal{E}} \|\mathcal{F}_N \mathcal{A}\|_1 + \lambda\|\mathcal{E}\|_1, \tag{27}$$

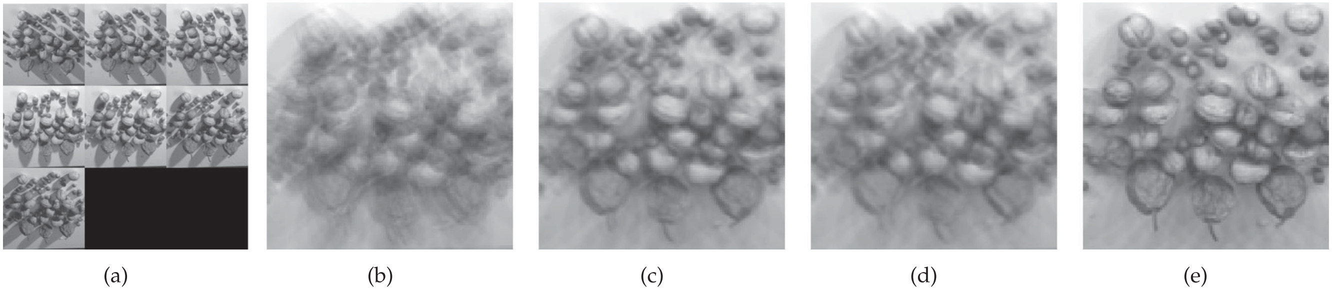


Fig. 3. Registration results on the “NUTS” data set: (a) the original “NUTS” images, (b) the average image of perturbed images, (c) the average image by RASL, (d) the average image by t-GRASTA, and (e) the average image by our method.

which is not hard to check $\|\mathcal{M}\|_\diamond \geq 0$, $\|\mathcal{M}\|_\diamond = 0 \Leftrightarrow \mathcal{M} = 0$, $\|t\mathcal{M}\|_\diamond = |t|\|\mathcal{M}\|_\diamond$ (for the linearity of Fourier transform and ℓ_1-norm). The triangle inequality follows from the convexity of the function $\|\mathcal{F}_N\mathcal{A}\|_1 + \lambda\|\mathcal{E}\|_1$. Let the transformation τ be represented by unknown variables x. Then, we can define $f(x) \equiv \nabla\mathcal{D} \circ \tau$. Hence, the batch mode of HSR in (11) can be regarded as solving the parameter x in (23) via iteration (24) and (25). Hence, since the map $x \mapsto f(x)$ is C^2, the result of [44] implies the quadratic convergence of the batch HSR. For the pair mode HSR (13), a similar analysis can also be applied. Although in general the manifolds of the transformed images are most likely not C^2, the transformed image $\mathbf{I}_i \circ \tau_i$ can be viewed as resampling transformations [45] of a perfect bandlimited reconstruction $\mathbf{I}_i$ from the digital image $\mathbf{I}_i$, in which case $f(x)$ is smooth. One practical concern is estimating the coefficient α in (26), which is still an open problem in the literature. We will leave this to future work and recommend [46] for more information.

5 Experimental Results

In this section, we validate our method on a wide range of applications. We compare our batch mode with RASL [15] and t-GRASTA [17], and compare our pair mode with RC [14] and SSD [2]. One of the most important advantages of our method is its robustness and accuracy on natural images under spatially-varying intensity distortions. As shown in [14] and Fig. 2, SAD [2], CC [36], CD2 [9], MS [37], MI [10] can easily fail in such cases. We do not include them in the following experiments. All experiments were conducted on a desktop computer with Intel i7-3770 CPU with 12 GB RAM.

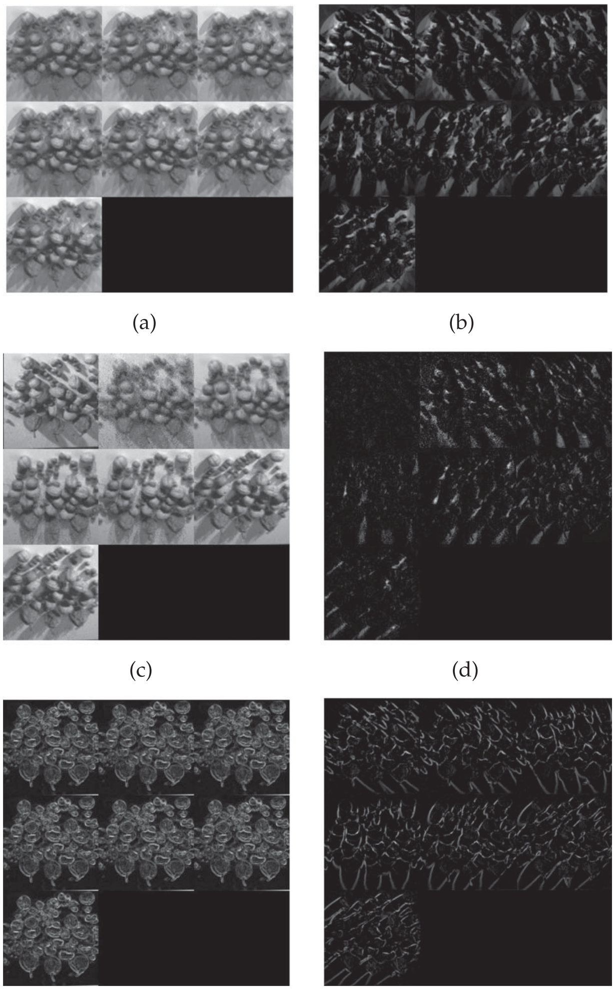
First, we verified the accuracy and robustness on two databases, both of which contain natural images captured under different illumination conditions. We then compared the accuracy of our method with RC on extensive synthetic examples. Finally, we tested our method on real-world multi-modal medical images and multi-sensor remotely sensed images. For the batch mode evaluation, one most important hyper-parameter is λ, which controls the balance between the tolerance of misalignment and the sparse error. Its value depends on the data. Specifically, it reflect our belief on the magnitude of the sparse error compared to the other term. We experiment various values for λ from $0.1/\sqrt{M}$ to $1000/\sqrt{M}$ and discover than the performance of the proposed approach is quite robust to different λ. Therefore, in all the following experiments, we always fix $\lambda = 1/\sqrt{M}$ without further tuning, where M is the number of pixels.

5.1 Batch Image Registration

To evaluate the performance of our batch mode, we used a popular database of naturally captured images in the VGG viewpoint dataset [47]. We chose the four data sets with the largest lighting variations: “NUTS”, “MOVI”, “FRUITS” and “TOY”. These data sets are very challenging to register, as they have up to 20 different lighting conditions and are occluded by varying shadows. Random translations of both directions were applied to the four data sets, which were drawn from a uniform distribution in a range of 10 pixels.

After registration on the “NUTS” data set, the average of perturbed images and results are shown in Fig. 3, where the average image by the proposed method has significantly sharper edges than those by the two existing methods. The two components of each algorithm are shown in Fig. 4. RASL [15] and t-GRASTA [17] failed to separate the shadows and large errors; however, we were able to successfully find the hierarchical sparse representation of the optimally registered images. Since our model successfully captured the variances in the variants like shadows (Fig. 4), we saw that the aligned images in $\mathcal{A}$ (in our case the gradient of the images) look more alike than the baseline approaches. The ability to capture the variances is actually the key to forming good alignments. The quantitative comparisons on the four data sets are listed in Table 1 over 20 random runs. The overall average of errors using our method was consistently lower than the average erros using RASL and t-GRASTA. More importantly, only our method can consistently achieve subpixel accuracy. For 20 images with a size of 128×128 pixels, the registration time is around 7 seconds for RASL and our method with the latter slightly faster (roughly 0.5 second faster). The t-GRAST registration time was around 27 seconds. This might seem contradictory to the previous complexity analysis. However, the running time does not depend solely on computational complexity but also depends on implementation skills that affect the constant factor in the complexity. Since the N is small in our case, it is reasonable that the proposed approach and RASL have similar running times.

We evaluate these three methods on the Multi-PIE face database [48]. This database contains 20 images of each subject captured under different illumination conditions. We randomly initialized the transformations with rotations in a range of 10° and translations in 10 pixels on the first 100 subjects from Session 1. After optimization, the resulted transformations are expected to be the same, because the original images are well aligned. As the optimal solution is not unique (e.g., all images shift by 1 pixel), we compared the standard derivation (STD) of the transformations after



(a) (b)

(c) (d)

(e) (f)

Fig. 4. Batch image registration on the NUTS data sets: (a) low rank component by RASL, (b) sparse errors by RASL, (c) subspace representation by t-GRASTA, (d) sparse errors by t-GRASTA, (e) visualization of $\mathcal{A}$ by our method and (f) sparse error $\mathcal{E}$ by our method.

registration. So ideally, the STD should be zero when all the initial perturbations have been exactly removed. Fig. 5 shows the average registration results after over 20 runs for each subject. We split each transformation parameter (i.e., rotation, x-axis translation and y-axis translation) to different plots (Figs. 5b and 5d). Our method was more accurate than RASL and t-GRASTA for almost every subject.

5.2 Pair Image Registration

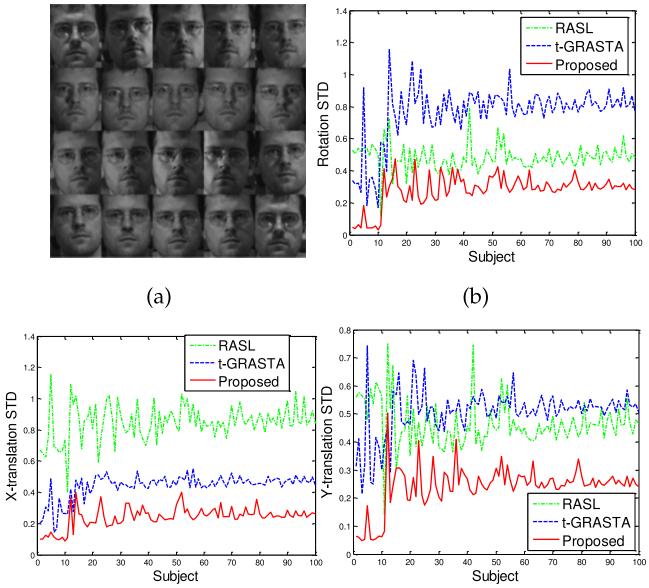
5.2.1 Simulations

For quantitative comparisons, we evaluated SSD, RC and the proposed method on the *Lena image* with random intensity distortions (Fig. 2) and random affine transformations (with a similar range as shown in the previous settings). The number of Gaussian intensity fields K is from 1 to 6. The reference image without intensity distortions was used as ground-truth. The root-mean-square error (RMSE) was used as the metric for error evaluation of both image intensities and transformations. We ran this experiment 50 times and the results are plotted in Fig. 6. It can be observed that the proposed method is consistently more accurate than SSD and RC, with different intensity distortions. The registration speed of our method is often faster than that of RC. The average speed for the pair mode is 6.5 seconds per registration on the brain image (216×180) while that of RC is 13.7 seconds per registration.

TABLE 1
Mean/Max Registration Errors in Pixels of RASL, t-GRASTA and Our Method on the Four Lighting Data Sets.

	RASL	t-GRASTA	Proposed
NUTS	0.670/2.443	1.153/3.842	**0.061/0.488**
MOVI	0.029/ 0.097	0.568/ 2.965	**0.007/0.024**
FRUITS	0.050/0.107	1.094/4.495	**0.031/0.076**
TOY	0.105/ 0.373	0.405/2.395	**0.038/0.076**

The first image is fixed to evaluate errors.



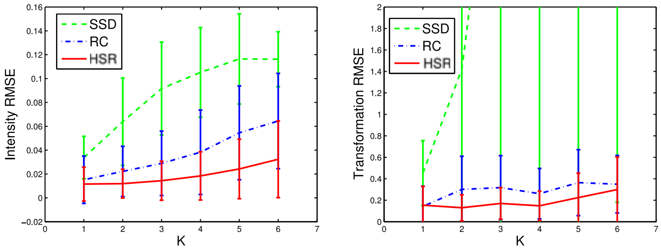
(a) (b)

(c) (d)

Fig. 5. (a) An example input of the Multi-PIE image database, (b) STD (in degrees) of rotations after registration, (c) STD (in pixels) of X-translation after registration, and (d) STD (in pixels) of Y-translation after registration.

5.2.2 Multisensor Remotely Sensed Image Registration

Multisensor image registration is a key preprocessing operation in remote sensing, e.g., for image fusion [49] and change detection. The same land objects may be acquired at different times, under various illumination conditions by different sensors. Therefore, it is very possible that the input



(a) (b)

Fig. 6. Registration performance comparisons with random transformation perturbations and random intensity distortions: (a) intensity RMSE on the Lena image and (b) transformation (affine) RMSE on the Lena image.

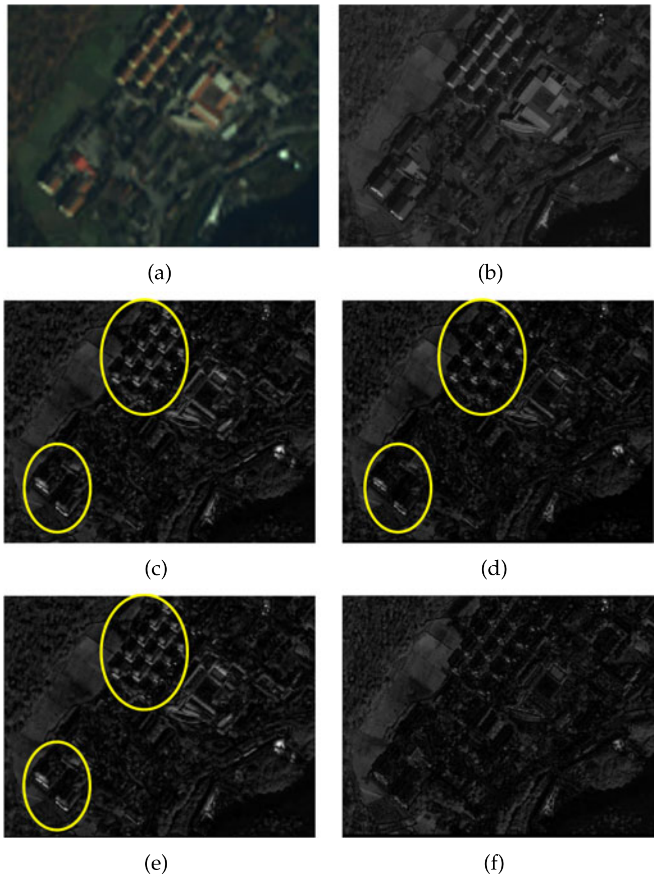


Fig. 7. Registration of a multispectral image and a panchromatic image: (a) reference image, (b) source image, (c) difference in image before registration, (d) difference in image by SSD, (e) difference in image by RC, and (f) difference in image by our method. Visible misalignments are highlighted by the yellow circles. Best viewed in ×2 sized color pdf file.

images have significant dissimilarity in terms of intensity values. Here, we register a panchromatic image to a multispectral image acquired by the IKONOS multispectral imaging satellite [50], which has been pre-registered at their capture resolutions. The multispectral image has four bands: blue, green, red and near-infrared, with four meter resolutions (Fig. 7a). The Pan image has a 1-meter resolution (Fig. 7b). The different image resolutions make this problem more difficult. From the difference image in Fig. 7c, we can observe that there exists misalignment in the northwest direction.

We compare our method with SSD [2] and RC [14], and the results are shown in Figs. 7d–f. It is assumed that the true transformation is formed by pure translation. Although we do not have the ground-truth, from the difference image, it can be clearly observed that our method can reduce the misalignment. The difference in images of the baseline approaches demonstrate some “bright” areas, which means the difference of these areas between the reference image and the source image is big. On the other hand, the difference in images of our approach looks darker, which means the errors are smaller. In contrast, SSD and RC were not able to find better alignments than the preregistration method. For this experiment, running time is 1.8 second for SSD, 5.2 second for RC and 3.4 second for the proposed HSR.

We registered an aerial photograph to a digital orthophoto. The reference image is the orthorectified MassGIS georegistered orthophoto [51]. The source image is a digital aerial photograph, which does not have any particular alignment or registration with respect to the earth. The input images and the results are shown in Fig. 8. MATLAB uses manually selected control points for registration, while RC and our registrations are automatic. At the first glance, all the methods obtained registration with good quality. A closer look shows that our method has a higher accuracy than the others. In the source image, two lanes can be clearly observed in streets A and B. After registration and composition, Street B in the result by MATLAB and Street A in the result by RC are blurry due to the misalignment. Our method is robust to the local mismatches of vehicles.

5.3 Face Alignment and Verification in the Wild

We evaluated the performance of SSD, RC and the proposed HSR on face images from the LFW dataset [52]. The faces for each subject were captured at different times and locations, with significant appearance inconsistency. In addition, the various expressions on the face make this problem more difficult. To handle such diversity, most existing methods require a batch of images as the inputs [15], [16], [53]. Then they can exploit the underlying structures of the image set, e.g., low rank. However, a few of them can be applied for the registration of only two images. A few methods could

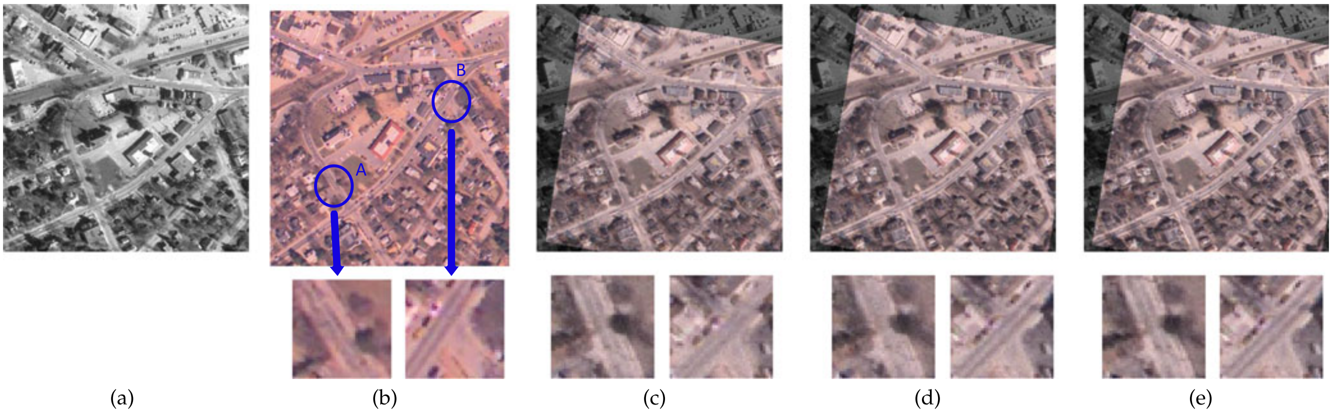


Fig. 8. Registration of an aerial photograph and a digital orthophoto. From left to right, the images are: the reference image, the source image, the overlay by MATLAB, the overlay by RC and the overlay by our method. The second row shows the zoomed-in areas of streets A and B. Best viewed in ×2 sized color pdf file.

TABLE 2
Unconstrained Face Verification Accuracy on View 1 of LFW Using Images Produced by Different Alignment Algorithms

	Original	SSD	RC	Proposed
Accuracy	0.742	0.726	0.755	**0.759**

be used to align these images, but they require a batch of images as the inputs [15], [16], [53]. Therefore, they cannot be applied here for two-image based registration.

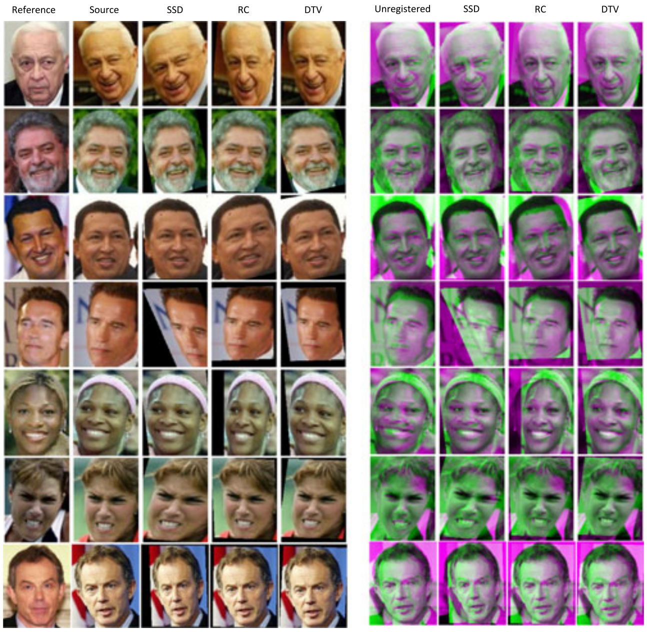
In order to provide quantitative analysis, we follow the setting of [53] and conduct an experiment on face verification. We measure the verification accuracy using View 1 of LFW. We implement a variant of Cosine Similarity Metric Learning (CSML) [54], which is also used in [53]. Each pair of images are aligned separately using pair alignment approaches. For preprocessing, whitening PCA is used to reduce the representation dimension to 500. Then the feature vector is normalize for each image. A linear SVM is applied to the each pair of image by combining the feature vectors using element-wise multiplication. For simplicity, we only used single image feature, the square root LBP features [53], [54] on 150 × 80 cropping of the full LFW images. Table 2 shows the results. The original column means the accuracy of CSML on unaligned images. Both RC and the proposed HSR have achieved better performance than the unaligned version, while the proposed HSR performs the best. SSD, however, achieves a suboptimal performance than the original one. This effect can be validate on some examples in Fig. 9, where SSD cannot capture the misalignment hence will lead to unreasonable distortion. We have also demonstrate this effect in Fig. 2. Here, we do not mean HSR is more accurate than the batch alignment methods [15], [16], [53]. However, when the number of images is limited, our method could be an alternative.

6 Non-Rigid Transformation

In many real world applications, the deformation is not always rigid. An example would be images with local motions or other smooth deformations, such as changes in facial expressions or organ motion in medical imaging. In these cases, non-rigid registration is required. For this problem, we use the free form deformation (FFD) transformation with B-spline control points [55], [56]. The basic idea behind the FFD is to model the underlying deformation as a mesh of control points. Suppose $\mathbf{I}_1$ is the reference image, and $\mathbf{I}_2$ is the source image to be registered. Let $\mathbf{I}(a, b)$ be the pixel value in the position (a, b). The problem is to find the non-rigid transformation $\mathbf{T} : (a, b) \mapsto (a', b')$. Also, let the domain of the image volume be $\Omega = \{(a, b) | 0 \le a < w, 0 \le b < h\}$ and let ϕ be the $n_1 \times n_2$ mesh of control points $\phi_{i,j}$ with uniform spacing δ. Then the transformation on the image can be formulated as

$$\mathbf{I} \circ \mathbf{T}(x, y) = \sum_{k=0}^{3} \sum_{l=0}^{3} B_k(u) B_k(v) \phi_{i+k, j+l}, \tag{28}$$

where $i = \lfloor a/n_1 \rfloor - 1, j = \lfloor b/n_2 \rfloor - 1, u = a/n_1 - \lfloor a/n_1 \rfloor, v = b/n_2 - \lfloor b/n_2 \rfloor$ and B_l represents the l uniform cubic B-Spline basis functions.



(a) (b)

Fig. 9. Face alignment results on the LFW data set [52]. Left to right: the input images, the warped results by SSD, RC and HSR, the overlays by SSD, RC and HSR.

The nonrigid B-spline FFD model is much more flexible than the affine transformation. The control points Φ are the parameters that need to be solved for the B-spline FFD. For a 10×10 mesh grid of control 2-D points, a B-Spline FFD has 200 degrees of freedom (DoF), which is much more than the affine transformation (DoF = 6) on the 2-D image. Therefore, with higher DoF, the non-rigid model is more flexible and can handle more subtle local motions. The price for this flexibility is computational complexity. The higher the DoF, the longer it takes to compute the solution. For all the experiments below, we set $\delta = 8$ and use coarse-to-fine hierarchical registration architecture.

With the FFD model, we can plug $\mathbf{T}$ into (13) and obtain our hierarchical sparse algorithm for non-rigid transformation. In order to avoid unnatural wraps, we impose an Euclidean loss upon the neighboring displacements of B-spline control points. The resulting formulation is listed below:

$$\min_{\mathbf{T}} \; \|\nabla \mathbf{I}_1 - \nabla \mathbf{I}_2 \circ \mathbf{T}\|_1 + \frac{\lambda}{2} \|\nabla \mathbf{T} - \nabla \mathbf{T}_0\|_F^2, \tag{29}$$

where T_0 is the initial configuration of the control points (i.e., the uniform spacing grid). Algorithm 2 can still be used to solve (29) with only a small modification of adding a term for the gradient regarding the regularization term.

7 Experiment Results for Non-Rigid Registration

7.1 BrainWeb Dataset

To evaluate the performance of the proposed method in non-rigid transformed cases, we first conducted a simulation on a brain MRI image from the BrainWeb dataset [57]. The source image is warped by a non-rigid transformation, perturbed from random zero-mean Gaussians with three pixels standard deviation. We added a few Gaussian

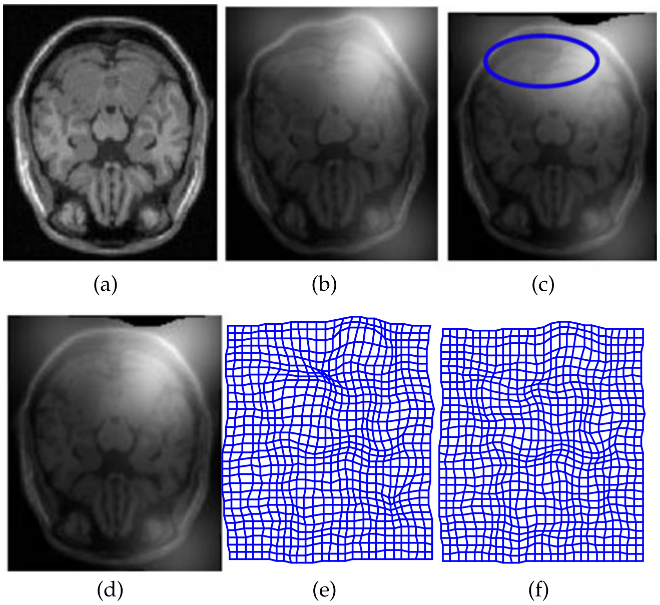


Fig. 10. Synthetic experiment with non-rigid transformation: (a) reference image, (b) source image with intensity distortion, (c) registration result by RC, (d) registration by our method, (e) transformation estimated by RC, and (f) transformation estimated by our method. Best viewed in a ×2 sized color pdf file.

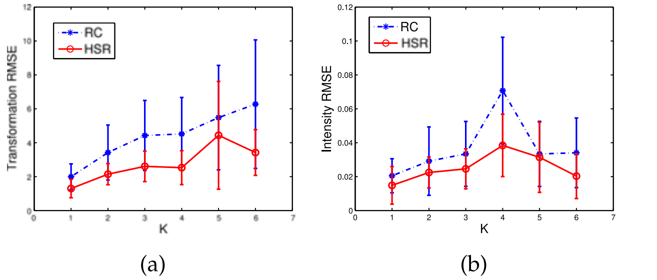


Fig. 11. Registration performance comparisons with random transformation perturbations and random intensity distortions: (a) intensity RMSE on the brain image and (b) transformation (non-rigid) RMSE on the brain image.

intensity fields to simulate the distortion and rescaled the images to [0,1]. Fig. 10 shows the input images and results by RC [14] and the proposed method. SSD is not compared in non-rigid registration, as it always failed although different settings were tried. Both results are very close to ground truth. A visible artifact can be observed in the image recovered by RC, which is highlighted by the blue circle. The estimated transformation by our method is smoother, and closer to the Gaussion perturbations. The corresponding gradient images are shown in the second row. Despite significant intensity differences, we were still able to find good similarity in the image gradients after registration. We also ran this experiment 50 times and show the RMSE in Fig. 11, where the proposed method is consistently outperform RC, with different intensity distortions (similar to Fig. 6). This figure successfully interprets our motivation of registration in the gradient domain. Under this severe intensity distortion, our method proved to be more accurate than RC for recovering image details.

7.2 Multimodal Medical Image Registration

We further validated the performance of different methods on real-world medical images. Temporal and multimodal registration procedures were performed on two retina images taken two years apart [58]. We call this *the Retina dataset*. The reference image and source image are shown in Figs. 12a and b. These retina images were quite difficult to register with intensities. In order to avoid local minimum, we used affine transformation for preregistration and the result is shown in Fig. 12c. From the overlay in Fig. 12d, we observed that misalignments still existed for the vessels at the bottom half of the overlay. A local error was found in the RC results, while our method eliminated the misalignments.

The proposed method is compared with RC on two images from a iris video sequence [14] (shown in Fig. 13). The deformation between the source image and reference image is highly nonlinear. The intensity artifact in the source image makes this problem more challenging. The overlay without registration is shown in Fig. 13c using green and magenta colors. The vessels are blurry due to the misalignment. After registration, both RC and the proposed method provided accurate alignments on the vessels. However, the image registered by RC was partially distorted due to severe intensity variance. For speed comparison, our approach ran 8 seconds in this experiment, while RC ran 12 seconds.

8 Conclusion and Discussion

In this article, we proposed a novel similarity measure for robust and accurate image registration. It was motivated by

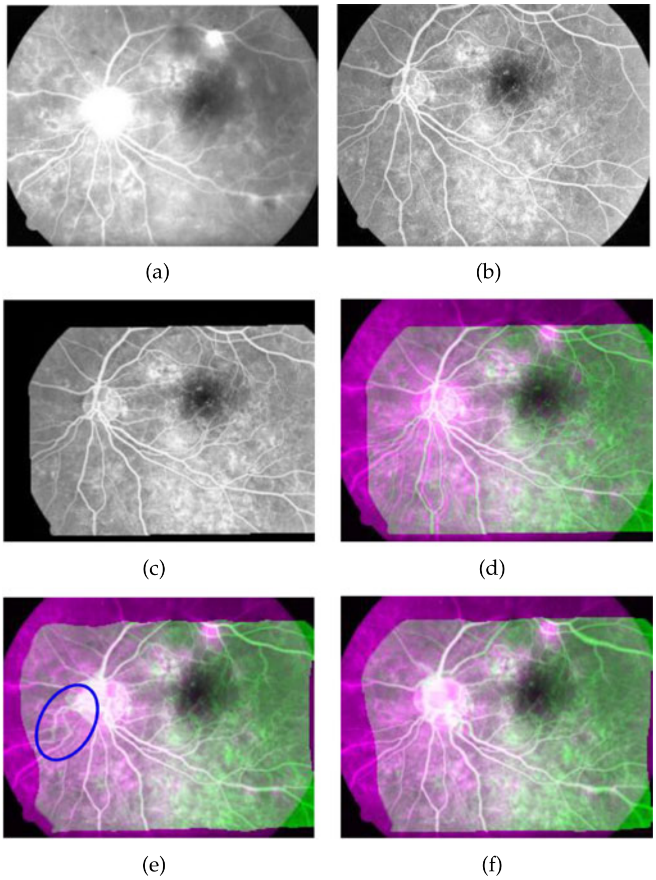
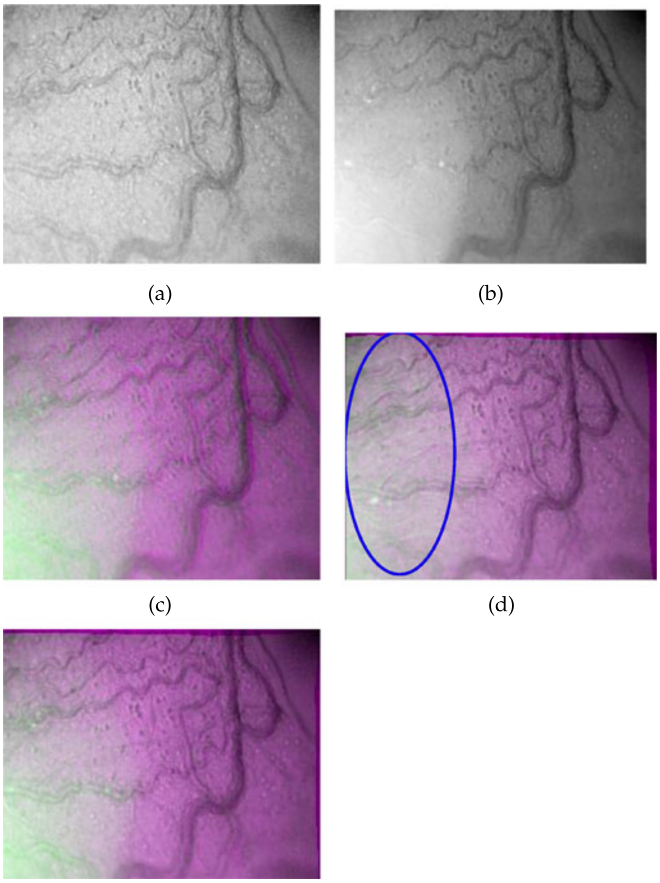


Fig. 12. Registration of two retina images [58]: (a) reference image, (b) source image, (c) source image after affine preregistration, (d) overlay before registration, (e) overlay after registration by RC, and (f) overlay after registration by our method. Visual artifact is highlighted by the blue circle.



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

Fig. 13. Registration of two iris images [14]: (a) reference image, (b) source image, (c) overlay before registration, (d) overlay after registration by RC, and (e) overlay after registration by our method. A visible artifact is highlighted by the blue circle. This is best viewed in ×2 sized color pdf file.

hierarchical sparse representation of optimally registered images. The benefit of the proposed method is three fold: (1) Compared with existing approaches, it can handle severe intensity distortions and partial occlusions simultaneously; (2) it can be used for registration of two images or a batch of images with various types of transformations; (3) its low computational complexity makes it scalable to large data sets. We conducted extensive experiments to test our method on multiple challenging data sets. The promising results demonstrate the robustness and accuracy of our method over the state-of-the-art batch registration methods and pair registration methods, respectively. We also show that our method can be used to reduce registration errors in many real-world applications.

Due to the local linearization in the optimization, our method as well as all the compared methods cannot handle large transformations. However, this is not a big issue for many real-world applications. For example, the remotely sensed images can be coarsely georegistered by their geographical coordinates. For images with large transformations, we can use the FFT-based algorithm [11] to coarsely register the images and then apply our method as a refinement. Therefore, we did not test the maximum amount of transformations that our method can handle. So far, the proposed method can only be used for offline registration. How to extend this method to the online mode has been targeted as an excellent topic for future research.

Acknowledgments

This work was partially supported by US National Science Foundation IIS-1423056, CMMI-1434401, CNS-1405985, IIS-1718853 and the NSF CAREER grant IIS-1553687

References

[1] B. Zitova and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[2] R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations Trends® Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, 2006.

[3] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.

[4] P. Blanc, L. Wald, and T. Ranchin, "Importance and effect of co-registration quality in an example of pixel to pixel fusion process," in *Proc. 2nd Int. Conf. Fusion Earth Data: Merging Point Measurements Raster Maps Remotely Sensed Images*," 1998, pp. 67–74.

[5] Y. Zheng, et al., "Landmark matching based retinal image alignment by enforcing sparsity in correspondence matrix," *Med. Image Anal.*, vol. 18, no. 6, pp. 903–913, 2014.

[6] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, Apr. 2014.

[7] J. Ma, W. Qiu, J. Zhao, Y. Ma, A. Yuille, and Z. Tu, "Robust L2E estimation of transformation for non-rigid registration," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1115–1129, Mar. 2015.

[8] J. Huang, X. Huang, and D. Metaxas, "Simultaneous image transformation and sparse representation recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[9] B. Cohen and I. Dinstein, "New maximum likelihood motion estimation schemes for noisy ultrasound images," *Pattern Recog.*, vol. 35, no. 2, pp. 455–463, 2002.

[10] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.

[11] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki, "Robust FFT-based scale-invariant image registration with image gradients," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1899–1906, Oct. 2010.

[12] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas, "Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change," *IEEE Trans. Med. Imag.*, vol. 25, no. 5, pp. 626–639, May 2006.

[13] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geoscience Remote Sensing*, vol. 46, no. 5, pp. 1301–1312, May 2008.

[14] A. Myronenko and X. Song, "Intensity-based image registration by minimizing residual complexity," *IEEE Trans. Med. Imag.*, vol. 29, no. 11, pp. 1882–1891, Nov. 2010.

[15] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, Nov. 2012.

[16] Y. Wu, B. Shen, and H. Ling, "Online robust image alignment via iterative convex optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 1808–1814.

[17] J. He, D. Zhang, L. Balzano, and T. Tao, "Iterative grassmannian optimization for robust image alignment," *Image Vis. Comput.*, vol. 32, no. 10, pp. 800–813, 2014.

[18] K. K. Wu, L. Wang, F. K. Soong, and Y. Yam, "A sparse and low-rank approach to efficient face alignment for photo-real talking head synthesis," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2011, pp. 1397–1400.

[19] W.-T. Tan, G. Cheung, and Y. Ma, "Face recovery in conference video streaming using robust principal component analysis," in *Proc. 18th IEEE Int. Conf. Image Process.*, 2011, pp. 3225–3228.

[20] R. Otazo, E. Candès, and D. K. Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components," *Magn. Resonance Med.*, vol. 73, no. 3, pp. 1125–1136, 2015.

[21] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[22] V. Hamy, et al., "Respiratory motion correction in dynamic MRI using robust data decomposition registration–application to DCE-MRI," *Med. Image Anal.*, vol. 18, no. 2, pp. 301–313, 2014.

[23] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Phys. Med. Biol.*, vol. 46, no. 3, 2001, Art. no. R1.

[24] Y. Li, C. Chen, F. Yang, and J. Huang, "Deep sparse representation for robust image registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4894–4901.

[25] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2011, Art. no. 11.

[26] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, 1992.

[27] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434, Nov. 2009.

[28] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, vol. 2, pp. 60–65.

[29] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imag. Vis.*, vol. 20, no. 1, pp. 89–97, 2004.

[30] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations Trends® Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[31] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[32] L. Condat, "A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms," *J. Optimization Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.

[33] I. Bayram and M. E. Kamasak, "Directional total variation," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 781–784, Dec. 2012.

[34] A. Toma, B. Sixou, L. Denis, J.-B. Pialat, and F. Peyrin, "Higher order total variation super-resolution from a single trabecular bone image," in *Proc. IEEE 11th Int. Symp. Biomed. Imag.*, 2014, pp. 1152–1155.

[35] R. H. Chan, H. Liang, S. Wei, M. Nikolova, and X.-C. Tai, "High-order total variation regularization approach for axially symmetric object tomography from a single radiograph," *Inverse Problems Imag.*, vol. 9, no. 1, pp. 55–77, 2015.

[36] J. Kim and J. A. Fessler, "Intensity-based image registration using robust correlation coefficients," *IEEE Trans. Med. Imag.*, vol. 23, no. 11, pp. 1430–1444, Nov. 2004.

[37] A. Myronenko, X. Song, and D. J. Sahn, "Maximum likelihood motion estimation in 3D echocardiography through non-rigid registration in spherical coordinates," in *Proc. Int. Conf. Functional Imag. Modeling Heart*, 2009, pp. 427–436.

[38] J. Huang, S. Zhang, and D. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," *Med. Image Anal.*, vol. 15, no. 5, pp. 670–679, 2011.

[39] J. Huang, S. Zhang, H. Li, and D. Metaxas, "Composite splitting algorithms for convex optimization," *Comput. Vis. Image Understanding*, vol. 115, no. 12, pp. 1610–1622, 2011.

[40] Y. Li, C. Chen, J. Zhou, and J. Huang, "Robust image registration in the gradient domain," in *Proc. Int. Symp. Biomed. Imag.*, 2015, pp. 605–608.

[41] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Proc. Eur. Conf. Comput. Vis.*, 1992, pp. 237–252.

[42] A. S. Lewis and S. J. Wright, "A proximal method for composite minimization," *Math. Programm.*, Springer, vol. 158, no. 1-2, pp. 501–546, 2016.

[43] K. Jittorntrum and M. Osborne, "Strong uniqueness and second order convergence in nonlinear discrete approximation," *Numerische Math.*, vol. 34, no. 4, pp. 439–455, 1980.

[44] L. Cromme, "Strong uniqueness: A far-reaching criterion for the convergence analysis of iterative procedures," *Numerische Math.*, vol. 29, no. 2, pp. 179–193, 1978.

[45] A. A. Goshtasby, "Image resampling and compositing," in *Image Registration*. Berlin, Germany: Springer, 2012, pp. 401–414.

[46] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in *Proc. IEEE Int. Symp. Inf. Theory*, 2010, pp. 1518–1522.

[47] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, Springer, vol. 60, no. 1, pp. 63–86, 2004.

[48] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[49] C. Chen, Y. Li, W. Liu, and J. Huang, "Image fusion with local spectral consistency and dynamic gradient sparsity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2760–2765.

[50] Space-Imaging, "IKONOS scene po-37836," *Geoeye IKONOS Scene Data*, 2000. [Online]. Available: http://glcf.umd.edu/data/ikonos/

[51] [Online]. Available: http://www.mathworks.com/help/images/register-an-aerial-photograph-to-a-digita l-orthophoto.html

[52] 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, MA, USA, Tech. Rep. 07–49, 2007.

[53] G. B. Huang, M. A. Mattar, H. Lee, and E. G. Learned-Miller, "Learning to align from scratch," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 773–781.

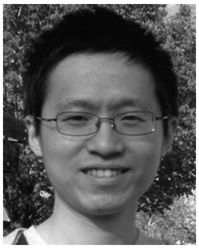
[54] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 709–720.

[55] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[56] S. Lee, G. Wolberg, and S. Y. Shin, "Scattered data interpolation with multilevel b-splines," *IEEE Trans. Vis. Comput. Graph.*, vol. 3, no. 3, pp. 228–244, Jul.–Sep. 1997.

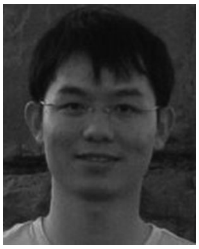
[57] C. A. Cocosco, V. Kollokian, R. K.-S. Kwan, G. B. Pike, and A. C. Evans, "Brainweb: Online interface to a 3D MRI simulated brain database," *NeuroImage*, vol. 5, 1997, Art. no. 425.

[58] F. Zana and J.-C. Klein, "A multimodal registration algorithm of eye fundus images using vessels detection and hough transform," *IEEE Trans. Med. Imag.*, vol. 18, no. 5, pp. 419–428, May 1999.



Yeqing Li received the BE degree in computer science and technology from Shantou University, China, in 2006, the ME degree from Nanjing University, Nanjing, China, in 2009 and the PhD degree from the Department of Computer Science, the University of Texas, Arlington, in 2015. His major research interests include machine learning, pattern recognition, medical image analysis, and computer vision. He is a member of the IEEE.

Chen Chen received the BE and the MS degrees both from Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2011, respectively, the second MS degree from the University of Texas, Arlington in 2015, and the PhD degree from the Department of Electrical and Computer Engineering, the University of Illinois, Urbana-Champaign since 2015. His major research interests include image processing, computer vision, and machine learning. He is a student member of the IEEE.



Fei Yang received the BE degree from Tsinghua University, Bejing, China, in 2003, the ME degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006, and the PhD degree in computer science from Rutgers University, New Brunswick, New Jersey, in 2013. He is a scientist in Facebook Inc. His research focuses on computer vision and machine learning, especially on face recognition, face tracking, and image classification etc. He is a member of the IEEE.

Junzhou Huang received the BE degree from Huazhong University of Science and Technology, Wuhan, China, in 1996, the MS degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003, and the PhD degree from Rutgers University, New Brunswick, New Jersey, in 2011. He is an associate professor in the Computer Science and Engineering Department, the University of Texas, Arlington. His research interests include machine learning, biomedical imaging informatics and computer vision, with focus on the development of sparse modeling, imaging, and learning for big data analytics. He is a member of the IEEE.