# Deep Sparse Representation for Robust Image Registration

Yeqing Li[1*], Chen Chen[1*], Fei Yang[2], and Junzhou Huang[1]
[1]University of Texas at Arlington. [2]Facebook Inc.

The definition of the similarity measure is an essential component in image registration. In many real-world applications of image registration, the images have significantly different appearances due to the intensity variations. Many existing intensity based methods may fail to solve these challenging problems. In this paper, we propose a novel similarity measure for registration of two or more images. The proposed method is motivated by that the optimally registered images can be deeply sparsified in the gradient domain and frequency domain, with the separation of a sparse tensor of errors. One of the key advantages of the proposed similarity measure is its robustness to severe intensity distortions, which widely exist on medical images, remotely sensed images and natural photos due to the difference of acquisition modalities or illumination conditions.

Unlike previous works that vectorize each image into a vector [2, 3, 4], we arrange the input images into a 3D tensor to keep their spatial structure. With this arrangement, the optimally registered image tensor can be deeply sparsified into a sparse frequency tensor and a sparse error tensor (see Fig. 1 for more details). Severe intensity distortions and partial occlusions will be sparsified and separated out in the first and second layers, while any misalignment will increase the sparseness of the frequency tensor (third layer). We propose a novel similarity measure based on such deep 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.

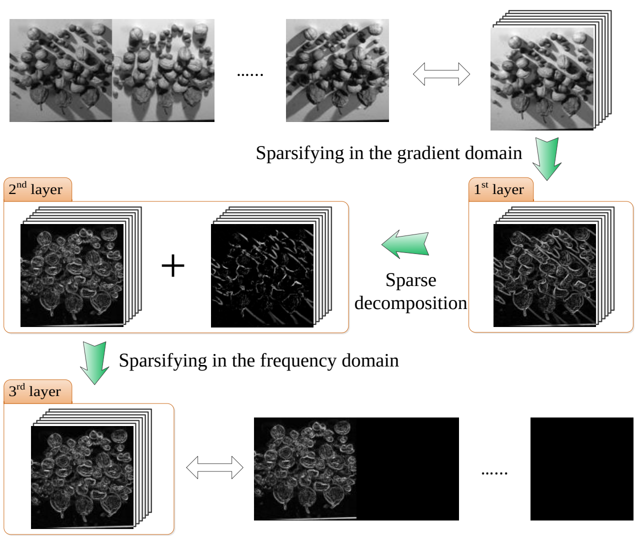


Figure 1: Deep sparse representation of the optimally registered images. First we sparsify the image tensor into the gradient tensor (1st layer). The sparse error tensor is then separated out in the 2nd layer. The gradient tensor with repetitive patterns are sparsified in the frequency domain. Finally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

We introduce our deep sparsity architecture in the inverse order for easy understanding. Suppose we have a batch of grayscale images $\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N \in \mathbb{R}^{w \times h}$ to be registered, where $N$ denotes the total number of images. First, we consider the simplest case that all the input images are identical and perturbed from a set of transformations $\tau = \{\tau_1, \tau_2, ..., \tau_N\}$. We arrange the input images into a 3D tensor $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{w \times h \times N}$, with

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

After removing the transformation perturbations, the slices show repetitive patterns. Such periodic signals are extremely sparse in the frequency domain. Ideally the Fourier coefficients from the second slice to the last slice should be all zeros. We can minimize the $\ell_1$ norm of the Fourier coefficients to seek the optimal transformations:

$$\min_{\boldsymbol{\mathcal{A}}, \tau} ||\mathcal{F}_N \boldsymbol{\mathcal{A}}||_1, \ s.t. \ \boldsymbol{\mathcal{D}} \circ \tau = \boldsymbol{\mathcal{A}}, \tag{2}$$

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

The above model can be hardly used on practical cases, due to the corruptions and partial occlusions in the images. Similar as previous work [3], we assume the noise is negligible in magnitude as compared to the error caused by occlusions. Let $\boldsymbol{\mathcal{E}}$ be the error tensor. We can separate it from the image tensor if it is sparse enough. Similar, we use the $\ell_1$ norm to induce sparseness:

$$\min_{\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \tau} ||\mathcal{F}_N \boldsymbol{\mathcal{A}}||_1 + \lambda ||\boldsymbol{\mathcal{E}}||_1, \ s.t. \ \boldsymbol{\mathcal{D}} \circ \tau = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{E}}, \tag{3}$$

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

The above approach requires that the error $\boldsymbol{\mathcal{E}}$ is sparse. However, in many real-world applications, the images are corrupted with spatially-varying intensity distortions. Existing methods such as RASL [3] and t-GRASTA [2] 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 almost keep the same, even under severe intensity distortions. Therefore, we register the images in the gradient domain:

$$\min_{\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \tau} ||\mathcal{F}_N \boldsymbol{\mathcal{A}}||_1 + \lambda ||\boldsymbol{\mathcal{E}}||_1, \ s.t. \ \nabla \boldsymbol{\mathcal{D}} \circ \tau = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{E}}, \tag{4}$$

where $\nabla \boldsymbol{\mathcal{D}} = \sqrt{(\nabla_x \boldsymbol{\mathcal{D}})^2 + (\nabla_y \boldsymbol{\mathcal{D}})^2}$ 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 layer architecture, which is shown in Fig. 1. We call it deep sparse representation of images. Comparing with existing popular low rank representation [3], 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.

This problem (4) can be solved by the augmented Lagrange multiplier (ALM) algorithm [1, 3]. We compare our method with 9 traditional and state-of-the-art algorithms on a wide range of natural image datasets, 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.

[1] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[2] Jun He, Dejiao Zhang, Laura Balzano, and Tao Tao. Iterative grassmannian optimization for robust image alignment. *Image and Vision Computing*, 32(10):800–813, 2014.

[3] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2233–2246, 2012.

[4] Yi Wu, Bin Shen, and Haibin Ling. Online robust image alignment via iterative convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1808–1814, 2012.

[*] indicates equal contributions. This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.