Image Registration Robust to Sparse Large Errors

Jing Liu, *EMBS Member,* Marian Chuang, Andrew Chisholm, Pamela Cosman, *Fellow, IEEE*

***Abstract*— Registration is difficult when images to be registered contain sparse but large-valued differences. We present a method for robust registration that ignores some fraction of large differences, while constraining the sparseness of these errors. We apply the method to stabilize microscopy videos of *C. elegans* tissues, in which bright moving filaments and tissue wounding appear as sparse large-valued differences. We demonstrate the advantage of the method on both synthetic and real data compared to state-of-the-art methods.**

I. Introduction

We consider the problem of image registration or video stabilization where the images to be registered have sparse but large-valued differences. Such differences arise in various ways, for example, registering face images where one image includes sunglasses, or registering images of blood vessels before and after administration of a contrast agent. In our experiments, we use the nematode *Caenorhabditis elegans* (*C. elegans*), focusing on subcellular dynamics in skin tissues during wound repair *in vivo*. A common problem with acquiring live images of *C. elegans*, and other living genetic model organisms, is the inability to completely immobilize the animal. This problem is further pronounced when we use a laser to wound the epidermis to visualize subcellular dynamics after injury, which often provokes sudden animal movement. To analyze subcellular processes such as microtubule dynamics or vesicle transport, we first need to robustly stabilize the video. Bright moving filaments and the wounding process can all be viewed as large-valued but sparse differences between frames.

Many robust image registration methods have been developed to handle outliers (e.g. occlusion, shadows). Rather than examining the similarity of actual intensity values, Kaneko et al. [1] proposed checking the consistency of intensity changes between two images. This was extended in [2] by using consistent pixels as a mask to calculate a selective correlation coefficient. However, many outliers are still used to compute the correlation coefficient. Also, these methods are not robust to relatively large noise.

Kim et al. [3] proposed the Robust Correlation Coefficients, and Arya et al. [4] proposed the M-estimator Correlation Coefficient. These aim to use some robust function as the weight to compute correlation coefficients. Similarly, robust error functions such as the Geman-McLure function [5-7] and Huber function [5, 8] are often used in intensity distance measures to cope with outliers. Ishikawa et al. [9] proposed a robust version of the Inverse Compositional (IC [10]) method; in each iteration, it ignores a fixed percentage of pixels that have relatively large errors, although it still uses the sum of squared differences (SSD) as the distance function.

The recent advances in sparse recovery and rank minimization inspired the work of robust alignment by sparse and low-rank decomposition (RASL [11]). They assume the matrix of transformed batch images (each column is a vectorized transformed image) can be decomposed as a low rank matrix and a sparse error matrix to handle corruption, occlusion, shadows, etc. However, they need the initial misalignment to be small, a problematic requirement for real applications. Also, their method cannot handle well additive noise that is not sparse.

Bernard et al. [12] also model the sparse errors (e.g. pixels belonging to moving players in field-sports video) in their frame-by-frame based Robust Video Registration (RVR) framework. However, their framework also lacks the consideration of noise. Also, the updating step of transformation parameters in both RASL and RVR is still influenced by the value of estimated sparse error pixels.

Our method also assumes sparse large-valued differences between the images. Our framework explicitly minimizes the L0-norm and can handle various additive noises as well as relatively large initial misalignment. We introduce our general framework and give a case study in Section II. In Section III, we focus on our particular application and demonstrate the value of our method. Discussion and conclusions are in Section IV.

II. Methods

A. Problem Formulation

Let us denote the first image as a fixed image $T \in R^{m\times n}$, and the second image (which will be registered to the first) is denoted $I \in R^{m\times n}$. We assume that these images differ in three ways: (1) an underlying 2D parametric transformation w between I and T, (2) some relatively small magnitude additive noise e, and (3) some sparse large magnitude differences E (which may include spatially clustered differences such as in the sunglasses example, and which may include some isolated large magnitude noise) between T and $I \circ w$ (transform I by w). More precisely, we model the relationship between T and I as $T = I \circ w$+e+ E. The errors E are sparse; only a small percentage of its entries are non-zero,

J. Liu and P. Cosman are with the Department of Electrical and Computer Engineering, University of California at San Diego, USA (e-mail: jil292@ucsd.edu; pcosman@ucsd.edu).

M. Chuang and A. Chisholm are with the Division of Biological Sciences, Section of Cell and Developmental Biology, University of California at San Diego, USA (e-mail: mchuang2004@gmail.com; chisholm@ucsd.edu).

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

but their magnitudes are large. The goal of the registration is to recover w.

We want to solve the following problem to estimate w:

$$\hat{w}, \hat{E} = \underset{w,E}{\operatorname{argmin}}\, dist(T, I \circ w + E) \quad (1)$$
$$s.t.\ ||E||_0 \leq k$$

where $||\cdot||_0$ denotes the cardinality (number of non-zero entries), k specifies the sparsity level of E (number of non-zero entries), and $dist(,)$ is a distance function. In our application, we use Mean of Absolute Difference (MAD):

$$dist(A, B) = \frac{1}{m \times n} \sum_{\boldsymbol{x}} |A(\boldsymbol{x}) - B(\boldsymbol{x})| \quad (2)$$

We express the problem (1) in a Lagrangian form:

$$\hat{w}, \hat{E} = \underset{w,E}{\operatorname{argmin}} J(w, E)$$
$$= \underset{w,E}{\operatorname{argmin}}\, dist(T, I \circ w + E) + \alpha||E||_0 \quad (3)$$

where the parameter α depends on the sparsity level k, distance function, and the data, as we will see below.

B. General Solution

The objective function in (3) is non-convex. For the second term in (3), most previous work relaxes the L0-norm to its convex surrogate (L1-norm of $vec(E)$), but we explicitly deal with the L0-norm. We use the coordinate gradient descent approach to minimize the cost function in (3). The overall procedure is summarized in Algorithm 1.

Algorithm 1 (General Solution)

Input: image T, image I, α

Initialization: $w^{(0)}$ as the solution of $\min_{w} dist_0(T, I \circ w)$;
$$E^{(0)} = \underset{E}{\operatorname{argmin}}\, dist\left(T, I \circ w^{(0)} + E\right) + \alpha||E||_0$$

While $J(w, E)$ not converged DO:

Step 1: fix E, update w by Δw, where

$$dist\left(T_{\delta(E)}, (I \circ w)_{\delta(E)} \circ \Delta w\right) \leq dist\left(T_{\delta(E)}, (I \circ w)_{\delta(E)}\right)$$

Step 2: fix w, update $E = \underset{E}{\operatorname{argmin}}\, dist(T, I \circ w + E) + \alpha||E||_0$

End While

Output: solution $\hat{\boldsymbol{w}}, \hat{\boldsymbol{E}}$ to problem (3)

In Algorithm 1, $\delta(E)(\boldsymbol{x}) = \begin{cases} 1, E(\boldsymbol{x}) = 0 \\ 0, otherwise \end{cases}$. $\delta(E)$ is used as a binary mask for valid pixels. $T_{\delta(E)}$ denotes the set of valid pixels obtained by applying the mask $\delta(E)$ to image T. The basic intuition in Step 1 is that, after we estimate the large magnitude sparse differences E, we should not consider the corresponding pixels when we update w.

In the initialization step, we use some distance function $dist_0(,)$ which is relatively robust to sparse large errors (e.g. L1-norm). One may choose $dist_0(,)$ to be different from $dist(,)$.

C. Determination of α

The parameter α balances between the value of $dist(T, I \circ w + E)$ and the sparsity level of the large errors E. If α is too large, no pixels will be considered as large sparse errors. If α is too small, many pixels will be considered as large "sparse" errors. The parameter α should depend on the sparsity level k and the data. The sparsity level k is equal to the percentage d% of non-zero entries in E times the total number of pixels. The parameter d% could be specified by the user or set to a nominal value (e.g. 0.1%).

From Step 2 of Algorithm 1, if the contribution of some pixel $\boldsymbol{x}$ to the overall distance $dist(T, I \circ w)$ is larger than α, this pixel will be considered as a sparse large error.

From the analysis above, based on the initial transformation $w^{(0)}$, we order the distance value contributed by each pixel in $dist\left(T, I \circ w^{(0)}\right)$. Then we set α_0 as the d%th largest distance value. We set $\alpha = \max(\alpha_0, c)$ (c depends on the distance function we use) to ensure α is not too small. This will also improve the robustness of our method in case there is no large sparse error at all.

Hence, no more than d% pixels will be considered as large sparse errors at the beginning. During the iteration, some other pixels may be reconsidered as sparse large errors, while some sparse large error pixels may be reconsidered as normal pixels.

D. Case Study

For the intensity-based image registration, one could use the common Lucas-Kanade algorithm [10, 13]. In [10], the transformation is modeled as a parameterized set of allowed warps $W(\boldsymbol{x}; \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, \ldots, p_n)^{\boldsymbol{T}}$ is a vector of transformation parameters. The warping function $W(\boldsymbol{x}; \boldsymbol{p})$ warps pixel location $\boldsymbol{x}$ in the fixed image T to the sub-pixel location $W(\boldsymbol{x}; \boldsymbol{p})$ in the moving image I.

Combining with the MAD distance function, the objective function in (3) now becomes:

$$J(\boldsymbol{p}, E) = \frac{1}{m \times n} \sum_{\boldsymbol{x}} \left|T(\boldsymbol{x}) - I\left(W(\boldsymbol{x}; \boldsymbol{p})\right) - E(\boldsymbol{x})\right| + \alpha\left||E|\right|_0 \quad (4)$$

This specific objective function is still non-convex. The distance function here is not differentiable for the gradient descent approach. We approximate the absolute value $|t|$ by $\sqrt{t^2 + \epsilon}$ as in [14], where ϵ is set to 10^{-5}.

The overall procedure to minimize this objective function is summarized in Algorithm 2. Please refer to the Appendix for the derivation of Step 1.

Algorithm 2 (Case Solution)

Input: image T, image I, α

Initialization:

$\boldsymbol{p}^{(0)}$ as the solution of $\min_{\boldsymbol{p}} \frac{1}{m \times n} \sum_{\boldsymbol{x}} \left|T(\boldsymbol{x}) - I\left(W(\boldsymbol{x}; \boldsymbol{p})\right)\right|$;

$$E^{(0)} = \underset{E}{\operatorname{argmin}} \frac{1}{m \times n} \sum_{\boldsymbol{x}} \left|T(\boldsymbol{x}) - I\left(W\left(\boldsymbol{x}; \boldsymbol{p}^{(0)}\right)\right) - E(\boldsymbol{x})\right| + \alpha||E||_0$$

While $J(\boldsymbol{p}, E)$ not converged DO:

Step 1: fix E, update $\boldsymbol{p} \leftarrow \boldsymbol{p} + \Delta\boldsymbol{p}$, where

$$\Delta\boldsymbol{p} = H^{-1}\sum_{\boldsymbol{x}\in\{\boldsymbol{x}|\boldsymbol{E}(\boldsymbol{x})=\boldsymbol{0}\}} R'(F(\boldsymbol{x})^2)\nabla I\frac{\partial W}{\partial \boldsymbol{p}}F(\boldsymbol{x})$$

$$F(\boldsymbol{x}) = T(\boldsymbol{x}) - I\big(W(\boldsymbol{x};\boldsymbol{p})\big)$$

$$H = \sum_{\boldsymbol{x}\in\{\boldsymbol{x}|\boldsymbol{E}(\boldsymbol{x})=\boldsymbol{0}\}} R'(F(\boldsymbol{x})^2)\left[\nabla I\frac{\partial W}{\partial \boldsymbol{p}}\right]^T\left[\nabla I\frac{\partial W}{\partial \boldsymbol{p}}\right]$$

$$R'(F(\boldsymbol{x})^2) = \frac{1}{2}(\epsilon + F(\boldsymbol{x})^2)^{-\frac{1}{2}}$$

Step 2: fix $\boldsymbol{p}$, update $E = \underset{E}{\operatorname{argmin}}\frac{1}{m\times n}\sum_{\boldsymbol{x}}\left|T(\boldsymbol{x}) - I\big(W(\boldsymbol{x};\boldsymbol{p})\big) - E(\boldsymbol{x})\right| + \alpha||E||_0$

End While

Output: solution $\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{E}}$ that minimize (4)

Setting α for this objective function is straightforward. Based on the initial transformation $W\big(\boldsymbol{x};\boldsymbol{p}^{(0)}\big)$, we order the distance value contributed by each pixel $\boldsymbol{x}$: $\frac{1}{m\times n}\left|T(\boldsymbol{x}) - I\left(W\big(\boldsymbol{x};\boldsymbol{p}^{(0)}\big)\right)\right|$. Then we set α_0 as d%th largest distance value. We set $\alpha = \max(\alpha_0, c)$ to ensure α is not too small. If we have normalized the intensity value of the image to [0,1], the maximum distance value one pixel can contribute to $\frac{1}{m\times n}\sum_{\boldsymbol{x}}\left|T(\boldsymbol{x}) - I\big(W(\boldsymbol{x};\boldsymbol{p})\big)\right|$ is $\frac{1}{m\times n}$. We set $c = 0.1\times\frac{1}{m\times n}$.

After α is set, it will be fixed. In the iteration process, updating E in Step 2 is also straightforward. If $\frac{1}{m\times n}\left|T(\boldsymbol{x}) - I\big(W(\boldsymbol{x};\boldsymbol{p})\big)\right|$ (the contribution of pixel $\boldsymbol{x}$) is larger than α, this pixel will be considered as a sparse large error, and we set $E(\boldsymbol{x}) = T(\boldsymbol{x}) - I\big(W(\boldsymbol{x};\boldsymbol{p})\big)$.

III. Application & Evaluation

In this section, we apply the proposed method to stabilize time-lapse imaging videos of subcellular dynamics in the *C. elegans* epidermis *in vivo*. *C. elegans* is a genetic model organism widely used by biologists to investigate questions in numerous fields, including development, cell biology, neurobiology, immunity, and other aspects of basic science and disease. Because of the large genetic toolkit available to manipulate gene expression and function, and the transparency of the animal, *C. elegans* has provided a great opportunity for scientists to study cell biological processes in a multicellular organism *in vivo*. The continual advancement in live microscopy techniques and development of fluorescent proteins have further enhanced the popularity to use *C. elegans* to study subcellular components and their dynamics, such as chromatin structure, organization of organelles, intracellular transport, and the cytoskeleton. The nematode epidermis is a multinucleate syncytium, making it amenable for imaging 3D and 4D dynamics of various organelles and macromolecules. We expressed different proteins tagged to green fluorescent protein GFP exclusively in the epidermis, and visualized their localization and dynamics using a spinning disk confocal microscope. There are various kinds of sparse large errors inside the video, such as the very bright but small moving puncta (spots) in the epidermal cytoplasm, and the changes due to wounding.

In Section III.A, we introduce the strategy to pre-initialize the transformation before the gradient descent based image registration. In Section III.B, we do simulations and comparisons to show the advantage of our approach. In Section III.C, we demonstrate the effectiveness of our method on real data, comparing with some state-of-the-art methods.

A. Pre-initialization of the Transformation

In every video, certain regions of the epidermis (e.g. organelles and other cell types not expressing the fluorescent proteins and thus appearing as silhouettes in the bright epidermal background, see red arrows in Fig.1) are relatively stable in both position and appearance for short periods of time (e.g. few seconds). The pixel intensities inside these regions are nearly uniform and different from that of the background. We use MSER (Maximally Stable Extremal Regions, [15]) to extract this kind of region. Then we adopt commonly used point-matching procedures to compute SURF [16] descriptors at these regions, and match them between the fixed and moving images. Finally, the initial transformation is estimated by MSAC [17] based on these matched feature points.

B. Simulation

In this section, we simulate the bright moving puncta as well as the wounding process on real images. The first image is the fixed image with some simulated bright puncta. To generate the moving image, we simulate some arbitrary Brownian motion (with a radius of 3 pixels) of the puncta as well as the wounding (white rectangle), and then apply a global rigid transformation with rotation angle θ and translation T_x and T_y. See Fig.1 for example.

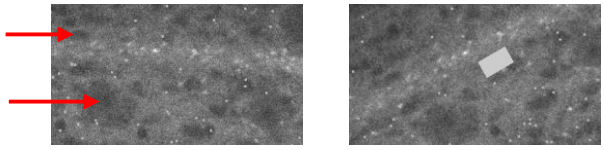


Figure 1. Example of organelles (red arrows) and simulated puncta, wounding and transformation. (left) fixed image; (right) moving image (θ=30°, T_x=40 pixels)

We simulate the above process with rotation angle θ varying from -30° to 30° (with a step size of 10°), T_x varying from -120 to 120 pixel (with a step size of 40 pixels) and T_y varying from -80 to 80 pixel (with a step size of 40 pixels). We repeat the above process on 44 different images.

We use the Inverse Compositional (IC) algorithm [10] for efficiency, which is a variant of the Lucas-Kanade algorithm [13]. We also combine it with different distance functions for comparison:

- Original distance function SSD used in IC.
- The approximated MAD.
- Geman-McLure (GM) function [5] (a robust error function): $\rho(x) = \frac{x^2}{x^2+\sigma_1^2}$

For the GM function, we set $\sigma_1 = 0.2$ instead of $1.4826\times median(\boldsymbol{x})$ [5] to get better results in this simulation.

We also compare with RIC [9]. For fair comparison, we do not estimate α from the data (otherwise our method will likely have a perfect estimation). We set $\alpha = 0.5/(m \times n)$ for our method. We set the equivalent threshold for RIC. The RIC method now ignores pixels with squared difference value greater than a certain threshold (set to 0.5^2) during the gradient descent. We also implement a variant of RIC for comparison, which uses MAD instead of the original SSD (called RIC-MAD). For RIC-MAD, we ignore pixels whose absolute difference value is greater than a certain threshold (set to 0.5) during the gradient descent.

As RASL [11] and RVR [12] require relative good initial transformation, we give them the pre-initial transformation estimated in III.A. We also compare the results of other methods by using that pre-initial transformation. The estimated angle and translation are compared with ground truth. Table 1 summarizes the average absolute angle error and translation error by each method.

Clearly, although the pre-initial transformation is not perfect, it does benefit the later method in most cases. We can also see the L1-norm based distance function (MAD) outperforms the traditional L2-norm based distance function (SSD) in our simulation. It is more robust to the sparse large errors in our simulated data.

TABLE I. AVERAGE ABSOLUTE ANGLE ERROR AND TRANSLATION ERROR OF EACH METHOD

Method	Average Angle Error (degrees)	Average Translation Error (pixel)
IC (SSD) [10]	3.31	11.1
RIC (SSD) [9]	3.11	27.5
IC-MAD	1.74	8.4
RIC-MAD	1.74	8.4
GM [10]	1.92	13.5
Pre-initial	9.73	51.7
Pre-initial →IC (SSD)	1.71	21.6
Pre-initial →RIC (SSD)	1.47	20.4
Pre-initial →IC-MAD	0.34	2.1
Pre-initial →RIC-MAD	0.34	2.1
Pre-initial →GM	0.57	8.3
Pre-initial →RASL [11]	8.75	48.6
Pre-initial →RVR [12]	8.85	49.8
Proposed (MAD)	**0.30**	**1.8**

The RIC seems slightly better than IC when using their original distance function (SSD). While using MAD, the RIC-MAD performs nearly the same as IC-MAD. GM performs better than original RIC, while worse than RIC-MAD.

The performances of RASL and RVR are worse than others. As the misalignment of the estimated pre-initial transformation is usually not very small, RASL and RVR often converge to some local minima very close to that initial point. Finally they are only slightly better than the initial transformation.

C. Results on Real Data

In this section, we test our method on 5 different videos with 25~400 frames. The first two videos (EBP-2-GFP1&2) have many bright puncta moving in the epidermal cytoplasm. EBP-2 is a worm homolog of the mammalian End Binding proteins, which bind to growing microtubule plus ends and give the appearance of flying comets in our video. The third video (GFP-Utrophin(CH)) contains images of the worm epidermis pre- and post- wounding. Utrophin is a protein which binds to the filaments of actin. We express a fragment of the protein, the calponin homology (CH) domain sufficient for actin binding, to visualize actin dynamics. The Utrophin-actin filaments can grow over time. After wounding, there is a gradual breakdown of filaments around the wound site, leaving an expanding hole in the epidermis empty of fluorescent protein. The fourth video (GFP-RAB-5) displays transport of endosomes in the epidermis. RAB-5 associates with early endosomes and is an accepted marker for these vesicles. GFP-RAB-5 forms small moving puncta, large stationary aggregates, and sometimes filaments. This video also has some artifacts during imaging. The fifth video (IRIS) is taken from [18], and shows severe eye jitter and pupil deformation. The motion blur and compression of this video make the problem even harder.

We stabilize these video sequences through frame-by-frame registration for our method, RIC [9], GM [10], and RVR [12], while RASL [11] simultaneously registers all the frames. For fair comparison, we use pre-initialization for all methods. For RIC and our method, we set the percentage of large sparse errors as 0.1% for all videos without fine tuning.

For visual comparison, we show the fixed image, registered image, and the difference between them. Table 2 has the average MSE and MAD of the residuals in overlapped regions in each video by the different methods.

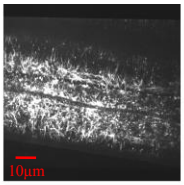
TABLE II. AVERAGE MSE/MAD OF RESIDUALS IN EACH VIDEO BY DIFFERENT METHODS

Video	RIC[9]	GM[10]	RASL[11]	RVR[12]	Ours
1	0.0061 /0.0600	0.0055 /0.0576	0.0059 /0.0600	0.0059 /0.0596	**0.0055 /0.0575**
2	0.0099 /0.0780	0.0094 /0.0760	0.0094 /0.0760	0.0098 /0.0777	**0.0094 /0.0759**
3	0.0377 /0.1175	0.0394 /0.1210	0.0435 /0.1284	0.0558 /0.1517	**0.0372 /0.1163**
4	**0.0029 /0.0290**	0.0040 /0.0324	0.0053 /0.0367	0.0181 /0.0770	0.0030 /0.0294
5	0.0617 /0.1926	0.0576 /0.1856	0.0705 /0.2073	0.0846 /0.2305	**0.0523 /0.1762**

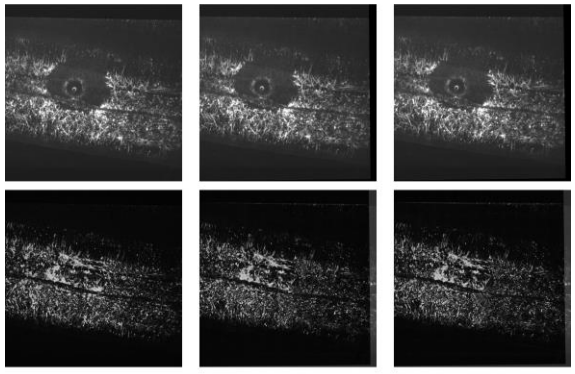
Fig. 2 shows the 1st frame and 301st frame of the GFP-Utrophin(CH) video before and after registration by different methods. The worm is wounded during this sequence and the appearance of the wound continues to change. Besides that, the filaments also have some complicated movements. Although the residuals of different methods are still high, they are often smaller than the original residual without registration. Our method robustly corrects the global movement of the worm body and achieves the smallest residual.

Fig. 3 shows the 1st frame and 101st frame of the GFP-RAB-5 video before and after registration by different methods. There are some artifacts (horizontal lines) at the top of the image. The bottom of Fig. 3c~g shows the residual of the registered 101st frame with respect to the 1st frame by different methods. We can see both RIC and our method significantly reduce the residuals on the worm body compared to other methods (the overlapped region to compute the residual includes the artifacts).

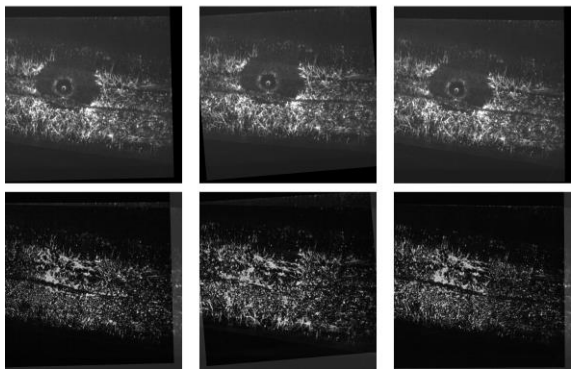
From Fig.2-Fig.3 and Table 2, we can see our method outperforms others both visually and in terms of MSE and MAD in most videos. The fifth video (results in Table 2) is a microscopic iris video, which shows the generality of our proposed method.



(a) 1st frame and scale bar (red)



MSE:0.0542 MAD:0.1475 (b) Original
MSE:0.0446 MAD:0.1308 (c) RIC
MSE:0.0448 MAD:0.1315 (d) GM

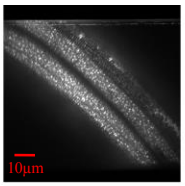


MSE:0.0464 MAD:0.1355 (e) RASL
MSE:0.0596 MAD:0.1591 (f) RVR
MSE:0.0440 MAD:0.1292 (g) Ours

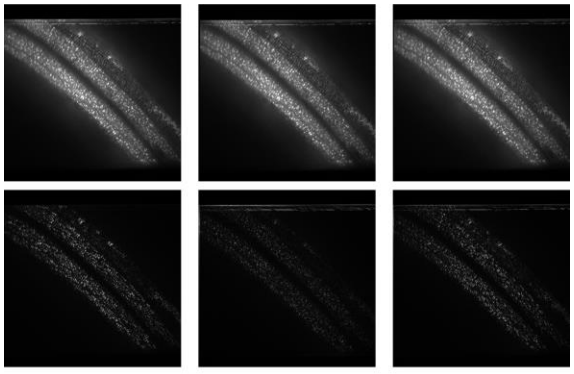
Figure 2. Example (GFP-Utrophin(CH)) registration results by different methods. (a) 1st frame; (b)~(g) 301st frame (top) and its residual (bottom) with respect to 1st frame: (b) original image (c) RIC (d) GM (e) RASL (f) RVR (g) Ours.

IV. Discussion & Conclusion

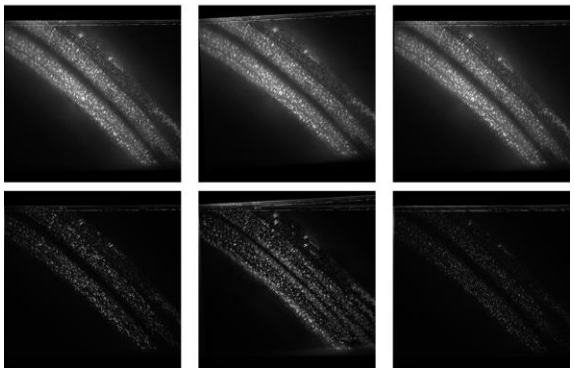
Although RIC [9] also ignores outliers (here we assume outliers are equivalent to sparse large errors) during gradient descent, it is different from ours in several aspects. First, the RIC ignores outliers in every step from the beginning. As the images may not be well aligned at the beginning, many inliers may also be considered as outliers by RIC. On the contrary, we do not reject outliers until our initialization step is done. Second, RIC always rejects a fixed percentage of pixels as outliers. Our method rejects a certain percentage of pixels at the beginning, but then the percentage of outliers will be automatically adjusted in the following steps (some former outliers may be reconsidered as inliers, and vice versa). Third, when there are no outliers, RIC will still reject a fixed percentage of pixels, so it loses information. Our method can still work well when there are no outliers (owing to the max function on α).



(a) 1st frame and scale bar (red)



MSE:0.0059 MAD:0.0391 (b) Original
MSE:0.0029 MAD:0.0292 (c) RIC
MSE:0.0047 MAD:0.0354 (d) GM



MSE:0.0054 MAD:0.0371 (e) RASL
MSE:0.0150 MAD:0.0722 (f) RVR
MSE:0.0030 MAD:0.0299 (g) Ours

Figure 3. Example (GFP-RAB-5) registration results by different methods. (a) 1st frame; (b)~(g) 101st frame (top) and its residual (bottom) with respect to 1st frame: (b) original image (c) RIC (d) GM (e) RASL (f) RVR (g) Ours.

While we utilize the same concept of sparse errors as RASL and RVR, the registration framework and overall formulation are quite different. Our framework can better handle various kinds of additive noise in real data and can deal with much larger misalignment.

Also, we explicitly minimize the L0-norm of the sparsity term instead of the extensively used L1-norm (e.g. [19] directly uses the L1-norm to constrain the sparsity of the boundary outside the template during image segmentation).

With the objective function in (4), where the distance function is based on the L1-norm, if we also use the L1-norm for sparse large errors E, all pixels will be considered as E in the optimal solution (as long as noise exists and $\alpha < \frac{1}{m \times n}$) ! This shows the advantage of using the original L0-norm in the objective function.

Section II.D shows a special case of our framework. Our framework could incorporate other deformable transformations (e.g. Free-Form Deformations [20]) as well as other distance functions/ similarity measures. Our method is not limited to biomedical images/video. It can be applied to various kinds of images/video with sparse large errors, such as field-sports video with moving players as in [12].

APPENDIX

Derivation [10, 21] of Step 1 in Algorithm 2: In Step 1 we want to minimize the following (ignore the constant) with respect to $\boldsymbol{p}$:

$$\sum_{\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}\}} |T(\boldsymbol{x}) - I(W(\boldsymbol{x}; \boldsymbol{p}))| \approx \sum_{\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}\}} R\left(\left(T(\boldsymbol{x}) - I(W(\boldsymbol{x}; \boldsymbol{p}))\right)^2\right) \quad (5)$$

where $R(t) = \sqrt{t + \epsilon}$. To minimize the expression in (5), the Lucas-Kanade algorithm assumes that a current estimate of $\boldsymbol{p}$ is known and solves for update $\Delta \boldsymbol{p}$. Thus the algorithm tries to minimize the following expression with respect to $\Delta \boldsymbol{p}$:

$$\sum_{\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}\}} R\left(\left(T(\boldsymbol{x}) - I(W(\boldsymbol{x}; \boldsymbol{p} + \Delta \boldsymbol{p}))\right)^2\right) \quad (6)$$

Approximating $I(W(\boldsymbol{x}; \boldsymbol{p} + \Delta \boldsymbol{p}))$ by first order Taylor expansion gives:

$$\sum_{\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}\}} R\left(\left(T(\boldsymbol{x}) - I(W(\boldsymbol{x}; \boldsymbol{p})) - \nabla I \frac{\partial W}{\partial \boldsymbol{p}} \Delta \boldsymbol{p}\right)^2\right) \quad (7)$$

where ∇I is the gradient of image I evaluated at $W(\boldsymbol{x}; \boldsymbol{p})$. $\frac{\partial W}{\partial \boldsymbol{p}}$ is the Jacobian of the warp.

Let $F(\boldsymbol{x}) = T(\boldsymbol{x}) - I(W(\boldsymbol{x}; \boldsymbol{p}))$, expanding (7) gives:

$$\sum_{\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}\}} R\left(F(\boldsymbol{x})^2 - 2F(\boldsymbol{x}) \nabla I \frac{\partial W}{\partial \boldsymbol{p}} \Delta \boldsymbol{p} + \Delta \boldsymbol{p}^T \left[\nabla I \frac{\partial W}{\partial \boldsymbol{p}}\right]^T \left[\nabla I \frac{\partial W}{\partial \boldsymbol{p}}\right] \Delta \boldsymbol{p}\right) \quad (8)$$

Performing a first order Taylor expansion gives:

$$\sum_{\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}\}} \{R(F(\boldsymbol{x})^2) + R'(F(\boldsymbol{x})^2)\left(-2F(\boldsymbol{x}) \nabla I \frac{\partial W}{\partial \boldsymbol{p}} \Delta \boldsymbol{p} + \Delta \boldsymbol{p}^T \left[\nabla I \frac{\partial W}{\partial \boldsymbol{p}}\right]^T \left[\nabla I \frac{\partial W}{\partial \boldsymbol{p}}\right] \Delta \boldsymbol{p}\right)\} \quad (9)$$

The closed-form solution is:

$$\Delta \boldsymbol{p} = H^{-1} \sum_{\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}\}} R'(F(\boldsymbol{x})^2) \nabla I \frac{\partial W}{\partial \boldsymbol{p}} F(\boldsymbol{x}) \quad (10)$$

where: $H = \sum_{\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}\}} R'(F(\boldsymbol{x})^2) \left[\nabla I \frac{\partial W}{\partial \boldsymbol{p}}\right]^T \left[\nabla I \frac{\partial W}{\partial \boldsymbol{p}}\right]$ (11)

$$R'(F(\boldsymbol{x})^2) = \frac{1}{2}(\epsilon + F(\boldsymbol{x})^2)^{-\frac{1}{2}} \quad (12)$$

REFERENCES

[1] S. Kaneko, I. Murase, and S. Igarashi, "Robust image registration by increment sign correlation," *Pattern Recognition,* vol. 35, pp. 2223-2234, 2002.

[2] S. Kaneko, Y. Satoh, and S. Igarashi, "Using selective correlation coefficient for robust image registration," *Pattern Recognition,* vol. 36, pp. 1165-1173, 2003.

[3] J. Kim and J. A. Fessler, "Intensity-based image registration using robust correlation coefficients," *IEEE Transactions on Medical Imaging,* vol. 23, pp. 1430-1444, 2004.

[4] K. V. Arya, P. Gupta, P. K. Kalra, and P. Mitra, "Image registration using robust M-estimators," *Pattern Recognition Letters,* vol. 28, pp. 1957-1968, 2007.

[5] P. J. Huber, *Robust Statistics.* 1981.

[6] M. J. Black and A. D. Jepson, "EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation," *International Journal of Computer Vision,* vol. 26, pp. 63-84, 1998.

[7] H. S. Sawhney and S. Ayer, "Compact representations of videos through dominant and multiple motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, pp. 814-830, 1996.

[8] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. on Pattern Anal. and Machine Intelligence,* vol. 20, pp. 1025-1039, 1998.

[9] T. Ishikawa, I. Matthews, and S. Baker, "Efficient Image Alignment with Outlier Rejection," Technical Report CMU-RI-TR-02-27, Robotics Institute, Carnegie Mellon University, 2002.

[10] S. Baker and I. Matthews, "Lucas-Kanade 20 Years On: A Unifying Framework," *Intl. J. of Computer Vision,* vol. 56, pp. 221-255, 2004.

[11] 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 Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 2233-2246, 2012.

[12] B. Ghanem, T. Zhang, and N. Ahuja, "Robust Video Registration Applied to Field-Sports Video Analysis," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2012.

[13] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the Intl. Joint Conference on Artificial Intelligence,* 1981, pp. 674-679.

[14] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, "High Accuracy Optical Flow Estimation Based on a Theory for Warping," in *Proceedings of the European Conference on Computer Vision,* 2004, pp. 25-36.

[15] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proceedings of British Machine Vision Conference,* 2002, pp. 384-396.

[16] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding,* vol. 110, pp. 346-359, 2008.

[17] P. H. S. Torr and A. Zisserman, "MLESAC: A New Robust Estimator with Application to Estimating Image Geometry," *Computer Vision and Image Understanding,* vol. 78, pp. 138-156, 2000.

[18] A. Myronenko and X. Song, "Intensity-Based Image Registration by Minimizing Residual Complexity," *IEEE Transactions on Medical Imaging,* vol. 29, pp. 1882-1891, 2010.

[19] P. Shah and M. D. Gupta, "Simultaneous registration and segmentation by L1 minimization," *Machine Learning in Medical Imaging,* pp. 128-135, 2012.

[20] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging,* vol. 18, pp. 712-721, 1999.

[21] S. Baker, R. Gross, T. Ishikawa, and I. Matthews, "Lucas-Kanade 20 Years On: A Unifying Framework: Part 2," Technical Report CMU-RI-TR-03-01, Robotics Institute, Carnegie Mellon University, 2003.

[22] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," IEEE International Symposium on Information Theory Proceedings, 2010, pp. 1518-1522.