# Multi-modal and Multi-spectral Registration for Natural Images

Xiaoyong Shen[1], Li Xu[2], Qi Zhang[1], and Jiaya Jia[1]

[1] The Chinese University of Hong Kong, China
[2] Image & Visual Computing Lab, Lenovo R&T,
Project Website, Hong Kong, China
`http://www.cse.cuhk.edu.hk/leojia/projects/multimodal`

**Abstract.** Images now come in different forms – color, near-infrared, depth, etc. – due to the development of special and powerful cameras in computer vision and computational photography. Their cross-modal correspondence establishment is however left behind. We address this challenging dense matching problem considering structure variation possibly existing in these image sets and introduce new model and solution. Our main contribution includes designing the descriptor named robust selective normalized cross correlation (RSNCC) to establish dense pixel correspondence in input images and proposing its mathematical parameterization to make optimization tractable. A computationally robust framework including global and local matching phases is also established. We build a multi-modal dataset including natural images with labeled sparse correspondence. Our method will benefit image and vision applications that require accurate image alignment.

**Keywords:** multi-modal, multi-spectral, dense matching, variational model.

## 1 Introduction

Data captured in various domains, such as RGB and near-infrared (NIR) image pairs [35], flash and no-flash images [22,1], color and dark flash images [18], depth and color images, noisy and blurred images [38], and images captured under changing light [24], are used commonly now in computer vision and computational photography research. They are multi-modal or multi-spectral data generally involving natural images. Although there are rigid and nonrigid methods developed for multi-modal medical image registration [23,21,17,2]. In computer vision, quite a few prior methods still assume already aligned input images, making them readily usable in applications to generate new effects.

For example, the inputs in [18,22,1,38,26,8] are produced from the same or calibrated cameras. The dynamic scene images used in [24] are aligned before HDR construction. In [35], a multi-spectral image restoration method was developed based on correctly relating pixels. It is clear when alignment is not a satisfied condition in prior, registering input images considering camera motion,

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 309–324, 2014.

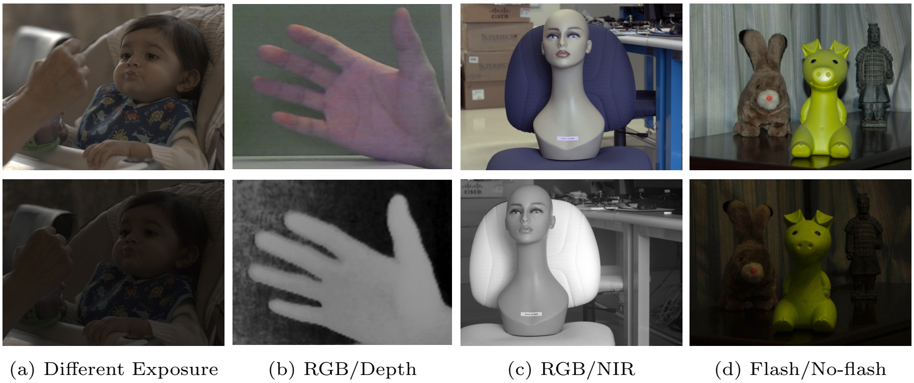


**Fig. 1.** Multi-modal images that need alignment. (a) Images from [24] captured under different exposure settings in dynamic scene. (b) RGB and depth image pair. (c) RGB and NIR images. (d) Flash and no-flash images captured at different time.

object deformation, and depth variation will be inevitable. It is challenging when large intensity, color, and gradient variation presents.

For images taken continuously from nearby cameras, or containing similar structure, state-of-the-art matching methods such as nonrigid image registration [7,36,16,29], optical flow estimation [13,5,6,41,5,3], and stereo matching [11,27] can help align them. But multi-modal images, like those in Fig. 1, cannot be easily dealt with. Color, gradient, and even structure similarity, which are commonly considered to establish constraints, are not applicable anymore, as detailed later in this paper. Moreover, the image pairs shown in Fig. 1 are with nonrigid displacement due to depth variation and dynamic moving objects, which makes matching very difficult.

In medical imaging, multi-modal registration methods are based on global or local statistic information like mutual information to search for region correspondence. They are mostly limited to gray level medical images and do not suit rich-detail natural image matching. For general multi-spectral image matching, Irani et al. [15] proposed a framework for multi-sensor image global alignment. Cross correlation on the directional Laplacian energy map was used to measure patch similarity. Variational frameworks ([12] and [37]) can estimate small displacements in multi-modal images. These methods do not work similarly well on heavy outlier images or those with large nonrigid displacement. General matching tools, such as SIFT flow [19], also do not handle multi-spectral images and lack sub-pixel accuracy in computation.

We aim to match general multi-modal and multi-spectral images with significant displacement and obvious structure inconsistency. We analyze and compare possible measures, and propose a new matching cost, named robust selective normalized cross correlation (RSNCC), to handle gradient and color variation, and possible structure divergence caused by noise, inconsistent shadow and reflection from object surface. In solution establishment, we provide new parameterization to separate the original descriptor into a few mathematically meaningful terms that explain optimality. Our method contains global and local phases to remove large displacements and estimates residual pixel-wise correspondence respectively. To verify our system, we build a dataset containing different kinds of image pairs with labeled point correspondence.

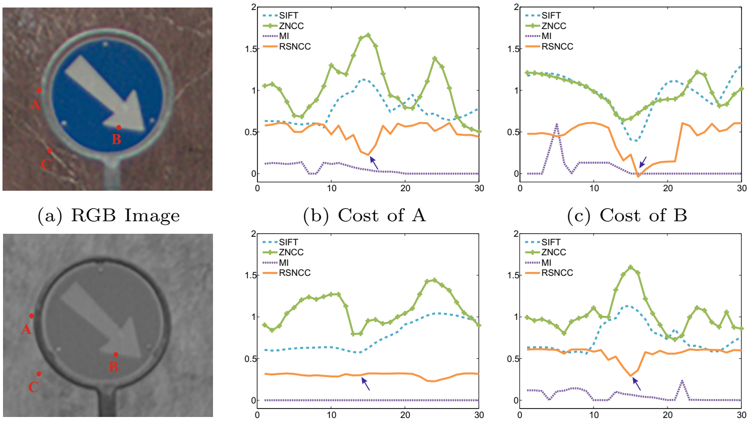
## 2 Related Work

Surveys of image matching were provided in [42,28,33]. We review in this paper related image registration methods and variational optical flow estimation.

The correspondence of images captured by different modalities is complex. The difference between multi-spectral images was analyzed in [15,35]. We coarsely categorize previous work into feature-based and patch-based methods. The feature-based methods extract multi-spectral invariant sparse feature points and then establish their correspondence for optimal transform. Hrkac et al. [14] aligned visible and infrared images by extracting corner points and getting the global correspondence via minimizing Hausdorff distance. Firmenichy et al. [9] proposed a multi-spectral interest points detection algorithm for global registration. Han et al. [10] used hybrid visual features like lines and corners to align visible and infrared images captured in controlled environment. These methods do not aim at very accurate dense matching due to feature sparseness.

Several methods employed local patch similarity to find correspondence. The effective measures include mutual information and cross correlation. Mutual information is robust for multi-modal medical image alignment, as surveyed in [23]. Hermosillo et al. [12] proposed a variational framework to match multi-modal images based on this measure. Zhang et al. [40] and Palos et al. [21] further enhanced the variational framework to solve the multi-modal registration problem. Yi et al. [37] adaptively considered global and local mutual information. As for cross correlation methods, Irani et al. [15] proposed the Laplacian energy map and computed cross correlation on it to measure multi-sensor image similarity. Cross correlation of gradient magnitude was used by Kolar et al. [17] to register autofluorescent and infrared retinal images. Recently, Andronache et al. [2] combined mutual information and cross correlation to match the multi-modal images. These measures are effective, but sometimes still suffer from outlier and large displacement influence during dense matching.

Our framework is related to modern optical flow estimation [13]. In modern methods, the data term usually enforces brightness or gradient constancy [5,6,41]. Robust functions, such as $L_1$ norm and Charbonnier function, were used by [5,3,31,39] in regularization. For large displacement handling, Xu et al. [34] improved the coarse-to-fine strategy by supplementing feature- and patch-based matching. We note optical flow methods cannot solve our problem since it relies on the brightness and gradient constancy constraints, which no longer hold for multi-spectral image matching. Based on the variational framework, Liu et al. [19] achieved general scene image matching using SIFT features.



(a) RGB Image (b) Cost of A (c) Cost of B

(d) NIR Image (e) Cost of C (f) Cost of A with Noise

**Fig. 2.** Matching cost comparison. (a) and (d) are the RGB and NIR images presented in [4]. Points $A$, $B$ and $C$ are inconsistent on structure/gradient. $A$ is with gradient reverse; $B$ has gradient magnitude variation; and $C$ is with gradient loss. (b), (c) and (e) are the matching costs under different descriptors along $A$, $B$ and $C$'s scanlines. (f) is the matching cost of $A$ with added noise on (a) and (d). The arrows point to the ground-truth matching points.

## 3 Problem Understanding

Images from different sensors are ubiquitous, as shown in Fig. 1. Their matching is thus a fundamental problem. We in what follows take the RGB and NIR image pairs as examples as they contain many different structures and intensity levels. We analyze the difficulties in dense image matching.

Let $I_1$ and $I_2$ be the two multi-spectral or multi-modal images, $p = (x, y)^T$ be pixel coordinates of the two images, and $w_p = (u_p, v_p)^T$ be the displacement of pixel $p$, which indicates $p$ in $I_1$ mapping to $p + w_p$ in $I_2$. $I_{1,p}$ and $I_{2,p}$ are the intensities (or color vectors) of $I_1$ and $I_2$ for pixel $p$ respectively.

For dense image matching, the cost for pixel $p$ between two input images can be generally expressed as

$$E^{\mathcal{D}}(p, w_p) = \mathrm{dist}\big(\mathcal{D}_1(p), \mathcal{D}_2(p + w_p)\big), \tag{1}$$

where $\mathcal{D}_1(p)$ and $\mathcal{D}_2(p + w_p)$ are matching descriptors for pixels $p$ and $p + w_p$ in $I_1$ and $I_2$ respectively. $\mathrm{dist}(\cdot)$ is a function to measure the descriptor distance.

**Color and Gradient.** As shown in Fig. 2(a) and (d), an RGB/NIR image pair captured by visual and NIR cameras contains structure inconsistency. Obviously, general color and gradient constancy between corresponding pixels that was used in many alignment methods under the Euler or robust Euler distance cannot be employed. Irani et al. [15] and Kolar et al. [17] computed similarity on gradient magnitude. Although it relaxes the color constancy condition, it is still not enough in many cases. Matching accuracy could reduce when only using the gradient correspondence.

**SIFT Features.** Another common type of matching costs are based on SIFT descriptors [20] that work well for images captured under similar exposures. We note SIFT may not be appropriate for multi-spectral matching with the following two reasons. First, SIFT is not invariant to gradient reversal existing in input images, as shown at point $A$ in Fig. 2(a) and (d). Although Firmenichy et al. [9] proposed gradient direction invariant SIFT, the performance is reduced compared with traditional SIFT. In (c), the minimum of SIFT descriptor difference does not correspond to the ground truth matching point. Second, SIFT descriptor is not that powerful to differentiate between true and false correspondences especially in featureless regions given its output scores.

**Mutual Information.** Mutual information (MI) is used popularly in medical image registration. However, for natural image with rich details, MI has its limitation. As shown in Fig. 2, the cost of MI in the 15 × 15 patch fails to find the correct correspondence. MI may also be sensitive to noise as shown in Fig. 2(f). The drawback of MI to measure small local patch similarity was explained by Andronache et al. [2]. For the variational frameworks [12,37] using local patch mutual information, only small displacements are computed.

# 4 Our Matching Cost

In order to handle structure inconsistency and notable gradient variation in multi-spectral and multi-modal images, we propose a matching cost given by

$$E^{\mathrm{RSNCC}}(p, w_p) = \rho\big(1 - |\Phi_I(p, w_p)|\big) + \tau\rho\big(1 - |\Phi_{\nabla I}(p, w_p)|\big). \tag{2}$$

This function is a robust selective normalized cross correlation (RSNCC) addressing a few of the concerns presented above. $\rho(x)$ is a robust function and weight $\tau$ is used to combine two terms defined respectively on color and gradient domains. We present details as follows.

## 4.1 *Φ* Definition

$\Phi_I(p, w_p)$ is the normalized cross correlation between the patch centered at $p$ in $I_1$ and patch $p + w_p$ in $I_2$ in the intensity or color space. $\Phi_{\nabla I}(p, w_p)$ is the one defined similarly in the gradient space. This definition is also extendible to other definitions. By generalizing $I$ and $\nabla I$ as feature $F \in \{I, \nabla I\}$, $\Phi_F(p, w_p)$ in feature space $F$ is given by

$$\Phi_F(p, w_p) = \frac{(F_{1,p} - \overline{F}_{1,p}) \cdot (F_{2,p+w_p} - \overline{F}_{2,p+w_p})}{\|F_{1,p} - \overline{F}_{1,p}\| \|F_{2,p+w_p} - \overline{F}_{2,p+w_p}\|}, \tag{3}$$

where $F_{1,p}$ and $F_{2,p}$ are pixels' feature vectors in patch $p$ in $I_1$ and patch $p+w_p$ in $I_2$ respectively. $\overline{F}_{1,p}$ and $\overline{F}_{2,p+w_p}$ are the means of $F_{1,p}$ and $F_{2,p+w_p}$ respectively.

The normalized cross correlation defined in Eq. (3) can represent structure similarity of the two patches under feature $F$ even if the two patches are transformed in color and geometry locally.

**Difference from Other Definitions.** Our cost definition in Eq. (2) has a robust function $\rho(x)$. It handles transform more complex than a linear one defined only using Pearson's distance $1 - \Phi_I(p, w_p)$.

In addition, our data cost models the absolute value of $\Phi_F(p, w_p)$ that minimizes the matching cost on either positive or negative correlation in Eq. (2), which is the major difference compared to other matching methods only working on similar-appearance natural images. This definition is effective to handle gradient reversal ubiquitous for NIR-RGB and positive-negative images, which produce negative correlation. This is why we call it a selective model.

An example is shown in Fig. 2(b) where point $A$ is with different gradient directions in the input images. Even in this challenging local correspondence problem that was seldom studied in previous work in natural image matching, optimizing our function can lead to reasonable results.

**Color and Gradient.** The combination of $\Phi_I(p, w_p)$ and $\Phi_{\nabla I}(p, w_p)$ is helpful to improve the stability in matching especially when intensity or color of the two patches differs a lot. For instance, point $B$ in Fig. 2(a) and (d) is with different edge magnitudes in the corresponding patches. Our method can find the correspondence while zero-mean normalized cross-correlation (ZNCC), SIFT and MI fail, as shown in (c). In addition, the combination makes matching more robust to noise, which is shown in Fig. 2(f) with more explanations in our experiment section.

However, the matching cost we defined is complex with respect to $w_p$. We linearize it by a two-order approximation. To achieve this, a robust function is carefully chosen and per-pixel Taylor expansion is employed.

### 4.2 Robust Function

$\rho(x)$ in Eq. (2) is a robust function to help reject outliers. The outliers include structure divergence caused by shadow, highlight, dynamic objects, to name a few. We show one example in Fig. 3.

$\rho(x)$ should also be robust to errors generated in $1 - |\Phi_F(p, w_p)|$, which is not continuous. This makes general robust functions, such as Charbonnier, not differentiable. To address this issue, we propose $\rho(x)$ as

$$\rho(x) = -\frac{1}{\beta}\log(e^{-\beta|x|} + e^{-\beta(2-|x|)}), \tag{4}$$

where $\beta$ is a parameter. To understand this function, we plot $\rho(x)$ and $\rho'(x)$ by varying $\beta$ in Fig. 4. A large $x$ does not cause an excessive penalty in $\rho(x)$. Note when $\beta \to \infty$, $\rho(x)$ becomes a nice approximation of the robust $L_1$ norm. Besides, it makes RSNCC continuous and solvable by continuous optimization.

This robust function is effective in image matching. For the inconsistent shadow structure in Fig. 3, our model handles it better than direct matching.

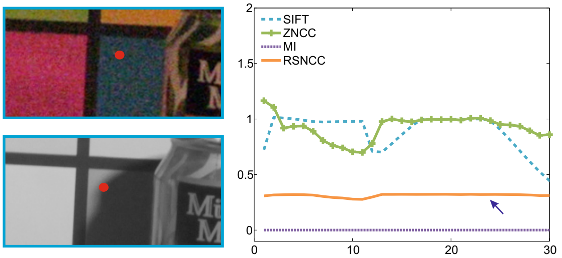
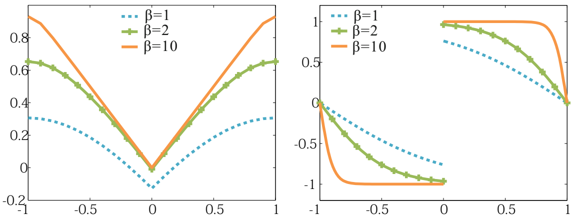


**Fig. 3.** Outlier example. The left two patches contain shadow only in one input. It should be regarded as an outlier in matching. The plots of matching costs in a scanline show that our method can safely ignore this outlier.



(a) $\rho(x)$ (b) $\rho'(x)$

**Fig. 4.** Robust function with different $\beta$.

### 4.3 Matching Cost and Derivation

After setting $\rho(x)$ as Eq. (4), the matching cost (2) is written as

$$E^{\mathrm{RSNCC}}(p, w_p) = -\frac{1}{\beta}\log\left(e^{-\beta(1-\Phi_I(p,w_p))} + e^{-\beta(1+\Phi_I(p,w_p))}\right) \\ -\frac{\tau}{\beta}\log\left(e^{-\beta(1-\Phi_{\nabla I}(p,w_p))} + e^{-\beta(1+\Phi_{\nabla I}(p,w_p))}\right). \tag{5}$$

In addition, the term $\Phi_F(p, w_p)$, which is the patch normalized cross correlation between $I_1$ and $I_2$ according to the feature space $F \in \{I, \nabla I\}$, is highly non-convex. We decompose it by linearization in Taylor expansion, which yields

$$\Phi_F(p, w_p + \delta w_p) \approx \Phi_F(p, w_p) + (\mathbf{A}_p^F)^T \delta\mathbf{w}_p + \frac{1}{2}\delta\mathbf{w}_p^T \mathbf{B}_p^F \delta\mathbf{w}_p, \tag{6}$$

where $\delta\mathbf{w}_p$ is the vector form of all $\delta w_p$ of patch $p$. $\mathbf{A}_p^F$ is the first-order approximation coefficient matrix and $\mathbf{B}_p^F$ is the second-order matrix that only includes diagonal elements. In this expansion and Eq. (5), local displacement field updated in iterations for patch $p$ is expressed as

$$\min\left((\omega_p^I \mathbf{A}_p^I + \omega_p^{\nabla I}\mathbf{A}_p^{\nabla I})^T \delta\mathbf{w}_p + \frac{1}{2}\delta\mathbf{w}_p^T(\omega_p^I \mathbf{B}_p^I + \omega_p^{\nabla I}\mathbf{B}_p^{\nabla I})\delta\mathbf{w}_p\right), \tag{7}$$

where $\omega_p^I$ and $\omega_p^{\nabla I}$ are weights coming from the derivative robust function. That is, $\omega_p^I = \rho'(1 - |\Phi_I(p, w_p)|)$ and $\omega_p^{\nabla I} = \tau\rho'(1 - |\Phi_{\nabla I}(p, w_p)|)$. $\mathbf{A}_p^F$ is given by

$$\mathbf{A}_p^F = \begin{bmatrix} \mathbf{S}_p^1 \circ \nabla_x \mathbf{F}_{2,p+w_p} \\ \mathbf{S}_p^1 \circ \nabla_y \mathbf{F}_{2,p+w_p} \end{bmatrix} \mathbf{1}, \tag{8}$$

and $\mathbf{B}_p^F = \mathrm{diag}(\widehat{\mathbf{B}}_p)$. $\widehat{\mathbf{B}}_p$ is in the following form:

$$\widehat{\mathbf{B}}_p = \begin{bmatrix} \mathbf{S}_p^1 \circ \nabla_x^2 \mathbf{F}_{2,p+w_p} + \mathbf{S}_p^2 \circ (\nabla_x \mathbf{F}_{2,p+w_p})^2 \\ \mathbf{S}_p^1 \circ \nabla_y^2 \mathbf{F}_{2,p+w_p} + \mathbf{S}_p^2 \circ (\nabla_y \mathbf{F}_{2,p+w_p})^2 \end{bmatrix} \mathbf{1}, \tag{9}$$

where $\circ$ represents element-wise multiplication, $\mathbf{F}_{1,p}$ is the updated $F_{1,p}$ denoted in Eq. (3), with each row being a feature vector for the pixel in patch $p$. $\mathbf{F}_{2,p+w_p}$ is defined similarly. $\mathbf{1}$ is an all-one vector whose length is the dimension of feature space $F$. $\nabla_x$ is an element-wise difference operator in $x$-direction and $\nabla_x^2$ is the second order one. $\nabla_y$ and $\nabla_y^2$ are corresponding operators in $y$-direction.

We denote $\mathbf{S}_p^1$ as the *first-order* normalized similarity and $\mathbf{S}_p^2$ as the *second-order* one. We explain their construction and effect.

**$\mathbf{S}_p^1$ and $\mathbf{S}_p^2$.** The matching cost defined in Eq. (3) comprises two parts. The *similarity* measure is given by

$$\mathcal{S}_p = (F_{1,p} - \overline{F}_{1,p}) \cdot (F_{2,p+w_p} - \overline{F}_{2,p+w_p}), \tag{10}$$

and the *confidence* term contains

$$\mathcal{C}_{1,p} = \|F_{1,p} - \overline{F}_{1,p}\|, \quad \mathcal{C}_{2,p} = \|F_{2,p+w_p} - \overline{F}_{2,p+w_p}\|. \tag{11}$$

Now coming to the definition of normalized similarity in two orders, $\mathbf{S}_p^1$ describes the confidence of matching for each pixel under the normalized cross correlation descriptor. It is normalized by the *similarity* and *confidence* as

$$\mathbf{S}_p^1 = \frac{1}{\mathcal{C}_{1,p}\mathcal{C}_{2,p}} \big(\mathbf{F}_{1,p} - \overline{F}_{1,p} - \frac{\mathcal{S}_p}{\mathcal{C}_{2,p}^2}(\mathbf{F}_{2,p+w_p} - \overline{F}_{2,p+w_p})\big). \tag{12}$$

To get $\mathbf{S}_p^2$, we first denote *normalized cross similarity* as

$$\mathbf{C}_p = \frac{1}{\mathcal{C}_{1,p}\mathcal{C}_{2,p}} \frac{(\mathbf{F}_{1,p} - \overline{F}_{1,p}) \circ (\mathbf{F}_{2,p+w_p} - \overline{F}_{2,p+w_p})}{\mathcal{C}_{2,p}^2}, \tag{13}$$

which describes correlation of the two patches. Given the two-order normalized descriptor of $\mathbf{F}_{2,p+w_p}$ as

$$\mathbf{D}_{p,2} = \frac{1}{\mathcal{C}_{1,p}\mathcal{C}_{2,p}} \frac{\mathcal{S}_p(\mathbf{F}_{2,p+w_p} - \overline{F}_{2,p+w_p})^2}{\mathcal{C}_{2,p}^4}, \tag{14}$$

$\mathbf{S}_p^2$ becomes the linear combination of $\mathbf{C}_p$ and $\mathbf{D}_{p,2}$ as

$$\mathbf{S}_p^2 = 3\mathbf{D}_{p,2} - 2\mathbf{C}_p - \frac{1}{\mathcal{C}_{1,p}\mathcal{C}_{2,p}} \frac{\mathcal{S}_p(\overline{N} - 1)^2}{\mathcal{C}_{2,p}^2 \overline{N}^2}, \tag{15}$$

where $\overline{N}$ is the number of pixels in the patch. The last (third) term is a bias imposed by the different similarity of the two patches.

Note that our two-order approximation is different from the form in [32], where the latter handles similar-exposure natural images for motion estimation and assumes that the displacement field is constant locally. Our approximation is pixel-wise with new expressions, thus modeling complex correspondence in multi-spectral and multi-modal images.

## 5 Matching Framework

To produce matching on challenging images, our solver contains phases for global transform and local dense matching respectively. Global matching estimates large position transform caused by camera position variation or scene motion. Then the local phase estimates residual errors and compensates them considering pixel-wise correspondence.

### 5.1 Global Matching

The global phase estimates a homography matrix $H$ for image-wise translation, rotation and scaling. The corresponding function is written as

$$E(H) = \sum_p E^{\mathrm{RSNCC}}(p, w_p), \tag{16}$$

where $w_p = (u_p, v_p)$ is under the homography constraint for every pixel. It is further expressed as

$$[u_p, v_p, 1]^T = [x_p, y_p, 1]^T (H - I)^T, \tag{17}$$

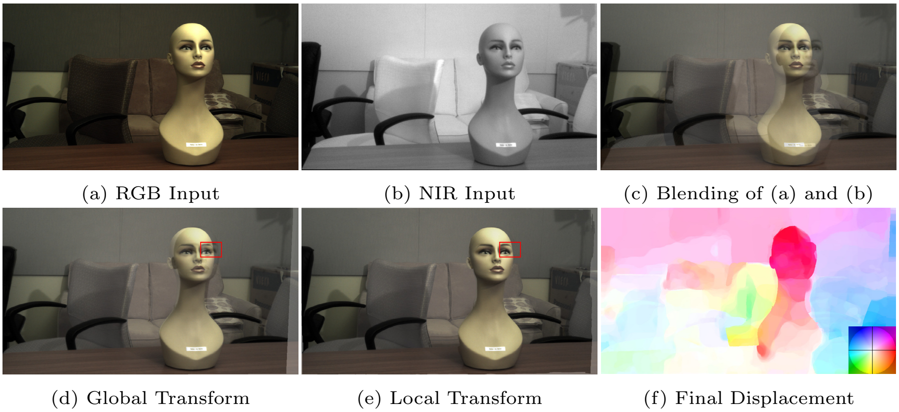
where $I$ is the identity matrix. We apply gradient decent to get optimal $H$. The first and second order derivatives of $E(H)$ are obtained following the chain rule. For quick and robust computation, we employ the coarse-to-fine scheme and estimate $H$ increment in each layer.

The RSNCC matching cost used here can robustly find similar structures and reject outliers. As shown in Fig. 5, our method estimates background transform despite large structure inconsistency in shadow and noise. Due to depth variation, a few pixels in Fig. 5(d) still contain errors. They are further refined in what follows.

### 5.2 Local Dense Matching

After global transform, we perform pixel-wise residual displacement estimation incorporating regularization terms. The function is written as

$$E(\mathbf{w}) = \sum_p E^{\mathrm{RSNCC}}(p, w_p) + \lambda_1 \sum_p \psi(\|\nabla w_p\|^2) + \lambda_2 \sum_{q \in N(p)} \|w_p - w_q\|, \tag{18}$$


(a) RGB Input

(b) NIR Input

(c) Blending of (a) and (b)

(d) Global Transform

(e) Local Transform

(f) Final Displacement

**Fig. 5.** Two-phase matching. (a) and (b) RGB/NIR pair. (c) Blending result. (d) Blending result of (a) and the globally transformed (b). (e) Blending result of (a) and locally transformed (b). (f) Final displacement estimate from (b) to (a) coded in color. Structures are aligned.

where $\mathbf{w} = (\mathbf{u}^T, \mathbf{v}^T)^T$ is the vector form of $w_p$. $\mathbf{u}$ and $\mathbf{v}$ are vectors of $u_p$ and $v_p$ respectively. For simplicity's sake, we denote the three terms as $E_D(\mathbf{w})$, $E_S(\mathbf{w})$, and $E_{NL}(\mathbf{w})$. $\lambda_1$ and $\lambda_2$ are two parameters.

The robust regularization term $E_S(\mathbf{w})$ is common for enforcing spatial smoothing. $\psi(x)$ is the robust penalty function in the Charbonnier form $\psi(x^2) = \sqrt{x^2 + \epsilon^2}$ with $\epsilon$ setting to $1E-4$ in all our experiments. This function is a differentiable variant of $L_1$ norm, availing optimization. $E_{NL}(\mathbf{w})$ is a nonlocal median filter. It can efficiently remove noise, as described in [25].

**Optimization.** Local dense matching is performed in a coarse-to-fine manner for high accuracy to optimize $E(\mathbf{w})$. In each level, $E(\mathbf{w})$ is updated and propagated to the next level for variable initialization. To handle the non-convex $E(\mathbf{w})$ in each level, we decompose it into two sub-functions both finding optimal solutions by the scheme of variable-splitting [30]. The two functions are

$$E(\mathbf{w}, \widehat{\mathbf{w}}) = E_D(\mathbf{w}) + \lambda_1 E_S(\mathbf{w}) + \frac{1}{\theta}\|\mathbf{w} - \widehat{\mathbf{w}}\|^2, \tag{19}$$

$$E(\widehat{\mathbf{w}}, \mathbf{w}) = \frac{1}{\theta}\|\widehat{\mathbf{w}} - \mathbf{w}\|^2 + \lambda_2 E_{NL}(\widehat{\mathbf{w}}), \tag{20}$$

where $\widehat{\mathbf{w}}$ is an auxiliary variable. When $\theta \to 0$, the decomposition approaches the original $E(\mathbf{w})$.

Our method minimizes Eqs. (19) and (20) respectively. The minimum of Eq. (20) can be obtained by the method of [25]. We solve Eq. (19) based on the variational configuration using iterative reweighted least squares. In each step, we update the result by a small $\delta\mathbf{w}$ after optimizing $E(\mathbf{w} + \delta\mathbf{w}, \widehat{\mathbf{w}})$. It is done by setting $\frac{\partial E(\mathbf{w}+\delta\mathbf{w}, \widehat{\mathbf{w}})}{\partial \delta\mathbf{w}} = 0$. Details are provided in the our project website (link in the title page). Our local matching improves pixel-wise alignment, as illustrated in Fig. 5(e) and (f).

## 6 Experiments and Evaluation

We implement our algorithm in MATLAB. The processing time of a 1200 × 800 image is less than three minutes on a 3.2GHz Core i7 PC. In our experiments, we set $\beta = 1.0$ in the robust function and patch size $9 \times 9$ to compute RSNCC. The weight $\tau$ is set to 1.0 in all our experiments. In local dense matching, $\lambda_1$ ranges from 0.1 to 0.5 and $\lambda_2$ is set to 0.01. In both the global and local matching phases, we employ five scales with down-sampling rate 0.8 during the coarse-to-fine optimization. More details are provided in our project website.

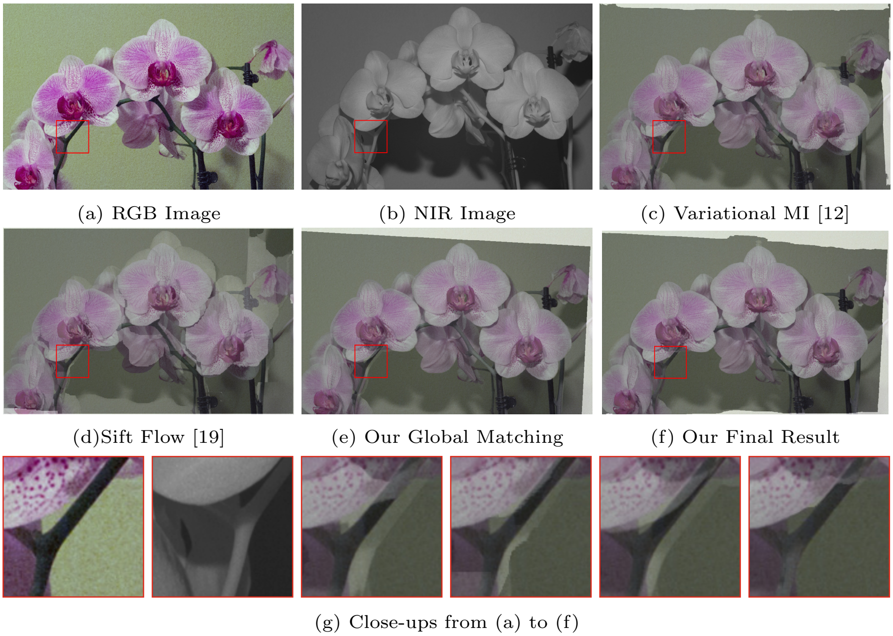
### 6.1 Evaluation

We build a dataset including four typical kinds of image pairs – RGB/NIR, RGB/Depth, different exposure, and flash/no-flash. The RGB/NIR images are captured by RGB and NIR cameras while the RGB/Depth images are captured by Microsoft Kinect. The different exposure image pairs and the flash/no-flash pairs are captured by the same camera with exposure and camera pose variation. These images contain depth variation or dynamic moving objects, needing rigid and nonrigid transformation estimation. To get the ground truth displacement, we select 100 corner points and label their correspondence. The images are shown in our website (link in the title page). In total, these images provide $2K$ ground truth correspondences and we employ them to evaluate our method.

Evaluation of our method and other state-of-the-arts is reported in Table 1. We compare the general scene matching SIFT Flow [19] and the modified SIFT Flow using the gradient direction invariant SIFT [9]. We implement the variational mutual information method [12]. SIFT Flow does not handle gradient reversal. The gradient invariant SIFT produces a level of errors for matching as well. Variational mutual information does not handle large displacement and correspondingly yields relatively large errors. Our method does not have these problems. As our matching cost is flexible to incorporate other features, we evaluate employing features proposed in [15] and of color, gradient, and the

**Table 1.** Evaluation of methods on our dataset. The quantities in each column are the mean errors on one labeled image pair. * denotes the SIFT implementation in [9].

| | RGB/NIR | RGB/Depth | Flash/No-flash | Different Exposure | All |
|---|---|---|---|---|---|
| SIFT Flow [19] | 10.11 | 18.32 | 8.76 | 10.03 | 11.47 |
| SIFT Flow* | 8.03 | 16.17 | 8.90 | 11.67 | 10.56 |
| Method by [12] | 12.03 | 15.19 | 16.57 | 13.24 | 13.81 |
| Ours with [15] | 2.34 | 4.57 | 6.87 | 3.68 | 3.96 |
| Ours with color | 2.55 | 4.83 | 6.64 | 3.43 | 4.00 |
| Ours with gradient | 2.28 | 4.46 | 6.03 | 3.02 | 3.61 |
| **Ours** | **1.89** | **4.17** | **4.56** | **2.25** | **2.95** |


(a) RGB Image

(b) NIR Image

(c) Variational MI [12]

(d)Sift Flow [19]

(e) Our Global Matching

(f) Our Final Result

(g) Close-ups from (a) to (f)

**Fig. 6.** Matching of RGB and NIR images with structure variation and nonrigid transform by different methods. (c)-(f) Blending result by warping (b) to (a). (g) Close-ups.

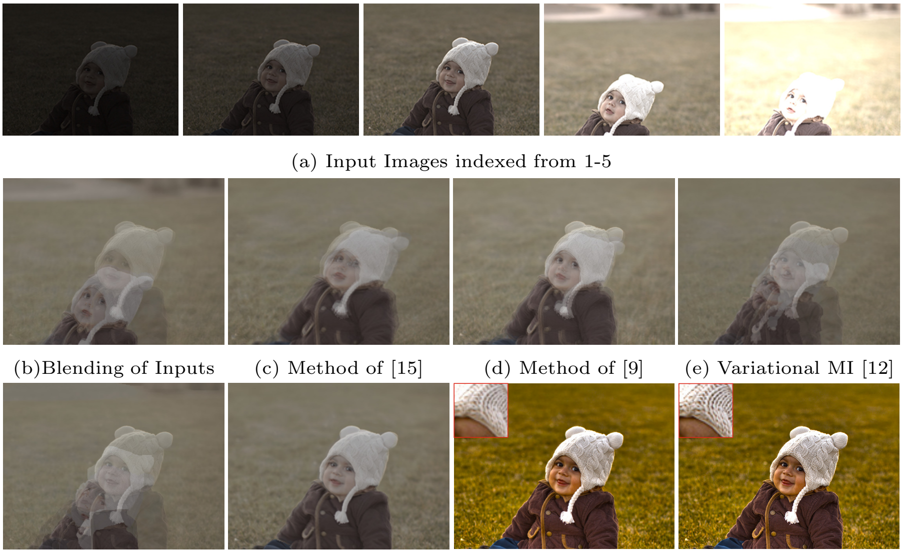
combination of color and gradient. The result in Table 1 proves that our current feature is the best among the four for this matching task.

Examples in Figs. 6 and 7 compare our method to others. The inputs in Fig. 6 are the RGB and NIR images with significant gradient, noise and shadow variation. Fig. 7 is an example to match a series of different exposure images. Both examples are with nonrigid transform and large displacements. Our results are with good quality thanks to the new matching cost and robust two-phase matching framework.

## 6.2 Applications

Our framework benefits computer vision and computational photography applications that need to align multi-spectral and multi-modal images. We apply it to HDR construction and multi-modal image restoration.

**HDR Image Construction.** Our method can match different exposure images for restoration of high dynamic range images. As shown in Fig. 7, Our results are with high quality. We employ the method proposed in [24] to merge low dynamic range images into a HDR one, where the tone mapping result is shown in (i). Our method yields rich details compared to that of [24].


(a) Input Images indexed from 1-5

(b)Blending of Inputs (c) Method of [15] (d) Method of [9] (e) Variational MI [12]

(f) Sift Flow [19] (g) Ours (h) HDR by [24] (i) Our HDR

**Fig. 7.** Different exposure images matching and HDR construction example. (a) Inputs from [24]. (b) Blending result of the inputs. (c)-(g) Blending results by warping all images to image 3 by different methods. (h) HDR Tone mapping result of [24]. (i) Our tone mapping result using the matched (g). More results and complete images are contained in our website.

**Multi-Modal Image Restoration.** We show an example of depth and RGB image matching in Fig. 8. Depth images captured by Kinect or other equipments are not accurately aligned with the corresponding RGB images as shown in Fig. 8(c). The depth image is also with noise and missing values. Simple smoothing by filter might damage original structures. Our method matches the smoothed depth image to the RGB one. It not only aligns structure but also helps restore it damaged by filtering as shown in Fig. 8(d) and (e).

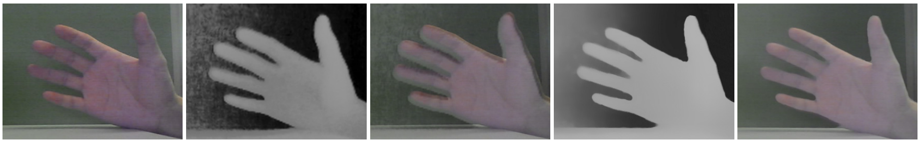
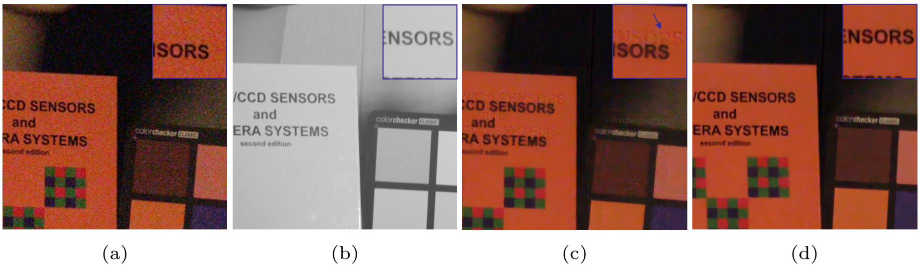

(a) (b) (c) (d) (e)

**Fig. 8.** RGB and depth matching and restoration. (a)-(b) are the RGB and depth raw data captured by Microsoft Kinect. (c) shows the blending result. (d) is the result using the framework of [35] but applying our matching method. (e) shows the blending of our depth and RGB images, which are aligned well.



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

**Fig. 9.** Multi-spectral image restoration. (a) and (b) are input noisy RGB image and NIR image with displacements. (c) and (d) are the restoration results without and with matching respectively.

NIR image is also a good guidance to restore noisy RGB image as described in [35]. Since RGB and NIR images are often captured by different cameras, they need to be aligned before restoration. The alignment is very challenging due to their nonrigid transformation. Our method handles this problem, and produces the result shown in Fig. 9.

Our matching framework can also be employed to enhance flash/no-flash images that require alignment. Several examples are contained in the project website.

## 7 Conclusion and Limitation

We have presented an effective dense matching framework for multi-spectral and multi-modal images. Unlike other methods working on natural or medical images under various constraints, we address more challenging issues, including structure inconsistency and existence of strong outliers caused by shadow and highlight. We proposed a robust matching scheme, optimized in two phases.

Our method inevitably has several limitations. First, if the two images contain quite different structures, the estimated displacement field could be wrong completely. Second, our method may cause large errors on regions that do not contain necessarily informative edges or textures for credible structure matching.

**Acknowledgements.** The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. 412911).

## References

1. Agrawal, A.K., Raskar, R., Nayar, S.K., Li, Y.: Removing photography artifacts using gradient projection and flash-exposure sampling. ToG 24(3), 828–835 (2005)
2. Andronache, A., von Siebenthal, M., Székely, G., Cattin, P.C.: Non-rigid registration of multi-modal images using both mutual information and cross-correlation. Medical Image Analysis 12(1), 3–15 (2008)

3. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. CVIU 63(1), 75–104 (1996)
4. Brown, M., Susstrunk, S.: Multi-spectral sift for scene category recognition. In: CVPR, pp. 177–184 (2011)
5. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J.(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)
6. Bruhn, A., Weickert, J.: Towards ultimate motion estimation: Combining highest accuracy with real-time performance. In: ICCV, pp. 749–755 (2005)
7. Chui, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration. Computer Vision and Image Understanding 89(2), 114–141 (2003)
8. Fattal, R., Lischinski, D., Werman, M.: Gradient domain high dynamic range compression. ToG 21(3), 249–256 (2002)
9. Firmenichy, D., Brown, M., Süsstrunk, S.: Multispectral interest points for rgb-nir image registration. In: ICIP, pp. 181–184 (2011)
10. Han, J., Pauwels, E.J., de Zeeuw, P.M.: Visible and infrared image registration in man-made environments employing hybrid visual features. Pattern Recognition Letters 34(1), 42–51 (2013)
11. Heo, Y.S., Lee, K.M., Lee, S.U.: Robust stereo matching using adaptive normalized cross-correlation. PAMI 33(4), 807–822 (2011)
12. Hermosillo, G., Chefd'Hotel, C., Faugeras, O.D.: Variational methods for multimodal image matching. IJCV 50(3), 329–343 (2002)
13. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. 17(1-3), 185–203 (1981)
14. Hrkać, T., Kalafatić, Z., Krapac, J.: Infrared-visual image registration based on corners and hausdorff distance. In: Ersbøll, B.K., Pedersen, K.S. (eds.) SCIA 2007. LNCS, vol. 4522, pp. 383–392. Springer, Heidelberg (2007)
15. Irani, M., Anandan, P.: Robust multi-sensor image alignment. In: ICCV, pp. 959–966 (1998)
16. Jian, B., Vemuri, B.C.: Robust point set registration using gaussian mixture models. PAMI 33(8), 1633–1645 (2011)
17. Kolár, R., Kubecka, L., Jan, J.: Registration and fusion of the autofluorescent and infrared retinal images. International Journal of Biomedical Imaging (2008)
18. Krishnan, D., Fergus, R.: Dark flash photography. ToG 28(3) (2009)
19. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: Sift flow: Dense correspondence across different scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 28–42. Springer, Heidelberg (2008)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
21. Palos, G., Betrouni, N., Coulanges, M., Vermandel, M., Devlaminck, V., Rousseau, J.: Multimodal matching by maximisation of mutual information and optical flow technique. In: IEEE International Conference on Engineering in Medicine and Biology Society, pp. 1679–1682 (2004)
22. Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M.F., Hoppe, H., Toyama, K.: Digital photography with flash and no-flash image pairs. ToG 23(3), 664–672 (2004)
23. Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: Mutual information based registration of medical images: A survey. IEEE Transaction on Medical Imaging 22(8), 986–1004 (2003)
24. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based hdr reconstruction of dynamic scenes. ToG 31(6), 203 (2012)

25. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: CVPR, pp. 2432–2439 (2010)
26. Sun, J., Kang, S.B., Xu, Z., Tang, X., Shum, H.Y.: Flash cut: Foreground extraction with flash and no-flash image pairs. In: CVPR (2007)
27. Sun, J., Zheng, N., Shum, H.Y.: Stereo matching using belief propagation. PAMI 25(7), 787–800 (2003)
28. Szeliski, R.: Image alignment and stitching: A tutorial. Foundations and Trends in Computer Graphics Vision 2(1), 1–104 (2006)
29. Tsin, Y., Kanade, T.: A correlation-based approach to robust point set registration. In: Pajdla, T., Matas, J.(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 558–569. Springer, Heidelberg (2004)
30. Wang, Y., Yang, J., Yin, W., Zhang, Y.: A new alternating minimization algorithm for total variation image reconstruction. SIAM Journal on Imaging Sciences 1(3), 248–272 (2008)
31. Wedel, A., Rabe, C., Vaudrey, T., Brox, T., Franke, U., Cremers, D.: Efficient dense scene flow from sparse or dense stereo data. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 739–751. Springer, Heidelberg (2008)
32. Werlberger, M., Pock, T., Bischof, H.: Motion estimation with non-local total variation regularization. In: CVPR, pp. 2464–2471 (2010)
33. Xiong, Z., Zhang, Y.: A critical review of image registration methods. International Journal of Image and Data Fusion 1(2), 137–158 (2010)
34. Xu, L., Jia, J., Matsushita, Y.: Motion detail preserving optical flow estimation. PAMI 34(9), 1744–1757 (2012)
35. Yan, Q., Shen, X., Xu, L., Zhuo, S., Zhang, X., Shen, L., Jia, J.: Cross-field joint image restoration via scale map. In: ICCV (2013)
36. Yang, J., Blum, R.S., Williams, J.P., Sun, Y., Xu, C.: Non-rigid image registration using geometric features and local salient region features. In: CVPR, pp. 825–832 (2006)
37. Yi, Z., Soatto, S.: Nonrigid registration combining global and local statistics. In: CVPR (2009)
38. Yuan, L., Sun, J., Quan, L., Shum, H.Y.: Image deblurring with blurred/noisy image pairs. ToG 26(3) (2007)
39. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. Pattern Recognition, 214–223 (2007)
40. Zhang, Z., Jiang, Y., Tsui, H.: Consistent multi-modal non-rigid registration based on a variational approach. Pattern Recognition Letters 27(7), 715–725 (2006)
41. Zimmer, H., Bruhn, A., Weickert, J., Valgaerts, L., Salgado, A., Rosenhahn, B., Seidel, H.-P.: Complementary optic flow. In: Cremers, D., Boykov, Y., Blake, A., Schmidt, F.R. (eds.) EMMCVPR 2009. LNCS, vol. 5681, pp. 207–220. Springer, Heidelberg (2009)
42. Zitová, B., Flusser, J.: Image registration methods: a survey. Image and Vision Computing 21(11), 977–1000 (2003)