Pattern Recognition 106 (2020) 107377

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Non-rigid infrared and visible image registration by enhanced affine transformation

Chaobo Min[a,*], Yan Gu[b], Yingjie Li[c], Feng Yang[b]

[a] College of Internet of Things Engineering, HoHai University, Changzhou, 213000, China
[b] North Night Vision Technology Co., Ltd, Nanjing, 211106, China
[c] North Information Control Research Academy Group Co., Ltd, Nanjing, 211153, China

ARTICLE INFO

Article history:
Received 20 August 2019
Revised 30 March 2020
Accepted 12 April 2020
Available online 11 May 2020

Keywords:
Registration
Non-rigid transformation
Infrared image
Image fusion

ABSTRACT

Image registration is a prerequisite for infrared (IR) and visible (VIS) image fusion. In practical application, most scenes are not planar and there is significant distinctness between IR and VIS cameras. Therefore, for non-rigid IR and VIS image registration, non-linear transformation is more applicable than affine transformation. Typically, non-linear transformation is modeled with point feature. However, this can degrade the generalization ability of transformation model and increase computational complexity. Aim at this problem, we propose an enhanced affine transformation (EAT) for non-rigid IR and VIS image registration. In this paper, image registration is transformed into point set registration and then the optimal EAT model constructed by global deformation is estimated from local feature. At first, a Gaussian-fields-based objective function is established and simplified by using the potential correspondence between an image pair. With the combination of affine and polynomial transformation, the EAT model is then proposed to describe the regular pattern of non-rigid and global deformation between an image pair. Finally, a coarse-to-fine strategy based on quasi-Newton method is designed and applied to determine the optimal transformation coefficients from edge point feature of IR and VIS images, in order to accomplish non-rigid image registration. The qualitative and quantitative comparisons on synthesized point sets and real images demonstrate that the proposed method is superior over the state-of-the-art methods in the accuracy and efficiency of image registration.

1. Introduction

Multi-sensor image fusion is widely applied and researched in recent years, including pixel-level fusion [1,2], feature-level fusion [3,4] and decision-level fusion [5,6]. The complementary information can offer more varied and comprehensive scene representations so that it is very helpful for human visual perception, object detection, tracking and recognition. However, due to the complementary of multi-sensor images, there is less mutual information between them, such as infrared (IR) and visible (VIS) images, resulting in multi-sensor image registration being a challenging task [7,8]. Successful image fusion must be on the basis of image pairs correctly aligned pixel by pixel. Therefore, IR and VIS image registration is the primary focus of this work.

It should be pointed out that the IR images involved in this work are thermal IR images. Theoretically, image registration can be implemented by the common-path optical system based on dichroic mirrors [9] or catadioptric imaging [10]. However, owing to elevated cost and low feasibility, it is difficult to use the common-path optical system for IR and VIS image registration. At present, an economical and practical optical structure with parallel optical axes is widely applied in IR and VIS image fusion systems, as shown in Fig. 1. It does not require any complex optical system, but software-based image registration algorithm must be introduced to eliminate the displacement between IR and VIS images. Therefore, we mainly study software-based image registration method in this paper.

1.1. Related work

Image registration has been widely applied in many fields including machine vision [11], medical imaging [12], remote sensing [13] and military reconnaissance [14]. However, IR and VIS image registration is still not straightforward because of the different spectral response of the sensors. IR cameras are sensitive to IR radiation, while VIS cameras (using CCD or CMOS) capture reflected light. The texture in VIS images may generally disappear in the

* Corresponding author.
E-mail address: chaobomin@outlook.com (C. Min).

https://doi.org/10.1016/j.patcog.2020.107377
0031-3203/

Fig. 1. The examples of IR and VIS image fusion systems.

corresponding IR images and vice versa, so that IR and VIS image registration is a type of multimodal registration.

In many survey papers [15–17], multimodal registration methods were generally classified into two categories: intensity-based method and feature-based method. However, we introduce the related work in another way. In recent years, many approaches prefer to formulate the registration problem as an optimization problem. The popular principle is that a transformation model and an optimization procedure are applied to obtain the best registration result by minimizing an objective function that can measure registration accuracy. Hence, key points many registration methods focus on are as follows: transformation model, optimization procedure and objective function. The most important thing is how to get the best transformation coefficients. Thus, transformation model is an essential component of image registration.

Affine transformation is a fundamental and widely used linear transformation model. Various registration algorithms have good performance with it. G.A. Bilodeau et al. [18] constructed the objective function by using the trajectories of moving objects in IR and VIS videos to measure registration accuracy. Random sample consensus (RANSAC) was then employed to determine the optimal coefficients of affine transformation. In [19], the correspondence point pairs extracted from IR and VIS images by visual saliency and RANSAC-based optimization procedure were applied to estimate the affine transformation model coefficients. Xiangzeng Liu et al. [20] proposed an affine and contrast invariant descriptor, which was deemed to an objective function to identify the correspondence point pairs by maximally stable phase congruency and RANSAC-based optimization procedure. However, affine transformation cannot produce accurate alignment when there exists the anisotropy of deformation between point pairs in numerous applications, especially in multi-sensor image fusion. To address this problem, many registration methods with non-linear transformation have been proposed in recent years.

Thin-plate spline (TPS) is a good non-linear transformation model [21]. It is commonly applied for representing flexible coordinate transformations. There are many methods which achieve non-rigid image registration and point matching via TPS. In [22], an objective function of fuzzy linear assignment-least squares was proposed to search the optimal TPS transformation coefficients between two point sets through a deterministic annealing process. Changcai Yang et al. [23] proposed the adaptive weighted objective function based on mixture model and used expectation-maximization (EM) algorithm to estimate the TPS model. In [24], an interpolation penalty function based on speed-up robust features was developed to determine the optimal coefficients of the TPS model.

Because B-splines are able to preserve the smoothness of non-linear deformation, they have been commonly utilized to construct non-linear transformation model. Suicheng Gu et al. [25] proposed the B-spline affine transformation and used the iterative closest point (ICP) method to achieve the registration of three-dimensional (3D) volumetric computed tomography (CT) data. In [26], the improved B-spline transformation model was constructed by control point parameterization and then, gradient optimization was applied to obtain the optimal B-spline basis coefficients. Wei Sun et al. [27] utilized the lower-order B-spline basis functions and the random perturbation technique for efficient registration.

Jiayi Ma et al. [28] proposed a non-linear transformation model within a vector-valued reproducing kernel Hilbert space (RKHS). The Gaussian-fields-based objective function, which measures the accuracy of image registration by the edge maps of an IR and VIS face image pair, was employed to determine the optimal transformation coefficients by the quasi-Newton method based on deterministic annealing. This method was applied successfully to VIS and IR face registration. Then, on the basis of the non-rigid transformation model within RKHS, some methods of image registration or point matching have been put forward. In [29], the L_2-minimizing estimator was presented to estimate the RKHS transformation coefficients by building robust sparse and dense correspondences between point sets. [30] introduced an objective function established by Gaussian Mixture Models (GMMs) to measure the registration performance under different transformation coefficients and employed EM algorithm to find the optimal solution.

Free-form deformation (FFD) model is able to deform an object by manipulating an underlying mesh of control points. In [31], the similarity measure of plant silhouettes extracted from IR and VIS images was minimized by the limited memory BGFS algorithm to get the optimal FFD coefficients.

There appears much significant difference between IR and VIS images in terms of principle of imaging, specification of sensor, optical system, etc. Moreover, most scenes are not planar in practical application such as target tracking and military reconnaissance. In this case, non-linear transformation is more applicable for IR and VIS image registration than rigid or affine transformation. As mentioned above, there exist many non-linear transformation methods for multimodal image registration. However, the non-linear transformation models used in these methods depend heavily on point feature. For instance, the TPS model is written as,

$$\mathbf{f}(\mathbf{x},\ \mathbf{d},\ \mathbf{w}) = \mathbf{x}\mathbf{d} + \mathbf{\Phi}(\mathbf{x})\mathbf{w}\ , \tag{1}$$

where $\mathbf{x}$ is certain point in a point set (or an image), $\mathbf{d}$ is the affine transformation, $\mathbf{w}$ is a warping coefficient matrix representing the non-affine transformation, each entry of TPS kernel $\mathbf{\Phi}(\mathbf{x})$ is defined by $\mathbf{\Phi}(\mathbf{x}) = \mathbf{x} - \mathbf{x_n}^2 \log \mathbf{x} - \mathbf{x_n}$, $\mathbf{x_n}$ can be considered as a control point selected from an image. Another example is non-rigid transformation model within RKHS, which is given by

$$\mathbf{f}(\mathbf{x}) = \sum_{n=1}^{N_0} \Gamma(\mathbf{x},\ \mathbf{x}_n)\mathbf{c}_n, \tag{2}$$

where $\Gamma(\ \cdot\ )$ is a diagonal Gaussian kernel, $\mathbf{c}_n$ is a transformation coefficient vector related to a control point $\mathbf{x}_n$, N_0 is the number of control points.

It can be seen from the above that the non-linear transformation models, which are commonly used to existing approaches, describe non-rigid deformation between point pairs by using local

feature in the neighborhoods of control points. Because transformation coefficients are mainly optimized in the neighborhoods of control points, the mapping of every point in an image is determined by nearby control points. Thus, the registration accuracy of the points which are relatively far away from control points may be degraded greatly. In other words, transformation models with control points rely heavily on local feature so that their generalization ability is not strong. Meanwhile, since the displacement vector of a point needs to be calculated by both transformation coefficients as well as feature point sets, the performances of the methods using non-linear transformation models with control points are limited by low efficiency and poor applicability for practical application.

1.2. Our contributions

At present, an optical structure with parallel optical axes is widely used in IR and VIS image fusion systems, as shown in Fig. 1. In this case, the displacement between pixels in image pairs results from the differences between IR and VIS cameras, such as spatial position, lens distortion, pixel area, resolution of sensor, etc. These factors independently have regular effect on image deformation. In other words, the deformation between an image pair could be described by a mixture of many different regular models. Therefore, we believe that the deformation between an IR and VIS image pair should exhibit a regular pattern in global scale.

This paper develops an enhanced affine transformation (namely EAT) for non-rigid IR and VIS image registration. The proposed method transforms image registration into point set registration. Then, the optimal EAT model, which can be used to global image registration, is estimated from local feature. At first, a Gaussian-fields-based objective function is established and simplified by using the potential correspondence between IR and VIS images to be aligned. Secondly, the EAT model is proposed to describe the regular pattern of global deformation between an image pair. Finally, a coarse-to-fine strategy based on quasi-Newton method is designed and applied to determine the optimal transformation coefficients from edge point feature of IR and VIS image pairs, in order to achieve non-rigid image registration.

The primary contributions of this work are as follows:

(1) The potential correspondence extracted from an image pair is used to decrease the computational complexity of the Gaussian-fields-based objective function.
(2) The proposed EAT model is able to reduce the dependence of non-rigid image transformation on local feature and has high generalization ability.
(3) A coarse-to-fine strategy based on quasi-Newton method is designed to achieve optimal registration on global scale and increases the accuracy of image registration.

Compared with the state-of-the-art methods, our method increases the accuracy of non-rigid registration and can easily be implemented. Thus, it has capability of improving the reliability of IR and VIS image fusion systems.

The rest of the paper is organized as follows. Section 2 presents the details of our method. In Section 3, the proposed method is evaluated on synthesized point set registration and then its performance is tested on real images under various scenarios with comparisons to the state-of-the-art approaches. Finally, Section 4 presents the concluding remarks for our work.

2. Methodology

Essentially, our method can be regarded as a gradient-based optimization process for non-rigid IR and VIS image registration. Hence, this section mainly focuses on objective function, transformation model, optimization procedure and how they are applied to non-rigid image registration.

2.1. Gaussian-fields-based objective function for non-rigid registration

Given two point sets $\mathbf{U} = \{\mathbf{u}_m\}_{m=1}^{M}$ and $\mathbf{V} = \{\mathbf{v}_n\}_{n=1}^{N}$, $\mathbf{u}_m, \mathbf{v}_n \in \mathbb{R}^2$. $\mathbf{U}$ is considered as 'model' point set and $\mathbf{V}$ is 'data' point set. The purpose of registration is to align the mode point set $\mathbf{U}$ to the data point set $\mathbf{V}$.

Let a map function $\mathbf{f}$ denote the non-rigid transformation. A point $\mathbf{u}_m$ is mapped to a new location $\hat{\mathbf{u}}_m = \mathbf{f}(\mathbf{u}_m)$. Thus $\hat{\mathbf{U}} = \{\hat{\mathbf{u}}_m\}_{m=1}^{M}$ represents the result of spatial transformation on $\mathbf{U}$.

The objective function for registration is essentially a criterion that can quantify the performance of registration with much accuracy. Thus, on the basis of Gaussian fields, an objective function for non-rigid registration is established as follow,

$$\min_{\mathbf{f}} E(\mathbf{f}) = \min_{\mathbf{f}} - \sum_{m=1}^{M}\sum_{n=1}^{N} \mathbf{C}_{mn} exp\left\{-\frac{\|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_m)\|^2}{2\sigma_d^2}\right\} + \lambda S(\mathbf{f}) \tag{3}$$

where $\|\cdot\|$ denotes the L_2 norm, σ_d is a range parameter, $\mathbf{C}_{mn}$ indicates the correspondence between $\mathbf{u}_m$ and $\mathbf{v}_n$. $\mathbf{C}$ is considered as the binary correspondence matrix. If a point $\mathbf{u}_m$ corresponds to a point $\mathbf{v}_n$, $\mathbf{C}_{mn} = 0$, otherwise $\mathbf{C}_{mn} = 1$. The first term of Eq. (3) is used to measure the Euclidean distance between corresponding points. The second term is employed to prevent the non-rigid transformation function $\mathbf{f}$ from excessive changing. λ is a regularization constant that balances these two terms.

There is a problem in the Gaussian-fields-based objective function. For image registration, $\mathbf{U}$ and $\mathbf{V}$ is considered as feature point sets. The time complexity of Eq. (3) is $O(MN)$ at least. If the iteration number of optimization is I, the total time complexity is $O(MNI)$ at least. The sizes of $\mathbf{U}$ and $\mathbf{V}$ must be small enough to avoid high computational complexity for solving optimization problem. However, the global accuracy of image registration cannot be correctly measured by too few feature points. Therefore, we have to simplify the objective function to improve the performance with low computational complexity.

The correspondence matrix is very important for the objective function, because the Euclidean distances between non-corresponding points could make the measure results of $E(\mathbf{f})$ inaccurate. However, it is very difficult to obtain a real corresponding matrix unless manual annotation is used. Thus, we design a procedure to extract point sets $\mathbf{U}$ and $\mathbf{V}$ from IR and VIS images to be aligned, in order to determine the corresponding matrix with high precision.

2.2. Feature point extraction

In this work, edge map is applied as the feature. Due to computational complexity, it is impossible and unnecessary to choose all edge points as feature points. Therefore, as with the other edge-based registration method [28], our method also requires a procedure which can discretize the edge maps of IR and VIS images into two feature point sets $\mathbf{U}$ and $\mathbf{V}$. However, we hope the feature points extracted from the edge maps are able to indicate the potential correspondence between an image pair and can be used to simplify the objective function.

Shape context (SC) feature descriptor [32], which can describe the neighborhood structures of points well, is used as the attribute of edge points in this work. For an IR edge point $\mathbf{b}_i^r$ and a VIS edge point $\mathbf{b}_j^v$, the SC-based similarity measure between two points is

Algorithm 1
Feature point sets extraction from an image pair.

Input: IR image I_r, corresponding VIS image I_v, parameters g, d and c
Output: Feature point set $\{\mathbf{pb}_k^r\}_{k=1}^K$ of an IR image and feature point set $\{\mathbf{pb}_k^v\}_{k=1}^K$ of a VIS image
1 Edge maps B_r and B_v are extracted from I_r and I_v by the Canny edge detector.
2 Construct edge point sets $\{\mathbf{b}_i^r\}_{i=1}^{N_r}$ and $\{\mathbf{b}_j^v\}_{j=1}^{N_v}$ from edge maps B_r and B_v, $\mathbf{b}_i^r$, $\mathbf{b}_j^v \in \mathbb{R}^2$.
3 Initialize $\{\mathbf{pb}_k^r\}_{k=1}^{N_m}$, $\{\mathbf{pb}_k^v\}_{k=1}^{N_m}$ and $k=1$, where N_m is the minimum of N_r and N_v.
4 *for* $i = 1$ *to* N_r by g
  Initialize Sc_{min} and $\mathbf{b}_{min}^v$;
  for $j = 1$ *to* N_v by 1
  if (the Euclidean distance between $\mathbf{b}_i^r$ and $\mathbf{b}_j^v \le d$)
  Compute $\mathbf{Sc}_{ij}$ between these two edge points by Eq. (4);
  if ($\mathbf{Sc}_{ij} \le Sc_{min}$)
  $\mathbf{b}_{min}^v = \mathbf{b}_j^v$; $Sc_{min} = \mathbf{Sc}_{ij}$;
  end
  end
  end
  if ($Sc_{min} \le c$)
  $\mathbf{pb}_k^r = \mathbf{b}_i^r$; $\mathbf{pb}_k^v = \mathbf{b}_{min}^v$; $k++$;
  end
end
5 Find out repeated points from $\{\mathbf{pb}_k^v\}_{k=1}^{K'}$, where K' is the final value of k.
6 The point pairs without the minimum SC-based similarity in repeated points are deleted from $\{\mathbf{pb}_k^r\}_{k=1}^{K'}$ and $\{\mathbf{pb}_k^v\}_{k=1}^{K'}$.
7 Final feature point sets of IR and VIS images are represented as $\{\mathbf{pb}_k^r\}_{k=1}^K$ and $\{\mathbf{pb}_k^v\}_{k=1}^K$.

defined as

$$\mathbf{Sc}_{ij} = \frac{1}{2}\sum_{t=1}^{T}\frac{\left(S_t(\mathbf{b}_i^r) - S_t(\mathbf{b}_j^v)\right)^2}{\left(S_t(\mathbf{b}_i^r) + S_t(\mathbf{b}_j^v)\right)} \tag{4}$$

where $S_t(\mathbf{b}_i^r)$ and $S_t(\mathbf{b}_j^v)$ represent the T-bin normalized histogram at $\mathbf{b}_i^r$ and $\mathbf{b}_j^v$, respectively. The smaller the similarity measure between two edge points is, the more similar they are. The simple procedure for extracting feature point sets from images can be seen in Algorithm 1.

In this paper, IR images are considered as 'mode' images and VIS images are considered as 'data' images. Thus, $\mathbf{U} = \{\mathbf{u}_m\}_{m=1}^K = \{\mathbf{pb}_k^r\}_{k=1}^K$ and $\mathbf{V} = \{\mathbf{v}_n\}_{n=1}^K = \{\mathbf{pb}_k^v\}_{k=1}^K$. K is the size of $\mathbf{U}$ and $\mathbf{V}$, namely the number of feature point pairs. In Algorithm 1, parameter d determines the search range of point correspondence. Parameter c is the threshold of preliminary screening of point correspondence. Parameter g determines the sizes of feature point sets $\mathbf{U}$ and $\mathbf{V}$. The above parameters are mainly employed to adjust the time and space complexity of feature point extraction. For the reason that the points in point set $\mathbf{U}$ are corresponding with the points in $\mathbf{V}$ one by one, the correspondence matrix $\tilde{\mathbf{C}}$ between these two point sets is a $K \times K$ dimensional identity matrix.

Fig. 2 illustrates the results of feature points extraction with $g = 30$ on the six pairs of edge maps of real IR and VIS images. We see that the one-to-one matches between $\mathbf{U}$ and $\mathbf{V}$ are not completely true, but almost. As a result, $\tilde{\mathbf{C}}$ is considered to be a potential correspondence matrix. With $\tilde{\mathbf{C}}$, the objective function (3) becomes the following simplified form:

$$\min_{\mathbf{f}} \boldsymbol{E}(\mathbf{f}) = \min_{\mathbf{f}} -\sum_{m=n=1}^{K} \exp\left\{-\frac{\|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_m)\|^2}{2\sigma_d^2}\right\} + \lambda \boldsymbol{S}(\mathbf{f}) \tag{5}$$

2.3. EAT model

Affine transformation model can describe the regular pattern of linear deformation between two images. In order to identify the regular pattern of non-linear deformation, we define the map function $\mathbf{f}$ of the EAT model as follows. $\mathbf{u}_m = [x_m,\ y_m]$ is a 1×2 dimensional coordinate vector of certain point in $\mathbf{U}$. $\mathbf{f}(\mathbf{u}_m) = [\hat{x}_m,\ \hat{y}_m]$ is the coordinate vector of $\mathbf{u}_m$ after transformation. The EAT model can be written as follow,

$$\mathbf{f}(\mathbf{u}_m) = [x_m,\ y_m,\ 1]\mathbf{A}^{\mathrm{T}} + [G(x_m,\ y_m,\ \mathbf{K}),\ G(x_m,\ y_m,\ \mathbf{P})] \tag{6}$$

where the affine transformation matrix $\mathbf{A} = [s_x\cos\theta,\ -\sin\theta,\ t_x;\ \sin\theta,\ s_y\cos\theta,\ t_y]$, θ is angle of rotation, s_x and s_y are scaling coefficients, t_x and t_y are translation coefficients, $G(\cdot)$ is the non-linear part of transformation and defined as

$$G(x,\ y,\ \mathbf{R}) = \sum_{i=2}^{5}\sum_{j=0}^{i} w_i r_{i,j} x^j y^{i-j} \tag{7}$$

where $\mathbf{R}$ is a 18×1 dimensional polynomial transformation coefficient vector, $r_{i,j}$ is an element of $\mathbf{R}$ and w_i is a weight value. From Eq. (7) we can see that the non-linear part consists of quadratic, cubic, quartic and quintic polynomials. Substituting Eq. (7) into Eq. (6), the EAT model becomes the following matrix form:

$$\mathbf{f}(\mathbf{u}_m) = \mathbf{g}_m\mathbf{A}_{\mathrm{Q}} \tag{8}$$

where the 1×21 dimensional polynomial vector $\mathbf{g}_m = [x_m,\ y_m,\ 1,\ x_m^2,\ y_m^2,\ x_m y_m,\ x_m^3,\ y_m^3,\ x_m^2 y_m,\ x_m y_m^2,\ x_m^4,\ y_m^4,\ x_m^3 y_m,\ x_m^2 y_m^2,\ x_m y_m^3,\ x_m^5,\ y_m^5,\ x_m^3 y_m^2,\ x_m^2 y_m^3,\ x_m^4 y_m,\ x_m y_m^4]$, $\mathbf{A}_{\mathrm{Q}}$ is called as the EAT matrix and defined as

$$\mathbf{A}_{\mathrm{Q}} = \left[\mathbf{A}\ [|\ \mathbf{K}_w\ |\ \mathbf{P}_w]^{\mathrm{T}}\right]^{\mathrm{T}} \tag{9}$$

where $\mathbf{K}_w$ and $\mathbf{P}_w$ are the weighted polynomial transformation coefficient vectors defined as $\mathbf{K}_w = diag(\mathbf{K}^{\mathrm{T}}\mathbf{w})$ and $\mathbf{P}_w = diag(\mathbf{P}^{\mathrm{T}}\mathbf{w})$ respectively. The 1×18 dimensional vectors $\mathbf{K} = [k_{2,2},\ k_{2,0},\ k_{2,1},\ k_{3,3},\ k_{3,0},\ k_{3,2},\ k_{3,1},\ k_{4,4},\ k_{4,0},\ k_{4,3},\ k_{4,2},\ k_{4,1},\ k_{5,5},\ k_{5,0},\ k_{5,3},\ k_{5,2},\ k_{5,4},\ k_{5,1}]$ and $\mathbf{P} = [p_{2,2},\ p_{2,0},\ p_{2,1},\ p_{3,3},\ p_{3,0},\ p_{3,2},\ p_{3,1},\ p_{4,4},\ p_{4,0},\ p_{4,3},\ p_{4,2},\ p_{4,1},\ p_{5,5},\ p_{5,0},\ p_{5,3},\ p_{5,2},\ p_{5,4},\ p_{5,1}]$. The weight vector $\mathbf{w} = [w_2,\ w_2,\ w_2,\ w_3,\ w_3,\ w_3,\ w_3,\ w_4,\ w_4,\ w_4,\ w_4,\ w_4,\ w_5,\ w_5,\ w_5,\ w_5,\ w_5]$. $diag(\cdot)$ is used to return a column vector of the main diagonal elements of a matrix. As can be seen from above, $\mathbf{K}_w$ and $\mathbf{P}_w$ are both 18×1 dimensional vectors and $\mathbf{A}_{\mathrm{Q}}$ is a 21×2 dimensional matrix.

Principally, the EAT model consists of two parts: affine transformation and polynomial transformation. Affine transformation model is constructed by the first three rows of $\mathbf{A}_{\mathrm{Q}}$ and the first three elements of $\mathbf{g}_m$. The polynomial transformation model is formed with the rest of $\mathbf{A}_{\mathrm{Q}}$ and $\mathbf{g}_m$. The affine part of the EAT model is employed to deal with linear deformation between two images. Meanwhile, the polynomial part is used to deal with nonlinear deformation. The weight vector $\mathbf{w}$ restrains the range of the non-rigid transformation coefficients in $\mathbf{K}$ and $\mathbf{P}$, used for improving the flexibility of the EAT model as well. With only $w_5 = 0$, the nonlinear part of the EAT model becomes the mixture of quadratic, cubic and quartic polynomials. When $w_5 = w_4 = 0$, the nonlinear part becomes quadratic and cubic polynomials. The EAT model with $\mathbf{w} = 0$ is the same as the traditional affine transformation.

The transformation coefficient vector, which includes affine and polynomial transformation coefficients, is defined as a 41×1 dimensional vector $\mathbf{Q} = [\theta,\ s_x,\ s_y,\ t_x,\ t_y\ |\ \mathbf{K}\ |\ \mathbf{P}]^{\mathrm{T}}$ Substituting the solution (9) into the objective function (5), it becomes

$$\min_{\mathbf{Q}} \boldsymbol{E}(\mathbf{Q}) = \min_{\mathbf{Q}} -\sum_{m=n=1}^{K} \exp\left\{-\frac{\|\mathbf{v}_n - \mathbf{g}_m\mathbf{A}_{\mathrm{Q}}\|^2}{2\sigma_d^2}\right\} + \lambda tr\left((\mathbf{Q}-\mathbf{Z})(\mathbf{Q}-\mathbf{Z})^{\mathrm{T}}\right) \tag{10}$$

where $\mathbf{v}_n$ is a 1×2 dimensional coordinate vector of any point in $\mathbf{V}$, $\mathbf{Z} = [0,\ 1,\ 1,\ 0,\ \cdots,\ 0]^{\mathrm{T}}$ is a 41×1 dimensional vector, $tr(\cdot)$ denotes the trace. The stabilizer in Eq. (10) prevents from great variation of the transformation coefficients during optimization.

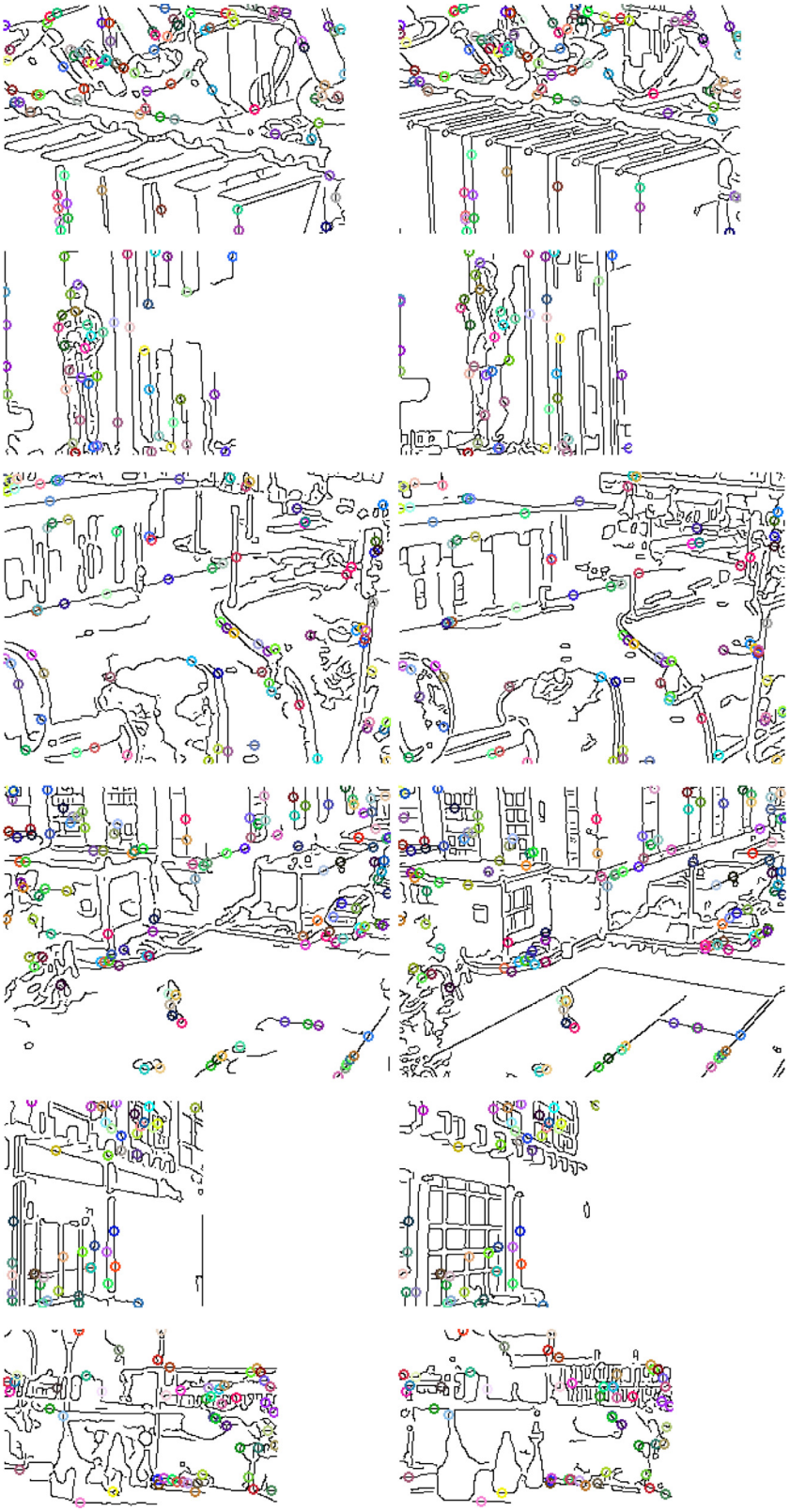


Fig. 2. The results of feature points extraction with $g = 30$ on the six pairs of edge maps of IR and VIS images. The IR edge maps with the feature points (colored '○') are shown on the left. The corresponding VIS edge maps with the feature points (colored '○') are shown on the right. The feature points of the same color are corresponding to each other in an edge map pair.

Algorithm 2
Non-rigid registration using the EAT model.

Input: Two point sets **U** and **V**, parameters σ_d and λ
Output: Optimal EAT matrix $\mathbf{A}_Q$
1 Construct the polynomial set $\mathbf{G} = \{\mathbf{g}_m\}_{m=1}^{K}$ of the mode point set **U**;
2 Initialize the transformation coefficient vector $\mathbf{Q}$ to 0 and set $w_2 = w_3 = 2 \times 10^{-4}$, $w_4 = 1 \times 10^{-4}$, $w_5 = 0$;
3 By using the derivative (11), optimize the objective function (10) by quasi-Newton method to obtain the optimal coefficient vector $\mathbf{Q}^c$ of the coarse optimization;
4 Initialize the transformation coefficient vector to $\mathbf{Q}^c$ and set $w_2 = w_3 = 2 \times 10^{-4}$, $w_4 = 1 \times 10^{-4}$, $w_5 = 1 \times 10^{-7}$;
5 By using the derivative (11), re-optimize the objective function (10) by quasi-Newton method to obtain the optimal coefficient vector $\mathbf{Q}^f$ of the fine optimization;
6 The optimal EAT matrix $\mathbf{A}_Q$ is generated from $\mathbf{Q}^f$ by Eq. (9).

2.4. Optimization

The objective function with the EAT model is always continuously differentiable with respect to the transformation coefficient vector $\mathbf{Q}$ so that the derivative of Eq. (10) is given by

$$\frac{\partial E(\mathbf{Q})}{\partial \mathbf{Q}} = \sum_{m=n=1}^{K} \frac{1}{\sigma_d^2} \frac{\partial \mathbf{g}_m \mathbf{A}_Q}{\partial \mathbf{Q}} (\mathbf{g}_m \mathbf{A}_Q - \mathbf{v}_n)^T exp\left\{ -\frac{\|\mathbf{v}_n - \mathbf{g}_m \mathbf{A}_Q\|^2}{2\sigma_d^2} \right\} + 2\lambda(\mathbf{Q} - \mathbf{Z}) \tag{11}$$

In Eq. (11), $\partial \mathbf{g}_m \mathbf{A}_Q / \partial \mathbf{Q}$ is a 41×2 dimensional matrix and can be written as follow:

$$\frac{\partial \mathbf{g}_m \mathbf{A}_Q}{\partial \mathbf{Q}} = \begin{bmatrix} -s_x x_m \sin\theta - y_m \cos\theta & x_m \cos\theta - s_y y_m \sin\theta \\ x_m \cos\theta & 0 \\ 0 & y_m \cos\theta \\ 1 & 0 \\ 0 & 1 \\ diag\left(\tilde{\mathbf{g}}_m^T \mathbf{w}\right) & 0 \\ 0 & diag\left(\tilde{\mathbf{g}}_m^T \mathbf{w}\right) \end{bmatrix} \tag{12}$$

where $\tilde{\mathbf{g}}_m = [x_m^2, y_m^2, x_m y_m, x_m^3, y_m^3, x_m^2 y_m, x_m y_m^2, x_m^4, y_m^4, x_m^3 y_m, x_m^2 y_m^2, x_m y_m^3, x_m^5, y_m^5, x_m^3 y_m^2, x_m^2 y_m^3, x_m^4 y_m, x_m y_m^4]$.

On the basis of the derivative in Eq. (11), gradient-based numerical optimization technique can be employed to determine the optimal transformation coefficients in $\mathbf{Q}$. In this work, quasi-Newton method is introduced to solve the optimization problem. However, the optimization procedure is limited by local convergence, because of the following reasons: 1) the Gaussian-field-based objective function is convex only in the neighborhood of the optimal solution; 2) the convergence of quasi-Newton method is susceptible to the choice of the initial value. Therefore, to improve optimization performance, a strategy with coarse-to-fine registration is designed and outlined in Algorithm 2.

In Algorithm 2, the coarse optimization is shown in the 2nd step and the 3rd step. The fine optimization is shown in the 4th step and the 5th step. It can be seen that the coarse optimization is mainly based on quartic polynomial, while the fine optimization is mainly based on quintic polynomial. The optimal result of the coarse optimization is used as the initial value of the fine optimization.

The optimal EAT matrix $\mathbf{A}_Q$ can be considered as the regular pattern of global deformation between a pair of IR and VIS images. After the optimal EAT matrix $\mathbf{A}_Q$ is obtained by Algorithm 2, the polynomial set $\mathbf{G}$ is computed for all pixels in an image and then the transformation result of an image is achieved by Eq. (8). It's worth mentioning that after solving the optimal EAT matrix, image registration is implemented without the feature point sets **U** or **V**. This can decrease the runtime of image registration and save storage space. Because there may be some blank areas in an image after transformation, an image interpolation algorithm such as bilinear interpolation is required.

2.5. Computational complexity

From the objective function (10) and its derivative (11), we can see that their time complexity are both $O(K)$ because the sizes of $\mathbf{g}_m$ and $\mathbf{A}_Q$ are fixed. In quasi-Newton method, we use Armijo criteria to calculate the optimal search step and determine when to stop the optimization procedure. The numbers of iterations in the coarse optimization and in the fine optimization are represented as I_c and I_f, respectively. Therefore, the total time complexity for solving the optimal EAT matrix is $O(K(I_c + I_f))$.

The space complexity for solving the optimal EAT matrix is $O(K)$ due to the requirements of storing the $K \times 21$ dimensional polynomial matrix $\mathbf{G}$. Meanwhile, the space complexity for image registration is $O(WH)$ where $W \times H$ is the size of the mode image. It's also worth mentioning that the algorithm of feature point extraction (Algorithm 1) is designed with pipeline architecture and that it has no requirement of storing large scale attribution matrix.

2.6. Implementation details

The state-of-the-art methods of non-rigid registration, such as TPS-RPM [33], CPD [34] and RGF [28], require data normalization so that it has zero means and unit covariance. However, floating point arithmetic can increase the difficulty of hardware implement, especially when dealing with the problem of precision and overflow. The integer coordinates of an image can be used for our method directly without any normalization. This can simplify the calculation steps and make our method implemented easily.

Our method requires five parameters to be set: d, c, g, λ and σ_d. Parameters d, c and g are used for feature point sets extraction. Parameter d controls the search range of point correspondence. Parameter c is used to eliminate the point pairs with over high $\mathbf{Sc}_{ij}$ (4) from feature point sets. Parameter g controls the sizes of feature point sets. The above parameters are mainly employed to adjust the time and space complexity of feature point extraction. We set $d = 20$ pixels, $c = 50$ pixels and $g = 5$ pixels throughout this work in order to make Algorithm 1 run efficiently on our computer. Using the above parameters, the runtime of feature point extraction with $N_r = 1000$ and $N_v = 1000$ is about 12.9 seconds. Parameters λ and σ_d are used for estimation of transformation. Parameter λ controls the trade-off between the closeness of two point sets and the smoothness of the solution. Parameter σ_d determines the scale of the closeness measure. In this work, $\lambda = 0.02$ and $\sigma_d = 6$ are determined through multiple experiments. In addition, we tuned the various weight vectors $\mathbf{w}$ to find the optimal settings: $w_2 = w_3 = 2 \times 10^{-4}$, $w_4 = 1 \times 10^{-4}$ and $w_5 = 1 \times 10^{-7}$.

3. Experiment

Firstly, we performed point set registration to test the performance of the proposed method. Our method was then compared with the state-of-the-art methods on real IR and VIS image pairs. At last, we measured the performance of the EAT model with various orders and registration strategies to demonstrate the efficiency of our technique. The experiments were completed by the computer with 3.9GHz Intel Core CPU, 4GB memory and Matlab code.

3.1. Dataset

Our dataset containing synthesized point sets and real image pairs was utilized to show the results of the proposed method. The synthesized dataset shown in Fig. 3 is constructed by Chui and Rangarajan [33]. It consists of three different point sets which are

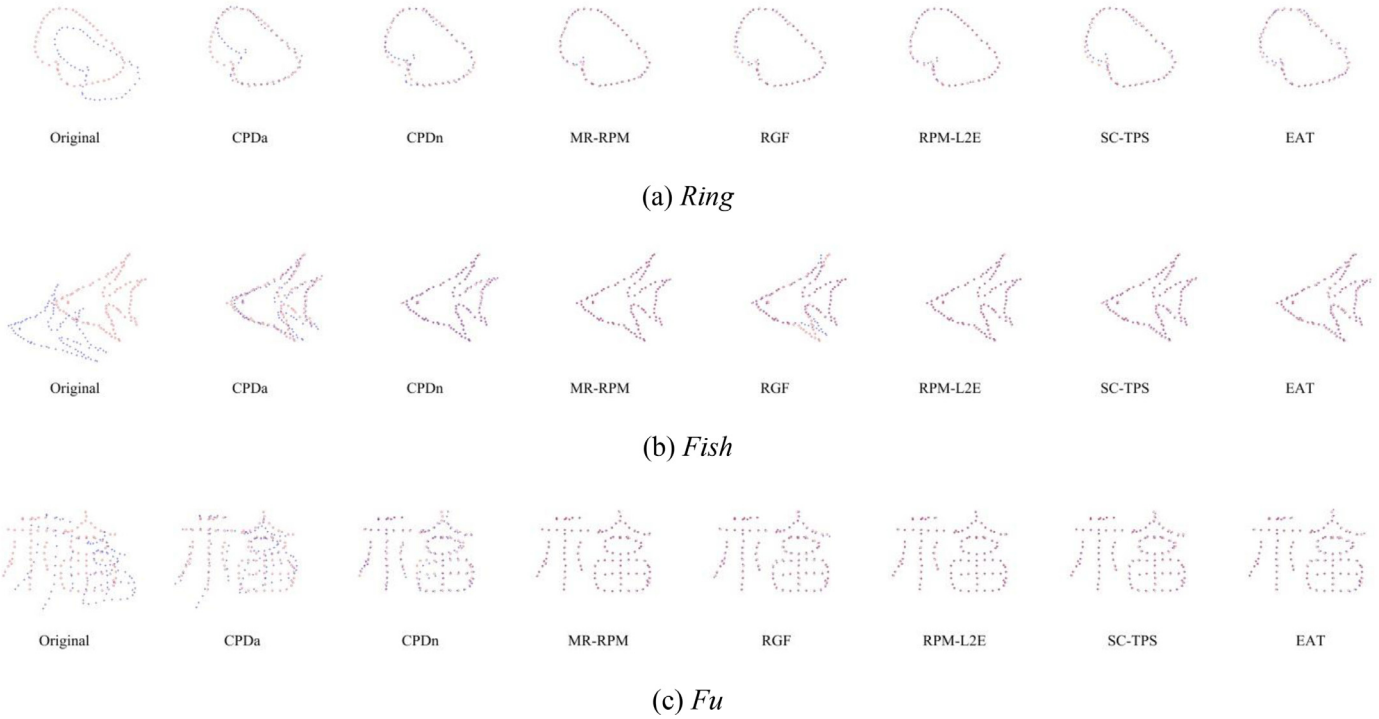


Fig. 3. Qualitative comparisons on the synthesized dataset. The registration results of the point sets of *Ring, Fish* and *Fu* are shown in the first, second and third row respectively, where the mode points are blue '+' and the data points are red '○'. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
The resolutions of the six image pairs.

Image pair	*Equipment*	*Lab*	*Campus 1*	*Campus 2*	*Window*	*Square*
Resolution	283 × 189	193 × 169	320 × 240	320 × 240	166 × 170	227 × 151

respectively named by *Ring, Fish* and *Fu* in this paper with 50, 96 and 108 points, separately.

Our method was compared with the state-of-the-art methods on the six pairs of IR and VIS images, i.e. *Equipment, Lab, Campus 1, Campus 2, Window* and *Square*, as shown in Fig. 5. The image pair of *Lab* is captured by our IR and VIS image fusion camera. The image pairs of *Campus 1* and *Campus 2* are from OTCBVS datasets [35]. The original image pairs of *Campus 1* and *Campus 2* have been already registered by manual operation. We distorted the original IR images randomly by non-rigid transformation to generate unregistered image pairs. The image pairs of *Window* and *Square* are from CVC datasets [20]. The resolutions of the six image pairs are shown in Table 1. To be fair, the same feature point sets were used for all the methods involved in this experiment on an individual image pair.

3.2. Comparison evaluation on synthesized dataset

In this section, the proposed method was tested on the synthesized datasets and compared with the state-of-the-art methods: CPD with the affine model (CPDa) [34], CPD with the non-rigid model (CPDn) [34], MR-RPM [36], RGF [28], RPM-L_2E [29] and SC-TPS [32]. All of these methods are able to estimate transformation model from point feature. Meanwhile, the transformation models used in these methods include the affine model, the RKHS-based model and the TPS model, etc. Because Algorithm 1 is an approach for extracting feature points from edge maps, it cannot be used to point set registration. Thus, the Hungarian method with SC [32] was employed for our method and MR-RPM to achieve point correspondence. In addition, the MATLAB codes of the above state-of-the-art methods were provided by their authors. The parameters of these methods in this work are also the same as those set by their authors.

Table 2
The average runtimes of RGF, RPM-L2E and EAT (Unit: second).

	T_s	T_i	Total
RGF	33.7440	0.0448	33.7888
RPM-L_2E	20.1889	0.3706	20.5595
EAT	**5.2052**	**0.1095**	**5.3147**

Fig. 3 shows the qualitative results on the synthesized datasets, where the mode points are blue '+' and the data points are red '○'. We aligned the blue '+' to the red '○'. Apparently, the registration results of CPDa are the worst of all. This demonstrates that the non-linear transformation model is superior over the linear transformation model on non-rigid registration. It also can be seen that MR-RPM and RPM-L_2E are better than the others on point set registration. The performance of the proposed method is very close to those of MR-RPM and RPM-L_2E.

The quantitative evaluation implemented by root-mean-square error (RMSE) is reported in Fig. 4. It can be seen that the quantitative results coincide with the results of subjective assessment well. The total average errors of CPDa, CPDn, RGF, SC-TPS, RPM-L_2E and MR-RPM are about 5.32, 2.29, 1.01, 0.82, 0.52 and 0.49 respectively, while that of the proposed EAT is about 0.7. Our method is superior over CPDa, CPDn, RGF and SC-TPS, while the accuracy of

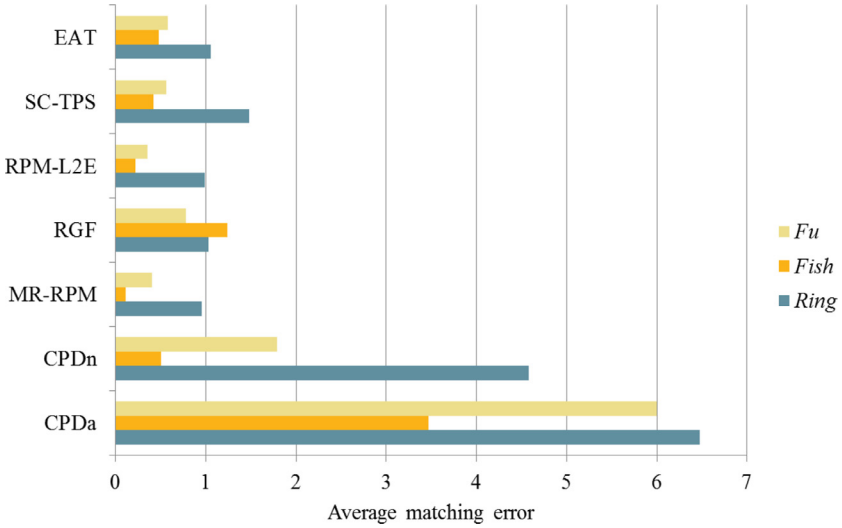


Fig. 4. Quantitative comparisons on the synthesized dataset.

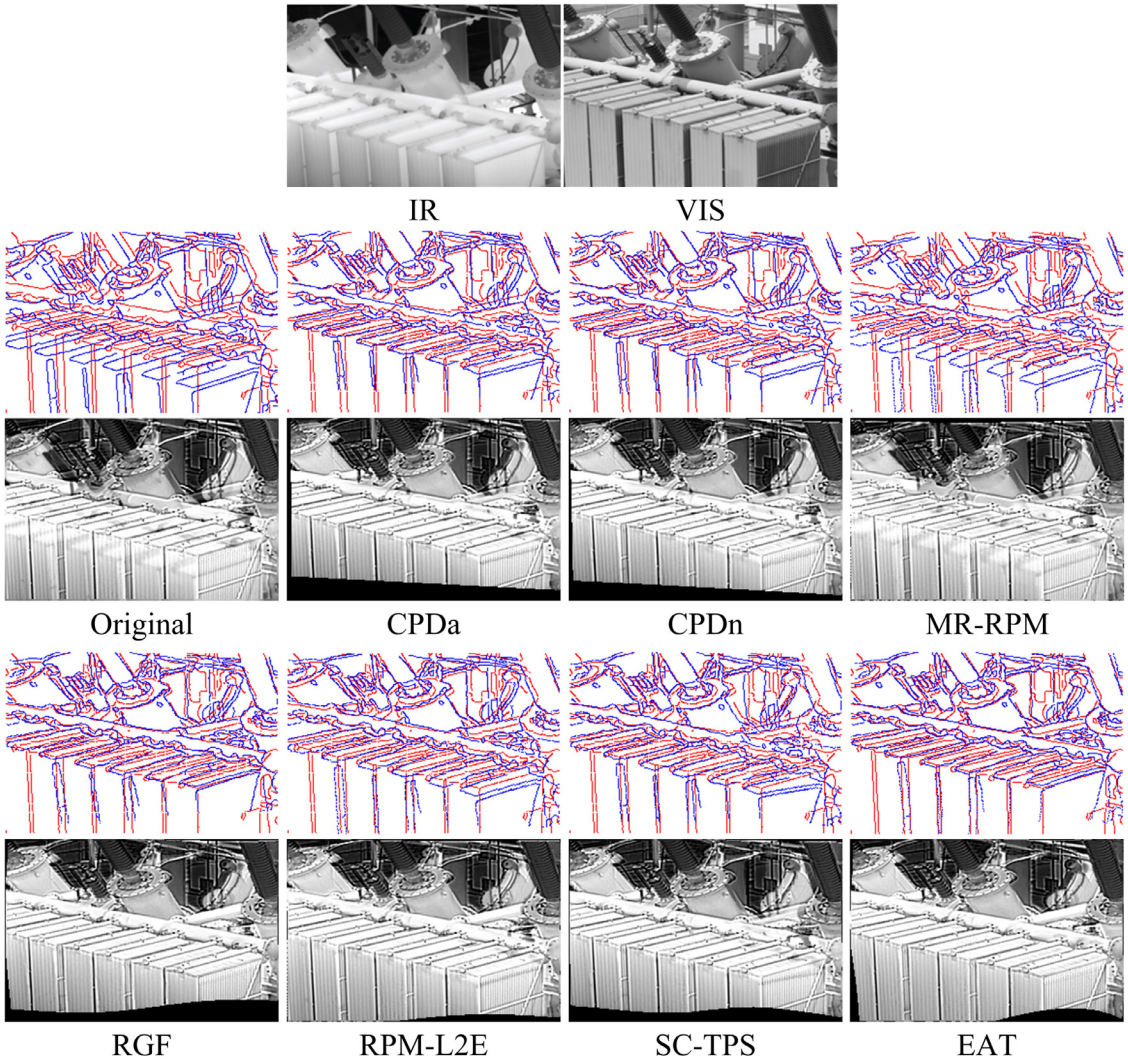
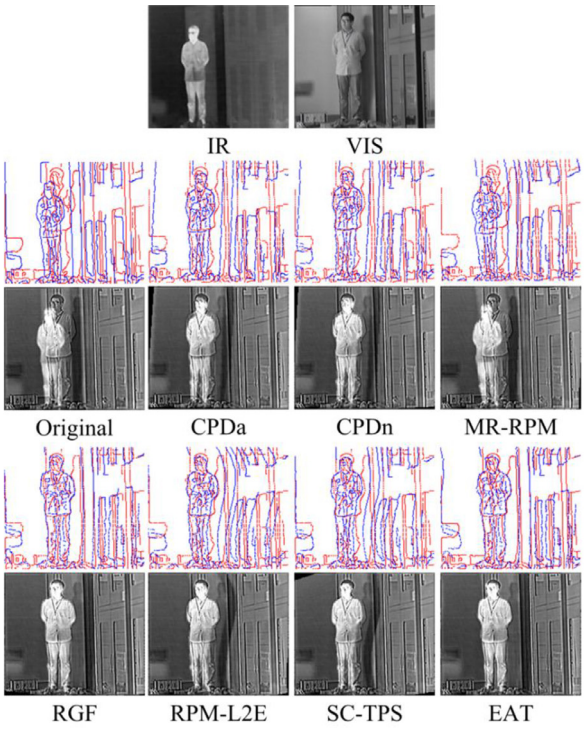
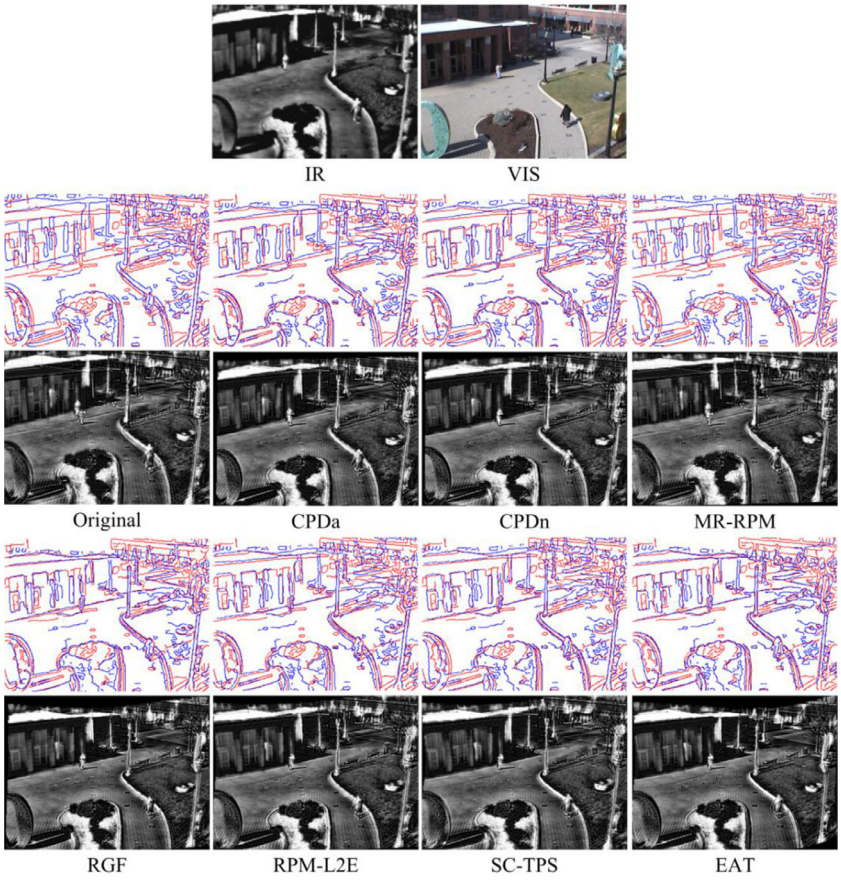


Fig. 5. Qualitative registration results of CPDa, CPDn, MR-RPM, RGF, RPM-L_2E, SC-TPS and EAT on the different image pairs.

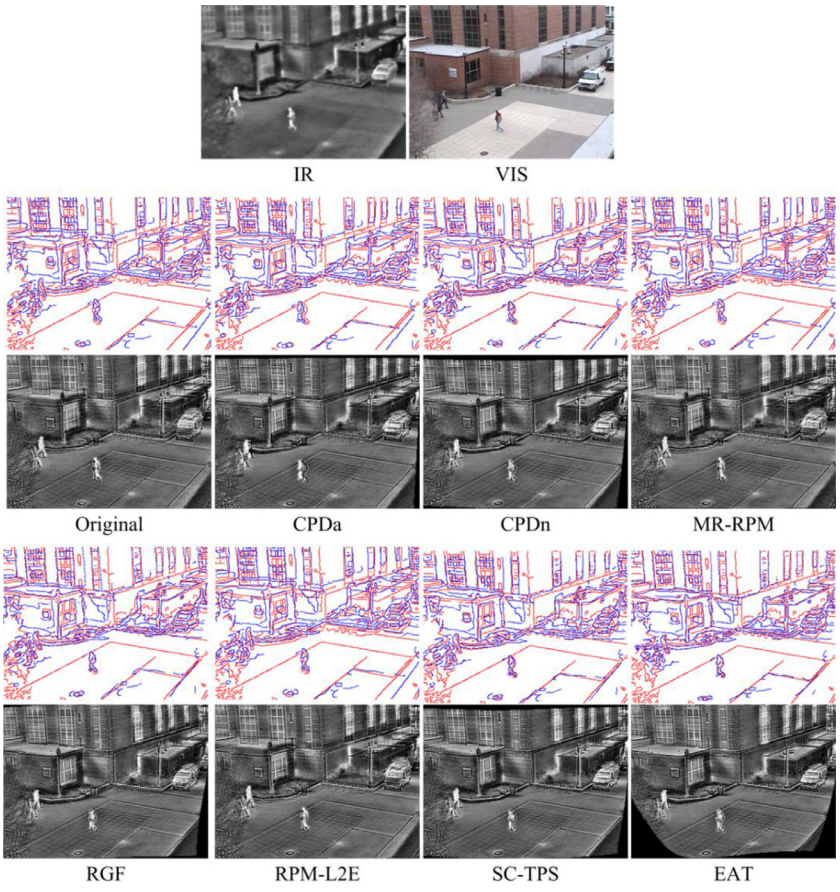


(b) *Lab*

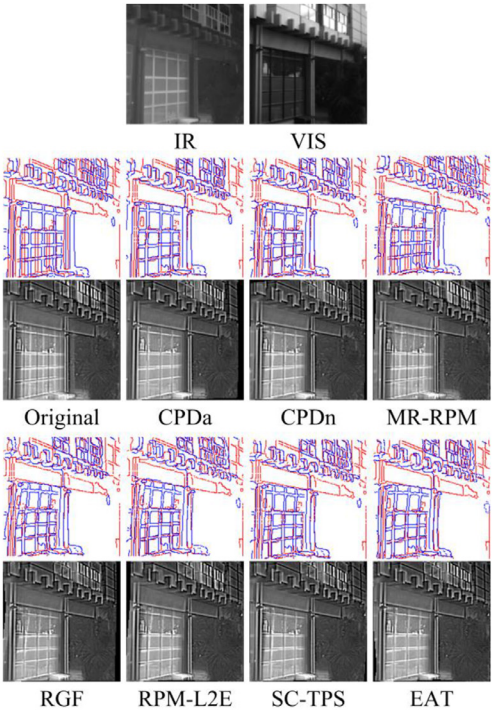


(c) *Campus 1*

Fig. 5. Continued

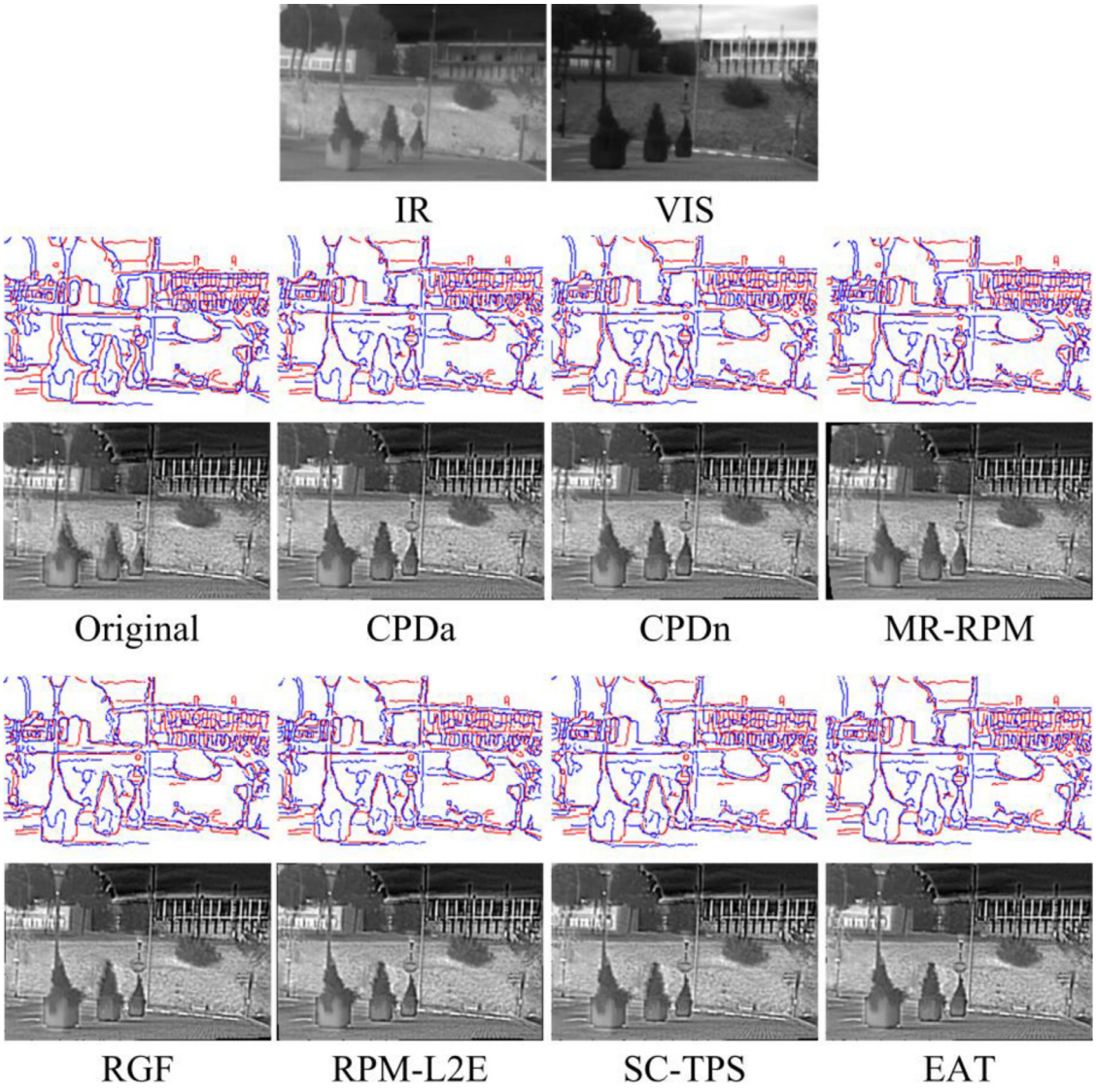


(d) *Campus 2*



(e) *Window*

Fig. 5. Continued



(f) *Square*

Fig. 5. Continued

our method approaches to those of RPM-L_2E and MR-RPM. Thus, in general, our method performs well on non-rigid registration of point sets. But, on the other hand, the performance of our method on point set registration is limited by the EAT model preferring to describe the regular pattern of global deformation, since point set registration emphasizes the estimation of local deformation model.

3.3. Comparison evaluation on real images

In this section, by using the dataset containing six pairs of IR and VIS images, the performance of the proposed method was compared with those of the state-of-the-art methods: CPDa, CPDn, MR-RPM, RGF, RPM-L_2E and SC-TPS. To present registration results visually, an image fusion method with bilateral filter was employed in this experiment. The detail layer is extracted from a VIS image and fused with the corresponding IR image by an average fusion algorithm. The qualitative results of various methods are shown in Fig. 5, including the results of edge map registration and image registration. As shown in Fig. 3, MR-RPM has excellent performance of point set registration. However, it fails in image registration because the transformation models estimated from local feature are not able to achieve global image registration accurately. The registration results of CPDa, CPDn, RPM-L_2E, RGF and SC-TPS are good but not good enough, while RPM-L_2E and RGF are better than the other methods, since the regularized Gauss field criterion is able to measure the accuracy of global registration from feature points. In comparison, the registration results of the proposed method have less distortion. Moreover, they are more precise and robust than the results of the other approaches. This demonstrates that the proposed EAT model has excellent generalization ability and is especially facilitated to achieve global image registration based on local feature.

For each IR and VIS image pair, we constructed two feature point sets as ground truth by using a semi-manual procedure. The feature point sets were extracted by Algorithm 1 with $g = 2$ and then we manually deleted the feature point pairs which are not corresponding to each other in fact. The ground truth contains more point pairs than feature point sets and can be employed to compute the recall of registration results for quantitative evaluation. Fig. 6 shows the quantitative comparisons of CPDa, CPDn,

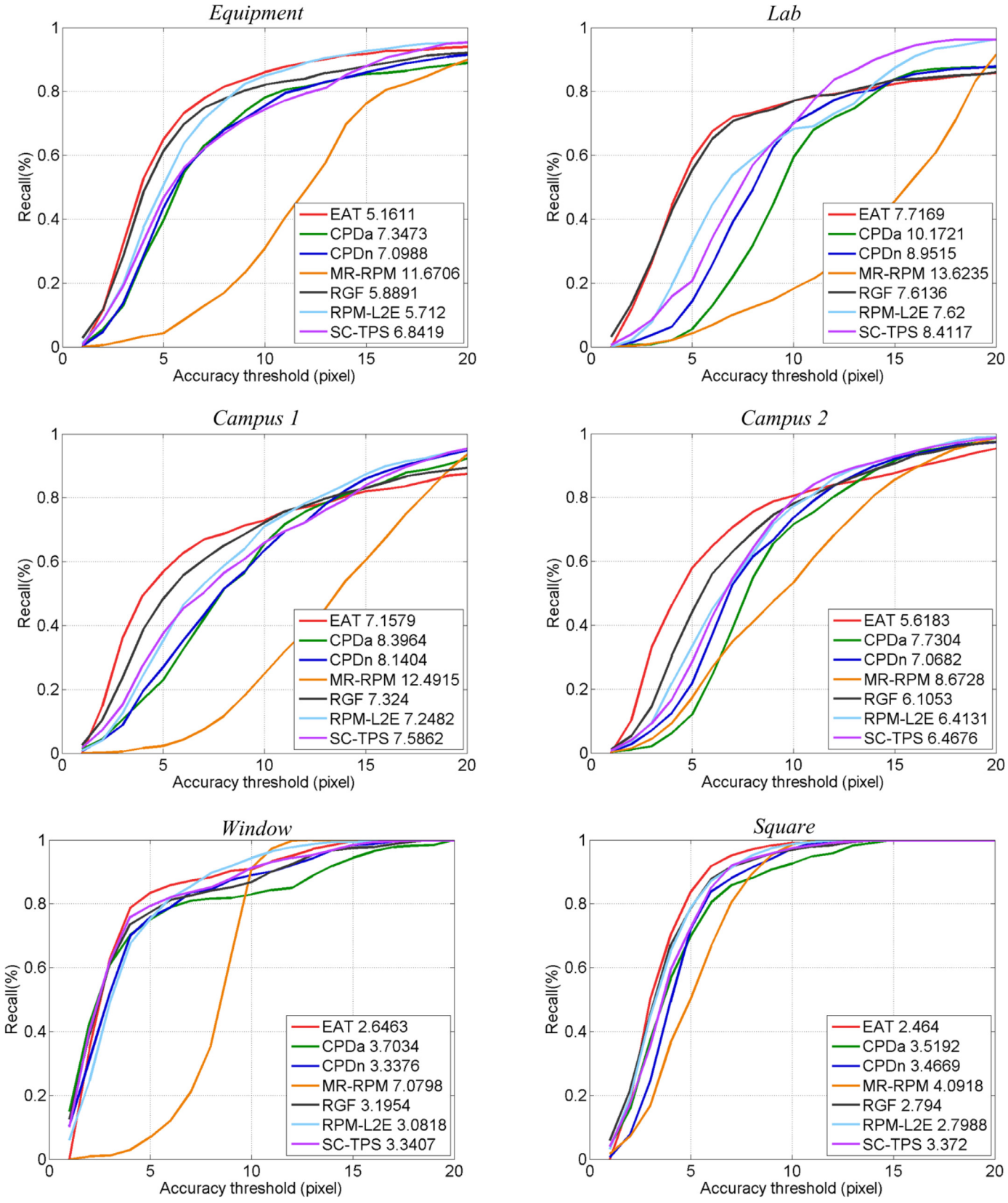


Fig. 6. Quantitative comparisons of EAT, CPDa, CPDn, MR-RPM, RGF, RPM-L^2E and SC-TPS on the various image pairs. The numbers in the legend are the average matching errors.

MR-RPM, RGF, RPM-L_2E, SC-TPS and EAT. We find that the quantitative results are in good agreement with the qualitative comparisons and the curves of EAT are above those of the other approaches in most cases. Compared with RGF and RPM-L_2E, the proposed method reduces the average matching error by about 6.5%. It demonstrates that our method has better accuracy and stability than the other methods. In addition, the registration methods with non-linear transformation, such as CPDn, RGF, RPM-L_2E, SC-TPS and EAT, are superior over CPDa in average matching error. Thus, it can be seen that non-linear transformation is more suitable for IR and VIS image registration than affine transformation.

We also tested the runtimes of the proposed method and compare it with the runtimes of RGF and RPM-L_2E which have second-best registration precisions. Table 2 reports the average runtimes, including the average runtimes of solving optimization (T_s) and the average runtimes of global image registration with known transformation coefficients (T_i). As we can see, the proposed method has the fastest speed for solving optimization thanks to the objective function which is simplified by the potential correspondence between IR and VIS image pairs. Meanwhile, the T_i of our method is more than that of RGF. The reason for this is that the coefficients of the EAT is more than that of RGF. The number of EAT coefficients is 41. According to [28], we set the number of control points N_0 to 15 in RGF. The total runtime of the proposed method is the least among the three non-rigid registration methods. It demonstrates that the proposed method is able to achieve fast registration on the basis of potential correspondence.

The non-rigid registration methods, such as CPDn, MR-RPM, RGF, RPM-L_2E and SC-TPS, have requirement of storing large scale matrix. The numbers of the feature point pairs used in CPDn and SC-TPS must be less than 60, otherwise CPDn and SC-TPS cannot run on our computer due to out of memory. In MR-RPM, RGF and RPM-L_2E, the increasing of control points results in oversized attribution matrix. This gives rise to difficulty in algorithm implementation. Because the EAT model is not constructed by local feature and image transformation can be achieved without feature points, there is not too much data for storage in our method. Thus, compared with the state-of-the-art methods, the proposed method is more easily to be applied in practical engineering.

3.4. Evaluation of the EAT model

In this section, we measured the performance of the EAT model with various orders and registration strategies on the dataset containing real images. Fig. 7 reports the curves of total average recall and the total average matching errors of the registration results with different EAT models on the dataset. In the legend of Fig. 7, EAT represents the complete version of our method, while the rest represent the simplified versions of our method. The simplified version is the proposed method with the various weight vector $\mathbf{w}$ and only the coarse optimization. EATa denotes the simplified version with $\mathbf{w} = 0$ (i.e. the affine model). EAT2 denotes the simplified version with $w_2 = 2 \times 10^{-4}$ and $w_3 = w_4 = w_5 = 0$. EAT3 denotes the simplified version with $w_2 = w_3 = 2 \times 10^{-4}$ and $w_4 = w_5 = 0$. EAT4 denotes the simplified version with $w_2 = w_3 = 2 \times 10^{-4}$, $w_4 = 1 \times 10^{-4}$ and $w_5 = 0$. EAT5 denotes the simplified version with $w_2 = w_3 = 2 \times 10^{-4}$, $w_4 = 1 \times 10^{-4}$ and $w_5 = 1 \times 10^{-7}$. EAT6 denotes the simplified version with the non-linear part of the EAT model $G(x, y, \mathbf{R}) = \sum_{i=2}^{6}\sum_{j=0}^{i} w_i r_{i,j} x^j y^{i-j}$, where $w_2 = w_3 = 2 \times 10^{-4}$, $w_4 = 1 \times 10^{-4}$, $w_5 = 1 \times 10^{-7}$ and $w_6 = 1 \times 10^{-11}$. Thus, the highest orders of the EAT models in EAT2~EAT6 are 2, 3, 4, 5 and 6 respectively.

Firstly, because the matching error of EATa is greatly more than those of the others and the recall curve of EATa is worst of all, we find that the non-linear part of the EAT model can significantly improve the performance of non-rigid image registration. Secondly,

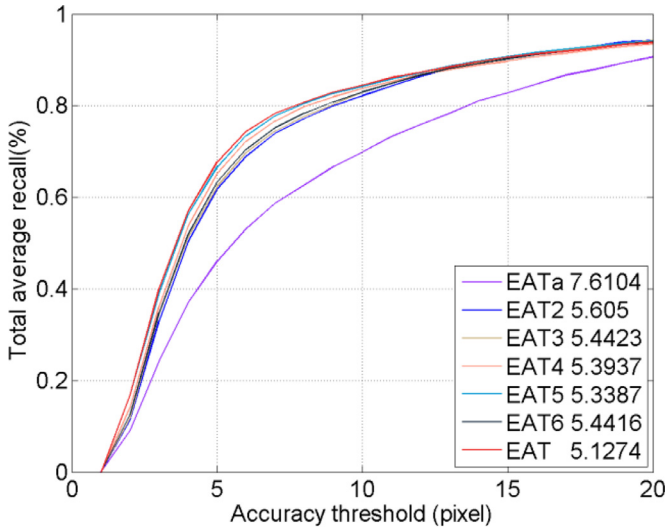


Fig. 7. Quantitative comparisons of the EAT models with various orders and registration strategies on the dataset of real images. The numbers in the legend are the total average matching errors.

the results show that the higher order of the EAT model is not always better. When the highest order of the EAT model ≤ 5, the registration accuracy is getting better with higher order. However, the registration accuracy is decreased when the highest order is set to 6. These are caused by two reasons mainly: 1) the EAT model with low non-linearity is not accurate enough to describe the non-rigid deformation between IR and VIS images; 2) the gradient of the objective function established by the EAT model with higher order is more susceptible to the initial error between IR and VIS images, as shown in Eq. (11). These also can result in performance deterioration of optimization procedure. Meanwhile, this is the reason why the highest order of the EAT model is determined to be 5 in this work. Thirdly, our method has the best performance in recall and matching error. Compared with EAT5 which is achieved by only the coarse registration, the strategy with coarse-to-fine registration reduces the matching error by about 4%. This shows clearly that the registration strategy used in this work improves the chance of reaching the global optima.

The total average matching errors of CPDn, RGF, RPM-L_2E and SC-TPS are about 6.34, 5.49, 5.48 and 6.00, respectively. Using the errors given above, it can be seen that the accuracies of our methods with only coarse registration, such as EAT3~EAT5, are better than those of the state-of-the-art methods. In other words, compared with the existing non-rigid transformation models, such as the TPS transformation model (1) or the non-rigid transformation model within RKHS (2), the proposed EAT models with the highest order ≥ 3 are able to describe the regularity of non-rigid deformation between IR and VIS images more accurately. This also demonstrates that the EAT model improves the accuracy of non-rigid IR and VIS image registration.

The EAT model is different from the existing non-linear transformation models: it does not require control points to achieve non-linear image transformation. In the existing non-linear transformation models Eqs. (1) or (2), control points can be regarded as the key coefficients of transformation models. However, the selection of control points is difficult to be optimized for image registration. Registration accuracy may be limited with the low quantity of control points, while computation complexity is significantly increased with the large quantity of control points. Moreover, the distribution of the locations of control points also influences the performance of image registration achieved by using transformation models with control points. Because the optimization of trans-

formation coefficients mainly focuses on the neighborhoods of control points, the mapping of every point in an image is determined by nearby control points. The registration accuracy may be degraded in the points which are relatively far from control points. This can result in low generalization ability of transformation models estimated from feature point sets. Thus, the optimal set of control points is hard to be determined so that transformation models with control points are not be fully optimized. In other words, transformation models with control points rely heavily on local feature so that the precision of non-rigid image registration is limited.

Because the EAT model consisting of affine and polynomial transformations does not require control points, there is none of the above issues in our method. From Eq. (8) we can see that except for the weight vector **w**, all coefficients of the EAT model can be optimized for image registration. This ensures that the proposed EAT model is more accurate than transformation models with control points. In addition, the EAT model prefers to describe the regularity of global deformation rather than local pattern in the neighborhood of feature points. The coefficients of the EAT model are optimized to correctly describe the regular pattern of global deformation between IR and VIS images. Therefore, with the EAT model, the mappings of all points in an image can be determined by a regular pattern (8). Although the EAT model is estimated from feature point sets, non-rigid image transformation can be implemented without feature point sets in our method. Hence, the EAT model is able to reduce the dependence of non-rigid image transformation on local feature. In addition, it can be seen from Eq. (8): the distances between non-feature and feature points do not affect the spatial transformation with the EAT model in non-feature points. Furthermore, experiment results in this work show that the EAT model improves the accuracy of non-rigid IR and VIS image registration. Thus, the EAT model has high generalization ability, especially compared with transformation models with control points.

4. Conclusion

Within this work, the affine transformation is generalized from linear to non-linear case. On this basis, we propose a method with the EAT model for non-rigid IR and VIS image registration. By using the simplified Gaussian-fields-based objective function, the optimal coefficients of the EAT model are determined from the edge maps of IR and VIS image pairs. The experiment results on the synthesized point sets show that the performance of the proposed method is limited on non-rigid point set registration. However, experiment results on the real images show that in contrast with the state-of-the-art methods, our method is able to achieve more accurate registration in a shorter time. The matching error of our method on non-rigid image registration is improved by 6.5% and the runtime is decreased by 74.2%. This demonstrates that the proposed EAT model describing the regularity of global deformation is helpful to improve the accuracy and efficiency of non-rigid image registration. In addition, the effectiveness of the strategy with coarse-to-fine registration is also proved by experiment.

There are mainly two weaknesses in our method. Firstly, our method has no superiority in non-rigid point set registration because the EAT model prefers to describe global deformation rather than local deformation. Secondly, the extraction of feature points which are one-to-one correspondence spends a long time. Thus, the future work mainly focuses on as follows: 1) improve the performance of the proposed EAT model on describing local deformation property; 2) further reduce the computation complexity of feature point extraction.

Declaration of Competing Interest

None.

Acknowledgments

The authors gratefully acknowledge the financial supports from the National Natural Science Foundation of China (no. 61901157) and the Science Foundation of Yunnan Electric Power Research Institute (Group) Co., Ltd. in China (no. YNKJXM20180244).

References

[1] Kai Zhang, Min Wang, Shuyuan Yang, Convolution structure sparse coding for fusion of panchromatic and multispectral images, IEEE Trans. Geosci. Remote Sens. 57 (2019) 1117–1130 https://doi.org/10.1109/TGRS.2018.2864750.
[2] Huafeng Li, Xiaoge He, Dapeng Tao, Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning, Pattern Recognit. 79 (2018) 130–146 https://doi.org/10.1016/j.patcog.2018.02.005.
[3] Ke Li, Changqing Zou, Shuhui Bu, Multi-modal feature fusion for geographic image annotation, Pattern Recognit. 73 (2018) 1–14 https://doi.org/10.1016/j.patcog.2017.06.036.
[4] Hao Chen, Youfu Li, Dan Sun, Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 86 (2019) 376–385 https://doi.org/10.1016/j.patcog.2018.08.007.
[5] Ce Zhang, Isabel Sargent, Xin Pan, VPRS-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images, IEEE Trans. Geosci. Remote Sens. 56 (2018) 4507–4521 https://doi.org/10.1109/TGRS.2018.2822783.
[6] Tong Tong, Gray Katherine, Qinquan Gao, Multi-modal classification of Alzheimer's disease using nonlinear graph fusion, Pattern Recognit. 63 (2017) 171–181 https://doi.org/10.1016/j.patcog.2016.10.009.
[7] VA Zimmer, Ballester MAG, multimodal image registration using Laplacian commutators, Inf. Fusion 49 (2019) 130–145 https://doi.org/10.1016/j.inffus.2018.09.009.
[8] P.A. Legg, P.L. Rosin, D. Marshall, Feature neighbourhood mutual information for multi-modal image registration: an application to eye fundus imaging, Pattern Recognit. 48 (2015) 1937–1946 https://doi.org/10.1016/j.patcog.2014.12.014.
[9] Jun Liu, Yang Cheng, Visible/infrared dual-band large field shared-aperture and parfocal optical system, J. Xi'an Technol. Univ. 34 (2014) 87–93 https://doi.org/10.3969/j.issn.1673-9965.2014.02.001.
[10] Aiping Sun, Yangyun Gong, Youpan Zhu, Optical system design of low-light-level and infrared image fusion hand-held viewer, Infrared Technol. 35 (2013) 712–719 https://doi.org/10.11846/j.issn.1001_8891.201311008.
[11] Chenqiang Gao, Lan Wang, Yongxing Xiao, Infrared small-dim target detection based on Markov random field guided noise modeling, Pattern Recognit. 76 (2018) 463–475 https://doi.org/10.1016/j.patcog.2017.11.016.
[12] A. Ghaffari, E. Fatemizadeh, Image registration based on low rank matrix: Rank-regularized SSD, IEEE Trans. Med. Imag. 37 (2018) 138–150 https://doi.org/10.1109/TMI.2017.2744663.
[13] Ziquan Wei, Yifeng Han, Mengya Li, A small UAV based multi-temporal image registration for dynamic agricultural terrace monitoring, Remote Sens. 9 (2017) 904 https://doi.org/10.3390/rs9090904.
[14] Lihui Chen, Xiaomin Yang, Lu Lu, An image fusion algorithm of infrared and visible imaging sensors for cyber-physical systems, J. Intell. Fuzzy Syst. 36 (2019) 4277–4291 https://doi.org/10.3233/JIFS-169985.
[15] Xunwei Xie, Yongjun Zhang, Xiao Ling, A new registration algorithm for multimodal remote sensing images, in: IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 7011–7014. https://doi.org/10.1109/IGARSS.2018.8517853.
[16] Yuanxin Ye, Jie Shan, Lorenzo Bruzzone, Robust registration of multimodal remote sensing images based on structural similarity, IEEE Trans. Geosci. Remote Sens. 55 (2017) 2941–2958 https://doi.org/10.1109/TGRS.2017.2656380.
[17] Qinglei Du, Aoxiang Fan, Yong Ma, Infrared and visible image registration based on scale-invariant PIIFD feature and locality preserving matching, IEEE Access 6 (2018) 64107–64121 https://doi.org/10.1109/ACCESS.2018.2877642.
[18] G.A. Bilodeau, A. Torabi, F. Morin, Visible and infrared image registration using trajectories and composite foreground images, Image Vision Comput. 9 (2011) 41–50 https://doi.org/10.1016/j.imavis.2010.08.002.
[19] Gang Liu, Zhonghua Liu, Sen Liu, Registration of infrared and visible light image based on visual saliency and scale invariant feature transform, EURASIP J. Image Video Process. 45 (2018) 45 https://doi.org/10.1186/s13640-018-0283-9.
[20] Xiangzeng Liu, Yunfeng Ai, Juli Zhang, A novel affine and contrast invariant descriptor for infrared and visible image registration, Remote Sens. 10 (2018) 658 https://doi.org/10.3390/rs10040658.
[21] F.L. Bookstein, Principal warps: thin-plate splines and the decomposition of deformations, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 567–585 https://doi.org/10.1109/34.24792.
[22] Haili Chuia, Anand Rangarajan, A new point matching algorithm for non-rigid registration, Comput. Vision Image Understanding 89 (2003) 114–141 https://doi.org/10.1016/S1077-3142(03)00009-2.

[23] Changcai Yang, Yizhang Liu, Xingyu Jiang, Non-rigid point set registration via adaptive weighted objective function, IEEE Access 6 (2018) 75947–75960 https://doi.org/10.1109/ACCESS.2018.2883689.
[24] Aoshuang Dong, Benqiang Yang, Danyang Zhao, Research of medical image non-rigid registration based on TPS-semisurf algorithm, Adv. Mater. Res. 791–793 (2013) 2112–2117 https://doi.org/10.4028/www.scientific.net/AMR.791-793.2112.
[25] Suicheng Gu, Xin Meng, Frank C. Sciurba, et al., Bidirectional elastic image registration using B-spline affine transformation, Comput. Med. Imaging Graph. 38 (2014) 306–314 http://dx.doi.org/10.1016/j.compmedimag.2014.01.002.
[26] Pingge Jiang, James A. Shackleford, B-spline registration of neuroimaging modalites with map-reduce framework, Brain Inform. Health 9250 (2015) 285–294 http://dx.doi.org/10.1007/978-3-319-23344-4_28.
[27] Wei Sun, Wiro J. Niessen, Stefan Klein, Randomly perturbed B-splines for non-rigid image registration, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1401–1413 http://dx.doi.org/10.1109/TPAMI.2016.2598344.
[28] Jiayi Ma, Ji Zhao, Yong Ma, et al., Non-rigid visible and infrared face registration via regularized Gaussian fields criterion, Pattern Recognit. 48 (2015) 772–784 https://doi.org/10.1016/j.patcog.2014.09.005.
[29] Jiayi Ma, Weichao Qiu, Ji Zhao, Robust L^2E estimation of transformation for non-rigid registration, IEEE Trans. Signal Process. 63 (2015) 1115–1127 https://doi.org/10.1109/TSP.2014.2388434.
[30] Jiayi Ma, Ji Zhao, L Alan, Yuille, Non-rigid point set registration by preserving global and local structures, IEEE Trans. Image Process 25 (2016) 53–64 https://doi.org/10.1109/TIP.2015.2467217.
[31] Shan-e-Ahmed Raza, Victor Sanchez, Gillian Prince, Registration of thermal and visible light images of diseased plants using silhouette extraction in the wavelet domain, Pattern Recognit. 48 (2015) 2119–2128 https://doi.org/10.1016/j.patcog.2015.01.027.
[32] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 509–522 https://doi.org/10.1109/34.993558.
[33] H. Chui, A. Rangarajan, A new point matching algorithm for non-rigid registration, Comput. Vis. Image Understand. 89 (2003) 114–141 https://doi.org/10.1016/S1077-3142(03)00009-2.
[34] A. Myronenko, X. Song, Point set registration: coherent point drift, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 2262–2275, doi:10.1109/TPAMI.2010.46.
[35] J. Davis, V. Sharma, Background-subtraction using contour-based fusion of thermal and visible imagery, Comput. Vision Image Understanding 106 (2007) 162–182 https://doi.org/10.1016/j.cviu.2006.06.010.
[36] Jiayi Ma, Jia Wu, Ji Zhao, Nonrigid point set registration with robust transformation learning under manifold regularization, IEEE Trans. Neural Netw. Learn. Syst. (2018) https://doi.org/10.1109/TNNLS.2018.2872528.

Chaobo Min received the Ph.D. degree from the School of Electronics and Optical Technology, Nanjing University of Technology, Nanjing, China, in 2014. He is now an lecturer with the College of Internet of Things Engineering, HoHai University. His current research interests include image processing and computer vision.

Yan Gu received the M.D. degree from the School of Electronics and Optical Technology, Nanjing University of Technology, Nanjing, China, in 2009. She is now a senior engineer with North Night Vision Technology Co., Ltd. Her current research interests include pattern recognition and image fusion.

Yingjie Li received the Ph.D. degree from the School of Electronics and Optical Technology, Nanjing University of Technology, Nanjing, China, in 2017. He is now an engineer with North Information Control Research Academy Group Co., Ltd. His current research interests include image processing and computer vision.

Feng Yang received the M.D. degree from the School of Electronics and Optical Technology, Nanjing University of Technology, Nanjing, China, in 2014. He is now an engineer with North Night Vision Technology Co., Ltd. His current research interests include image processing and machine learning.