Adaptive enhanced affine transformation for non-rigid registration of visible and infrared images

Chaobo Min | Yan Gu | Yingjie Li | Feng Yang

Abstract
Non-rigid registration, performing well in all-weather and all-day/night conditions, directly determine the reliability of visible (VIS) and infrared (IR) image fusion. On account of non-planar scenes and differences between IR and VIS cameras, non-linear transformation models are more helpful to non-rigid image registration than the affine model. However, most of non-linear models usually used on non-rigid registration are constructed by control points at present. Aiming at the issue that the adaptiveness and generalization of the control-point-based models are limited, adaptive enhanced affine transformation (AEAT) is proposed for image registration, generalizing the affine model from linear to non-linear case. Firstly, Gaussian weighted shape context, measuring the structural similarity between multimodal images, is designed to extract putative matches from edge maps of IR and VIS images. Secondly, to implement global image registration, the optimal parameters of the AEAT modal are estimated from putative matches by a strategy of subsection optimization. Experiment results show that this approach is robust in different registration tasks and outperforms several competitive methods on registration precision and speed.

1 INTRODUCTION
Non-rigid image registration, as we know, is a prerequisite for infrared (IR) and visible (VIS) image fusion [1] applied widely in many applications, such as night vision [2], medical imaging [3] and remote sense [4] etc. The goal of registration is to estimate a spatial transformation model between two images to be aligned. Without image registration, IR and VIS image fusion could not be implemented [5]. However, the differences between acquisition systems and intensity distributions of multimodal images raise a tough challenge to image registration, making multimodal image registration a hot topic. On this account, non-rigid registration of IR and VIS images is our focus in this work.

Image registration could be considered as a fitting problem. Mutual features between VIS and IR images are regarded as known data points. Spatial transformation model is a fitting function. A fitting criterion, which is used to quantify the distance between two images to be aligned, is established by mutual features and transformation model. Then, image registration could be achieved by determining transformation parameters minimizing the fitting criterion [11]. Thus, mutual feature extraction and spatial transformation model are essential for image registration.

Three types of mutual features would be introduced here: point, intensity and structure. Point feature is prevalent so that many approaches of point matching are used for mutual feature extraction [6]. Although some point-feature-based registration algorithms do not require any explicit set of point matches, the feature descriptors quantifying the degree of point correspondence still control the precision of registration directly. Also, intensity-based feature descriptors are widely employed for image registration, such as scale invariant feature transform (SIFT) [7], corner feature [8], histogram of gradient (HOG) [9] and speeded up robust feature (SURF) [10] etc. Even so, for IR and VIS image registration, intensity-based features could not produce accurate results due to the great difference between intensity distributions of VIS and IR images. Compared to intensity-based feature descriptors, structural feature descriptors are affected slightly by intensity distributions. Moreover, owing to the similarity between the global structures of VIS and IR images to be aligned [11], mutual features could be extracted accurately by structural feature descriptors, one of
which called shape context (SC) [12] has been successfully used to measure the similarity between the structure of point sets. In recent year, coherent distance SC (CDSC) [13], inner distance SC (IDSC) [14], rotation invariant SC (RISC) [15] and normalized weighted SC (NWSC) [16] have evolved from SC. However, these feature descriptors are developed for point set registration or shape matching. Besides, structures in IR images are likely to be missed or deformed in VIS images. Thus, a more powerful feature descriptor measuring the structural similarity between VIS and IR images to be aligned is required for image registration.

For image registration, spatial transformation models are used to represent the pattern of deformation between images to be aligned. Among these models, affine model [17] is a typical and widely-used linear model. However, it could not produce accurate alignment when there is deforming anisotropy between images in numerous applications, especially in multimodal image fusion. To address this problem, many control-point-based transformation models have been proposed, such as B-splines model [18], thin-plate spline (TPS) model [19] and a model within reproducing kernel Hilbert space (RKHS) [20].

Above-mentioned models could accurately describe the pattern of non-linear deformation between images to be aligned but with the weakness of heavy reliance on control points. In that the transformation parameters of control-point-based models are optimized in neighbourhoods of control points, the distribution and quantity of control points both have influence on the precision of global registration. Furthermore, control points or the ranges of selecting control points are regarded as preset parameters in general. But it is unpractical to select different optimal control points for various images in real applications. Therefore, the control-point-based models are limited by low adaptability and robustness.

In this work, our contributions include following two aspects. Firstly, we develop Gaussian weighted shape context (GWSC) to quantify the structural similarity between IR and VIS images. Secondly, by generalizing the affine model from linear case to non-linear case, adaptive enhanced affine transformation (AEAT) model is proposed to adaptively determine the regular pattern of global deformation between VIS and IR images to be aligned. In brief, the key idea of our method (GWSC-AEAT) is to estimate the AEAT model from putative matches extracted by GWSC. The experiment results indicate that compared to the state-of-the-art methods, GWSC-AEAT has better performance on non-rigid VIS and IR image registration. Consequently, it could be employed to raise the performance of IR and VIS image fusion applications.

The remainder of this paper is structured as follows. Section 2 reviews the most important works related to our work. Section 3 describes how to extract putative matches from edge maps by GWSC. Section 4 presents the estimation of the AEAT model and how to apply it to global image registration. Section 5 shows comparison results of GWSC-AEAT and state-of-the-art methods on real datasets, followed by analyses and discussions of experiment results. Finally, concluding remarks and weaknesses of our work are presented in Section 6.

## 2 RELATED WORKS

If considering image registration as a fitting problem, the distance measure quantifying registration performance is a fitting criterion. According to various distance measures, registration methods could be classified into intensity-based methods, spectral methods and feature-based methods.

As for intensity-based techniques, the distance measures, such as mutual information (MI) [21] and normalized mutual information (NMI) [22], are defined directly by image intensity. Also, MI-based measures can be formulated by image gradient [23] and image patch [24]. The assumption MI-based measures mainly rely on is that the statistical regularities of intensity are similar between images to be aligned. However, there is significant difference between intensity distributions of multimodal images. Therefore, intensity-based measures are not suitable for multimodal image registration [25]. What is more, intensity-based measures are generally formulated as non-convex functions, thereby it is difficult to minimize them quickly and efficiently [26].

When it comes to spectral methods, which represent image data structures by spectral decomposition, they are the upgrading of intensity-based measures. Typically, the first embedding coordinate of diffusion maps and Laplacian Eigen maps were used for feature representation of multimodal images [27]. Then, the joint graph obtained by spectral decomposition [25] and Laplacian commutativity [11] were developed to measure the similarity between multimodal images in high dimensional Eigen spaces. Because spectral-based measures established by $L_1$ or $L_2$ distance measures are convex functions, the solution of spectral-based measures is more concise than MI-based measures. However, the derivatives formulas of spectral-based measures are rather complex and difficult to be deduced. Thus, the development and real application of spectral methods are limited.

Concerning feature-based methods, they measure the performance of image registration by using the distance between match features of images to be aligned [28]. Image features representing the correspondence between images, such as points, curves and surfaces [29], are typically used as match features. Actually, point feature is more popular because curves and surfaces could be discretized into point sets. Although feature-based image registration could be regarded as point set registration, there is a difference between image registration and point set registration. That is the transformation model estimated from two point sets is only used to align these points to achieve point registration. However, as for image registration, the transformation model aligning two feature point sets is used to align all of the points in images. Therefore, the transformation models used for image registration have to represent the pattern of global deformation between images to be aligned.

Some typical models have been employed to construct feature-based measures, including $L_2$ loss criterion [30], $L_2$-minimizing estimator ($L_2$E) [31], regularized Gaussian field criterion (RGF) [20] and Gaussian mixture model (GMM) [32]. As a result of the convexity of feature-based measures, these...
models could be minimized by gradient-based techniques, such as the gradient descent method and the quasi-Newton method. In addition, one more important issue for feature-based measures is how to find enough match features from multimodal images to be aligned, for the reason that the accuracy of point matching directly affects the precision of the feature-based measure.

Point matching used for feature-based measures could be divided into two types, soft-assignment and hard-assignment. Generally speaking, feature-based measures using soft-assignment are established by an attribute matrix indicating the degree of matches between all pairs of points. During the estimation of transformation parameters, explicit correspondence is iteratively solved by the optimization with soft-assignment as initial matching. The typical methods include iterated closest point (ICP) [33], robust point matching using TPS (TPS-RPM) [30], coherence point drift (CPD) [34] and face registration using RGF [20]. Because explicit correspondence and transformation parameters are solved simultaneously, it is very likely to fall into local convergence and huge computation for the optimization of feature-based measures with soft-assignment.

Unlike soft-assignment, hard-assignment is to determine explicit matches before transformation estimation. Thus, the likelihood and the speed of reaching the global minimum of feature-based measures with hard-assignment are increased. Hard-assignment is commonly used in feature-based registration methods, such as robust point matching using \( L_2E \) estimator (RPM-\( L_2E \)) [31], robust point matching using manifold regularization (MR-RPM) [15], point registration using spatially constrained Gaussian fields (SCGF) [35] and image registration using affine and contrast invariant descriptor [36]. In general, feature-based measures with hard-assignment require a highly accurate method of point matching, the keys of which are feature descriptors and bipartite graph matching. Feature descriptors representing the similarity between two point sets have been discussed in Section 1. In addition, there are many typical approaches for bipartite graph matching, such as the Hungarian algorithm [37], the deferred-acceptance algorithm [38] and the shortest augmenting path algorithm [39].

3 PUTATIVE MATCH EXTRACTION BY GWSC

This section mainly describes the definition of GWSC and how to extract putative match from edge maps by GWSC. It should be noted that VIS images are considered as reference images in this paper, and thus our purpose is to align IR images to VIS images.

3.1 Problem formulation

The set of putative matches is represented as \( F = \{ (r_k, s_k) \}_{k=1}^K \), where \( r_k \) is an edge point in IR images and \( s_k \) is the putative matching edge point in VIS images, \( r_k, s_k \in \mathbb{R}^2, k \in \mathbb{N}^+ \). \( K \) denotes the number of putative matches. In this work, canny edge descriptor is employed to extract the edge maps from IR and VIS images.

A transformation model with parameters \( c \) is represented as \( \varphi_c \), which indicates a mapping from IR image space to VIS image space. The parameters \( c \) are estimated by aligning point set \( \{ r_j \}_{j=1}^J \) to \( \{ s_j \}_{j=1}^J \). This process could be considered as an optimization procedure. On this account, an objective function quantifying registration accuracy is essential.

In this work, Gaussian-fields-based model is chosen to construct the objective function since it is continuously differentiable and can converge to a global optimal solution quickly [20]. Therefore, the objective function is written as:

\[
\min_c E( c ) = \min_c - \sum_{k=1}^K \exp \left\{ - \frac{ \| s_k - \varphi_c(r_k) \|^2 }{ 2 \sigma_c^2 } \right\} + \lambda S( c ) , (r_k, s_k) \in F,
\]

where \( \| \cdot \| \) represents the \( L_2 \) norm, \( \sigma_c \) is a parameter. The first term measures the distance between two point sets. The second term \( S( c ) \) is a stabilizer that establishes control over the transformation. \( \lambda \in \mathbb{R} \) is a normalized weight balancing the two terms.

From Equation (1) we can see that the performance of the objective function is largely determined by the accuracy of putative matches. The reason is that incorrect point matches could not produce accurate measurement of registration performance. Moreover, compared with the estimation of transformation parameters, the computational cost of point matching is higher. Furthermore, IR and VIS image pairs with highly different intensity distribution and missing structures increase the difficulty of point matching. Therefore, in the next section, we would introduce a feature descriptor GWSC to improve the speed and accuracy of point matching between VIS and IR images.

3.2 GWSC

SC describing the neighbourhood structures of points is always used on point matching in edge maps. The SC-based similarity measure between two edge points of IR and VIS images is defined as:

\[
C_{s_{ij}}( \varphi_c ) = \frac{1}{2} \sum_{i=1}^{J} \left( \frac{ S_j( r_i' ) - S_j( s_i' ) }{ S_j( r_i' ) + S_j( s_i' ) } \right)^2 ,
\]

where \( \{ r_i' \}_{i=1}^J \) and \( \{ s_i' \}_{i=1}^J \) are the numbers of the edge points of IR and VIS images respectively. \( J \) and \( J \) are the numbers of the edge points of IR and VIS images respectively. \( S_j( \cdot ) \) represents the \( J \)-bin normalized histogram of an edge point and it is given by:

\[
S_j( b_j ) = \# \{ q \neq b_j : (q - b_j) \in \text{bin}( j ) \} ,
\]
For addressing the above problems, the feature descriptor GWSC is proposed to improve the discrimination between the points which are similar but not corresponding. GWSC is improved from SC and the cost matrix of GWSC is written as:

$$C_g = (D_r^{-1}W'CsW'D_r^{-1}) \ast W'',$$

where $W'$, $W''$ and $W'''$ are the Gaussian weight matrices and defined as follow:

$$W_j = \exp \left\{ -\frac{\varepsilon_j}{\omega_j} \| r_j - r_i \|^2 \right\}, \qquad i, j \in \{1, 2, ..., I\}, \quad \varepsilon_i, \omega_i \in \mathbb{R}^+$$

$$W_v = \exp \left\{ -\frac{\varepsilon_v}{\omega_v} \| s_v - s_i \|^2 \right\}, \qquad j, \nu \in \{1, 2, ..., J\}, \quad \varepsilon_j, \omega_j \in \mathbb{R}^+$$

$$W'' = \exp \left\{ -\frac{\varepsilon''}{\omega''} \| r'' - r_i \|^2 \right\}, \quad i \in \{1, 2, ..., I\}, \quad j, \nu \in \{1, 2, ..., J\}, \quad \varepsilon'' \in \mathbb{R}^+$$

where $\varepsilon_i$ and $\varepsilon_j$ control the range of interaction between points respectively. $\omega_j$, $\omega_v$ and $\omega''$ are the normalized coefficients, where $\omega_j = \max\{\|r_i - r_j\|^2, i = 1, 2, ..., I\}$, $\omega_v = \max\{\|s_i - s_j\|^2, i = 1, 2, ..., J\}$ and $\omega'' = \max\{\|r'' - r_i\|^2, i = 1, 2, ..., I\}, j = 1, 2, ..., J\}$. $D_r = \text{diag}(\sum_{j=1}^J W_j)$ and $D_v = \text{diag}(\sum_{i=1}^I W_i)$ are the diagonal degree matrices of $W'$ and $W_v$, respectively. $\ast$ denotes the Hadamard product.

The original SC measures the similarity between two points via the neighbourhood structure whereas that the radius of the neighbourhood of SC-based measure is too small to obtain correct point matching as shown in the top of Figure 2. However, increasing the radius could result in high computational complexity, and moreover, the SC-based measure might be easily disturbed by outliers. In the proposed GWSC, the first term in the bracket of Equation (4) is used to calculate the weighted average of the SC-based similarity between the neighbourhoods of $r_i$ and $s_i$, as shown in the bottom of Figure 2. Accordingly, it is possible to enlarge the scope of SC without increasing the radius of the SC-based similarity measure Equation (2).

The second term $W''$ indicates the relative distances between all pairs of edge points in $\{r_i\}_{i=1}^I$ and $\{s_j\}_{j=1}^J$. The assumption $W''$ relies on is that the distance between two points of an actual match in multimodal images to be aligned is not too large. In real applications, this could be easily achieved by controlling the displacement between the optical axes of VIS and IR cameras. In addition, the Gaussian weight matrices in the GWSC are used to reduce the negative impact of outliers on similarity measuring.

Therefore, $C_{gij}$ indicates the degrees of correspondence between all pairs of edge points in VIS and IR images. The lower $C_{gij}$ is, the more similar $r_i$ and $s_j$ are.

### 3.3 Extraction of putative matches

On the basis of the cost matrix, point matching is considered as bipartite graph matching so that the Hungarian method and its similar approaches are employed to determine putative matches between point sets. However, these traditional approaches are not suitable for extracting putative matches from edge maps of VIS and IR images. In most cases, a large number of points in
edge maps of real images sharply increase the computational cost of bipartite graph matching. Moreover, for implementing VIS and IR image registration, it is unnecessary to find the maximum matching between all pairs of edge points.

The major purpose of point matching employed for image registration is to extract putative matches from all pairs of edge points. Thus, the GWSC-based approach is designed for the extraction of putative matches and outlined as follows.

It could be known from Algorithm 1 that the point pairs with the lowest GWSC-based measure are extracted as putative matches so that the set $F$ indicates a potential correspondence between VIS and IR images. In addition, the time complexity of extracting putative matches from the cost matrix (i.e., the steps 3-5) is $O(|F|)$. Thus, compared with bipartite graph matching, Algorithm 1 has an advantage on calculation speed.

4 ESTIMATION OF THE AEAT MODEL

Transformation model $\varphi_p$ is a map function, which represents the pattern of deformation between two images to be aligned. This section presents the AEAT model and how to apply it for global image registration.

4.1 The AEAT model

The framework of the parallel optical axis is widely employed for VIS and IR image fusion systems. As a result, the causes of the deformation between VIS and IR images to be aligned include the displacement between two cameras, the differences between the lens of VIS and IR cameras, the distinctions between VIS and IR sensors etc. Actually, most of them could be formulated by a regular pattern. As a result, in this work, we make an assumption that the deformation pattern between a VIS and IR image pair consists of various regular patterns. In other words, the global deformation could be formulated by a mixed regular pattern.

Typically, the affine model is able to describe the global regularity of linear deformation. The smoothness of the affine model can enforce the consistency of spatial transformations for all points. Moreover, due to the regularity of the affine model, control points are not required. Hence, we believe that the mixed regular pattern of non-rigid deformation between VIS and IR images can be represented by generalizing the affine model from linear to non-linear case.

Let $r_k = [x_k, y_k]_T$, which is a $2 \times 1$ dimensional coordinate vector. Meanwhile, let $\varphi_c^{(a)}(r_k) = [x_k, y_k]_T$ denote the mapping coordinate vector by spatial transformation. The AEAT model is defined by:

$$
\varphi_c^{(a)}(r_k) = \begin{bmatrix} x_k \\ y_k \\ 1 \end{bmatrix} + P^{(a)}(x_k, y_k, \alpha^{(a)}, \beta^{(a)})
$$

where $A = [t_x, s_x \cos \theta - t_y, s_x \sin \theta, t_y, s_y \cos \theta, t_x]$ is the affine transformation matrix, $s_x$ and $s_y$ are scaling coefficients, $t_x$ and $t_y$ are translation coefficients, $\theta$ is the angle of rotation. From Equation (6) we can see that the first term is the traditional affine model. The second term $P^{(a)}$ is the enhanced part of the AEAT model, which is given by:

$$
P^{(a)}(x_k, y_k, \alpha^{(a)}, \beta^{(a)}) = \sum_{j=1}^{n} \sum_{j=1}^{n} \beta_{i,j} x_k^i y_k^j, \quad \alpha_{i,j}, \beta_{i,j} \in \mathbb{R}, i, j \in \mathbb{Z}_+^n
$$

where $P^{(a)}$ is the $2 \times 1$ dimensional vector of mixed polynomial transformation. $\alpha^{(a)}$ and $\beta^{(a)}$ are the $1 \times n_p(n_p = n(n + 3)/2)$ dimensional parameter vectors containing all of $\alpha_{i,j}$ and $\beta_{i,j}$ respectively. $n$ is the order of the AEAT model.

The first term of Equation (6) is used to describe the linear deformation pattern between two point sets. The second term is the mixed polynomial model, used for representing the regular pattern of non-linear deformation. Hence, the AEAT model is actually evolved from the affine transformation by generalizing the affine model from linear to non-linear case. Furthermore, the order of the AEAT model is easily adjusted via the parameter $n$. In other words, the AEAT model can be adapted for various non-linear degrees of non-rigid deformation between two point sets. Meanwhile, as a result of the smoothness and regularity of the affine and polynomial models, the AEAT model is smooth and has no requirement of any control points.

Substituting Equation (7) into (6), the AEAT model becomes the following matrix form:

$$
\varphi_c^{(a)}(r_k) = H_c^{(a)} \pi_k^{(a)},
$$

where $\pi_k^{(a)} = [x_k, y_k, 1]_T$ is the polynomial vector of the point $r_k$, $\pi_k^{(a)}$ is the $1 \times n_p$ dimensional vector which contains all of $x_k^i y_k^j$ in (7). $H_c^{(a)}$ denotes the AEAT matrix with the parameter
vector \( c \), which is written as follows:

\[
H^{(a)}_c = \begin{bmatrix}
A_l \begin{bmatrix} (x^{(a)})^T & (\beta^{(a)})^T \end{bmatrix}^T
\end{bmatrix},
\]  

(9)

where \( H^{(a)}_c \) is a \( 2 \times (n_p + 3) \) dimensional matrix. In this work, the parameter vector \( c^{(a)} \) of the AEAT model is represented as \([\theta, s_x, s_y, t_x, t_y, s_x x, s_y y]^{(a)} \beta^{(a)} \)\, which is a \((2n_p + 5) \times 1\) dimensional vector.

With Equation (8), the objective function Equation (1) becomes,

\[
\min_{c^{(a)}} E(c^{(a)}) = \min_{c^{(a)}} \sum_{k=1}^{K} \exp \left\{ -\frac{\|s_k - H_c^{(a)} \pi_k^{(a)}\|^2}{2\sigma^2} \right\} + \lambda \text{tr} \left( (c^{(a)} - z^{(a)}) (c^{(a)} - z^{(a)})^T \right),
\]  

(10)

where \( z^{(a)} = [0, 1, 1, 0, ..., 0]^T \) is the \((2n_p + 5) \times 1\) dimensional vector and \( \text{tr}(\cdot) \) denotes the trace. The stabilizer (i.e. the second term of Equation (10)), which describes the variation of the AEAT model during the optimization process, is used to prevent the spatial transformation from being out of control.

### 4.2 Optimization

With Gaussian fields, Equation (10) is a continuously differentiable function with respect to the transformation parameters \( c^{(a)} \). As a consequence, we could give the matrix form of the corresponding derivative:

\[
\frac{\partial E(c^{(a)})}{\partial c^{(a)}} = \frac{1}{\sigma^2} \sum_{k=1}^{K} \frac{\partial H^{(a)}_c \pi_k^{(a)}}{\partial c^{(a)}} \left( H^{(a)}_c \pi_k^{(a)} - s_k \right)^T \times \exp \left\{ -\frac{\|s_k - H^{(a)}_c \pi_k^{(a)}\|^2}{2\sigma^2} \right\} + 2\lambda (c^{(a)} - z^{(a)})^T,
\]  

(11)

where the matrix form of \( \partial H^{(a)}_c \pi_k^{(a)}/\partial c^{(a)} \) is written as:

\[
\begin{bmatrix}
-\partial x \sin \theta - y \cos \theta & -s_x \sin \theta & x \cos \theta - s_y \sin \theta \\
-x \cos \theta & 0 & 0 \\
0 & y \cos \theta & 0 \\
0 & 1 & 0 \\
\pi_k^{(a)}^T & 0 \\
0 & \pi_k^{(a)}^T
\end{bmatrix},
\]  

(12)

where 0 is the \( n_p \times 1 \) dimensional zero vector. On the basis of the derivative Equation (11), the optimization of Equation (10) can be solved by the quasi-Newton method, before which we have to figure out how to determine the optimal order \( n \) of the AEAT model.

The purpose of optimization is to estimate the optimal AEAT model from the putative matches, which could be considered as a fitting problem in our work. The putative matches extracted by Algorithm 1 are the known data points of the spatial transformation model. The AEAT model is a fitting function based on polynomial. The objective function is an evaluation criterion. Therefore, the order of the AEAT model directly impacts on the performance of optimization. To be more specific, the AEAT model with low order could not represent the complex deformation pattern between multimodal images. But on the other hand, due to the mixed polynomial model \( P^{(a)} \), the AEAT model with high order is sometimes more accurate but suffers from Runge phenomenon. Figure 3 shows the examples of Runge phenomenon caused by the AEAT model with order 4. There are many distortions in the registration results by the AEAT model with excessively high order.

Therefore, to determine the optimal order and parameters of the AEAT model, we develop a strategy of subsection optimization. It can be seen in Algorithm 2.

As to be seen in Equation (6), the proposed AEAT model consists of two parts: the affine model and the mixed polynomial model. Since the affine model is a linear transformation, it is only used to coarsely describe the regular pattern of non-rigid deformation between images to be aligned. The mixed polynomial model with high order is able to finely represent the regular pattern of deformation due to high non-linearity. Therefore, in the estimation of the AEAT model (Algorithm 2), we use the affine model for coarse registration and then mainly use the mixed polynomial model for fine registration.

Although the mixed polynomial model with order 1 seems similar to the affine model, there is a difference between the parameters of the two models. It is apparent the parameters of the mixed polynomial model can be optimized without any constraint in the estimation of the AEAT model. If the affine model is considered as a linear polynomial, the polynomial parameters are the affine transformation matrix \( A = [s_x \cos \theta, -s_y \sin \theta; s_x \sin \theta, s_y \cos \theta; \sin \theta, s_y \cos \theta] \). We can see that except for \( t_x \) and \( t_y \), the polynomial parameters of the affine model are constrained within reasonable ranges. This is very helpful for avoiding over-fitting during optimization. Thus, the estimation of the affine model can be considered as a constrained optimization problem, while the estimation of the mixed polynomial model is an unconstrained optimization problem.

In the initial case, the distance between images to be aligned is large. The optimization with large initial error typically suffers from an over-fitting problem. Thus, as shown in Algorithm 2, we mainly use the affine model for constrained optimization in the first round, in order to avoid the over-fitting problem and achieve coarse registration. Then, by using the result of coarse registration as an initial value, the mixed polynomial model with the optimal order is estimated iteratively to achieve fine registration.
4.3 Computational complexity

Let $N = n_p + 3$. The time costs of the objective function Equation (10) and the derivative Equation (11) are both $O(KN)$. In general, the time cost of the quasi-Newton for solving the $(2n_p + 5) \times 1$ dimensional parameter vector $c^{(o)}$ is approximately $O(n^2_p)$ [20]. Thus, the total time cost of Algorithm 2 in one iteration is $O(KN + n^2_p)$ and the space complexity of Algorithm 2 is $O(NK + N)$ with the requirements of storing $\{\hat{r}^{(o)}_k\}_{k=1}^N$ and $H^{(o)}_c$.

By examining Equation (8), we could see that the time cost of non-rigid image transformation with the AEAT model is $O(NL)$. Meanwhile, the space complexity is $O(NL)$ since the size of $\{\hat{r}^{(o)}_k\}_{k=1}^L$ which needs to be stored is $N \times L$.

5 | EXPERIMENT

Firstly, GWSC was evaluated for point matching. An ablation study was then performed on the AEAT model. Finally, our method was tested on real IR and VIS images and compared with the state-of-the-art algorithms. In this work, the algorithms were implemented in Matlab and ran on a quad-core CPU (3.9 GHz) with 4 GB RAM.

5.1 Dataset

In our experiments, we used a dataset consisting of 21 pairs of real VIS and IR images. Examples of the dataset (selected from CVC datasets [17]) can be seen in Figure 4. The difference between the intensity distributions of the VIS and IR images raises difficulties for putative match extraction. Moreover, to make the non-rigid registration harder, we increased the deformation degree between the VIS and IR images. Thus, GWSC-AEAT can be effectively assessed on our dataset.

In addition, we manually constructed a set of point matches as ground truth for each VIS and IR image pair. By using ground truth, the metric such as recall could be employed to the quantitative evaluation in our work. Consider an actual match $(r_g, s_g)$ in the ground truth. The transformation result of $r_g$ is denoted as $\hat{r}'_g$. If the Euclidean distance $||r'_g - s_g||$ is less than a given threshold (e.g. 5 pixels), the actual match $(r_g, s_g)$ is considered to be aligned correctly. Hence, we define the recall as the ratio of the total numbers in actual matches aligned correctly to actual matches in the ground truth.
5.2 | Parameter setting

In our method, there are four parameters to be set: $\varepsilon_s$, $\varepsilon_{rv}$, $\sigma_e$ and $\lambda$. The parameters $\varepsilon_s$ and $\varepsilon_{rv}$ are used for GWSC, while $\sigma_e$ and $\lambda$ are employed in the objective function Equation (10). The parameter $\varepsilon_s$ is used to control the neighborhood size of the weighted average of SC-based similarity between two point sets. The range of the relative distances between two point sets is adjusted by the parameter $\varepsilon_{rv}$. $\sigma_e$ is a range parameter which controls the scale of the objective function and $\lambda$ is used to balance two terms of the objective function Equation (10). In this work, we set $\varepsilon_s = 80$, $\varepsilon_{rv} = 300$, $\sigma_e = 6$ and $\lambda = 0.02$. The influence of the parameter setting can be seen in Figure 5. These parameters are the optimal settings determined through multiple experiments, which are kept constant throughout this experiment.

In addition, the codes of the other approaches which are used to compare with the proposed method were provided by their authors. According to the original papers, the corresponding parameters were tuned to find the optimal settings.

5.3 | Evaluation of GWSC

In our work, GWSC is used to point matching in the edge maps of VIS and IR images. Thus, we did not compare GWSC with some SC-based descriptors developed for shape matching, such as CDSC [13], IDSC [14] and NWSC [16]. Here, GWSC was tested on point matching of ground truth and compared with the original SC [12] and RISC [15]. In addition, to show the differences among the various feature descriptors clearly, the Hungarian method [37] is uniformly used for SC, RISC and GWSC to achieve bipartite graph matching.

The illustrations of point matching by the original SC, RISC and GWSC are shown in Figure 6, where the coloured lines indicate the point matches. Meanwhile, the quantitative comparisons on our dataset are reported in Figure 7. The total average recalls of SC and RISC are 0.5411 and 0.4046 respectively, while that of GWSC is 0.8714. The qualitative and quantitative results both show that GWSC has a good performance on point matching between VIS and IR images. Furthermore, compared with the original SC, GWSC improves the accuracy of point matching by 61.04%. Hence, we can see that GWSC is successfully used to point matching between multimodal images.

In addition, we also tested the runtimes of the Hungarian method with SC, RISC and GWSC. The total average runtimes of SC and RISC are 2.33 and 2.68 min separately, while that of GWSC is 0.47 min. This shows that the augmenting paths of the bipartite graph constructed by GWSC are less compared to SC and RISC. Hence, it is proved again that GWSC is able to describe the similarity between two point sets more accurately and robustly.

Figure 8 shows the illustrations of putative match extraction by Algorithm 1, where the coloured lines indicate the putative matches. We can see that most of the putative matches are in line with the actual situation. Therefore, on the basis of the putative matches extracted by Algorithm 1, the objective function Equation (10) is capable of quantifying the precision of VIS and IR image registration accurately.

5.4 | Ablation study of the AEAT model

At first, for evaluating the AEAT model comprehensively, we constructed a new version of our method by replacing the AEAT model with the affine model. The qualitative results of the affine model and the AEAT model are shown in Figure 9, where the optimal orders of the AEAT models are 2, 3, 2 and 3, respectively. The figures on the top are the registration results of edge maps, where the blue and red lines denote the edges of VIS and IR images separately. The figures on the bottom are the fusion results of the VIS and IR images aligned by the algorithms. It could be concluded from the comparison results that
the AEAT model deals with the non-linear deformation which the affine model cannot handle. Meanwhile, the AEAT model with appropriate order preserves the advantages of the affine model so that it can describe linear deformation correctly and does not cause Runge phenomenon.

Secondly, the AEAT model with adaptive order was compared to that with fixed orders. Figure 10 shows the total average recalls and matching errors of the AEAT models with various orders, where AEAT-2, -3, -4 and -5 separately represent the AEAT models with order 2, 3, 4 and 5. From comparison results, we can see that the AEAT model with adaptive order raises the accuracy of non-rigid registration. This demonstrates that the strategy of subsection optimization with the AEAT model is able to improve the possibility of reaching a global optimal solution. It is also a critical factor in helping our method to make the robust alignment.

5.5 Comparison with state-of-the-art methods

The proposed GWSC-AEAT was further tested and compared with CPD [33], MR-RPM [15], RGF [20], RPM-L2E [31] and SC-TPS [12] which are the state-of-the-art methods belonging to point-feature-based registration algorithm. Hence they are similar with GWSC-AEAT on the basic framework. To make a fair comparison, the feature point sets extracted by Algorithm 1 are applied as feature points in all of these methods. The qualitative comparison can be seen in Figure 11 and the quantitative results are reported in Figure 12.

Firstly, because all of the registration results in Figure 11 are implemented with the feature point sets extracted by Algorithm 1, it is proved that Algorithm 1 can find enough point features for estimation of transformation. Meanwhile, this demonstrates again that the feature descriptor GWSC works fine on point matching between VIS and IR images.

Secondly, it is obvious in Figure 11 and Figure 12 that there are less distortion and mismatch in the registration results of GWSC-AEAT. Moreover, our method is more accurate on global image registration and has better recall curves in most
cases than the other methods. The total average matching errors of SC-TPS, RPM-L2E, RGF, MR-RPM, CPD are 10.08, 6.22, 3.91, 12.13 and 5.55 pixels separately, while that of GWSC-AEAT is 2.72 pixels. These show that the AEAT model estimated from the point feature improves the accuracy of the global registration of IR and VIS images by 30.27% at least. Meanwhile, it is proved that the AEAT model with adaptive order has higher generalization performance than the non-rigid transformation models used in the other approaches.

Thirdly, in [15], MR-RPM has better performance of point set registration than SC-TPS, RGF and CPD, but it fails in image registration in this work. This is caused by the difference between point set registration and image registration. The transformation model aligning two point sets is estimated from these points. But, the transformation model estimated from feature point sets is required to align all pixels of images. Hence, the transformation model used for image registration needs to represent the global deformation pattern between images to be aligned. Based on the above analyses, the failure of MR-RPM on image registration proves that the transformation model estimated by MR-RPM cannot describe global deformation between images to be aligned. Meanwhile, the comparison results demonstrate that the deformation pattern described by the AEAT model has highest accuracy and can be successfully used to global image registration.

5.6 Runtimes

The runtime of GWSC-AEAT is tested on real images. Image registration by GWSC-AEAT consists of three steps: putative match extraction by GWSC, estimation of the AEAT model and global image transformation. Hence, the runtime of our method can be divided into three parts.

When the Hungarian method with GWSC is used to extract putative matches from two point sets each containing 1000 edge points extracted from VIS and IR images, the average computation time is about 10 min. At the same conditions, the average runtime of Algorithm 1 is about 1.31 min. Compared with Hungarian method, the steps 3–5 of Algorithm 1 can reduce the runtime by 86.9%. Undoubtedly, the robustness of GWSC is the internal factor of improving the extraction speed.
FIGURE 11  The qualitative registration results by SC-TPS, RPM-L2E, RGF, MR-RPM, CPD and our method (GWSC-AEAT). In each group, the figures on the top are the registration results of edge maps of VIS and IR images, where the blue and red lines separately represent the edge maps of VIS and IR images. The figures on the bottom are the fusion results of VIS and IR images, which have been aligned by algorithms. The arrows highlight regions where there are significant differences between the registration results by various approaches.

Since RGF has the second-best registration accuracy, our method was compared with RGF on the runtimes of transformation estimation and image transformation. On our dataset, the total average runtime of estimation of the AEAT model is about 5.91 s, while that of transformation estimation in RGF is about 36.81 s. These test results agree very well with the analysis of computational complexity in Section 4.3. It is proved that the set of putative matches reduces the time cost of the objective function greatly.

The total average runtime of image transformation by the AEAT model is about 0.0544 s, while that of global image transformation in RGF is about 0.0483 s. According to the analysis of computational complexity in Section 4.3 and [20], we found that the time cost of image transformation is mainly determined by the length of the transformation parameter vector. In the AEAT model, the lengths of the parameter vectors $c^{(2)}$ and $c^{(3)}$ is 15 and 23. In this experiment, the number of control points used in RGF is set to be 15 so that the length of the transformation
FIGURE 12  The quantitative results of SC-TPS, RPM-L_E2, RGF, MR-RPM, CPD and our method (GWSC-AEAT). The numbers in the legend are the average matching errors.

FIGURE 13  IR images with low resolution from the TNO dataset and the work [40].

The parameter vector is 15. Thus, the global image transformation of our method is slightly slower than that of RGF. Overall, the average runtime of non-rigid image registration by GWSC-AEAT on our dataset is about 1.5 min.

6 CONCLUSION AND DISCUSSION

In this paper, GWSC-AEAT is proposed for VIS and IR image registration. At first, the feature descriptor GWSC is developed to measure the structural similarity between multimodal images, and to extract putative matches from edge maps. Then, the AEAT model is designed to represent the deformation regularity between VIS and IR images. Finally, the strategy of sub-section optimization is employed to efficiently determine the optimal AEAT model from putative matches. The experiment results demonstrate that GWSC-AEAT outperforms the state-of-the-art approaches and has the advantages of high precision and low cost.

Admittedly, GWSC-AEAT performs excellently in our experiments evaluated from many aspects, but it might still not work for some particular images since it utilizes image edge as the feature. When IR images with low resolution have less and blurred detail textures as Figure 13, GWSC is invalid so that it is difficult to find enough putative matches correctly for the estimation of the AEAT model, which is also the common problem of feature-based registration approaches [11].

One of the ideas to improve the above situation is that if semantic segmentation performs on IR and VIS images to be aligned, it would provide robust feature to putative match extraction and improve the efficiency of mutual feature extraction from low resolution images. Fortunately, many excellent methods [41–43] have been proposed for foreground object extraction in IR and VIS images. Hence, in future, we mainly focus on improving GWSC-AEAT via semantic segmentation techniques.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from the National Natural Science Foundation of China (No. 61901157).

ORCID

Chaobo Min https://orcid.org/0000-0002-6515-7063

REFERENCES

1. ArceSantana, E.R., et al.: Multimodal image registration based on the expectation-maximization methodology. IET Image Proc. 11(12), 1246–1253 (2017)
2. Yuan, X., et al.: Improved visual/infrared colour fusion method with double-opponency colour constancy mechanism. IET Image Proc. 12(9), 1560–1566 (2018)
3. Yelampalli, P.K.R., et al.: Daubechies wavelet-based local feature descriptor for multimodal medical image registration. IET Image Proc. 12(10), 1692–1702 (2018)
4. Li, H., et al.: Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing. Remote Sens. 9, 441 (2017)
5. Piella, G.: A general framework for multiresolution image fusion: From pixels to regions. Inf. Fus. 4(4), 259–280 (2003)
6. Borovec, J., et al.: ANHIR: Automatic non-rigid histological image registration challenge. IEEE Trans. Med. Imaging 39(10), 3042–3052 (2020)
7. Xiang, Y., et al.: OS-SIFT: A robust SIFT-LIKE algorithm for high-resolution optical-to-SAR image registration in suburban areas. IEEE Trans. Geosci. Remote Sens. 56(6), 3078–3090 (2018)
8. He, S., Medical image registration method based on tensor voting and Harris corner point detection. J. Med. Imaging Health Inf. 8(3), 583–589 (2018)
9. Rashwan, H.A., et al.: Using curvilinear features in focus for registering a single image to a 3D object. IEEE Trans. Imaging Process. 28(9), 4429–4443 (2019)
10. Cao, W., et al.: Multimodal medical image registration based on feature spheres in geometric algebra. IEEE Access 6, 21164–21172 (2018)
11. Zimmer, V.A., et al.: Multimodal image registration using Laplacian commutators. Inf. Fus. 49, 130–145 (2019)
12. Belongie, S., et al.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24(4), 509–522 (2002)
13. Rong, X., et al.: Hand shape recognition based on coherent distance shape contexts. Pattern Recognit. 45, 3348–3359 (2012)
14. Ling, H., Jacobs, D.W.: Shape classification using the inner-distance. IEEE Trans. Pattern Anal. Mach. Intell. 29(2), 286–299 (2007)
15. Ma, J., et al.: Nonrigid point set registration with robust transformation learning under manifold regularization. IEEE Trans. Neural Networks Learn. Syst. 30(12), 3584–3597 (2019)
16. Zhang, J., et al.: Normalized weighted shape context and its application in feature-based matching. Opt. Eng. 47(9), 097201 (2008)
17. Liu, X., et al.: Robust and fast registration of infrared and visible images for electro-optical pod. IEEE Trans. Ind. Electron. 66(2), 1335–1344 (2018)
18. Sun, W., et al.: Randomly perturbed B-splines for non-rigid image registration. IEEE Trans. Pattern Anal. Mach. Intell. 39(7), 1401–1413 (2017)
19. Chen, Y., et al.: 3D craniofacial registration using thin-plate spline transform and cylindrical surface projection. PLoS One 12(10), e0185567 (2017)
20. Ma, J., et al.: Non-rigid visible and infrared face registration via regularized Gaussian fields criterion. Pattern Recognit. 48, 772–784 (2014)
21. Viola, P., Wells, W.M.: Alignment by maximization of mutual information. Int. J. Comput. Vision 24, 137–154 (1997)
22. Studholme, C., et al.: An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognit. 32(1), 71–86 (1999)
23. Pradhan, S., Patra, D.: Enhanced mutual information based medical image registration. IET Image Proc. 10(5), 418–427 (2016)
24. Chen, Y., Lin, C.: PCA based regional mutual information for robust medical image registration. In: 8th International Symposium on Neural Networks. Guilin, China, pp. 355–362 (2011)
25. Bansal, M., Danilidis, K.: Joint spectral correspondence for disparate image matching. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Portland, pp. 2802–2809 (2013)
26. Haber, E., Moderitzki, J.: Intensity gradient based registration and fusion of multi-modal images. Methods Inf. Med. 46(3), 292–299 (2007)
27. Piella, G.: Diffusion maps for multimodal registration. Sensors 14(6), 10562–10577 (2014)
28. Feng, R., et al.: Robust registration for remote sensing images by combining and localizing feature- and area-based methods. ISPRS J. Photogramm. Remote Sens. 151, 15–26 (2019)
29. Zitová, B., Flusser, J.: Image registration methods: A survey. Image Vision Comput. 21, 977–1000 (2003)
30. Chuia, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration. Comput. Vision Image Understanding 89, 114–141 (2002)
31. Ma, J., et al.: Robust $L_2$E estimation of transformation for non-rigid registration. IEEE Trans. Signal Process. 63(5), 1115–1129 (2015)
32. Ma, J., et al.: Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration. Inf. Sci. 417, 128–142 (2017)
33. Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. IEEE Trans. Pattern Anal. Mach. Intell. 14(2), 239–256 (1992)
34. Myronenko, A., Song, X.: Point set registration: Coherent point drift. IEEE Trans. Software Eng. 32(12), 2262–2275 (2010)
35. Wang, G., et al.: Robust non-rigid point set registration using spatially constrained Gaussian fields. IEEE Trans. Image Processing 26(4), 1759-1769 (2017)
36. Liu, X., et al.: A novel affine and contrast invariant descriptor for infrared and visible image registration. Remote Sens. 10, 658 (2018)
37. Christos, H.: Combinatorial optimization: Algorithms and complexity. IEEE Trans. Acoust. Speech Signal Process. 32(6), 1258–1259 (1982)
38. Elhenaw, M., Rakha, H.: A heuristic for rebalancing bike sharing systems based on a deferred acceptance algorithm. In: 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems. Naples, Italy, pp. 188–193 (2017)
39. Jonker, R., Volgenant, A.: A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing 38, 325–340 (1987)
40. Doulamis, N., et al.: Human object detection using very low resolution thermal cameras for urban search and rescue. In: PETRA’17 Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments. Island of Rhodes, Greece, pp. 311–318 (2017)
41. Davis, J.W., Sharma, V.: Background-subtraction in thermal imagery using contour saliency. Int. J. Comput. Vision 71(2), 161–181 (2007)
42. Davis, J.W., Sharma, V.: Background-subtraction using contour-based fusion of thermal and visible imagery. Comput. Vision Image Understanding 106, 162–182 (2007)
43. Makantasis, K., et al.: Data-driven background subtraction algorithm for in-camera acceleration in thermal imagery. IEEE Trans. Circuits Syst. Video Technol. 28(9), 2090–2104 (2017)

**How to cite this article:** Min C., Gu Y., Li Y., Yang F. Adaptive enhanced affine transformation for non-rigid registration of visible and infrared images. *IET Image Process.* 2021;15:1144–1156. https://doi.org/10.1049/ipr2.12093