Pattern Recognition Letters 34 (2013) 42–51

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Visible and infrared image registration in man-made environments employing hybrid visual features ☆

Jungong Han *, Eric J. Pauwels, Paul de Zeeuw

Centrum Wiskunde & Informatica (CWI), Science Park 123, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Available online 4 April 2012

Keywords:
Image registration
Line detection
Geometric analysis
Local deformation

ABSTRACT

We present a new method to register a pair of images captured in different image modalities. Unlike most of existing systems that register images by aligning single type of visual features, e.g., interest point or contour, we try to align hybrid visual features, including straight lines and interest points. The entire algorithm is carried out in two stages: line-based global transform approximation and point-based local transform adaptation. In the first stage, straight lines derived from edge pixels are employed to find correspondences between two images in order to estimate a global perspective transformation. In the second stage, we divide the entire image into non-overlapping cells with fixed size. The point having the strongest corner response within each cell is selected as the interest point. These points are transformed to other image based on the global transform, and then used to bootstrap a local correspondence search. Experimental evidence shows this method achieves better accuracy for registering visible and long wavelength infrared images/videos as compared to state-of-the-art approaches.

1. Introduction

Recent advances in imaging, networking, data processing and storage technology have resulted in an explosion in the use of multimodality images in a variety of fields, including video surveillance, urban monitoring, cultural heritage area protection and many others. The integration of images from multiple channels can provide complementary information and therefore increase the accuracy of the overall decision making process. A fundamental problem in multimodality image integration is that of aligning images of the same scene observed from different positions and/or in different sensor modalities. This problem is known as image registration and the objective is to recover the correspondences between the images. Once such correspondences have been found, all images can be transformed into the same reference, enabling on augmenting of the information in one image with the information from the others.

1.1. Prior work on image registration

Several related survey papers for image registration have appeared over the years. Brown (1992), Zitova and Flusser (2003) and Xiong and Zhang (2010) have provided a broad overview of over two hundred papers for registering different types of sensors. Before embarking on a more in-depth discussion of some of the related prior work, we point out that in accordance with most of the literature, we also divide existing techniques into two categories: pixel-based methods and feature-based methods. Pixel-based methods first define a metric, such as the sum of squared differences or mutual information (Zitova and Flusser, 2003), which measures the distance of two pixels from different images. The registration problem is then recast as the total distance minimization between all pixels in one image and the corresponding pixels in the other image. In feature-based methods, interest points like Harris corners, scale invariant feature transform (SIFT), speed-up robust feature (SURF), etc., are first extracted from images. Subsequently, these features are matched based on metrics, such as cross correlation or mutual information. Once more than four feature correspondences are obtained, the projective transform can be computed. In principle, a pixel-based method is better than a feature-based method, because the former takes all pixels into account when minimizing the cost function, while the latter minimizes the cost function based on a part of pixels only. In practice however, feature-based method performs well in many applications, because interest points are supposed to be distinctive, thus leading to better matching. Moreover, pixel-based methods are much more expensive than feature-based algorithms in the sense that every pixel needs to be involved in the computation. Considering both accuracy and efficiency of the algorithm, we adopt the feature-based method in this paper. Therefore, we limit our review to feature-based registration methods, and pay special attention to visible (ViS) and infrared (IR) image registration.

☆ This work is supported by EU-FP7 FIRESENSE project.
* Corresponding author. Tel.: +31 205924209.
E-mail address: j.han@cwi.nl (J. Han).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.03.022

Many approaches have been proposed for automatically registering IR and ViS images. In (Hrkac et al., 2007), an approach developed for aligning IR and ViS images is presented, in which the corner points are used. The similarity between the ViS corners and corners from IR image is measured by directed partial Hausdorff distance. Firmenich et al. (2011) employ the *multispectral corner detector* which aims to improve the quality of interest point extraction. The new method generalizes the Harris detector by summing autocorrelation matrices per band. In (Jarc et al., 2007), the image is first processed by using laws texture coefficient, which combines four one-dimensional filters. Each filtered image is then converted to a sort of texture-like energy image. The image alignment is conducted based on measuring/optimizing mutual information of two quantized texture energy images. Edge/gradient information is a popular feature as their magnitudes (Lee et al., 2010) and orientations (Firmenich et al., 2011; Kim et al., 2008) may match between infrared and visible images. In (Coiras et al., 2000), authors first extract edge segments, which are then grouped to form triangles. The transform can be computed by matching triangles from source to destination images. Huang and Chen (2002) proposes a contour-based registration algorithm, which integrates the invariant moments with the orientation function of the contours to establish the correspondences of the contours in the two images. Normally it is difficult to obtain accurate registration by using contour-based method, because precisely matching all contours detected from two images is challenging. Moreover, this method drastically increases computation time compared to interest point-based registration. To improve this work, Han and Bhanu (2007) propose to find correspondences on *moving* contours. They extract silhouettes of moving humans from both images. Matching only the contours of human bodies significantly improves both the performance and the efficiency of the algorithm. An alternative (Caspi et al., 2006) is to make use of the object motion paths generated by object tracking algorithm. Finding correspondences between trajectories helps to align images. This type of algorithm works very well when moving objects can be precisely tracked from both channels. Unfortunately, the current tracking algorithm is not satisfactory in many applications.

1.2. Problem statement

Most publications in this area focus on solving three problems: (1) feature extraction that guarantees that the majority of features in both images are identical; (2) feature descriptors that ensure accurate feature matching across different images; (3) transformation model that considers both global consistency and local deformation. The first two problems are more challenging when dealing with images captured by different types of cameras. When matching images from the visible and infrared part of the spectrum, the properties of cameras are completely different due to the difference in the electromagnetic wavelengths. The pixel brightness in IR images is determined by the temperature and emissivity property of the objects in the scene. However, in the visible spectrum, the brightness of image pixels is mainly influenced by light reflected on the object. Therefore, the pixel intensities in IR and ViS images have in general no direct relationship, increasing the difficulty of extracting and matching identical feature points in the two images. To illustrate this statement, in Fig. 1(a), we extract equivalent number of interest points from both IR and ViS images exploring two popular algorithms, where SURF method enables a scale- and rotation-invariant interest point detection but Harris method focuses on detecting corner points on the single scale. As apparent from the results, the majority of extracted interest points is unfortunately shared among the two images. To explain this feature matching failure, we show statistics (see Fig. 1(b)) of image patches (15×15) surrounding two corresponding points. We compute the distribution (normalized histogram) of three image characteristics within the image patch, viz. intensity value, gradient magnitude and gradient orientation. They are all feature descriptors widely used for IR and ViS image registration in the literature. To obtain a good feature matching result, we expect the feature distributions around two corresponding points to be similar. Unfortunately, none of them is capable of highlighting the correspondence between the two points in this case, although the gradient orientation clearly outperforms the others. This example illustrates that comparing image patches may not be a reliable way to extract correspondences between long wavelength IR and ViS images.

1.3. Our contributions

In order to address the three problems mentioned above, we propose a new algorithm here, which differs from existing work in two aspects. Basically, we use a two-stage procedure in which we first use line features to establish a global but approximate transformation between the two images. This global transformation is then used to bootstrap a more accurate, locally adaptive transformation that is based on a windowed optimization of feature point matching. More precisely, straight lines derived from boundaries of objects can easily be extracted and matched as the orientations of corresponding lines in two images are more or less the same. Therefore, an initial global (projective) transformation based on aligning a small number of lines ($\geqslant 4$) can be quickly estimated. We then build on that to find more accurate feature correspondences, we extract interest points from one image, and transform them to another image using the initial transformation. Searching around the initial corresponding point enables to find *accurate* correspondence effectively. Hence, although we start from a global transformation estimation, we end up with a local transformation computation.

This approach has two advantages. On one hand, we can partially solve the problem that the global transformation cannot model the local deformation between images. On the other hand, local adaption initialized by a global optimization is more robust as it will avoid the risk of settling for a local minimum, a fate that often befalls local transformation starting from the scratch.

Like many approaches using a strong assumption, our approach also has its own limitation. We always assume that there are more than four corresponding straight lines extracted from both images, so that the approximation of an initial transformation can be obtained. However, this may be invalid when we deal with the natural scenes, such as forest, in which we probably cannot extract a sufficient number of lines. Therefore, we focus on images captured in man-made environments, in which line-like structures are usually plentiful.

In Section 2, we first outline our methodology and mathematical model. In Section 3, we present the implementation of our line-based global transformation computation, where several key algorithms, such as line reorganization, line initial matching and line-configuration computing are introduced. In Section 4, we describe the interest point-based local transformation estimation. The experimental results are provided in Section 5. Finally, Section 6 draws conclusions and addresses our future research.

2. Overview of our methodology

The goal of image registration is to match two or more images so that identical coordinate points in these images correspond to the same physical region of the scene being imaged. To make our explanation simple, we assume that there are only two images

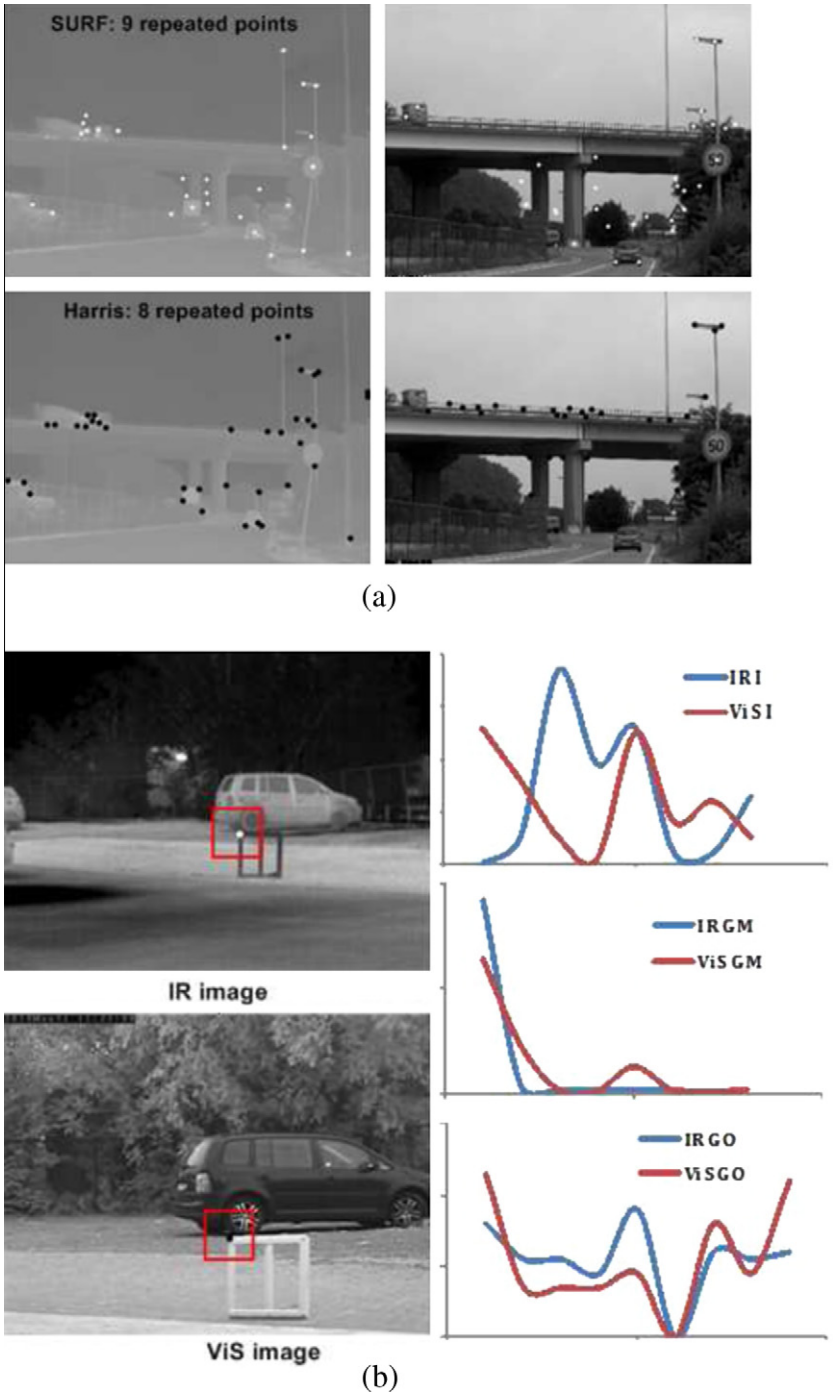


Fig. 1. Image statistics that imply problems for IR and ViS image registration. (a) Interest point detection for IR and ViS images. Two different methods: SURF and Harris corner detection are used. (b) Left: original images and manually labeled corresponding points (white dot and black dot). Right: from the top to the bottom, the distribution (normalized histogram) of Intensity (I) values, the distribution of Gradient Magnitudes (GM), and the distributions of Gradient Orientations (GO), respectively.

involved in the registration. In fact, the registration is to find a mathematical transformation model between the two image planes which minimizes the *energy* function of image matching. This optimization procedure can be described mathematically

$$\widetilde{\boldsymbol{H}} = \arg\min_{\boldsymbol{H}} \sum_i E(p_i, \boldsymbol{H}p_i'). \tag{1}$$

Here, p_i is the ith pixel in the image I and p_i' is its corresponding pixel in the image I'. The energy function is to measure the *distance* between I and the transformed version of I' based on $\boldsymbol{H}$. This transformation helps to establish a mapping between the two image planes, transforming a position p in one plane to the coordinate p' on another plane. In this paper, we assume a 2D projective transformation. Writing positions as homogeneous coordinates, the transformation $p = \boldsymbol{H}p'$ equals

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} u' \\ v' \\ w' \end{pmatrix}. \tag{2}$$

Homogeneous coordinates are scaling invariant, reducing the degrees of freedom for the matrix $\boldsymbol{H}$ to eight. In order to determine the eight parameters, at least four point-correspondences between the two images have to be found. In the literature, most publications rely on interest (corner) points for establishing point-correspondences.

If all detected interest points are involved in the computation of (1), there will be only one transformation between images, which is called *global* transformation. Alternatively, the image can also treated as a composition of patches, where each patch is matched to the corresponding patch in another image. The transformation $\boldsymbol{H}$ between two corresponding patches can also be estimated using (1), for which the involved interest points are restricted to lie within the patch. Hence, the overall transformation between the images is composed of many local transformations, each with different parameters. The global transformation has the advantage of having a relatively small number of parameters to be estimated, and the global nature of the model ensures a consistent transformation across the entire image, while the disadvantage is that one global mapping cannot properly handle images deformed locally. On the contrary, the local transformation is able to handle local deformation. However, its computational load is heavy due to the larger number of parameters that need to be estimated. Additionally, it is not easy to guarantee global consistency.

To own the benefits from both transformations, we try to combine them in our framework. The algorithm is carried out in two stages, as depicted in Fig. 2. In the first stage, we estimate a global perspective transformation by aligning straight lines derived from edges of the images. These lines strongly relate to boundaries of objects, which often appear in both images though IR sensor and ViS sensor have significantly different properties. We have chosen a perspective (projective) model for the transformation between the two images as this is the appropriate exact transformation whenever the cameras are observing a planar scene from different viewpoints and viewing angles. It is also an excellent approximation whenever the cameras are observing a 3D scene in which the depth difference between the objects is small compared to the distance to the cameras. Fortunately, this assumption usually holds in our application. In the second stage, we divide the entire image into cells of fixed size. The most salient interest point is extracted from each cell, and is transformed to another image based on the global transformation matrix. We allow the interest point to find a better correspondence within a window surrounding its initial corresponding point. By doing so, we can cope with the local geometric differences between images. Since the size of the searching window is limited, the estimated local transformation will not be significantly different from the global one, thus guaranteeing a global consistency.

3. The implementation of line-based global transformation

3.1. The mathematical model

As we mentioned before, the first stage of our work aims to obtain a global projective (perspective) transformation by aligning straight lines between two images. Theoretically, the objective is to compute a point-to-point transform matrix $\boldsymbol{H}$ explained by (2). However, due to the well-known principle of duality in projective geometry, it follows that the transformation can also be determined by specifying a sufficient number of line correspondences.

Let us now denote two corresponding lines (l and l') on both image coordinates as:

$$au + bv + w = 0 \quad \text{and} \quad a'u' + b'v' + w' = 0. \tag{3}$$

The above two lines can be expressed by their homogenous coordinates which are recorded by a linear transform $A : (a,b,1)^T$ and $A' : (a',b',1)^T$, respectively. Based on (2), we can describe the relation between these two lines by a transformation $\widehat{\boldsymbol{H}}$, which is specified by:

$$A = \widehat{\boldsymbol{H}}A'. \tag{4}$$

To clarify the relationship between $\widehat{\boldsymbol{H}}$ and $\boldsymbol{H}$ in (2), we rewrite the line equation to

$$A^T \begin{pmatrix} u \\ v \\ w \end{pmatrix} = 0 \quad \text{and} \quad (A')^T \begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} = 0. \tag{5}$$

If we substitute (4) into (5), it will become

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \widehat{\boldsymbol{H}}^{-T} \begin{pmatrix} u' \\ v' \\ w' \end{pmatrix}. \tag{6}$$

Comparing (6) and (2), we can deduce that $\boldsymbol{H} = \widehat{\boldsymbol{H}}^{-T}$. Obviously, it is possible to compute a point-to-point transformation $\boldsymbol{H}$ given a line-to-line mapping $\widehat{\boldsymbol{H}}$. Eq. (6) also confirms that four line-correspondences suffice to compute $\widehat{\boldsymbol{H}}$ (as expected from duality).

3.2. Algorithm implementation

Our line-based perspective transformation estimation consists of two modules, addressing line generation and line matching, respectively. The line generation module consists of line detection, line labeling and sorting. The line matching module includes initial matching and geometric matching of line composition. All steps are designed with an eye on efficiency.

3.2.1. Line generation

The input of our line detection algorithm is the edge pixel extracted by the Canny operator. Prior to the edge extraction step, we have a contrast enhancement step based on the histogram equalization, which helps to detect the blurred edge pixels. We utilize a RANSAC-like algorithm (Han et al., 2008, 2011) discussed in our previous work to detect the dominant line given the data-set. RANSAC is a randomized algorithm that hypothesizes a set of model parameters and evaluates the quality of the parameters. After several hypotheses have been evaluated, the best one is chosen. Specifically, we hypothesize a line by randomly selecting two edge pixels, from which we compute line parameter g. For this line hypothesis, we compute a score $s(g)$ as

$$s(g) = \sum_{(x,y)\in\Omega} max(\tau - d(g,x,y), 0), \tag{7}$$

where Ω is the set of edge pixels and $d(g,x,y)$ denotes the distance between (x,y) and the line g. This score effectively computes the

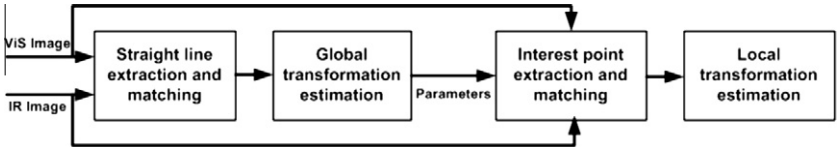


Fig. 2. Framework of our system, which composes of line-based global transformation estimation and interest point-based local transformation computation.

support of a line hypothesis as the number of edge pixels close (as determined by τ) to the line, weighted with their distance to the line. The score and the line parameters are stored and the process is repeated until about 25 hypotheses are generated randomly. At the end, the hypothesis with the highest score is selected. The output of this detection algorithm also includes the start- and end-point of each line. More precisely, it returns a line *segment*. We filter out some shorter line segments, and extend the line segments to the image borders. The reason for the step is that line segments extracted from both images vary dramatically, but most *major* segments with sufficient length are appeared in both images.

Next, lines are labeled as either "lying" or "standing". This label is determined by the parameter:

$$L_{ls} = |x_{end} - x_{start}|/D_{start \to end}, \tag{8}$$

where x_{start} and x_{end} refer to x coordinates of start point and end point of a line, respectively. $D_{start \to end}$ denotes the distance between these two points. If L_{ls} is larger than 0.7, the line is labeled as a lying line. Otherwise, the line is considered to be a standing line. Note that the value of this threshold is not that important in the sense that it may only influence the initial matching of one or two line, whose L_{ls} is very close to the threshold. For this extreme case, one line might be labeled as a "lying" line in one image but is labeled as a "standing" line in another image, resulting in a wrong initial line matching. However, this mistake of the individual line matching is not critical, because our algorithm is attempting to find the best line-configuration matching between images, which is a sort of optimization procedure based on many lines. After labeling lines, the set of standing lines are ordered left to right, the set of lying lines from top to bottom. Later, when we will search for correspondences between images, we will put the constraint on the assignment that the order must be preserved. This constraint is likely valid in case that our transform is either affine transform or perspective transform.

Finally, the line is modeled by three parameters, which are L_{ls}, sp and os. If the line is a lying line, sp is defined as the angle of the line to the x-axis, and os means the offset of the line on the y-axis. The roles of the x- and y-axis are reversed in case of a "standing" line. The definitions for sp and os are just inverse if line is a standing line. Fig. 3 shows the samples processed by our line generation algorithm.

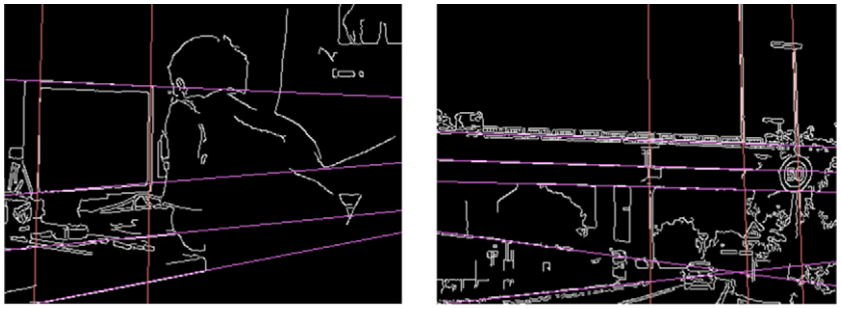
3.2.2. Line matching

As we can see from the problem statement part, feature initial matching schemes used by existing systems are in general not accurate enough. The main reason is that two images captured by different modalities are quite different at the pixel level. To solve this problem, our system enables a sort of one-to-many feature matching, which allows a line in one image to have several correspondences on another image. By doing so, we can increase the likelihood that several matching candidates must include the correct one. The basic idea for this initial matching is to check and compare three parameters of two lines located in two images. The first parameter is L_{ls}, where we assume that two corresponding lines should have similar L_{ls}. The assumption is valid for most applications, where modalities are mounted on the same platform. The second parameter is sp, where we assume that corresponding lines have similar slope to the axis. The last parameter is to compare distributions of the edge pixel surrounding the line. The surrounding area is the zone between two border lines, which have the same slope with the candidate line but with $\pm\epsilon$ offset shift, respectively. The distribution of the edge pixel within this area can be simply specified by the edge pixel percentage pec_{edge} of that area, equaling to N_{edge}/N_{total}. Here, N_{edge} refers to the number of edge pixels within that area, while N_{total} means the total number of pixels within that area. If we denote the parameters of two candidate lines as (L_{ls}, sp, pec_{edge}) and $(\widetilde{L}_{ls}, \widetilde{sp}, \widetilde{pec}_{edge})$, our matching score S can thus be formulated as:

$$S = K\left(\frac{L_{ls} - \widetilde{L}_{ls}}{\sigma_{L_{ls}}}\right) \cdot K\left(\frac{sp - \widetilde{sp}}{\sigma_{sp}}\right) \cdot K\left(\frac{pec_{edge} - \widetilde{pec}_{edge}}{\sigma_{pec}}\right). \tag{9}$$

The three terms in the equation implement the same logic: $K(\cdot)$ is the Epanechnikov kernel function and σ indicates the width of the kernel, which can be set manually. We compute the matching scores between a given line and all candidate lines. Instead of selecting the best one, we allow one line to have three candidate correspondences in terms of the ranking of the matching score.

After this initial line matching stage, we proceed by processing the alignment of geometric configurations comprising four lines. Earlier we pointed out that the matching result between two individual lines may not be reliable. However, the geometric configuration (layout) of different lines is much more consistent between images. This observation motivated us to align images by minimizing the discrepancy between two geometric configurations formed by lines. The basic idea is that we randomly choose four lines from the first image, thus creating a so-called *mini-configuration*. Depending on the initial matching result, we will have several corresponding mini-configurations in the second image. Each configuration-correspondence allows us to compute the parameters of a projective transformation by solving the system of linear Eq. (4). Using the transformation thus obtained, we project one image onto the other image. The match between two images is evaluated by computing the total distance between each line and its closest projected line. We iteratively search all possible configuration-correspondences and settle for the one that minimizes the total distance as the best configuration-correspondence. The transformation can thus be



(a) (b)

Fig. 3. Line generation results for two samples. Lying lines and standing lines are marked with different colors. The original edge pixels are labeled with the white color.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

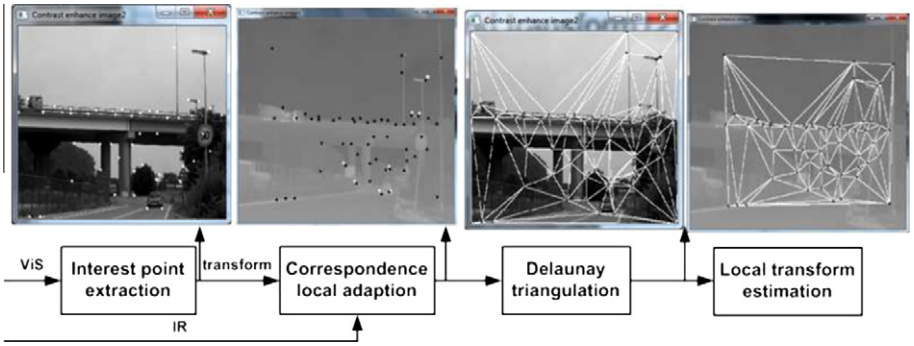


Fig. 4. Basic steps for the local transformation estimation.

estimated based on the best correspondence. From the mathematical perspective, finding the best configuration match can be recast as minimizing the matching error M_e,

$$M_e = \sum_{l \in \Phi} \min(\|l', \boldsymbol{H}l\|_2, e_m), \tag{10}$$

where Φ the collection of lines in the *image* 1 and l' is the closest line of the projected line $\boldsymbol{H}l$ in the *image* 2. The metric $\|,\|_2$ denotes the Euclidean distance between the two lines, and the error for a line is bounded by a maximum value e_m. More details about this procedure can be found in (Han et al., 2012).

4. Interest point-based local transformation estimation

In the preceding section we explained how we used line configurations to determine the initial global projective transformation between two images. In this section we will detail how we refine this first approximation by a local search for interest points. More precisely, we start by extracting interest point in one image. Subsequently, those interest points are transformed into the other image based on the initial global transformation. In the second step, we allow those projected interest points to find better correspondences within a window surrounding their initial estimates. In the last step, we use a Delaunay triangular mesh with an appropriate data structure to construct the local adaptation. Indeed, three neighboring points form a triangle, and its corresponding triangle in the other image can be found by connecting the corresponding points. For a pair of triangles, an affine model is employed to approximate the spatial relation between these two local areas. Fig. 4 illustrates the basic algorithmic steps for our local transformation estimation.

4.1. Interest point extraction

As already mentioned in the problem statement section, the majority of interest points extracted from both images by using conventional algorithms is not reproducible. The main reason is that conventional algorithms check the response value of a detector at each pixel position of an image. The pixel is labeled as an interest point if the response value at this position is larger than a pre-defined threshold. In fact, this procedure is to select the top n pixels according to the ranking of response values, where the number n can be controllable parameter. Since the IR image has noticeably less texture in some areas than the ViS image, it is very unlikely that the n interest points extracted from ViS image are identical to the n interest points extracted from an IR image. However, we have noticed that the pixels having maximal corner response within a local neighbourhood, do indeed match in both ViS and IR images. The reason is that those pixels are usually on the boundaries of objects, which are clearly visible in both images. Therefore, an interest point is found at the location $\tilde{u}$ if the response function returns the maximum value within a local neighborhood B, which is described by:

$$\tilde{u} = \arg\max_{u}(R(u)), \quad u \in B. \tag{11}$$

Here, $R(u)$ returns the corner response in the pixel u within a small window B in the image. In our algorithm, we use the Harris corner detector to compute $R(\cdot)$. For more details about the Harris corner response function, we refer to the original paper (Harris and Stephens, 1988). One common phenomenon in a relatively flat region is that the maximum value of $R(\cdot)$ might be very close to zero. For this case, we discard the detected interest point in this local area, because it is difficult to match it with other interest points in the next step. In our implementation, we divide the entire image into non-overlapping cells of 32×32 pixels, and apply our interest point extraction for each cell.

4.2. Correspondence local adaptation and Delaunay Triangulation

We transform the interest points detected in the first image to the second image based on the global transformation (determined earlier). These transformed points are used as an initial estimate for the corresponding points. We then establish a search window surrounding each of these initial estimates and proceed by searching for better correspondences within this local window. The pixel in this search window exhibiting the strongest corner response is labeled as the searched-for correspondence. Fig. 5 shows an example, where we extract interest points from the ViS image on the left and locally adapt their correspondences on the IR image. The white dot refers to the initial correspondence computed by the transformation, while the black dot represents the position of the interest point after the local adaptation. Clearly, most interest points highlighted by red circles[1] are re-assigned to a better match.

Following this adaptive local correspondence search we proceed by carving up the entire image into patches to which we can apply local transformations. Here, we have opted to describe the image as a set of adjacent triangles by using Delaunay Triangulation (DT) applied to the first image. The use of DT is particularly suited when we do not want to force any constraints on the set of points to be connected. For our case, detected interest points are connected by DT algorithm and three neighboring points form a triangle. Its corresponding triangle in another image can be easily constructed by using corresponding points of three vertices. Once we have obtained a pair of corresponding triangle (patch), the relation between them can be described

[1] For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.

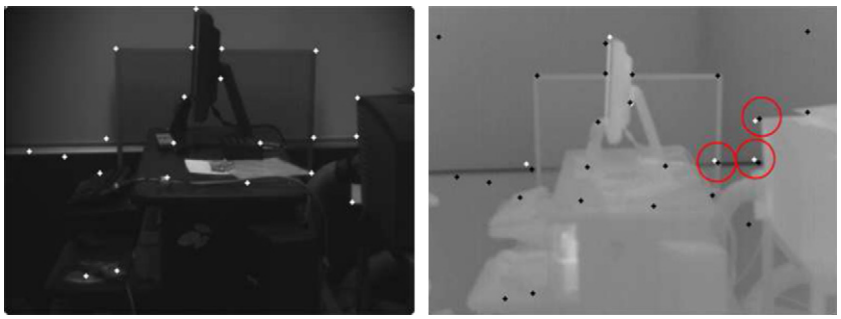


Fig. 5. Local adaptation of the interest point correspondence. White dots on the left image are extracted interest points. White dots on the right image are point correspondences using the initial transformation. Black dots represent the position of the point correspondences after adaptation.

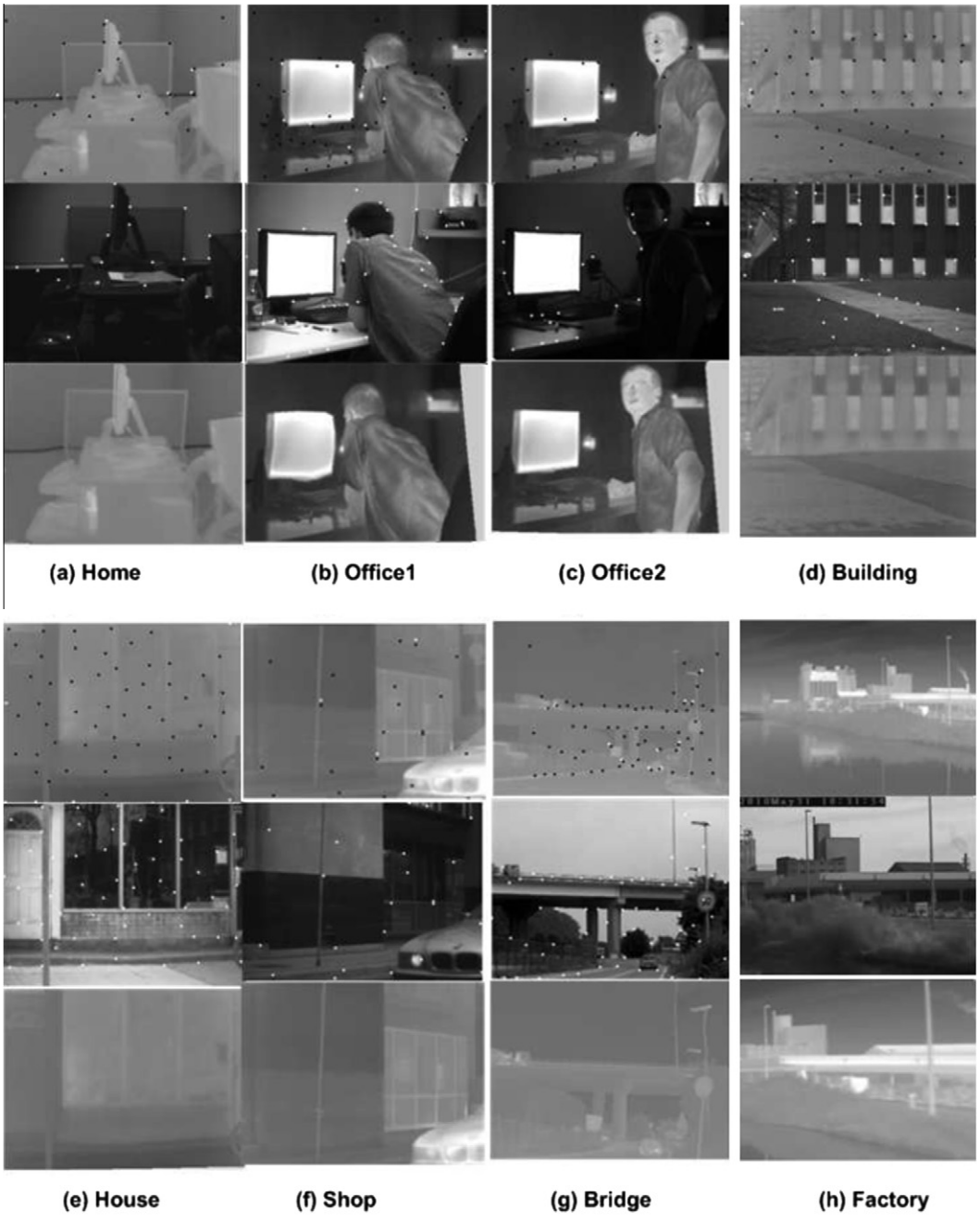


Fig. 6. Registration and intermediate results. For each subfigure, the top row shows the original IR image, and the middle row displays the original ViS image. The bottom row provides a warped image (from IR to ViS). We extract interest points from the ViS image (white dots), and initial correspondences on IR image are also marked by white dots. The final correspondence after the local adaption is marked as a black dot.

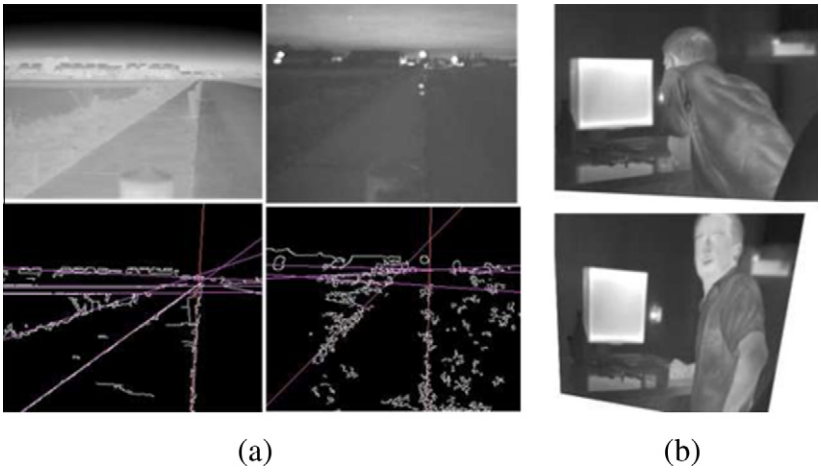


Fig. 7. (a) Image pair for which our algorithm fails. Top: original images. Bottom: the results for edge and line detection. (b) The registration results obtained by using gradient orientation based descriptor.

Table 1
The measurement for transform errors, in which the unit is pixel.

	"Home"	"Office 1"	"Shop"	"Bridge"
Initial transform ($\mu;\sigma$)	1.78; 1.82	6.40; 2.76	2.92; 2.73	5.28; 3.22
Local transform ($\mu;\sigma$)	0.76; 1.11	4.63; 2.43	1.50; 1.58	4.28; 4.01

by an affine model:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \tag{12}$$

where (x,y) and (x',y') are coordinates of two corresponding points. There are six parameters in the affine model, which can be estimated by using coordinates of three vertices and their correspondences.

5. Experimental results

Our proposed system is implemented in C++ on a Laptop PC platform (Dual core 2.53 GHz, 4 GB RAM) with a 64-bits operation system. We have tested our algorithm with nine pairs of IR and ViS images/videos, where six of them are outdoor scenarios and three of them are depicting indoor scenarios.[2] There are two video sequences in the testing dataset, called "Bridge" and "Factory", which contain 50 and 30 frames, respectively.

We have registered these images by using our algorithm. A key parameter is the minimum length of the accepted line, which we set to 40 pixels. In general, our algorithm can register all pairs of images except the last pair. The visual results of our registration are shown in Fig. 6, where the first two rows of each subfigure illustrate original IR and ViS images, and the last one shows a warped image from IR to ViS based on estimated transformation. To highlight the benefit of our local adaption for feature points matching, we also show the intermediate results on the original images. More specifically, we extract interest points from the ViS image that are indicated by white dots on the image, and then transform them to the IR image using the initial global transformation. The transformed points are also marked by white dots on the IR image. Afterwards, feature points are allowed to find the better correspondences within a local window, and the final correspondences are marked by black dots.

As can be observed from the results, most feature points have zero shift (white and black dot coincide) after the local adaption, which implies that our initial transformation is sufficiently accurate. However, it is also apparent that some feature points do indeed find better correspondences after the adaptation procedure, e.g., in home, office1, shop and bridge images. For this last pair of images, we only apply the initial global transformation and dispense with the adaptive local matching, as the result of the initial transformation is not sufficiently accurate. The reason is that this sequence is too challenging in the sense that two cameras have significantly different focal lengthes. Our algorithm fails to register one image pair depicted in Fig. 7(a), which were captured during the night. Seen from the edge maps, it is possible to extract enough straight lines from the IR image, but line detection does not work properly on the ViS image due to low level of illumination.

To evaluate our registration algorithm, we measure and report the transform errors in Table 1. The transform error is measured by the distance between one point and its transformed corresponding point. More specifically, we randomly choose eight salient points in the IR image, and transform them to ViS image using the estimated transformation models. We manually label the correspondences of those eight points, and use this as ground truth. The distance between the ground truth and the transformed point is proportional to the transform error. We calculate the average and the standard deviation of transform errors caused by initial global transformations as well as the local transformation. It can be revealed that our local adaptation indeed helps to reduce the transform error. All image pairs used for this experiment are with resolution of 384×288.

We also compared our algorithm with existing algorithms relying on *point* matching. Since gradient magnitude (Lee et al., 2010) and orientation (Kim et al., 2008) are widely used for IR and ViS image registration, our implementation explores statistics of gradient magnitude and orientation to describe the feature point extracted by SURF algorithm, where the gradient orientation is actually the main feature used by SIFT descriptor. Afterwards, nearest neighbor approach is applied for feature matching. Next, RANSAC is used for rejecting outliers. Finally, perspective transform matrix is computed based on a number of point correspondences between two images. We have tested these two feature descriptors for the same dataset. The gradient

[2] Videos and images are provided by XenICs NV (Belgium) and authors of paper Morris et al. (2007).

		GM-based descriptor	GO-based descriptor
IP number (76: 78)	Initial matching	14	17
	After RANSAC	4	5
	Correct matching	1	1
IP number (154:151)	Initial matching	18	20
	After RANSAC	7	6
	Correct matching	2	3
IP number (241:253)	Initial matching	25	27
	After RANSAC	6	9
	Correct matching	2	4

(a) Statistics for two exiting descriptors

(b) An example for feature point matching

Fig. 8. (a) Statistics by using two existing descriptors, where GM and GO refer to gradient magnitude and gradient orientation, respectively. (b) An example showing the initial correspondences obtained by the GO descriptor.

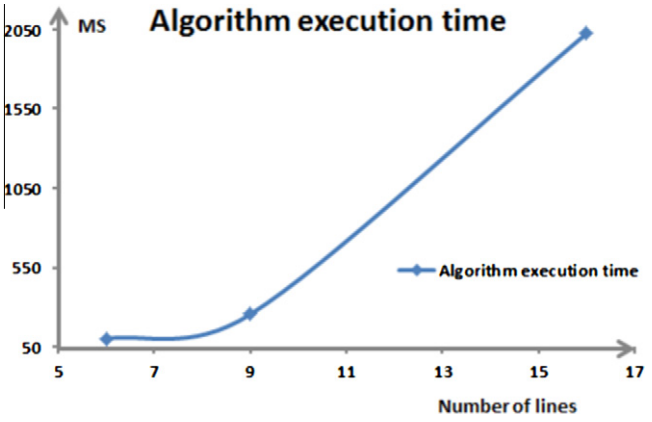


Fig. 9. The relationship between algorithm execution-time and scene complexity.

magnitude-based descriptor failed for all the pairs, and gradient orientation-based descriptor only *succeeded* in registering "office 1" and "office 2" image pairs. We show the warped images in Fig. 7(b), where we warp the IR image to ViS image based on computed transformation. Although the registration results are obtained, the accuracy is far from satisfactory. Additionally, we try to investigate the failure of this algorithm by means of analyzing the point matching results. In Fig. 8(a), we calculate the number of matched interest points for image pair "Bridge", given a number of interest points detected on both images. For example, the algorithm found 14 correspondences after the initial matching, among which four correspondences remained after RANSAC check. However, only one of four correspondences is correct after manually examining. It can be concluded that the gradient orientation-based descriptor (GO) is slightly better than the gradient magnitude-based descriptor (GM) in terms of the correct number of correspondences. Unfortunately, the majority of detected correspondences using both descriptors is not correct. In Fig. 8(b), we give an example, where the initial matching result of the GO descriptor is provided. Obviously, most correspondences were already wrong at this stage.

Our algorithm is designed for an industrial project, so that the real-time capability of the algorithm is desired. In Fig. 9, we report the computational cost of the algorithm, and show the relationship between the algorithm execution time and the scene *complexity*. The scene complexity is defined to be proportional to the number of straight lines extracted from two images. Three image pairs used for the experiment are "Home", "Office 1" and "Building", where the resolutions of IR and ViS images are 384×288 and 656×490, respectively. For these three pairs, the average numbers of lines involved in finding the feature correspondence between two images are 6, 9, and 16, respectively. From the result, we can see that the algorithm can process 4–10 image pairs per second when the scene is not extremely complex.

6. Conclusion

In this paper, we have examined the combination of line feature and interest point feature for registering IR (long wavelength) and ViS images. Alignment of straight lines between two images provides an initial estimation for a global transformation. Furthermore, we divide the entire image into cells, and extract interest point from each cell. We transform detected interest points to the other image based on the initial transformation, and allow them to find better correspondences locally, thus leading to a locally adaptive transformation model. The key element of our approach is the line-based perspective transformation estimation. In comparison to existing work that aligns feature points, lines derived from edge pixels delineate object boundaries and have a good reproducibility on images captured by different modalities. Our new algorithm provides significant advantages over state-of-the-art approaches. Future work will focus further on improving the quality of line detection algorithm. Such an improved line detection algorithm should be capable of handling noisy images, such as the ones in Fig. 7(a).

References

Brown, L., 1992. A survey of image registration techniques. ACM Comput. Surv. 24 (4), 325–376.
Caspi, Y., Simakov, D., Irani, M., 2006. Feature-based sequence to sequence matching. Internat. J. Comput. Vision 68 (1), 53–64.
Coiras, E., Santamaria, J., Miravet, C., 2000. Segment-based registration technique for visual-infrared images. Optical Eng. 39, 282–289.
Firmenich, D., Brown, M., Susstrunk, S., 2011. Multispectral interest points for RGB-NIR image registration. In: Proc. Internat. Conf. on Image Processing, Brussels, Belgium, pp. 181–184.
Han, J., Bhanu, B., 2007. Fusion of color and infrared video for moving human detection. Pattern Recognition 40, 1771–1784.
Han, J., Farin, D., de With, P., 2008. Broadcast court-net sports video analysis using fast 3-D camera modeling. IEEE Trans. Circ. Syst. Video Tech. 18 (11), 1628–1638.
Han, J., Farin, D., de With, P., 2011. A mixed-reality system for broadcasting sports video to mobile devices. IEEE MultiMedia 18 (2), 72–84.
Han, J., Pauwels, E., de Zeeuw, P., 2012. Visible and infrared image registration employing line-based geometric analysis. In: Proc. MUSCLE Internat. Work. on Comput. Internat. for Multi. Under., Pisa, Italy, pp. 114–125.
Harris, C., Stephens, M., 1988. A combined corner and edge detector. In: Proc. Alvey Vision Conference, pp. 147–151.
Hrkac, T., Kalafatic, Z., Krapac, J., 2007. Infrared-visual image registration based on corners and hausdorff distance. In: Proc. Scandinavian Conf. on Image Analysis, Aalborg, Denmark, pp. 263–272.
Huang, X., Chen, Z., 2002. A wavelet-based multisensor image registration algorithm. In: Proc. ICSP, Beijing, China, pp. 773–776.

Jarc, A., Pers, J., Rogelj, P., Perse, M., Kovacic, S., 2007. Texture features for affine registration of thermal and visible images. In: Proc. Computer Vision Winter Workshop, Graz University of Technology.

Kim, Y., Lee, J., Ra, J., 2008. Multi-sensor image registration based on intensity and edge orientation information. Pattern Recognition 41, 3356–3365.

Lee, J., Kim, Y., Lee, D., Kang, D., Ra, J., 2010. Robust CCD and IR image registration using gradient-based statistical information. IEEE Signal Process. Lett. 17 (4), 347–350.

Morris, N., Avidan, S., Matusik, W., Pfister, H., 2007. Statistics of infrared images. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Minnesota, USA, pp. 1–7.

Xiong, Z., Zhang, Y., 2010. A critical review of image registration methods. Internat. J. Image Data Fusion 1 (2), 137–158.

Zitova, B., Flusser, J., 2003. Image registration methods: A survey. Image Vision Comput. 21, 977–1000.