# A Multimodal Image Registration Technique for Structured Polygonal Scenes*

Siniša Šegvić
Faculty of Electrical Engineering and Computing
Department of Electronics, Microelectronics, Computer and Intelligent Systems
Unska 3, 10000 Zagreb, Croatia
sinisa.segvic@fer.hr

## Abstract

*A technique for multimodal image registration based on a hypothesize-and-test approach is presented. The technique is based on aligning edge elements from the two input images, since they often originate from physical discontinuities which are likely to be detected by both sensors. The design has been specifically adapted for robust operation on images of regular objects with few distinct structural axes, in the context of automated inspection. The hypotheses are formed by speculating a correspondence between the pairs of parallelograms from the two input images, and evaluated according to the quality of the match of a transformed edge image. The best hypothesis is finally refined by a nonlinear optimization algorithm. The technique has been tested on 42 pairs of 14 infrared and 3 RGB images of an unevenly heated metal prism, and the obtained results are reported.*

## 1. Introduction

The problem of registration is related to finding an alignment between the two images of the common scene [15]. The registration is mostly concerned with approximately planar scenes, for which the deviations from the planarity are small compared to the distance from the camera. In such conditions, the alignment between the images can be expressed as a planar projective transform [3], or one of its more special instances such as the affine [2] or the similarity transform [12]. The choice of the alignment model is the basic step towards the solving of the registration problem: less general models have less degrees of freedom and therefore require less elaborated hypotheses. On the other hand, more general models can compensate for larger deformations due to different viewpoints, which may show advantageous in a concrete application. The properties of the three most frequently used models are summarised in the table 1. The 4 degrees of freedom (DOF) similarity transform can account for rotation, translation and isotropic scaling, and is useful whenever the viewpoints of the two cameras are rather close. The affine transform introduces additional 2 DOF (anisotropic scaling, skew), which enables matching images from distant viewpoints, provided that the imaged plane is far from both cameras. Finally, the full projective transformation, homography, can capture all deformations of planar shapes, including the projective ones. Note that all previous considerations hold only for (approximately) planar scenes. A more general approach would involve solving for 3D positions of both cameras and all feature points for which the unambiguous correspondence has been established. There is a procedure for finding a solution to that problem when only 5 correspondences are available [9]. However, in such a general setup, the established set of correspondences can not easily be extrapolated to other points, which is usually one of the registration goals.

| model | DOF | points | condition |
|---|---|---|---|
| similarity | 4 | 2 | $\sphericalangle(\mathbf{z_1}, \mathbf{n}) \approx \sphericalangle(\mathbf{z_2}, \mathbf{n})$ |
| affinity | 6 | 3 | $\|z_{ik} - z_{il}\| \ll \|z_{ik}\|$ |
| homography | 8 | 4 | — |

$\mathbf{n}$ – the normal of the imaged plane;
$\mathbf{z_i}$ – the optical axis of the i-th camera;
$z_{ik}$ – distance from the k-th point to the camera along $\mathbf{z_i}$.

**Table 1. Alignment models for planar scenes.**

The presented work is concerned with registering multimodal images, for which a complex relationship between the brightness values of the corresponding pixels should be expected [4]. Pixels of the same colour in the visible spectrum can correspond to pixels with differing brightnesses in the infrared image, especially if the considered object has been artificially heated in order to make the deformations visble. Many of the previous multimodal registration techniques therefore focussed the processing on the edges obtained from the two input images. Some of these approaches rely on further groupings of edge elements into line segments [12, 2, 6], or more general contours [7], while others work directly on variations of the image gradient [4, 5], or

*This work has been supported by the Croatian Ministry of Science and Technology, as a part of the TEST (technological R&D) programme, administrative number #4046 (2004). I would particularly like to thank Slobodan Ribarić and Ivan Fratrić for helpful comments on preliminary versions of this manuscript. I would also like to thank Denis Vedrina and other project participants for acquiring images on which the presented experiments were performed.

the image Laplacian [14]. The matching is consequently expressed as a minimization of the objective function providing a numerical estimate of the candidate transform quality. There are two main approaches to finding the transform which minimizes the objective function: non-linear gradient descent local optimization procedures [4, 6, 13, 5], and the comprehensive evaluation of hypotheses determined following a certain set of assumptions [8, 2, 14]. The principal shortcoming of the former approach is the possibility of being attracted by a local optimum, which can be alleviated by a multiresolution refinement [5]. The latter approach has no such problems, but it may become computationally intractable if too many features are found in the input images. Further, the feature extraction step usually introduces additional errors, which makes the matching criterions depending on basic image features potentially more accurate.

The rest of the paper is organized as follows. The specific assumptions are described in Section 2, while Section 3 summarizes the performed preprocessing operations. Section 4 outlines the proposed solution, while the obtained experimental results are described in the Section 5. The properties of the proposed technique are summarised in the Section 6, together with the directions for the future work.

## 2. Assumptions

A multimodal image registration is considered, in the context of regular polygonal scenes. The two images are acquired with different sensor technologies: a conventional digital camera, and an industrial standard infrared camera. Obtaining accurate registrations is an important capability, since it would allow a superior performance in inferring the qualities of the tested object, by combining its visual and thermal properties. Certain defects of plastic or metal industrial products could be identified in a very robust fashion by fusing the information from the two complementary sources. Consequently, pairs of input images are acquired, while the object is artificially heated and then relinquished to become cool again.

It is assumed that the object contains large planar subsets and that the cameras are placed at a safe distance from the object in order to avoid perspective deformations so that (i) parallel lines in the scene map to parallel lines in the image, and (ii), the two images can be related through an affine transform. Additionally, it is assumed that the object has a regular polygonal structure with few (possibly only two) distinctive structural axes. Finally, it is presumed that there is a good mutual correspondence between the edges extracted from the two input images, although exceptions due to reflections and visible texture should be tolerated.

There is usually a sheer disproportion in the resolutions of the two input images, since the infrared technology is more sophisticated and expensive than the usual sensors for detecting visible radiation. Consumer grade digital cameras provide resolutions over $10^6$ pixels, while industrial infrared cameras offer a modest half VGA resolution of only $320 \times 240$. Thus, the design of a registration procedure must consider variations of the input images in the both realms of modality and resolution.

## 3. Preprocessing

As stated above, all the processing is based on edge elements which are extracted from the two input images. A manually tuned Canny edge detector has been employed in all experiments, but similar performance is expected from other algorithms that produce thin edges. Straight line segments are extracted from the input images using a previously developed procedure based on Hough transform [11]. Further processing relies both on the extracted line segments and the raw edges provided by the edge detector. An example of a visible-infrared (RGB-IR) image pair with the corresponding extracted edges and line segments is shown in fig.1. It can be seen that most (although not all) extracted

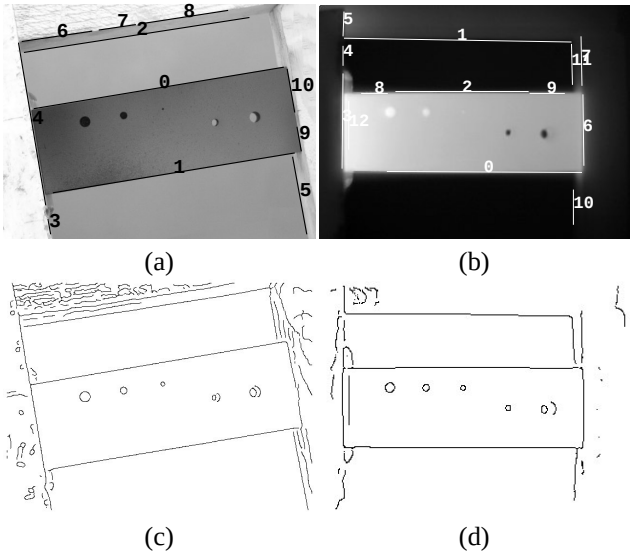



**Figure 1. RGB (a) and IR (b) input images with the extracted line segments, and the corresponding edges (c,d). For better presentation, the RGB images have been scaled down.**

segments correspond to object boundaries which are visible in both images. However, many parts of the visible boundaries have not been correctly identified, leading to cracks and shortened line segments. Additionally, several segments are only visible in one of the input images (e.g, IR:7[1], IR:11, IR:12, RGB:3, RGB:6, RGB:7, RGB:8, fig.1). This property is intrinsic in multimodal image registration: the same physical properties of the scene are often detected in different ways by different sensors. Consequently, the data obtained by preprocessing should be further refined in order to obtain robust features suitable for achieving stable results.

[1]IR:m refers to the line segment labeled *m* in the IR image of the referenced fi gure (the same notation is used for segments in RGB images).

## 4. The proposed technique

The proposed technique for multimodal registration tries to intertwine good features of the previous methods and apply them to the specific problem at hand. Specifically, it tries to avoid the uncertainties tied with being stuck on suboptimal alignments, while still achieving the superior accuracy of performing the registration at the pixel level. A mixed hypothesize-and test approach is therefore devised, in which the hypotheses are generated from speculated high-level feature correspondences (as in [2]), but evaluated by an objective function operating on raw edges directly (as in [5]).

The proposed technique can be subdivided into five processing stages. In the first stage, the line segments obtained by preprocessing are regularized in order to increase stability and minimality of the resulting set. The refined segments are used in the stage 2 to generate an exhaustive set of candidate transform hypotheses. Candidate transforms are evaluated as high-level features mappings (stage 3), and raw edges mappings (stage 4). The best candidate transform is finally further refined in stage 5, by a general purpose gradient optimization algorithm. Implementation details of the individual stages are described in the following subsections.

### 4.1. Regularizing the set of line segments

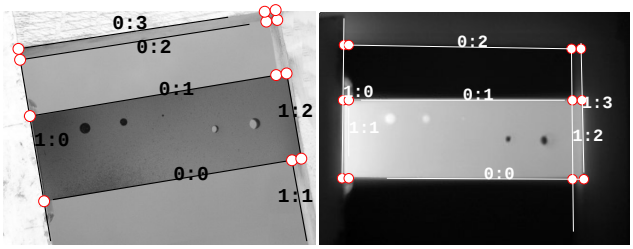
Due to the regularities in the observed scenes, straight line crossings can be employed as robust point features in the alignment procedure. A good property of such a feature is that it is quite well defined even in cases in which the segments have not been extracted over the entire boundary. However, multiple segments on the same line caused by cracks (e.g. IR:2, IR:8, fig.1) and occluded edges (e.g. IR:10, IR:11, fig.1) are problematic since they might give rise to multiple point features due to the same cause. In order to obtain the minimal but complete set of point features in images of the assumed environment, the set of input segments is regularized by the following procedure.

1. **Merging**: pairs of line segments nicely extending from each other's ends are iteratively merged (e.g, IR:8, IR:2, IR:9, and IR:10, IR:11, fig.1).

2. **Clustering**: the motivation for this step is to cluster segments originating from parallel 3D boundaries. If the projective deformations are expected in the image, line segments intersecting a hypothesized vanishing point could be clustered as well. Line segments are clustered according to their slope using a Hough-like approach with 36 accumulators covering the interval $[0, \pi)$. The clusters are formed by merging the most populated neighbouring accumulator pairs.

3. **Sorting**: line segments $s_{ci}$ from each cluster $c$ are sorted with respect to the line $l_c$ being orthogonal to the cluster direction and passing through the image center $I_C$; the sorting parameter $d(s_{ci})$ is calculated as distance along $l_c$ between the $I_C$ and the intersection of $l_c$ with the line $l_{s_{ci}}$ on which $s_{ci}$ is situated.

At this moment, point features can be formed by combining all pairs of line segments from different clusters. This is illustrated in fig.2, in which the regularized line segments and the extracted point features are shown for the same input images as in fig.1.




(a) (b)

**Figure 2. Regularized line segments and the corresponding point features in the RGB (a) and the IR (b) image.**

### 4.2. Generating hypotheses

The task of this processing stage is to build a comprehensive set of transform hypotheses which should contain the goal transform with high confidence. However, from the other perspective, the constructed set should be as small as possible, in order to be able to test all hypotheses in a reasonable time. The proposed procedure tries to stay on the tractable side by reducing the computational complexity whenever possible.

First, the two dominant directions of the scene structure in both images are detected as segment clusters having longest total length of the member segments. Then, the mutual correspondence between these directions is established by minimizing the total angular deviation. This is possible since rotations larger than $45°$ are not expected to occur. The last assumption could be avoided by hypothesizing all possible correspondences between the cluster directions, at the expense of somewhat increased execution times.

High-level features are extracted from both images as quadrangles (nearly parallelograms) formed by lines defined by combinations of line segment pairs from different clusters. Let $c_i$ be the $i$-th cluster of line segments, let $d(s_{ij})$ be the sorting parameter introduced in the previous subsection, and let $d_{min}$ be the quadrangle dimension threshold. Then each combination of the four line segments $(s_{00}, s_{01}, s_{10}, s_{11})$ must satisfy the following:

$$\begin{aligned} &s_{i0}, s_{i1} \in c_i, i = 0, 1 \\ &d(s_{i0}) - d(s_{i1}) > d_{min}, i = 0, 1 \end{aligned} \quad (1)$$

The dimension threshold has been introduced in order to reduce the computation complexity, by suppressing hypotheses involving small quadrangles having large positional un-

certainties. In all experiments, $d_{min}$ has been set to H/5, where H represents the smaller image dimension, in pixels.

Hypotheses are finally formed by speculating correspondences between all quadrangles from the two images satisfying (1). Each hypothesized correspondence induces a distinct projective transformation between the two images, which will be tested for support in the subsequent processing. The number of hypotheses is very high in theory, $O(n^8)$, but in practice this problem is alleviated by the precedent organization of the line segments into clusters. Note that choosing triangles as in [2] is not an option since the assumed environment has only two structural directions.

### 4.3. Evaluating hypotheses on feature alignment

At this point it is possible for each hypothesis to apply the induced transform to all point features from one image and check the alignment with the point features in the second one. This is a very attractive evaluation approach, since it remains in the computationally inexpensive symbolic domain (there are 12 point features, but more than 1000 edge elements in the image shown in fig.1). However the approach has several theoretical and practical shortcomings which severely limit its usefulness. Consider the problem of registering the two sketches in 3 which roughly correspond to images from fig.1, with the addition of one spurious line segment. It is clear that (A,B) is most probably in correspondence with (C,D), but chances are that, under the point feature evaluation approach, the hypothesis aligning (A,B) with (E,F) and (X,Y) with (C,D) would obtain a better score (in the sense of the sum of squared feature distances) than the correct hypothesis. Note that the point feature evaluation could not prefer the correct alignment even in the absence of the spurious segment, since the two hypotheses (A,B)→(C,D) and (A,B)→(E,F) would obtain nearly identical scores. The approach described in the next subsection

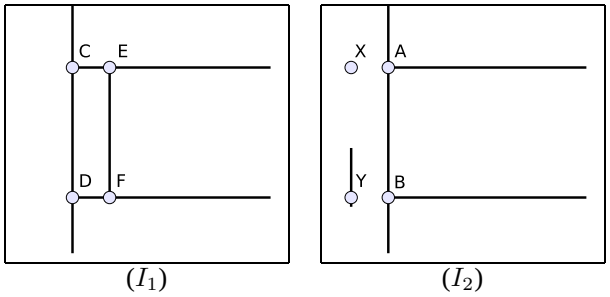



**Figure 3. Evaluating hypotheses based on point features promotes a false correspondence (X,Y,A,B)→(C,D,E,F) (see text).**

is more robust with respect to such problems, but is computationally more complex. It therefore might prove useful to use symbolic checking only as a gate for rejecting obvious blunders, such as hypotheses which does not succeed to map any additional point features, besides the four ones used for their construction.

### 4.4. Evaluating hypotheses on raw edges alignment

Following the discussion in the previous subsection, the final decision on the most accurate hypothesis describing the relationship between the two images is performed in the domain of raw edges. This approach has been also used in [5], where the transformation $T$ for aligning images $I_2$ and $I_1$ is found by optimizing the following functional, in which $g(I_2)$ represents the set of high gradient pixels in $I_2$:

$$\sum_{p_i \in g(I_2)} |\nabla I_1(T(p_i))|^2, \tag{2}$$

The proposed procedure for evaluating hypotheses follows the same basic idea: edges from one image should be aligned with edges in the another. However, experiments have shown that several modifications are needed in order to obtain best results from the idea in the context of evaluating hypotheses. Summing gradients at the transformed points as in (2) may cause high gradient regions which are present only in image $I_1$ to falsely attract parts of the boundary in $I_2$ which can be better aligned elsewhere (see fig.4). A

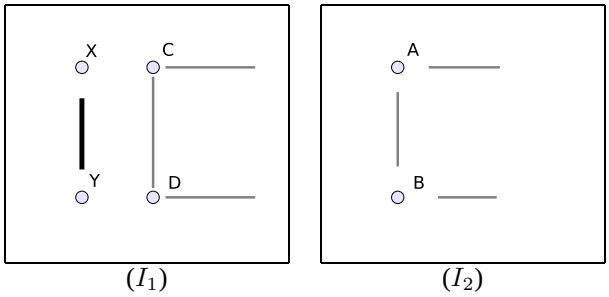



**Figure 4. Evaluating hypotheses by summing gradient contributions promotes a false correspondence (A,B)→(X,Y) (see text).**

different evaluation procedure is therefore proposed, which *counts* high gradient pixels from $I_2$, $p_i \in g(I_2)$, landing on high gradient pixels in $I_1$, $T(p_i) \in h(I_1)$:

$$\sum_{p_i \in g(I_2)} \delta_{I_1}(T(p_i)), \tag{3}$$

where

$$\delta_I(p) = \begin{cases} 1, & p \in h(I), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Different variations of the above criterion may be constructed by different choices for $g$ and $h$. The thinned output of an edge detector is a good candidate for $g$, since it decreases the number of pixels for which the candidate transform needs to be applied, without loosing important information. On the other hand, edges in $h$ must be thick in order to tolerate small deviations due to shadows, reflections and other kinds of noise which are differently expressed in the images obtained by the two sensors. After experimenting

with thresholded gradient magnitude and the smoothed output of the edge detector, it turned out that the latter provided somewhat better results for the utilized set of input images.

Note that the criterion (3) is similar to the partial Haussdorf distance $H_q(T(p_i), r_j)$, $p_i \in g(I_2)$, $r_j \in g(I_1)$. Although the latter approach is conceptually clearer, both alternatives suffer from ad-hoc thresholds: the width of the smoothing kernel and the quantile value $q$, respectively. In the end, the partial Haussdorf distance was not considered since it would incur high computational cost [13] to the hypothesized transformation evaluation, which is a part of the critical loop in the proposed procedure.

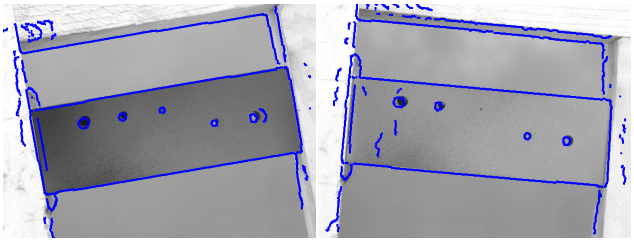
### 4.5. Refining the best hypothesis

After a quadrangle correspondence is found which maps most edge elements from $I_2$ onto smoothed edge elements of $I_1$, the induced transform can be further refined by nonlinear optimization. It is assumed that the rough alignment has already been achieved at this point, so that taking actual values into account does not raise risks described in the previous subsection and illustrated in fig.4. The gently sloped ridge of the smoothed edge map $h(I_1)$ usually makes it possible for an optimization procedure to achieve considerable improvements over the first approximation. Because of its availability, an implementation of the Levenberg-Marquardt's [10] optimization algorithm has been employed, which is a part of the Cephes library accessible from `http://www.netlib.org`. The optimization usually improves on the first approximation by mapping 5%–10% more edges from $g(I_2)$ onto $h(I_1)$.

Note that, alternatively, the distance transform of $g(I_1)$ [13, 1] could be used both for evaluating mappings from $g(I_2)$ and performing the optimization of the best hypothesis. This would allow for avoiding the ad-hoc smoothing parameter, as well as for decoupling the procedure from the edge detector implementation details. Unfortunately, these considerations have not been tested yet.

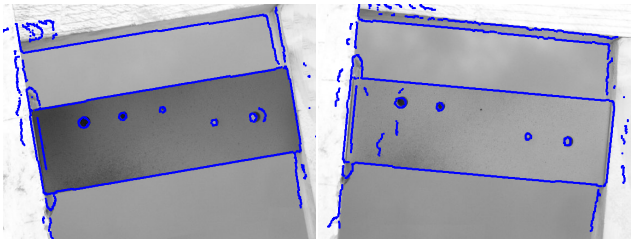
## 5. Experimental results

The experiment was performed on images of an artificially heated metal object with 5 holes. 16 infrared images of the object were taken, at different moments in time, in order to test the program's behaviour for different object temperatures. A registration procedure has been performed by combining each of the 16 IR images with 3 RGB images obtained for different rotations of the camera. In all experiments the RGB image resolution was $640 \times 480$ pixels, while the infrared image was $326 \times 244$. Before the preprocessing, the infrared image was normalized in order to make a better use of the 16 bit dynamic range. The evaluated transforms operated on edges from the smaller infrared image. Due to the simplicity of the considered scenes, the count of hypotheses generated in the processing stage 2 ranged between 50 and 100. Because of that, the theoretically most dangerous stage of hypotheses evaluation, required less processing time than nonlinear optimization and edge detection (100 ms vs 200 ms vs 150 ms, respectively). This would likely change for more involved scenarios, but such experiments have not been performed yet. Typical results, before and after applying the non-linear optimization are shown in figs. 5-6. It can be seen that even with optimization switched on, the procedure does not manage to find a perfect alignment for vertical edges. This is mainly caused by the imperfect IR imaging conditions: the wooden frame supporting the metal object reflects the infrared radiation from the metal making the object appear wider in the infrared spectrum, as seen in fig.1.



(a) (b)

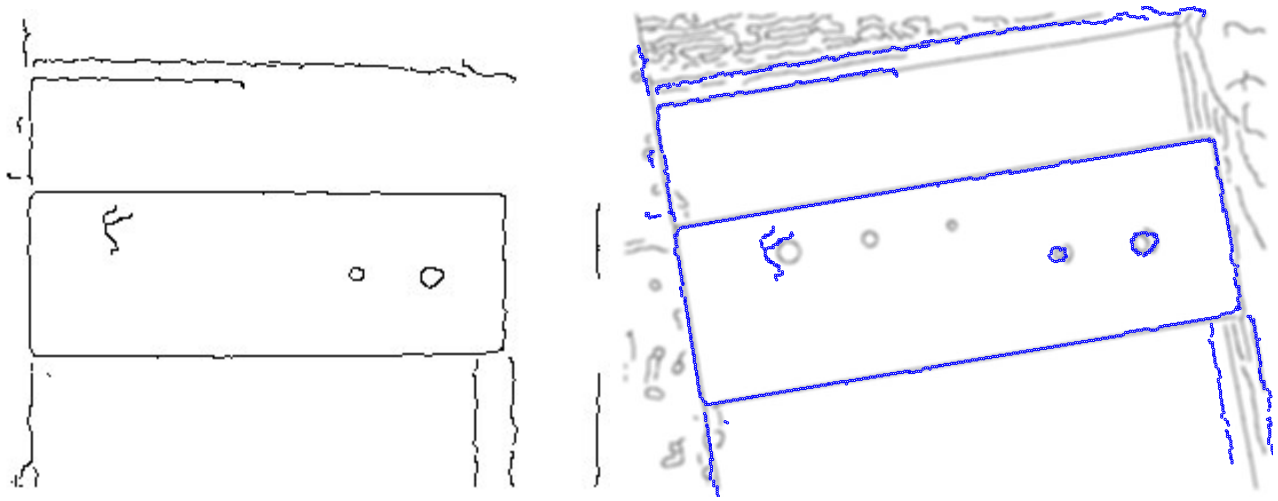
**Figure 5. The results obtained for two input pairs (a,b) without the optimization stage. For each pair, the transformed edges from the IR image are overlayed over the RGB image.**



(a) (b)

**Figure 6. The processing results on the same input as in fig.5, with optimization included.**

Finally it is interesting to see how the procedure performs in more problematic conditions, when the two images are taken from distant viewpoints, while the object is only lightly heated, producing noisy thermal images. The processing results are shown in fig.7. The registration procedure has succeeded to ignore the impossibility of matching the first and the third vertical edge from the right in the IR image, and perform the alignment based on other edges for which a planar correspondence can be found. The rightmost vertical IR edge is outside the field of view in the RGB image (not visible at all), while the third vertical IR edge from the right can not be related to the corresponding RGB edge by a planar transformation, due to the motion parallax. This is not a problem for other edges which lie in the common plane, or are parallel with the displacement vector between the two cameras.



(a) (b)

**Figure 7. The processing results for the image pair obtained from distant viewpoints: edges extracted from the original IR image (a), and their transformation onto the smoothed edges from the RGB image, used as the target during the transform evaluation (b).**

## 6. Conclusions

A technique for multimodal image registration has been presented, which is particularly suitable for aligning images of regular man-made objects in the context of automatic quality assessment. Although the technique assumes planar transformation between the two images, experiments have shown that good results can be obtained even for a non-planar scene, for moderate changes in the viewing direction. The obtained results suggest that a related procedure might be applicable in more involved scenarios as well. The principal problem in that direction would be to design a general interest operator which could be used to accurately locate and relate corresponding points in multimodal images. The future work will therefore be concentrated on more general ways to pose a tractable but complete set of hypothesized correspondences from which candidate mappings could be constructed and consequently evaluated.

## References


[1] H. S. Alhichri and M. Kamel. Image registration using virtual circles and edge direction. In *Proceedings of the International Conference on Pattern Recognition*, volume 2, pages 969–972, Quebec, Canada, Aug. 2002.

[2] E. Coiras, J. Santamaria, and C. Miravet. A segment-based registration technique for visual-ir images. *Optical Engineering*, 39(1):282–289, Jan. 2000.

[3] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993.

[4] M. Irani and P. Anandan. Robust multi-sensor image alignment. In *Proceedings of the International Conference on Computer Vision*, pages 959–966, Bombay, India, 1998.

[5] Y. Keller and A. Averbuch. Implicit similiarity: a new approach to multisensor fusion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 543–549, Madison, Wisconsin, USA, 2003.

[6] W. Krüger. Robust and effi cient map-to-image registration with line segments. *Machine Vision and Applications*, 13(1):38–50, 2001.

[7] H. Li, B. S. Manjunath, and S. K. Mitra. A contour-based approach to multisensor image registration. *IEEE Transactions on Image Processing*, 31(1):39–52, Jan. 1995.

[8] K. Nagao and W. Grimson. Affi ne matching of planar sets. *Computer Vision and Image Understanding*, 70(1):1–22, Nov. 1998.

[9] D. Nistér. An effi cient solution to the fi ve-point relative pose problem. *IEEE Transactions on Pattern recognition and Machine Intelligence*, 26(6):756–770, 2004.

[10] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK, 1993.

[11] S. Šegvić. Robust extraction of line segments from colour images by an iterative weighted polarised hough transform. In *Proceedings of 21th International Convention MIPRO '98*, volume 2, pages 35–38, Opatija, Croatia, May 1998.

[12] C. Shekhar, V. Govindu, and R. Chelappa. Multisensor image registration by feature consensus. *Pattern Recognition*, 31(1):39–52, Jan. 1999.

[13] Y. Sheng, X. Yang, L. Sévigny, and P. Valin. Robust multisensor image registration with partial distance merits. In A. Hyder, E. Shakbazian, and E. Waltz, editors, *Multisensor Fusion*, volume 70 of *NATO Sciences Series*, pages 593–609. Kluwer Academic Publishers, Netherlands, 2002.

[14] K. C. Walli. Automated multisensor image registration. In *Proceedings of the Applied Imagery Patern Recognition Workshop*, pages 103–110, Washington, DC, USA, 2003.

[15] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, Nov. 2003.