Silhouette-based features for visible-infrared registration

Guillaume-Alexandre Bilodeau, Pier-Luc St-Onge, and Romain Garnier
LITIV Lab., Department of computer engineering and software engineering
École Polytechnique de Montréal
P.O. Box 6079, Station Centre-ville, Montréal, (Québec), Canada, H3C 3A7
{guillaume-alexandre.bilodeau, pier-luc.st-onge, romain.garnier}@polymtl.ca

Abstract

We study the registration problem for infrared-visible stereo pairs. Given the properties of infrared and visible images that make them mostly similar near boundaries, we propose a method to extract keypoints on the boundary and on the skeleton of a region of interest (ROI). We show that our keypoints may be applied for partial image ROI and global registration either for videos or for still images given that the ROI silhouette is detected. In experiments, we show that our method gives better results than other classic keypoints and it gives results that are close to a state-of-the-art global registration trajectory-based method that uses temporal information.

1. Introduction

The image registration problem has now been studied for many years. The focus recently is to accelerate the calculations [10] or to improve the fine precision of disparity and registration calculations [9]. Because infrared cameras are becoming more affordable, using visible-infrared camera pairs is now attracting more attention. Visual surveillance applications, including some in the medical world to monitor a patient and its temperature, benefit from the combination of the information from these two sensors because they both perform well in complementary situations. For example, detecting humans is easier in infrared, but visible information is better to get a discriminative model of that human. Finding the transformation that maps an object of interest in infrared to the other image in the visible spectrum allows improving detected object boundaries and clarifying boundaries between objects, while also allowing depth estimation.

However, corresponding information in visible and infrared images for registration or disparity calculations is challenging. The application of visible stereo methods for visible-infrared camera configurations is not straightforward since infrared and visible images are manifestations of two different phenomena [15]. Visible cameras measure reflected light on objects, while infrared cameras measure principally infrared radiations emitted by objects. A texture or an edge in a visible image is often missing in the infrared image because texture seldom influences the heat emitted by an object. Furthermore, the way clothes fit the body of a person gives rise to different clothes surface temperature, thus creating heat-based textures.

To solve this challenging problem, we propose a new method for finding correspondences between pairs of visible and infrared images for their registration. Since human or living creatures are often the subjects of interest in visual surveillance, and because image regions are not very reliable, our intuition to solve this problem is to consider the boundary and the skeleton of binary silhouettes to find corresponding points. Our method extracts keypoints on the boundary of the silhouette using Discrete Curve Evolution (DCE), and junctions and terminal points on the silhouette skeleton. Our method may be applied for global image registration or partial image regions of interest (ROI) registration depending on the scene [15]. Our assumption about the stereo pair configuration is that the cameras are co-located and roughly parallel.

Our results show that our method performs partial image ROI registration better than classic feature points such as edges. We can obtain global registration results with a precision close and sometime better than trajectory-based methods given sets of points that are not too collinear. The advantages of our proposed method are that it does not rely on texture, so it allows correspondence for regions with different textures (as it happens often for objects in visible-infrared sensor pairs) and it does not rely on temporal information. Thus, it may be applied to images, not just videos, and for partial image ROI registration. It just requires a method to obtain the object of interest silhouette, which is now feasible with the latest people detection methods [6, 19].

The paper is structured as follows. Section 2 explores related works. Sections 3 and 4 present our proposed method. Section 5 presents validating experiments and section 6 concludes the paper.

2. Related works

To register visible-infrared images or videos, previous works have considered either region matching or feature points correspondence. Region matching methods are usually based on image region correlation or mutual information [3, 15, 16, 20]. Because the textural content is quite different in visible and infrared, the mutual information or correlation is often good only on a small portion of the images. Thus, mutual information is often used only on a selected region of an image [3], on regions with similar edge density [16], or on a detected foreground [15]. Region-based methods are not reliable for our task because they rely on similarity of textures in both images of the stereo pair. Visible-infrared stereo pairs do not always respect this assumption. This is why feature points on boundaries are also often considered. Using edges is one of the most popular method as their magnitudes and orientations may match between infrared and visible for some object boundaries [5, 8, 13]. Raw edges alone are not very reliable, so they may be considered as connected groups for correspondence [5, 8]. Another alternative is to extract object trajectories from tracking, and do the correspondence between the trajectories [2, 12, 21]. The drawback of this approach is that it relies on trajectories and requires the information from many frames, thus excluding the possibility to do partial image ROI registration.

Since humans may be segmented using motion for videos, or using people detectors followed by a classifier for still images, the human silhouette may be obtained to abstract textures. Furthermore, the orientation of the points on the contour of this silhouette may be used for correspondence, abstracting edge orientation differences between visible and infrared. Our proposed method capitalizes on silhouette information for point correspondence. Our method is related to the work of [5] and [8] that use edge groups.

We now describe how we extract and describe our proposed keypoints.

3. Keypoints extraction and description

From an object silhouette, we find two types of keypoints: 1) the vertices of an approximated skeleton and 2) the significant points obtained with an adapted implementation of the DCE [18]. Then, keypoints are further described in order to make correspondences. Pairs of points are validated based on silhouette correspondences. For the rest of the paper, the left and right images are the visible and infrared images respectively.

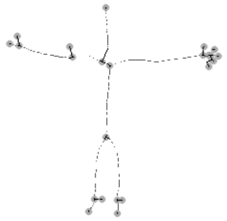

Figure 1. Keypoints on a skeleton. They are indicated by the gray circles. The black lines bind the keypoints to their neighbors.

3.1. Skeleton-based keypoints

In a skeleton interpreted as a tree, the keypoints are the vertices that are not connected by two tree edges (see Figure 1). They are extracted using the following steps [1]:

1. Smooth the contours with a closure, i.e. a morphological transformation dilating and eroding the blob (object's silhouette). The kernel used is an ellipse in a matrix of dimensions 3 by 3;

2. Apply a distance transform on the silhouette;

3. With the convolution product, apply the following Laplacian operator on the distance transform result: $\begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}$;

4. Apply a threshold to first get skeleton feature points. We have found that a value of 4 keeps enough vertices while getting rid of most of the smallest peaks in the distance transform image;

5. Use Prim's algorithm [4] to get the minimum spanning tree of the complete graph made of all skeleton feature points;

6. In the minimum spanning tree, keep all vertices that do not have two edges. These are our keypoints.

Figure 1 gives an example of keypoints found. Keypoint descriptors are required for matching using a correspondence matrix. For a skeleton-based keypoint, we use a feature vector with four components $\left(d_s, \vec{p}, n_{se}, \vec{\theta_n}\right)$:

- d_s: Its distance from the boundary of silhouette based on the distance transform. By normalizing linearly the distance-transformed image to values from zero to one, each keypoint is described according to its relative distance from itself to the silhouette's contour.

- $\vec{p}$: Its normalized position in the referential of the silhouette.

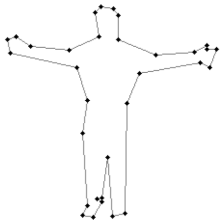

Figure 2. DCE contour (line) and DCE significant points (black dots).

- n_{se}: Its number of skeleton edges. In the minimum spanning trees (figure 1), each keypoint has one or many edges.
- $\vec{\theta_n}$: Its angles in the skeleton. For each keypoint, all keypoint's skeleton edges are modeled as vectors that are oriented from the keypoint to its neighboring vertices in the minimum spanning tree. If a keypoint has many neighbors, it has one angle for each neighbor.

3.2. Discrete Curve Evolution keypoints

We use the DCE algorithm described in [17] to keep the most significant points of the silhouette contour. At each iteration of the DCE, instead of removing the complete set $V_{min}(P^i)$ from $Vertices(P^i)$ (see [17]), we remove only one vertex $v \in V_{min}(P^i)$. In this way, we can control how many vertices we want to keep in the final contour. We did not implement the topology preserving DCE as the final contours usually have no bad loops.

The contours end with 32 vertices at most (by default). Figure 2 gives an example of the keypoints found in each image. For a DCE keypoint, we use a feature vector with three components $\left(\vec{p}, K, \vec{\theta_c}\right)$:

- $\vec{p}$: Its normalized position in the referential of the silhouette.
- K: Its relevance measure. As defined in [17], the relevance measure is

$$K(\beta, l_1, l_2) = \frac{\beta l_1 l_2}{l_1 + l_2} \ , \qquad (1)$$

  where β is the external angle, and l_i is the length of the i^{th} edge normalized by the total length of the contour.
- $\vec{\theta_c}$: Its angles on the contour. For each keypoint, all keypoint's vertices are modeled as vectors that are oriented from the keypoint to neighboring keypoints on the DCE contour. Each keypoint has two angles.

4. Keypoint matching metrics

For skeleton-based keypoints matching, the following score is maximized:

$$S_{skel} = S_{dist} + S_{eucl} + S_{edge} + S_{angle} \ , \qquad (2)$$

where:

- S_{dist}: This first metric enforces the criterion that corresponding keypoints should be positioned similarly inside a pair of matching silhouettes.

$$S_{dist} = -|d_s^l - d_s^r| \ , \qquad (3)$$

  where d_s^l and d_s^r are the d_s feature components for the left and right images.
- S_{eucl}: This second metric is based on the hypothesis that corresponding keypoints should be positioned similarly in the respective referential of a pair of matching silhouettes. It is possible to evaluate the distance between the two keypoints. The second metric is the sigmoid

$$S_{eucl} = \frac{1}{1 + e^{-3+6d}} \ , \qquad (4)$$

  where d is the Euclidean distance in pixels between the two points' coordinate $\vec{p}$. The sigmoid parameters have been determined experimentally to obtain a smooth transition between the maximum and minimum score.
- S_{edge}: This third metric enforces the criterion that the number of skeleton edges should be similar between corresponding keypoints. The third metric is

$$S_{edge} = \begin{cases} 2 & \text{if } (n_{se}^l \geq 3) = (n_{se}^r \geq 3) \\ 0 & \text{if not} \end{cases} \ , \qquad (5)$$

  where n_{se}^l and n_{se}^r are the n_{se} feature components for the left and right image. The first condition is true if both keypoints have one edge or both keypoints have three or more edges. Thus, this criterion rejects matches between a keypoint with one edge and a keypoint with three and more edges.
- S_{angle}: This fourth metric also enforces skeleton edges similarity. Given the set of corresponding edges P_e, the fourth metric for a candidate pair of keypoints is

$$S_{angle} = \frac{\sum_{n \in P_e} \cos(\vec{\theta_n^l} - \vec{\theta_n^r})}{\max(n_{se}^l, n_{se}^r)} \ . \qquad (6)$$

Similarly, for DCE-based keypoints, the following score is maximized:

$$S_{dce} = S_k + S_{eucl} + S_{angle} \ , \qquad (7)$$

where

$$S_k = \frac{-|K^l - K^r|}{\left(\begin{cases} \sqrt{(K^l)^2 + (K^r)^2} & \text{if } \sqrt{(K^l)^2 + (K^r)^2} > 0 \\ 1 & \text{if not} \end{cases}\right)}, \tag{8}$$

where K^l and K^r are the relevance measures of the left and right keypoints respectively. Two corresponding keypoints should have the same relevance. The metrics S_{eucl} and S_{angle} are formulated similarly to the same metrics for skeleton-based keypoints.

4.1. Silhouette pair filter

The previous metrics matched points by comparing all keypoints of the left image to all the keypoints of the right image using a correspondence matrix. Of course, there are outliers that match keypoints from two silhouettes that do not correspond. In order to avoid these outliers, we classify each pair of keypoints in bins representing all possible pairs of corresponding silhouettes; this is another correspondence matrix, and the score is the number of pairs of keypoints. Then, it is quite easy to identify the most probable pairs of silhouettes. Finally, those pairs of silhouettes are used to compute and refine the pairs of keypoints.

5. Experimental validations

To test the validity of our proposed keypoints for correspondence between infrared and visible images, we used two planar homography applications. Keypoints are matched using our proposed metrics, and the matching pairs are used to calculate the homography. In the first application, the silhouettes are large and the scene is not planar. We perform partial image ROI registration on each silhouette. We compare results with DCE keypoints alone, skeleton keypoints alone, a combination of DCE and skeleton keypoints, and edge keypoints found using Phase congruency [14] as previously done for infrared image pairs [11]. The edge keypoints are matched using their gradient orientation and equation 4.

In the second application, the silhouettes are small and the scene assumed to be planar. We perform global image registration. This time, in addition, we also compare our results with a state-of-the-art method using trajectory point matching [21]. We use two publically available datasets, which are the OTCBVS dataset [7] and the LITIV dataset [21]. The homography is calculated using openCV with the RANSAC method. In the first application, the RANSAC distance is 4, in the second one, the distance is 2. In both applications, the silhouettes were obtained using background subtraction.

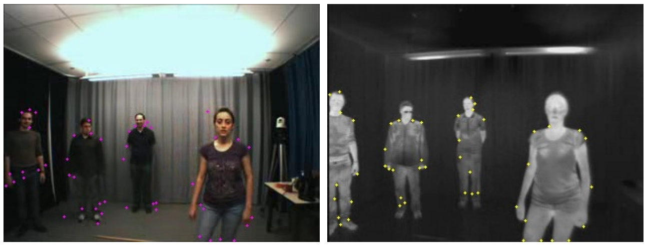

Figure 3. Stereo frames sample for the first experiment with DCE keypoints detected.

Table 1 shows the mean registration errors for 8 randomly selected frames (216, 456, 560, 724, 960, 988, 1247, and 1949) of a video pair (2494 frames with 320 x 240 resolution at 25 frames per second, Xvid codec) showing actors moving in a scene (see Figure 3). The table also shows the number of keypoint pairs selected to get these results from the initial set of keypoint pairs inputed to the homography calculation. There are three criteria that are important to qualify the quality of the keypoint pairs: 1) The number of keypoints found, 2) the number of keypoints that are good matches, and 3) the accuracy of their location on matching object structures. The DCE keypoints obey the best these three criteria. The registration error is the lowest of the tested methods, a good number of keypoint pairs are found and a large portion of these are good matches and are used for the homography calculation. Skeleton keypoints are lower in number, and unfortunately do not give a small registration error as we would have wished. In fact, their location on image structure is not stable because the skeleton can be sensitive to object segmentation error. Thus, even combined with DCE keypoints, the result is not as good as DCE alone. DCE keypoints may also be affected by segmentation errors, but the errors remain local, while the skeleton is affected more globally. For edge keypoints, the results are a little worse than for the skeleton in term of registration accuracy. Many keypoints are found, but many are rejected as bad correspondence. This reflects the fact that matching edge keypoints are not on similar location in the visible and in the infrared images.

Table 1. Partial image ROI registration

Keypoints	E_x	E_y	P_{init}	P_{used}
DCE	1.05	3.11	14.30	7.21
Skeleton	5.69	21.06	8.25	5.74
DCE+Skeleton	1.66	6.81	23.43	9.21
Edges	11.24	17.71	42.13	7.62

E_x,E_y: Mean registration error in x and y of all the ground-truth points to register. P_{init}: Number of keypoints initially, P_{used}: Number of keypoints selected for the homography calculation.

Table 2 gives results for global image registration. This time, we compare our method with the results reported by Torabi *et al*. [21] that uses a method based on trajectories of

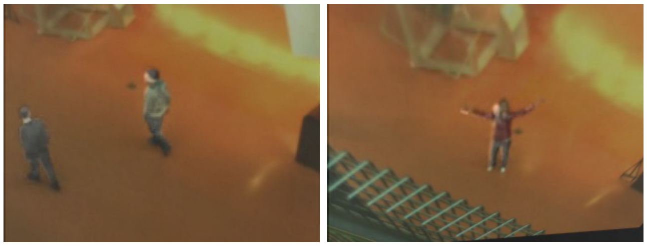

Figure 4. Registration results. Left: best result for Seq. 1, right: best result for Seq. 2.

moving objects. To contrast our results with only trajectory-based registration, we used their results without fusion for comparison. Because Torabi *et al.* use a matrix selection method to refine registration, our results are compared with theirs by using the minimum registration error that we find. Figure 4 shows the best registration results obtained for Seq. 1 and Seq. 2. Registration is less accurate for Seq. 2 and Seq. 3 because there is only one person in the scene (less keypoints that are not near collinear). Torabi *et al.* method gives more accurate results by 2 or 3 pixels in x and y, except for Seq. 1. In that case, we get better results because in that video two persons are moving nearby with parallel trajectories. The transformation is thus not well constrained by the trajectories. In the case of our method, we get points that are better position in the images giving a better estimation of the transformation.

In fact, our method is design to work on a single pair of images, while trajectory-based methods need sequence of images. Given the fact that we just calculate the homography using simple point correspondence in two images, we believe that our propose methodology is a valuable alternative. As for DCE, skeleton, DCE+skeleton and edges results, the same comments as previously applies. The hierarchy of the method is about the same. For some videos, edge keypoints perform very well because they are always in large number. However, the performance is less consistent than DCE keypoints, as exemplify by Seq 2.

These experiments allowed us to verify in which conditions our keypoints may be used for correspondence between an infrared image and a visible image. For good and numerous correspondences, the silhouette should be complete in both images and have a similar shape, that is, with matching convexities because of the DCE process. However, our proposed method can still work in the case of occlusions if, from the two views, parts of the contour have similar matching convexities. These parts of contour should be in a large enough number to obtain a reasonable number of corresponding keypoints. The same observation applies for contour deformation caused by segmentation. The skeleton keypoints are much more sensitive to the shape of the silhouette. In the end, it seems that DCE keypoints are sufficient, as the keypoints added by the skeleton method are not well enough localized for application requiring precision.

Table 2. Global image registration

$Seq.$	Method	E_x	E_y
1	DCE keypoints	3.23	2.90
	Skeleton keypoints	3.77	8.39
	DCE+Skeleton keypoints	3.36	2.34
	Edges	9.89	37.55
	Torabi *et al.* [21]	6.03	11.00
2	DCE keypoints	5.93	6.49
	Skeleton keypoints	26.06	22.14
	DCE+ Skeleton keypoints	5.95	7.16
	Edges	47.51	17.77
	Torabi *et al.* [21]	4.55	3.99
3	DCE keypoints	6.86	11.43
	Skeleton keypoints	12.29	22.71
	DCE+ Skeleton keypoints	9.19	16.21
	Edges	4.28	10.30
	Torabi *et al.* [21]	4.09	8.45
4	DCE keypoints	3.85	2.72
	Skeleton keypoints	26.11	15.95
	DCE+ Skeleton keypoints	4.93	2.79
	Edges	0.97	1.19
	Torabi *et al.* [21]	1.90	1.40

Seq.1-3, videos from LITIV dataset (dataset 01a)[21] and Seq. 4, videos from the OTCBVS dataset (dataset 03, seq. 5)[7]. E_x,E_y: Mean registration error in x and y of all the ground-truth points to register.

For successful registration using our keypoints, the following are required. In general:

- the silhouette should be complete in both images and have a similar shape.

if the objects of interest are small (< 2500 pixels of area):

- only global registration may be performed because more than one silhouette is required: 1) to avoid having only collinear keypoints, and 2) to have a reasonable number of keypoints pairs since smaller contour give points that not localized as well. Many keypoints are nearby.

if the objects of interest are relatively large (> 2500 pixels of area):

- one silhouette is sufficient and partial image ROI registration may be performed.

These observations mean that for small objects, partial image ROI registration is not possible because keypoints are nearby, but this is true for any keypoint-based registration method. Thus, for small objects, the image should be registered globally, but for larger objects partial image ROI registration can be performed.

6. Summary and conclusions

We have presented an alternative to region-based and trajectory-based correspondence methods that is suitable for visible-infrared stereo pairs. It accounts for the visual properties and differences between visible and infrared by extracting keypoints on the boundary and on the skeleton of an ROI silhouette. We tested our method for partial image ROI registration and show that it gives better results than other classic keypoints. We also tested our method for global registration and show that it gives results that are close and sometime better than a state of the art trajectory-based method, while having the benefit of not being dependent on temporal information.

Future work: Although we focused on defining new keypoints for visible-infrared correspondence, it would be useful to improve registration performance by including iterative refinement and an additional registration metric, like ROI silhouette overlap, to improve keypoints matching. Tracking of keypoints should also be considered in the case of video registration.

References

[1] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344–371, 1986. 69

[2] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. *Int. J. Comput. Vision*, 68(1):53–64, 2006. 69

[3] H. Chen, P. K. Varshney, and M.-A. Slamani. On registration of regions of interest (roi) in video sequences. In *AVSS '03: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 313–318, Washington, DC, USA, 2003. IEEE Computer Society. 69

[4] D. Cheriton and R. E. Tarjan. Finding minimum spanning trees. *SIAM Journal on Computing*, 5(4):724–742, 1976. 69

[5] E. Coiras, J. Santamaria, and C. Miravet. Segment-based registration technique for visual-infrared images. *Optical Engineering*, 39:282–289, jan 2000. 69

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893, 2005. 68

[7] J. W. Davis and V. Sharma. Fusion-based background-subtraction using contour saliency. In *CVPR '05: IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 11–19, 2005. 71, 72

[8] M. I. Elbakary and M. K. Sundareshan. Multi-modal image registration using local frequency representation and computer-aided design (cad) models. *Image Vision Comput.*, 25(5):663–670, 2007. 69

[9] D. Gallup, J. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1418 –1425, 2010. 68

[10] S. Gehrig and C. Rabe. Real-time semi-global matching on the cpu. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 85 –92, jun. 2010. 68

[11] K. Hajebi and J. S. Zelek. Sparse disparity map from uncalibrated infrared stereo images. In *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, pages 17–24, 2006. 71

[12] J. Han and B. Bhanu. Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40(6):1771–1784, 2007. 69

[13] X. Huang and Z. Chen. A wavelet-based multisensor image registration algorithm. In *Signal Processing, 2002 6th International Conference on*, volume 1, pages 773–776 vol.1, 2002. 69

[14] P. Kovesi. Phase congruency: A low-level image invariant. *Psychological Research*, 64(2):136–148, 2000. 71

[15] S. J. Krotosky and M. M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Comput. Vis. Image Underst.*, 106(2-3):270–287, 2007. 68, 69

[16] S. K. Kyoung, H. L. Jae, and B. R. Jong. Robust multi-sensor image registration by enhancing statistical correlation. In *Information Fusion, 2005 8th International Conference on*, volume 1, page 7, 2005. 69

[17] L. J. Latecki and R. Lakamper. *Polygon Evolution by Vertex Deletion*. : Scale-Space Theories in Computer Vision: Second International Conference, Scale-Space'99, Corfu, Greece, September 1999. Proceedings. 1999. 70

[18] L. J. Latecki and R. Lakamper. Shape similarity measure based on correspondence of visual parts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(10):1185–1190, 2000. 69

[19] O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 709 –716, jun. 2010. 68

[20] A. Roche, G. Malandain, and X. Pennec. The correlation ratio as a new similarity measure for multimodal image registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI98)*, volume 1496, pages 1115–1124, 1998. 69

[21] A. Torabi, G. Masse, and G.-A. Bilodeau. Feedback scheme for thermal-visible video registration, sensor fusion, and people tracking. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 15 –22, jun. 2010. 69, 71, 72