Online Multimodal Video Registration Based on Shape Matching

Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau
LITIV lab., Dept. of Computer & Software Eng.
Polytechnique Montréal
Montréal, QC, Canada
{pierre-luc.st-charles,gabilodeau}@polymtl.ca

Robert Bergevin
LVSN - REPARTI
Université Laval
Québec City, QC, Canada
robert.bergevin@gel.ulaval.ca

Abstract

The registration of video sequences captured using different types of sensors often relies on dense feature matching methods, which are very costly. In this paper, we study the problem of "almost planar" scene registration (i.e. where the planar ground assumption is almost respected) in multimodal imagery using target shape information. We introduce a new strategy for robustly aligning scene elements based on the random sampling of shape contour correspondences and on the continuous update of our transformation model's parameters. We evaluate our solution on a public dataset and show its superiority by comparing it to a recently published method that targets the same problem. To make comparisons between such methods easier in the future, we provide our evaluation tools along with a full implementation of our solution online.

1. Introduction

Although automatic multimodal image and video registration has been intensively studied over the years [25, 15], most methods still rely on dense feature matching through area-based similarity measures computation [17, 18, 11, 5]. While this approach has the advantage of being able to register non-planar scenes, it is generally very computationally expensive, thus making it unsuitable for emerging mobile and distributed computer vision applications where information fusion might be required. Many multimodal surveillance systems capture images at medium/long distances from their targets, meaning that in those cases, planar models can be assumed without excessively compromising registration quality. In this context, lightweight approximate registration solutions can be adopted to replace their more complex counterparts. Instead of densely searching for local correspondences between images, lightweight solutions rely on sparse correspondences taken from common salient features, which are fitted to a parametric model in order to find a frame-wide rigid or projective transformation (homography).

The main problem behind this simplified approach is finding features that are shared between the studied image modalities and that are easy to automatically identify and match. As presented in [1], traditional keypoint detectors and invariant descriptors are not well suited to multimodal imagery, and require important tuning to achieve decent results. Specialized detection and description methods have been proposed to address this problem [1, 14, 23], but these methods do not work for image modalities where the relation between the appearance of objects is not easily defined. These methods are also inadequate when image resolution is too low or when there is a lack of similar textural content between the images. Therefore, more robust means of finding matches between images have to be considered.

In this paper, we propose an automatic video registration method that relies on correspondences found via shape-of-interest matching. Instead of independently analyzing each video frame to extract salient features or edge maps to use for correspondences, we rely on shape contours found using continuous foreground-background video segmentation. While this restricts our approach to applications where the targets of interest can be automatically segmented, it is not affected by the difference in pixel color characteristics of the studied image modalities. This means that our method can be used with any type of sensor, as the appearance of the targets does not matter (as long as they can be segmented).

Our first contribution is a strategy to preserve good shape contour matches throughout the analyzed sequences, which makes our transformation estimation approach robust to continuously imperfect segmentation and small, static targets that do not contribute useful correspondences. Instead of temporally accumulating correspondences in a first-in, first-out buffer for RANSAC-based fitting [9], we use a buffer in which correspondences that are identified as persistent outliers based on a voting scheme are randomly replaced. Our second contribution is a method for smoothing transitions between transformations estimated at different times based on the approximate overlap of the analyzed

foreground shapes. This prevents our solution from locking on to a global registration transformation that might not adequately reflect the nature of the studied scene (*e.g.* when the scene is not truly planar). Under the assumption that the foreground shapes are the real targets of interest in the scene, this smoothing approach allows improved overall registration through the continuous update of the transformation model's parameters.

We evaluate our solution by comparing it to a recently proposed method using a public dataset, and show that our overall strategy is superior. To make future comparisons on this dataset and with our method easier, we have made our source code public[1], and we provide the video segmentation masks and the evaluation tools we used[2].

2. Related Work

As described in [11], a planar model can be assumed for image or video registration in two cases: 1) when the sensors are nearly collocated and the alignment targets are far from them (infinite homographic registration), or 2) when all alignment targets lie on the same plane in the scene (planar ground registration). In both cases, the parallax effects caused by the camera baseline distance are assumed to be negligible. Registration can then be achieved by having a human expert select various keypoints in one image modality and search for their equivalent in the other, and by solving the parametric transformation model using these correspondences. Manual registration is however troublesome when large datasets containing many different sensor configurations have to be processed, as it is extremely time-consuming. Automatic video registration thus has to solve the multimodal keypoint detection and matching problem, and provide a robust way to identify the best homography within a temporal window.

Target silhouette information and edge maps have been used before to find correspondences between multimodal image sets [7, 16, 14, 21]. Coiras *et al*. [7] estimated transformations in outdoor urban environments by extracting straight lines from edge maps and using them to find correspondences between sets of polygons. More recently, Pistarelli *et al*. [16] proposed to use Hough space as the search domain to find multimodal correspondences between segments under similar conditions. Mouats and Aouf [14] opted to use a keypoint detector based on phase congruency instead of edge points directly, but relied on local edge histograms to describe and match them between modalities. Tian *et al*. [21] also used edge maps for thermal-visible face registration, but described point sets using shape context [2]. In their case, the infrared images were very contrasted, meaning that many strong edges were easily identifiable. Strategies based on silhouettes and edge maps are not always adequate, as contours representing region boundaries are not guaranteed to be shared between all modalities. For example, when using thermal-infrared sensors, objects with uniform temperatures might not display any edges while some are identifiable in other spectra.

The advantage of video registration over image registration is that methods can rely on target motion to find correspondences more easily. Shape contours obtained via temporal foreground-background segmentation [10, 3, 24, 19] or shape trajectories [6, 4, 22] are viable strategies, as long as they can properly distinguish targets from the background. Correspondences found between shape contours or trajectories are used to find the transformation model parameters, but their quality highly depends on the accuracy of the segmentation algorithm. In [10], a hierarchical genetic algorithm is adopted to quickly find an optimal transformation while avoiding local maxima. In [24], the authors simplified their transformation model using calibration priors; they found optimal parameters through a 1D-scan approach based on heuristics. The works of [19, 6, 22] all rely on a similar strategy: correspondences from a temporal window are added to a global potential match buffer (or "reservoir"), which is then analyzed using a random sample consensus (RANSAC)-based method [9] to find the parameters which best fit the transformation model. In [6], correspondences from all frames are used at once, meaning that it cannot estimate the registration transformation in an online fashion. In [19, 22], a small first-in, first-out (FIFO) buffer is used to accumulate and analyze a few seconds worth of correspondences. Our own strategy presented in Section 3.3 uses a similar RANSAC approach for online transformation estimation, but accumulates contour matches using a random sampling and persistence voting approach (meaning that these matches do not have a predefined lifetime in the buffer).

The shape contour description and matching strategy that most closely resembles ours is that of [19]: they used the Discrete Curve Evolution (DCE) algorithm originally proposed in [13] to prune foreground shapes into hexadecagons with visually similar boundary parts. We believe that more accurate registration can be achieved by describing and matching all shape contour points without pruning, leaving the filtering responsibility to RANSAC. Therefore, for our own contour description and matching needs, we use shape context [2], as detailed in Section 3.2.

3. Proposed method

Our method can be split into several parts, as shown in Fig. 1. First, foreground-background segmentation is used on each video frame to obtain shape contours from targets present in the scene. Contour points are then described and matched using the iterative shape context approach of [2].

[1] https://bitbucket.org/pierre_luc_st_charles/multimodal-vid-reg
[2] http://www.polymtl.ca/litiv/vid/index.php

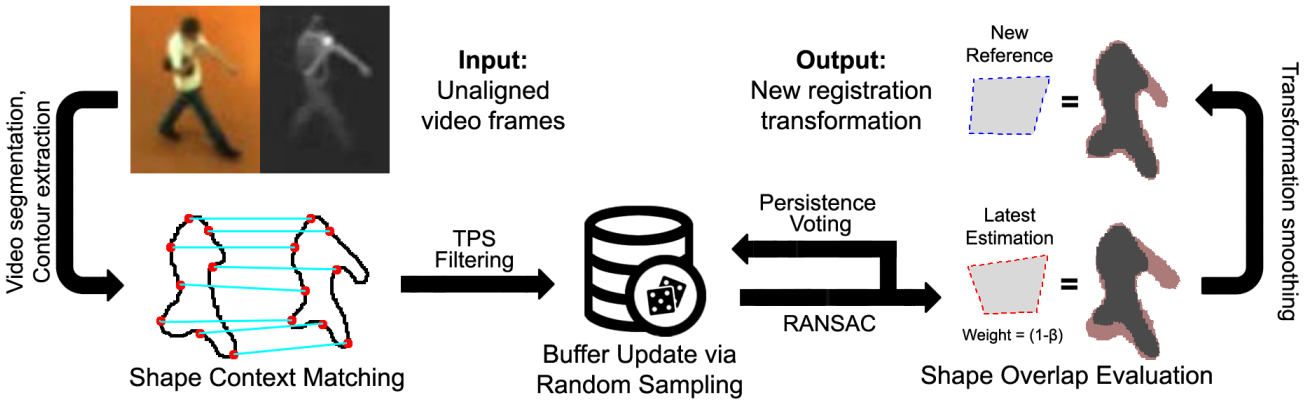

Figure 1: Overview of our proposed method's principal processing stages.

Following that, all matches are added to a correspondence reservoir (*i.e*. a temporal buffer), which itself is analyzed by a RANSAC algorithm to identify inliers and outliers and to estimate ideal transformation parameters. Finally, the identified inliers are used for persistence voting in the reservoir, and the estimated model parameters are used to update the reference (or "best-so-far") registration homography. In the following subsections, we detail each step of this process.

3.1. Shape extraction

The initial step in our method is identifying targets of interest in each video sequence using foreground-background segmentation. Since the dataset we use in Section 4 only contains sequences with static cameras, we opted for an approach based on change detection via background modeling (commonly referred to as "background subtraction"). The method we use is the one in [20]: it builds a statistical model of the observed scene using color and binary features, and uses feedback mechanisms to dynamically adapt to changing conditions. Getting shape contours from the binary image masks provided by this method is trivial, and no pruning or extra post-processing is done to simplify these foreground shapes. We chose this segmentation method due to its ease-of-use and because it provides good segmentation results in image modalities inside and outside the visible spectrum.

3.2. Contour points description and matching

We address the shape contour matching problem as the task of establishing correspondences in a bipartite graph where the disjoint sets are composed of noisy polygon vertices taken from different image modalities. As mentioned earlier, we follow the approach introduced in [2] to compare foreground object contours. Our ultimate goal is to find, for each contour point in the first image modality, the contour point in the second modality that offers the best match given their respective position within their original shapes.

This approach is straightforward: first, contour points are all assigned a shape context descriptor which expresses the relative disposition of other contour points in the same modality using a uniform log-polar histogram (as shown in Fig. 2). These descriptors are then exhaustively compared using a χ^2 test to determine similarity scores, and the correspondence problem is solved using the Hungarian method [12]. These three steps are repeated multiple times for each frame in order to eliminate outlying matches from the contours, which are identified after solving the correspondence problem based on their low similarity scores. Between each iteration, a Thin Plate Spine (TPS) model [8] is used to determine the optimal elastic transformation that aligns the filtered contours, and new descriptors based on the transformed shapes are generated. Given a predetermined maximum number of iterations to run, this approach helps identify which correspondences will be used to find the frame-wide registration transformation detailed in Section 3.3. Note that the transformation estimated by the TPS method cannot be used for frame-wide image registration, as its elastically fitted solution might cause important distortions in regions far from the analyzed contour points. Therefore, it is only used as a temporary solution, and the contour point matches later added to the correspondence buffer contain their original coordinates.

Unlike the DCE approach of [19] that directly relies on Euclidean distances between multimodal contours, our approach is completely invariant to translations and scaling since all distances in shape context descriptors are relative and normalized. Furthermore, it is not restricted to a constant number of points per contour and it does not consider boundary convexity as a shape attribute, meaning it is more robust to noisy shapes caused by inadequate segmentation. The correspondence problem is also solved optimally, which is better than using the greedy matching algorithm of [19]. Besides, note that due to the unknown relation between the analyzed image modalities, we do not consider local appearance when computing the similarity scores between shape contours.

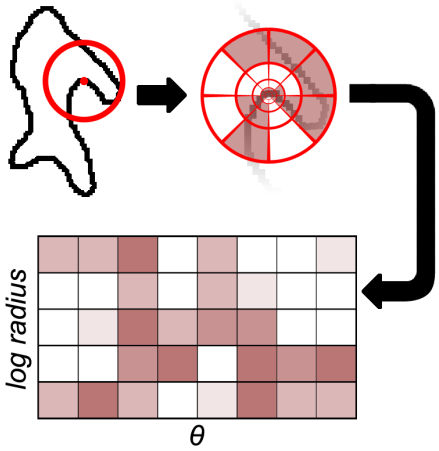

Figure 2: Example of shape context description on a human shape contour point using 5 radius bins and 8 angle bins.

3.3. Correspondence reservoir and voting

Using contour matches from a single frame pair for scene registration would likely result in noisy transformations that disregard large planar scene areas (which might be of interest to some applications). Therefore, the homography we are looking for has to be computed using correspondences taken from multiple frames to ensure accurate scene-wide registration. To address this problem, we use a temporal buffer (or reservoir) to accumulate enough correspondences so that a robust model fitting algorithm, RANSAC [9], can estimate a proper global homography.

In [19], a first-in, first-out (FIFO) circular buffer strategy was adopted to keep 100 frames worth of contour point pairs. While easy to implement, the primary disadvantage of this approach is that if the targets of interest remain static (or do not move much) during those 100 frames, or if the segmentation is continuously inaccurate, the buffer will be filled with correspondences that are not representative of the sought frame-wide transformation. To solve this problem, we use an identically sized buffer, but instead of following a FIFO rule, we replace the correspondences it contains using a random policy reminiscent of conservative sample consensus models.

Simply put, given a reservoir $R = \{p_1, p_2, \ldots, p_N\}$ containing N previously found point pairs, for each new point pair p found by multimodal contour matching, we will randomly pick one of the reservoir pairs and replace it, but only if it is considered a "persistent outlier". These persistent outliers can be identified based on the number of times they were omitted by the RANSAC algorithm during the estimation of the homography parameters for each new frame. To keep track of this, we define a voting map, noted $V = \{c_1, c_2, \ldots, c_N\}$, which accumulates the inlier/outlier counts of each point pair (following RANSAC fitting) based on this logic:

$$c_i = \begin{cases} c_i + 1 & \text{if } p_i \text{ is an inlier according to RANSAC} \\ c_i - 1 & \text{otherwise} \end{cases}$$

For a new frame, all pairs p_i which have negative c_i values are considered persistent outliers. In practice, we determined that with this approach, the proportion of such outliers is always around 50%, which means that the reservoir never saturates and new correspondences can always be swapped in. We also determined that due to the presence of our reservoir, even a small shape moving across a limited portion of the sensor's field of view can provide enough contour matches to estimate a good global homography; this is discussed further in Section 4.

3.4. Homography smoothing

Following the planar ground assumption to simplify video registration tasks might not always faithfully reflect the reality of the observed scenes. Therefore, the goal of our method is not to simply find a good frame-wide homography for the entire analyzed sequence, but to make sure that this transformation adequately aligns the targets of interest, even when non-planar transformations are involved. Basically, we are looking for a middle ground between estimating a timeless, global homography and estimating one that only focuses on aligning currently visible targets of interest without regard to previously found homographies or to the rest of the scene. To achieve this, while processing a pair of video sequences, we continuously update the homography which results in the best registration seen thus far (determined heuristically) using the model parameters found via RANSAC for each new frame. Under the assumption that segmented foreground shapes are truly objects of interest, this allows for slightly improved contour alignment in "almost planar" scenes while minimally affecting the registration quality of other frame regions. When the analyzed scene fully respects the planar assumption, the results provided by this middle ground approach are identical to those of the global, timeless approach.

First, we describe how the registration quality of a homography is appraised at run time. The only data which can be used to assess if a transformation is appropriate are the foreground shapes obtained by the segmentation step. Thus, the appraisal metric we use is the overlap error, defined for two foreground shapes S_i and S_j (both in the same coordinate space) as

$$E(S_i, S_j) = 1 - \frac{\#(S_i \cap S_j)}{\#(S_i \cup S_j)}, \qquad (1)$$

where $\#(S)$ returns the pixel count of foreground region S. Given perfectly segmented targets in a truly planar scene, a null overlap error would indicate an ideal registration transformation (both shapes are perfectly aligned).

Algorithm 1 Homography smoothing for each frame.

if $E_{new} < E_{curr}$
  if $E_{new} < E_{ref} \ \vee \ E_{new} < E_{curr} \cdot 2$
    $\alpha \leftarrow 2$
  else
    $\alpha \leftarrow \alpha + 1$
  end if
  $\beta \leftarrow \frac{\alpha - 1}{\alpha}$
  $E_{ref} \leftarrow E_{ref} \cdot \beta \ + \ E_{new} \cdot (1 - \beta)$
  $H_{ref} \leftarrow H_{ref} \cdot \beta \ + \ H_{new} \cdot (1 - \beta)$
end if

As for the smoothing step itself, given a smoothing factor α (by default, α=2), a newly estimated homography H_{new}, its calculated foreground shape overlap error E_{new}, a reference (or "best-so-far") homography H_{ref}, its reference (or "best-so-far") overlap error E_{ref}, and finally H_{ref}'s current overlap error on the latest foreground shapes E_{curr}, we follow the strategy detailed in Algorithm 1 to smooth the homography transition between two frames. In summary, if the new homography H_{new} produces an overlap error E_{new} smaller than what the reference homography H_{ref} produced on the current foreground shapes (E_{curr}), then the reference needs to be updated. In that case, if the disparity between the two errors is large enough, or if the new error is simply smaller than the reference error (E_{ref}), the reference homography (H_{ref}) and error value (E_{ref}) will be replaced by the straight average between them and their newly found counterparts. Otherwise, they will be replaced by a weighted mean which depends on the value of the smoothing factor (α).

The role of α is to control the weight given to the reference homography when it is combined with a new one. It is responsible for automatically balancing our method between directly using new homographies for each frame which focus on aligning matched contour points, and converging to a global homography which fits the entire scene more adequately. The latter case can be achieved when $\alpha \rightarrow \infty$, but in practice, this is unlikely to happen; as we will see in Section 4, our method never truly stops adapting to newly estimated homographies. Overall, as noted earlier, this smoothing strategy allows for better alignment of contours when considering "almost planar" scenes, and it helps quickly stabilize the registration when processing the first video frames with contour matches.

4. Evaluation

Evaluating how well a method behaves in terms of registration quality for "almost planar" scenes is not trivial. In the case of real planar scenes, the sought homography can be found manually, and registration quality can be evaluated by calculating the distance between points projected using this homography and the automatically estimated one. For scenes that do not fully respect the planar assumption, results have to be qualitatively evaluated, or a criterion based on the degree of overlap of manually identified scene structures has to be used. Such a criterion is proposed by Torabi *et al*. in [22]: for their own visible-infrared registration dataset, they manually selected points throughout the frames of their sequence pairs which were easily identifiable and matchable, and connected them to create polygons sets (an example of this operation is shown in Fig. 3). Once a polygon set is projected into the other's coordinate space, the overlap error can be used as the criterion to judge the image registration quality. By choosing points on scene elements that do not respect the planar assumption, one can highlight part of the non-rigid transformation that needs to be modeled by the automatic approach. Using these polygons instead of the segmented foreground shapes to calculate the overlap error eliminates the uncertainty caused by inaccurate video segmentation, and it allows a better coverage of the observed scene.

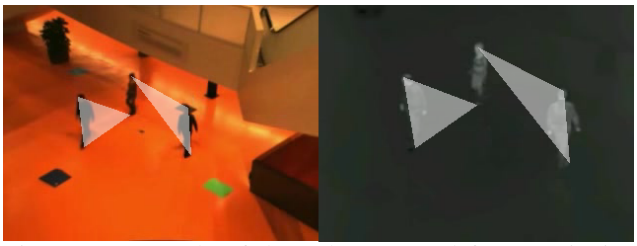

Figure 3: Example of the polygons used for quantitative evaluation formed with manually identified keypoints in the ninth sequence pair of the LITIV dataset.

For our own tests, we also use the LITIV dataset of [22]. However, since the polygons they manually drew for their quantitative evaluations could not be obtained, we had to draw our own. To make future quantitative comparisons between video registration methods easier, we have made these new polygon sets, the segmentation masks we obtained from [20] as well as our evaluation tools available online, along with a C++ implementation of our method[3].

In total, nine visible-infrared sequence pairs of lengths varying between 200 and 1200 frames were analyzed. These were taken with different sensor baselines at various orientations from the ground plane. Homographies found by matching manually identified points are provided in the dataset, and are used in the following figures to illustrate ground truth global registration results. The authors of [19] provided us with an implementation of their own method, which we use as our basis for comparison. For fairness, both methods rely on the same segmentation results, both are evaluated using the overlap error defined in (1), and both use a single parameter set for all sequences.

[3] Links are given at the end of Section 1.

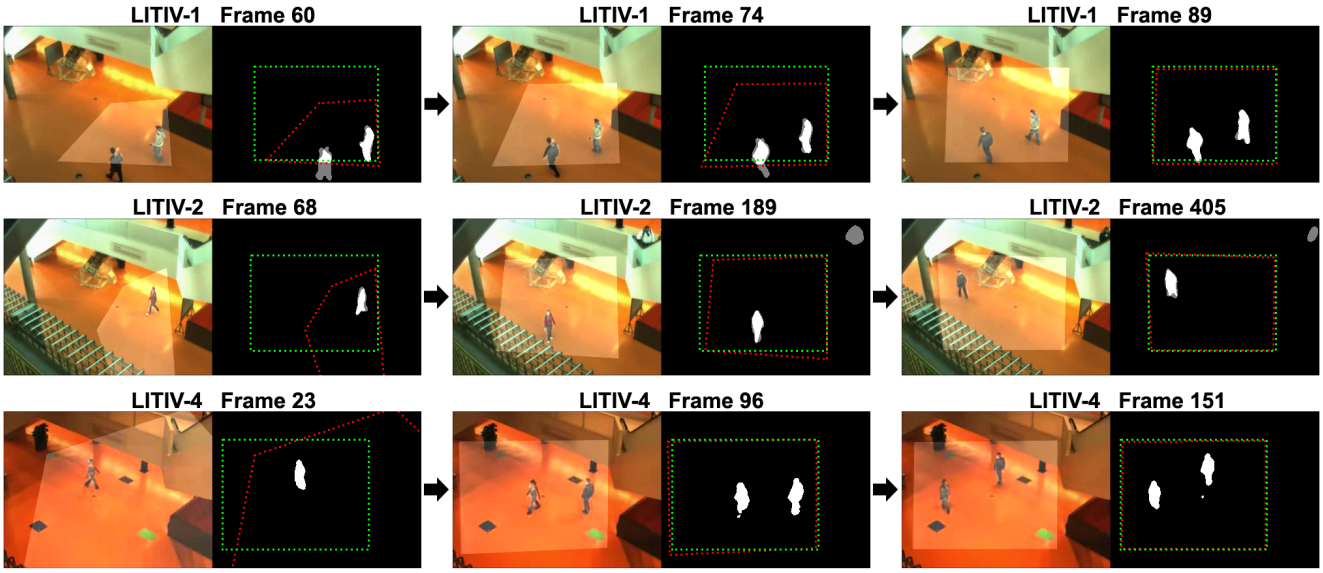

Figure 4: Registration results obtained at various moments of the first, second and fourth sequence pairs of the LITIV dataset using our proposed method. The left image in each pair shows the estimated frame-wide registration, and the right shows foreground shape registration at the same moment. The red dashed polygon shows the estimated transformation applied to the infrared image boundary, and the green one shows the ground truth transformation applied to this same boundary.

We show in Fig. 4 how our method performs in the first, second and fourth sequence pairs of the studied dataset. We can see that for various sensor placements, an acceptable alignment of foreground shapes is found soon after a target first becomes visible (this happens at different moments in each sequence; our earliest results are shown in the left column). Over time, this alignment is refined to make the registration of other scene elements possible. For the first sequence pair (top row), even though the detected targets only travel in a small portion of the sensor's field of view, a very good transformation is found less than 30 frames after the first appearance of foreground shapes.

For quantitative evaluation, since online video registration takes time to stabilize, it makes little sense to compare average error measures that might be affected by important aberrations present early in the analyzed sequences. Previous works [19, 22] addressed this issue by either considering only the minimum errors achieved for each sequence pair, or by arbitrarily picking time intervals where the method is considered "stable", and computing average metrics from those. In our case, in order to present a global view of how our method adapts to each newly estimated homography, we present error-to-time curves for our method and compare them to those of [19] in Figs. 5 and 6. The overlap errors of Fig. 5 are calculated using (1) with the ground truth polygonal shapes, and the Euclidean distance errors of Fig. 6 are calculated with the vertices of these polygons. In both cases, the transformation is applied on the infrared set, and the common coordinate space is in the visible image.

From the results shown in Fig. 5, we can note that our method reaches lower overlap errors faster than [19], stabilizes at those levels more often, and manages to outperform the ground truth homography in five out of nine sequence pairs (LITIV-4 and LITIV-6 through LITIV-9). Outperforming the ground truth is possible because its homography reflects a global transformation only ideal for a planar scene, and the manually drawn polygons, just like the rest of the scene, do not fully respect the planar assumption. Since these polygons are partially based on points found on the targets of interest, transformations that focus on the alignment of these targets are more likely to get smaller overlap errors.

Besides, our method had no trouble estimating frame-wide registrations for LITIV-7 and LITIV-8, unlike [19], which was unable to find adequate homographies through the entire lengths of these sequences. In LITIV-7, we can see a strong temporary increase in overlap error near the end of the sequence: this is due to the matching of a single small shape with a strong shadow which produces outliers for a long period of time. This error quickly fades as better homographies are estimated after the target starts moving in the following frames.

As shown in Table 1, our method reached much lower minimum errors than [19] in all but one sequence pair (LITIV-5), where the difference between the two is very small. In all cases except LITIV-4 and LITIV-7, an homography resulting in an overlap error of less than 50% was

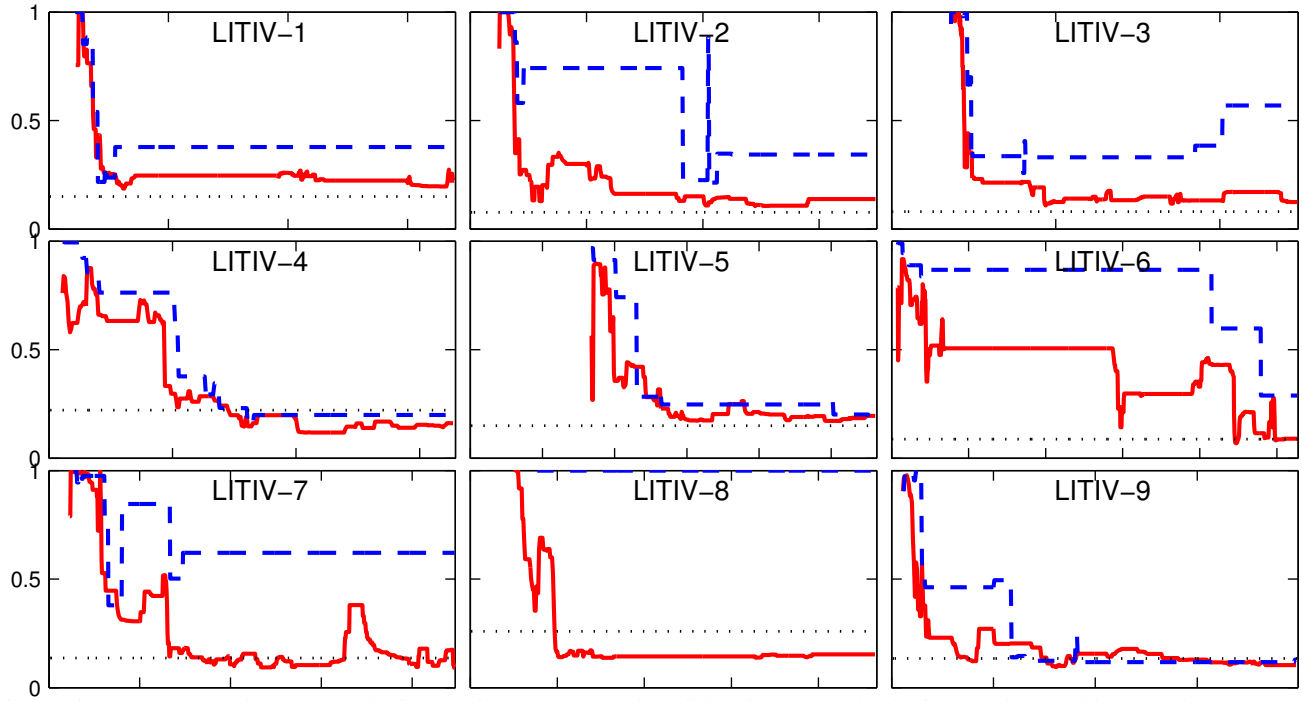

Figure 5: Polygon overlap errors obtained using our method (solid red), the method of [19] (dashed blue), and the ground truth homography (dotted gray) for the full lengths of all sequence pairs of the LITIV dataset.

Sequence pair	Proposed	Sonn *et al.* [19]
LITIV-1	**0.187**	0.217
LITIV-2	**0.106**	0.214
LITIV-3	**0.108**	0.258
LITIV-4	**0.118**	0.152
LITIV-5	0.172	**0.167**
LITIV-6	**0.069**	0.289
LITIV-7	**0.091**	0.379
LITIV-8	**0.137**	1.000
LITIV-9	**0.095**	0.117

Table 1: Minimum overlap errors achieved for all video sequence pairs of the LITIV dataset (bold entries indicate the best result).

found after processing less than 30 frames (after the first appearance of foreground) containing at least one shape visible in both fields of view.

The curves illustrating polygon vertices registration errors shown in Fig. 6 generally depict the behavior observed in Fig. 5, but with a larger gap between our method and [19] for LITIV-6 through LITIV-9. In three of those cases, the curves of [19] are mostly outside the 15 pixels error range of the graphs, but our method reaches 2 or 3 pixels errors by the end of each sequence pair. We can also notice in the last graph of this new figure (LITIV-9) that our smoothing approach prevents our solution from locking onto a "decent" homography, and instead continuously refines one to achieve extremely small registration errors at the end of the sequence.

As for the computation time, when operating directly on the foreground shapes provided by the video segmentation algorithm, our proposed registration method processed video sequences at speeds varying between 15 and 150 frames per second, depending on the number of targets in the scene (we used C++ code on a laptop's 4th generation Intel i7 CPU at 2.8 GHz).

5. Conclusion

In this paper, we presented an online multimodal video registration method that relies on the matching of shape contours to estimate the parameters of a planar transformation model. We showed that randomly sampling a correspondence buffer and adding temporal smoothing between estimated homographies can quickly lead to stable results, and even allow "almost planar" scenes to be registered adequately. Our solution outperforms a recently published method. It manages to align manually annotated polygon sets based on scene structures better than the ground truth homography could in the majority of sequences we tested. Given adequate target segmentation, this approach could be used to register image sequences from cameras which are slowly moving, but still have overlapping fields of view.

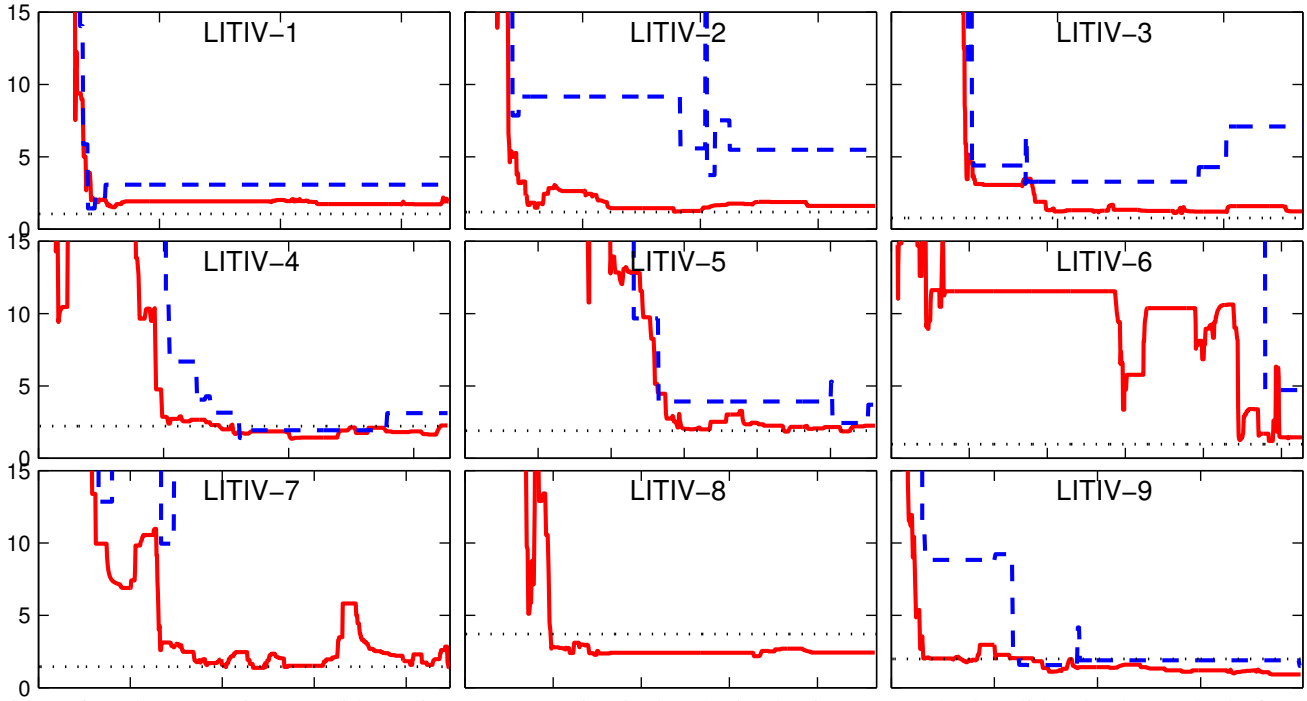

Figure 6: Polygon vertices Euclidean distance errors (in pixels) obtained using our method (solid red), the method of [19] (dashed blue), and the ground truth homography (dotted gray) for all sequences of the LITIV dataset. Note that the Y axis has been cropped similarly for all graphs.

It could also be generalized to non-planar registration if transformations were continuously estimated for each foreground shape.

6. Acknowledgements

This work was supported by FRQ-NT team grant No. 2014-PR-172083 and by REPARTI (Regroupement pour l'étude des environnements partagés intelligents répartis) FRQ-NT strategic cluster.

References

[1] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–12672, 2012.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, Apr 2002.

[3] G.-A. Bilodeau, P. St-Onge, and R. Garnier. Silhouette-based features for visible-infrared registration. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 68–73, June 2011.

[4] G.-A. Bilodeau, A. Torabi, and F. Morin. Visible and infrared image registration using trajectories and composite foreground images. *Image and Vis. Comp.*, 29(1):41–50, 2011.

[5] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi. Thermalvisible registration of human silhouettes: A similarity measure performance evaluation. *Infrared Physics & Technology*, 64(0):79–86, 2014.

[6] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. In *Proc. European Conf. Comput. Vis. Workshops*, 2002.

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

[8] J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive Theory of Functions of Several Variables*, volume 571, pages 85–100. Springer Berlin Heidelberg, 1977.

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

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

[11] S. J. Krotosky and M. M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Comput. Vis. and Image Understanding*, 106(23):270–287, 2007.

[12] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[13] L. J. Latecki and R. Lakämper. Shape similarity measure based on correspondence of visual parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1185–1190, 2000.

[14] T. Mouats and N. Aouf. Multimodal stereo correspondence based on phase congruency and edge histogram descriptor. In *Proc. 16th Int. Conf. on Inf. Fusion*, pages 1981–1987, July 2013.

[15] F. P. Oliveira and J. M. R. Tavares. Medical image registration: a review. *Comput. Meth. in Biomech. and Biomed. Eng.*, 17(2):73–93, 2014.

[16] M. Pistarelli, A. Sappa, and R. Toledo. Multispectral stereo image correspondence. In *Computer Analysis of Images and Patterns*, pages 217–224. Springer Berlin Heidelberg, 2013.

[17] J. Pluim, J. Maintz, and M. Viergever. Image registration by maximization of combined mutual information and gradient information. *IEEE Trans. Med. Imag.*, 19(8):809–814, 2000.

[18] J. Pluim, J. Maintz, and M. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Trans. Med. Imag.*, 22(8):986–1004, 2003.

[19] S. Sonn, G.-A. Bilodeau, and P. Galinier. Fast and accurate registration of visible and infrared videos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 308–313, June 2013.

[20] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin. A self-adjusting approach to change detection based on background word consensus. In *Proc. IEEE Winter Conf. Appl. Comp. Vis.*, pages 990–997, Jan 2015.

[21] T. Tian, X. Mei, Y. Yu, C. Zhang, and X. Zhang. Automatic visible and infrared face registration based on silhouette matching and robust transformation estimation. *Infrared Physics & Technology*, 69(0):145–154, 2015.

[22] A. Torabi, G. Massé, and G.-A. Bilodeau. An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116(2):210 – 221, 2012.

[23] Y. Ye and J. Shan. A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences. *J. Photogrammetry and Remote Sens.*, 90(0):83–95, 2014.

[24] J. Zhao and S.-C. S. Cheung. Human segmentation by geometrically fusing visible-light and thermal imageries. *Multimedia Tools and Appl.*, 73(1):61–89, 2014.

[25] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vis. Comp.*, 21(11):977–1000, 2003.