Non-planar Infrared-Visible Registration for Uncalibrated Stereo Pairs

Dinh-Luan Nguyen
Faculty of Information Technology
University of Science, VNU-HCMC
Ho Chi Minh city, Vietnam
1212223@student.hcmus.edu.vn

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

Abstract

Thermal infrared-visible video registration for non-planar scenes is a new area in visual surveillance. It allows the combination of information from two spectra for better human detection and segmentation. In this paper, we present a novel online framework for visible and thermal infrared registration for non-planar scenes that includes foreground segmentation, feature matching, rectification and disparity calculation. Our proposed approach is based on sparse correspondences of contour points. The key ideas of the proposed framework are the removal of spurious regions at the beginning of videos and a registration methodology for non-planar scenes. Besides, a new non-planar dataset with an associated evaluation protocol is also proposed as a standard assessment. We evaluate our method on both public planar and non-planar datasets. Experimental results reveal that the proposed method can not only successfully handle non-planar scenes but also gets state-of-the-art results on planar ones.

1. Introduction

The problem of thermal infrared and visible spectrum (TIR-Vis) video content registration is a familiar task in computer vision. The fundamental idea of registration is finding correspondences from video frame pairs to allow scenes and objects to be represented in a common coordinate system. Some works proposed to use dense feature matching for high registration quality [4, 12, 15] while others [17, 16, 5] use sparse correspondences taken from common salient features for fast registration. Although these systems have some contributions in this area, they still have drawbacks that need to be solved. We address their three main disadvantages as follows.

First, dense correspondence methods that use area-based measurement to match correspondences from two frame pairs are too slow to be applied on videos [12, 4]. Thus, there is a need for a lightweight method to boost the speed of this registration process. Furthermore, these methods need rectified video frames which are not always readily available when tackling non-planar scenes (i.e. scenes in which objects appear on different depth planes). Some authors have proposed their own dataset [4] along with rectified videos created by calibration as inputs. These works cannot adapt to raw input video captured from different cameras. Besides, in video applications, the registration quality can be lower. As a result, in this paper, we address the problem of sparse feature correspondence for fast registration.

Second, existing sparse correspondence methods [16, 5] can only deal with planar scenes. Their frameworks assume that all captured scenes are approximately planar. Thus, this assumption limits their applicability to planar scenes only.

Third, since most sparse methods [16, 5] rely on brute force matching strategies, their computational complexity is usually quite high. They are thus unsuited for mobile or distributed video surveillance applications.

The typical structure of current existing frameworks used for sparse registration comprises three main steps, which are feature extraction, feature matching and image warping. In feature extraction and matching, traditional feature descriptors are exploited using sparse correspondence [16] between multimodal imagery [1]. Other technique has been proposed [13] to get more meaningful features from two types for TIR-Vis registration. However, these techniques are not always successful because of the differences in texture and resolution of TIR-Vis image pairs. In the image warping step, with the assumption that all captured scenes are nearly planar, a homography transformation is applied to maximize overlap area between objects. It should be noted that no existing framework uses unrectified videos as inputs for TIR-Vis non-planar scene registration. In this paper, we address the drawbacks of current existing systems in the TIR-Vis video registration problem for both planar and non-planar scenes.

Main contribution. There are four significant contributions presented in this paper. First, a novel method for align-

978-1-5090-1437-8/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPRW.2016.48

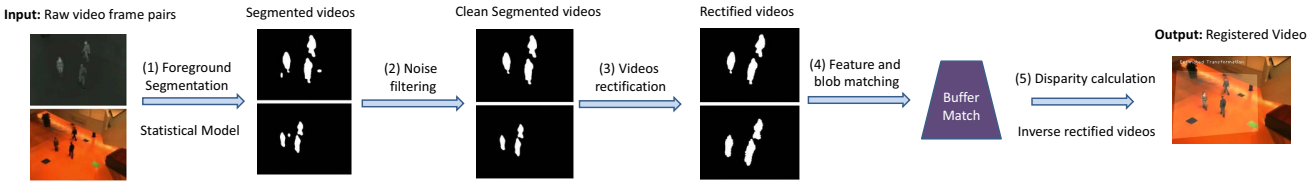

Figure 1. Proposed framework overview. First, raw input videos are segmented to get foreground objects using a statistical model [18]. Second, a noise filtering strategy is applied to eliminate spurious noisy blobs in segmented videos to reduce unnecessary computation. Third, videos are rectified using fundamental matrix estimation. Fourth, disparities are calculated from corresponding blobs of pair of video frames. Finally, videos are derectified to restore frames to raw input condition.

ing TIR-Vis blobs using sparse correspondences for raw input videos is proposed to deal with non-planar scenes. Experimental results show that the proposed framework also outperforms the state-of-the-art on planar scenes.

Second, a segmentation noise filtering strategy is presented to eliminate spurious blobs at early processing stages, which reduces unnecessary calculations afterward.

Third, a corresponding blob preservation algorithm is introduced to approximate correspondence between blobs in every frame without using a brute force method. The advantage is that the correspondence list needs only to be updated when there is a change in the order of objects in video pairs.

Fourth, we created a new public dataset for TIR-Vis registration with raw input video[1]. Groundtruth together with an evaluation protocol are also presented to simplify comparisons between different frameworks in the future.

2. Related work

To get features from both TIR-Vis videos, the first works on the topic [6, 8] used edge maps and silhouette information. Shape skeletons have also been exploited as features to estimate homographic registration transformations [3]. Besides, blob tracking [20] was also utilized to find correspondences. The above methods are good only in special cases. More specifically, their accuracy mainly depends on captured video quality. Thus, we can use [6] only for infrared video with large contrast between foreground and background information. Furthermore, although skeletons and edge information are handy for general estimation, they do not give precise corresponding features to match because they roughly represent objects as simple polygons.

The idea of foreground segmentation before processing [10, 21] has been proposed to increase accuracy in finding object features. However, this approach simply exploits shape contours and treats frames separately. Thus, there is nearly no connecting information between frames. As a result, noise in the segmentation step has a big effect in the accuracy of the registration system. To decide if a feature match is good or not, a temporal correspondence buffer can also be constructed [16, 17]. Several kinds of buffer filling strategies can be used, such as first-in, first-out (FIFO) [16], or RANSAC combined with persistence voting [17]. Nonetheless, these methods are just applicable for planar scenes since they assume all input videos as planar ones. There is still no lightweight method available to solve the non-planar, unrectified video registration problem.

Since all recent sparse correspondence methods [16, 17] are designed for planar videos, only one transformation matrix is applied to register entire frames. This method of registration cannot be adapted to non-planar scenes where each object has its own disparity (or lies on its own depth plane). Our framework proposed in Section 3.3 addresses this common limitation. We treat each object as separate blobs so that many transformations matrices may be used in a single frame.

The work of St-Charles *et al*. [17] is the most closely related to ours. Their work uses PAWCS segmentation [18] to extact the foreground of TIR-Vis videos. Contour extraction together with shape context matching are used to get correspondence between blobs. Besides, they also create a random sampling buffer with voting scheme to filter inliers and outliers. Transformation smoothing is used to improve resilience to noise. However, their work is designed for planar scene registration while ours is designed to deal with non-planar scenes, which is more general. We build upon the merit of their work by proposing: (1) a new segmentation noise filtering method in the early processing stage, (2) a fast blob matching strategy, (3) a keypoint matching strategy that accelerates the framework by avoiding exhaustive searches, and (4) a video rectification and disparity calculation method to register non-planar scenes.

As far as we know, our proposed framework is the first to register non-planar TIR-Vis videos with sparse correspondences. There is no public dataset and evaluation protocol suitable for this problem. Although Bilodeau *et al*. [4] created a public dataset for non-planar video registration, the input video frames are rectified. Thus, it is not general. As a result, we also create a new dataset, an extended version of Bilodeau's work [4], and provide our evaluation protocol as a standard one beside the overlap assessment metric.

[1] https://github.com/luannd/TIRVisReg

3. Framework Architecture

Our proposed framework is shown in Figure 1. We consider all input frame pairs as from a non-planar scene. Thus, each object has its own disparity. To the pair of frames, we apply PAWCS method [18] for segmentation, which perform background subtraction with a statistical model. The resulting foreground segmentation however is still noisy and unfit for the following blob matching step. To filter noise, we propose a new way to remove spurious blobs based on a coarse warping of images. Warped blobs that do not have a correspondence in the other image of the pair are removed, as explained in Section 3.1.

This new cleaned version of foreground segmentation is used for feature matching. Contours are extracted from object blobs and shape context matching is applied to get correspondences between each pair of frame. Besides, RANSAC algorithm [9] is also applied to filter outliers in order to increase transformation accuracy between object blobs. Instead of using a brute force method to get best match for each blob, a preservation matching strategy is proposed to increase the processing speed and eliminate wrong matches during early processing stages. This preservation matching strategy consists of a correspondence match list to keep track of match pairs throughout the analyzed video sequences. The match list is updated only when spatial relationships between objects are changed. The details of feature matching is discussed in Section 3.2.

Then, input video frames are rectified to reduce the disparity search space from 2D to 1D. In Section 3.3, the method to register non-planar scenes is described. The disparity for each object in every frame is calculated using the corresponding blob pairs obtained from previous stage. Based on these disparities, a transformation is applied in each object and video is unrectified to give the output as the same format as raw input.

3.1. Segmentation and noise filtering

Similarly to [17], we use background subtraction based on a statistical model using colors, binary features and an automatic feedback mechanism to segment object foreground blobs from the scene's background. We use the PAWCS method [18]. The resulting segmentation contains spurious blobs from background. Eliminating these spurious blobs makes our framework more robust. As shown in Figure 2, from raw segmentation returned by PAWCS, we computed a coarse transformation to estimate a homography of the whole scene. This transformation is then used to overlap the frame pair. We remove the blobs in the frame pair that do not overlap after the transformation.

Algorithm 1 describes our strategy in details. In the algorithm, $B^{"(F_i)}$ represents all blobs in the other frame of the i^{th} frame pair, n and $m^{(F_i)}$ are the number of frames and number of blobs in frame F_i respectively. There are situations where blobs do not have correspondences in the other frame of a frame pair due to the position of each camera (homography does not explain perfectly the non-planar scene). We handle this case by applying a voting scheme instead of computing a scene-wide homography. A coarse transformation matrix $M(B_k^{(F_i)})$ is computed for each blob, and each matrix votes for overall scene transformation. $M(B_k^{(F_i)})$ is computed by extracting the contour and general shape of each blob $B_k^{(F_i)}$ in frame F_i. From these shapes, we compute the best match for each blob based on point matching strategy described in Section 3.2. Because this is a coarse registration to eliminate noise in early stage, we just compute a homography transformation instead of calculating the disparity of each blob to reduce computation costs. Based on the obtained correspondence list, if one blob does not have correspondence in the other modality, it does not take

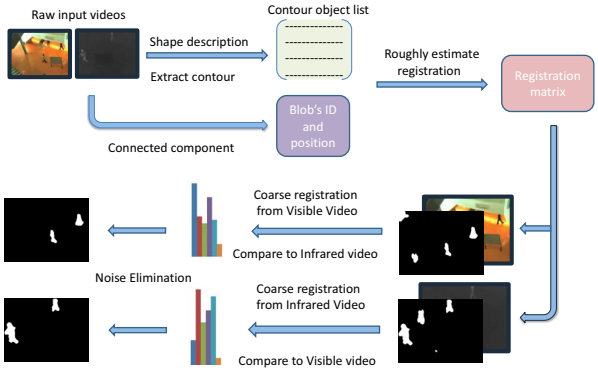

Figure 2. Segmentation and Spurious blobs filtering strategy

Algorithm 1 Noise elimination by rough registration

Input: TIR-Vis frame pairs
Output: Cleaned videos version without spurious blobs

procedure NOISEELIMINATION
  for $F_i \in F_1, F_2, ..., F_n$ **do**
    $M_{F_i} = \frac{1}{m^{(F_i)}} \sum_{k=1}^{m^{(F_i)}} M(B_k^{(F_i)})$;
    Let $B^{(F_i)_{new}} = \emptyset$
    for $B_k^{(F_i)} \in B_1^{(F_i)}, B_2^{(F_i)}, ..., B_m^{(F_i)}$ **do**
      $B'^{(F_i)}_k$ = Apply matrix M_{F_i} for blob $B_k^{(F_i)}$
      $B'^{(F_i)}_k$ = Expand $B'^{(F_i)}_k$ by α percentage
      if $B'^{(F_i)}_k \cap B^{"(F_i)} = \emptyset$ **then**
        Eliminate $B_k^{(F_i)}$
      else
        $B_{new}^{(F_i)} = B_{new}^{(F_i)} \cup B_k^{(F_i)}$
      end if
    end for
    Save new cleaned frame $B_{new}^{(F_i)}$
  end for
end procedure

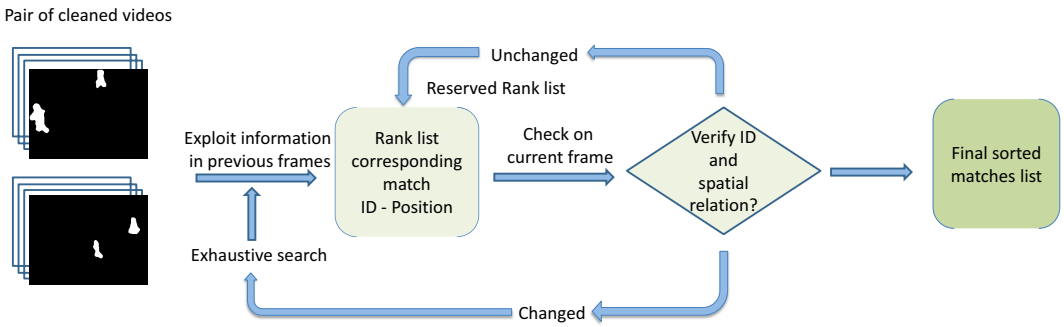

Figure 3. Blob correspondence matching strategy

part in the voting scheme. Then, the final coarse transformation M_{F_i} for the current frame pair is the mean transformation of all voting blobs.

Finally, we use this scene transformation to verify the overlap between blobs in the frame pair. Blobs are expanded by $\alpha = 120\%$ from their original size to decide whether they have overlap with any blobs in the other frame. Blobs with a corresponding overlapping blob are kept, the other removed. We filter blobs in the visible video with infrared video as a reference and vice versa.

3.2. Feature matching

In TIR-Vis registration, keeping track of blobs to find correspondence is one of many challenges. Indeed, corresponding features should be found only on corresponding blobs. St-Charles *et al*. [17] use a brute force method to find feature correspondences in each frame pair. In their method, a feature is a contour point extracted and described using the shape context descriptor [2]. χ^2 tests are used to calculate similarity scores and find matches. For each iteration, to verify the optimal transformation between blob features, the Thin Plate Spine (TPS) model [7] is applied. We inherit the merit of this strategy to find correspondences. The key difference is that we do not exhaustively consider all possible feature matches and we do not treat frames separately. As such, we propose a new method for faster computing of correspondences. Our main idea is that we preserve the correspondences from the previous frame pair and apply them to the new one. This gives rise to two situations: the easier case, where the same number of blobs appears in consecutive frame pairs, and the harder one when this is not the case.

To deal with the first case, we exploit useful information from the previous frame pair. More specifically, each blob has a unique ID and a center position. A buffer for temporarily saving correspondences in each frame pair is constructed. The consecutive frame pairs are captured after a very short time interval. Based on that observation, it is clear that spatial relationships between objects are mostly preserved. We exploit this characteristic by accumulating

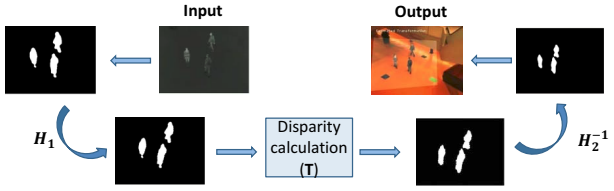

Figure 4. Non-planar scene registration

blobs with ID and position into the buffer based on a ordering on their position from left to right and top to bottom. To get the correspondence for one blob, we just need to look up blob ID in the sorted list. Besides, we also use a buffer of previous frames to guarantee that a blob still exists in current frame by comparing position with blob ID. When two blob pairs in consecutive frame are associated, correspondences are just search on the new blob pairs , instead of everywhere in the image. Figure 3 details our strategy to find correspondences.

Although consecutive frame pairs have a large proportion of unchanged number of blobs, there are situations in which objects are visible in one video but they are invisible in the other. To handle these situations, we use a brute force method to get blob correspondences based on shape context [2]. However, our brute force method is not similar to [17]. We treat blobs by their positions so that for each processing blob, we only keep blobs in the other frame of frame pair whose position is similar to the current one. Thus, search space is reduced to position based area. Blob order is also updated and used as reference for later frames.

To sum up, by using this new strategy, we only need to update the frame buffer if there are missing or new objects in either frame of the video pair. Therefore, the correspondence search speed is significantly increased.

3.3. Non-planar registration

Our framework for non-planar registration comprises three steps. Schematic diagram of our framework is presented in Figure 4. General formulation of our framework

is described by

$$D = H_1 * T * H_2^{-1} \quad (1)$$

where D is the registration matrix to register objects in non-planar scene, T is the disparity transformation for each blob in current frame, H_1 and H_2 are the rectification matrix to transform from raw video into rectified one of input and output video respectively.

3.3.1 Frame rectification strategy

First, we address two challenges with video rectification. As it can be seen from (1), to get the correct transformation for each object in video pairs, we need to estimate H_1 and H_2 correctly. However, the main difficulty is calculating the fundamental matrix. If the fundamental matrix is far from the real one, the result, of course, is affected. Thus, H_1 and H_2 matrices are not correct and it leads to wrong disparity calculation. Some existing techniques such as [8, 11, 14] are useful for the image rectification problem, not for the video one. Given this observation, we propose a new technique for robust rectification using spatial and temporal frame information.

The first part of this technique is treating each frame as a single image. The fundamental matrix is calculated using the correspondence buffer. To clarify, since our segmentation frames are free of spurious blobs, features for each blob in Vis and TIR frame are accumulated to get corresponding feature lists. Then we calculate fundamental matrix from these feature lists of Vis and TIR frame. Because a noisy fundamental matrix FM_{cur} will be obtained by only using a single frame, we create a global fundamental matrix FM_g as an optimal value by using temporal information. Equation (2) describes the relationship between current fundamental matrix and the global one.

$$FM_g = \beta * FM_g + (1 - \beta) * FM_{cur} \quad (2)$$

where β is an adaptation factor. We used a fixed value for the whole dataset used in our experiments.

The second part of our technique is adaptive decision for updating FM_g. Since not all fundamental matrices are good enough to take part in the updating scheme, we apply a coarse registration to validate the quality of the new matrix. Specifically, from FM_{cur}, we calculate H_1 and H_2 value. In disparity calculation step, which is described in Section 3.3.2, we approximate disparity values by using average blob disparities. The reason for this approximation is to reduce running time and estimate the fundamental matrix without redundant calculations.

After disparities for whole scene are estimated, we roughly use them for coarse registration as in Section 3.1. Besides, the error threshold ϕ_{cur} for computing registration is also used to decide whether FM_{cur} is qualified to update or not. If current registration error E_{cur} is lower than mean of registration errors of $\tau = 30$ recent frames (ϕ_{cur}), it is kept and used for update; otherwise, we eliminate FM_{cur}. Algorithm 2 describes our technique for rectifying videos.

Algorithm 2 Rectification Video strategy

Input: Pair of segmented videos
Output: Value $H_{1_{best}}$ and $H_{2_{best}}$

procedure RECTIFY
  $E_{min} = \infty$,$FM_{best} = \emptyset$,$H_{1_{best}} = \emptyset$,$H_{2_{best}} = \emptyset$
  for $F_{cur} \in F_1, F_2, ..., F_n$ **do**
    $H_1, H_2 \leftarrow FM_{cur}$
    Estimate coarse registration T_{cur}
    $E_{cur} = register_{\tau}(H_1, T_{cur}, H_2)$
    if $E_{cur} < E_{min}$ **then**
      $FM_g = \beta * FM_g + (1 - \beta) * FM_{cur}$
      $E_{cur} = E_{min}$, $FM_{best} = FM_g$
      $H_{1_{best}}, H_{2_{best}} \leftarrow FM_{best}$
    end if
  end for
end procedure

3.3.2 Disparity calculation

Finding disparity is one of the most important parts of our framework. At this stage, the two videos are rectified so that we only need to find disparity in one dimension for each object in each frame.

As mentioned in Section 3.1, each object is represented by its foreground blob following the segmentation step. Thus, calculating disparity is equivalent to calculating the translation between two blobs. There are two steps to do this. First, to reduce unnecessary computation, we roughly estimate translation of two corresponding blobs by subtracting their centroids. After that, the disparity search range is set to 150% of the blob size to find a correct match. For instance, let us suppose that we have a blob whose position is α and rough disparity is η, the real range for finding disparity is $[\alpha + \eta - \theta * \gamma, \alpha + \eta + \theta * \gamma]$, where γ is blob's width and θ is equal to 0.5. This approach allows the search for an optimal match to be completed more quickly.

However, there is still one problem we need to address, which is the registration evaluation criteria. Thus, we propose a new formula to estimate registration quality. The work of Bilodeau *et al*. [4] already proposed a criterion for planar scene registration, which we adapt to individual blob registration instead of whole scene. Specifically, let $B_{i,k}^{(1)}$ and $B_{i,k}^{(2)}$ be the i^{th} blob taken from the k^{th} frame of the first video and second video, respectively; registration error

Criteria	Framework	LITIV1	LITIV2	LITIV3	LITIV4	LITIV5	LITIV6	LITIV7	LITIV8	LITIV9
Min	Sonn *et al*. [16]	21.67	21.38	25.81	15.17	**16.78**	28.90	37.94	100	11.73
	St-Charles *et al*. [17]	18.74	**10.63**	10.81	**11.80**	17.24	6.94	**9.13**	13.77	9.51
	Proposed method	**17.43**	20.50	**9.42**	12.98	18.25	**6.62**	15.15	**10.59**	**9.18**
Mean	Sonn *et al*. [16]	39.92	53.78	42.25	39.94	33.87	78.45	66.79	100	24.09
	St-Charles *et al*. [17]	32.77	27.57	31.08	31.81	41.45	34.28	32.73	29.67	20.94
	Proposed method	**17.54**	**21.25**	**12.71**	**13.94**	**18.43**	**6.87**	**18.50**	**11.59**	**9.90**

Table 1. Mean and minimum registration error (%) comparisons with St-Charles *et al*. [17] and Sonn *et al*. [16] in planar scenes

for i^{th} blob in k^{th} frame is calculated as follows:

$$E_{i,k} = 1 - \frac{B_{i,k}^{(1)} \cap B_{i,k}^{(2)}}{B_{i,k}^{(1)} \cup B_{i,k}^{(2)}} \tag{3}$$

Based on this error evaluations strategy, disparity for registration of blob $B_{i,k}$ is chosen to have the lowest error value. Furthermore, we also propose overall video registration error as follows:

$$E_{Vid} = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{m_k} \sum_{i=1}^{m_k} argmin(E_{i,k}) \tag{4}$$

where m_k and n are the number of objects in frame k and number of frames in video respectively.

In our framework, objects are treated separately so that each object has its own disparity. Thus, we apply disparity translation and multiply with H_2^{-1} as in (1) to obtain the final registered scene.

4. Experiments

Although the final purpose of our framework is TIR-Vis non-planar video scene registration, we conducted experiments on both planar and non-planar scenes to show the superiority and generalization of our framework in both cases.

4.1. Experimental method

Dataset for planar scenes. We use LITIV dataset [19] for fair comparison with other state-of-the-art methods on planar scenes. This dataset contains 9 videos covering several kinds of planar scene situations. We used the polygon overlap evaluation protocol of the dataset to evaluate registration errors.

Dataset for non-planar scenes. As far as we know, there is no public dataset general enough to cover all kinds of situations for TIR-Vis video registration. Work of Torabi *et al*. [19] proposed a dataset for planar scenes, while the work of Bilodeau *et al*. [4] provided non-planar scenes for registration. However, the input of [4]'s dataset is different from ours. It provides rectified input videos, which is less general. Therefore, we provide a new dataset for non-planar scene whose inputs are raw (unrectified) videos and public evaluation protocol for easy comparison.

By contacting the authors of [4], we obtained the unrectified videos. With the raw videos, we created our own ground truth by using manually corrected PAWCS segmentation [18]. Furthermore, our new evaluation is based on overlap between one frame of the pair with the other after transformation. The equation used to evaluate the overall video registration error is the same as (4).

4.2. Results for planar scenes

Figure 5 and Table 1 show the comparison with Sonn *et al*. [16] and St-Charles *et al*. [17]. Results reveal that the proposed method has often a higher accuracy than two other state-of-the-art methods. With the idea of dealing with planar scenes as non-planar scenes, we get very small registration errors. As it can be seen from these results, the proposed method is superior to the others by reducing errors at the beginning of the videos. Our mean registration error is thus always lower. This error reduction is the result of accurately estimating the fundamental matrix. Since LITIV dataset scenes are not perfectly planar, the groundtruth provided with this dataset, which is a scene-wide homography, also results in higher errors in comparison with the proposed method in Videos 4, 6, 8, and 9. Besides, in Video 8 where there is a big blackboard in the camera view, our result has low registration error (11.59 %) while the others, St-Charles *et al*. (29.67 %) and Sonn *et al*. (100 %), still have high error because we filter out this blackboard as background and as a result it does not take part in registration step.

Furthermore, our results show that our method is robust to noise. It can be seen from Figure 5 that in Video 4, our framework has a small rise in the error in the middle of video because it has several frames which people are occluding. As a result, the proposed framework treats these occluded people as a single person and gives out one disparity for two people. This situation is very challenging for a sparse registration method. However, our registration error is still not only lower than Sonn *et al*. and St-Charles *et al*. work, but also lower than the homographic groundtruth for some videos.

4.3. Results for non-planar scenes

We applied our method and the methods of Sonn *et al*. and St-Charles *et al*. on our new dataset. We used their suggested default parameters and the authors' C++ imple-

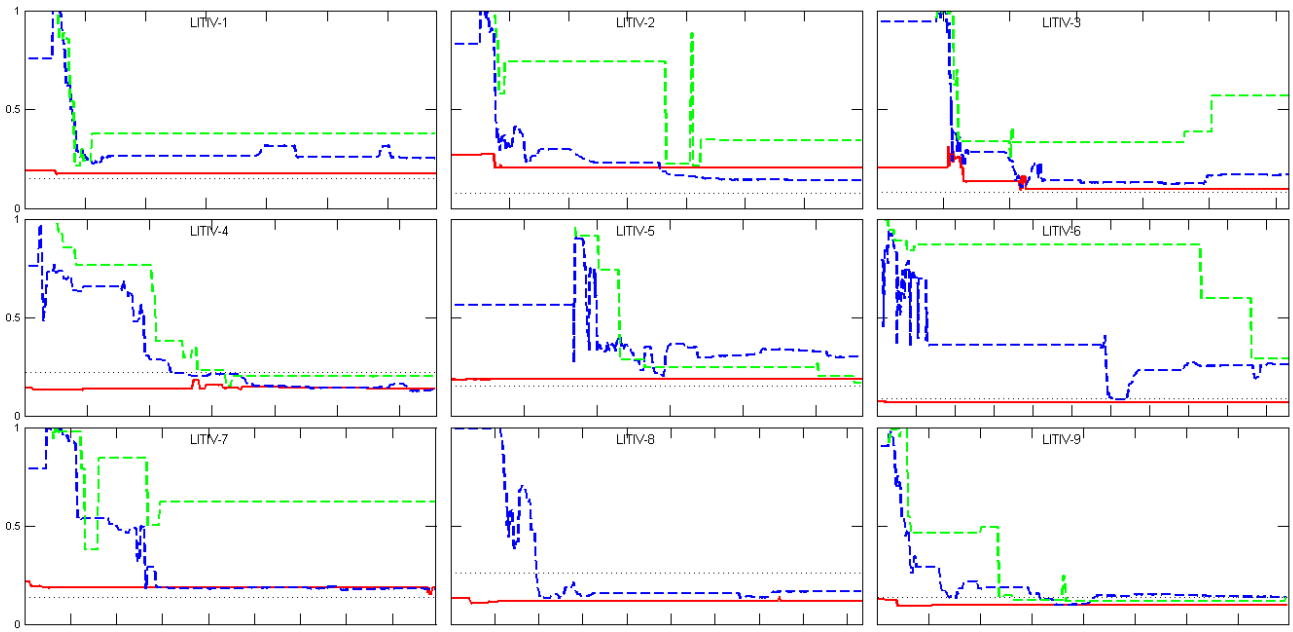

Figure 5. **Comparison with state-of-the-art techniques in planar scenes.** Red curve: proposed method, blue dash curve: St-Charles *et al*. work [17], green dash curve: Sonn *et al*. work [16]. Proposed method outperforms them for nearly all videos with low error rate at beginning of videos.

Framework	Video 1	Video 2	Video 3	Video 4
Sonn *et al*. [16]	82.72	39.58	38.68	51.19
St-Charles *et al*. [17]	78.19	34.27	32.61	41.27
Proposed method	62.77	21.20	20.18	23.93

Table 2. Mean registration error (%) using blob overlap metric with groundtruth segmentation

Framework	Video 1	Video 2	Video 3	Video 4
Sonn *et al*. [16]	91.97	48.02	47.41	59.00
St-Charles *et al*. [17]	87.51	42.10	39.30	50.16
Proposed method	65.90	26.85	22.42	27.54

Table 3. Mean registration error (%) using blob overlap metric with noisy segmentation

mentations that are publicly available. Results are shown in Table 2. We can observe that our proposed method significantly reduces the mean registration error by up to 17.34% (from 41.27% to 23.93%) in Video 4, and by at least 12.43% (from 32.61% to 20.18%) in Video 3.

One question might arise when studying these results, i.e. why is the registration error still important? The answer is simple: since input videos are of non-planar scenes, people in videos have different sizes according to their depth. Thus, sometimes people near the camera occlude in part of the ones behind. As a result, two different depths are merged in a single blob and registration is thus not accurate for those cases. Furthermore, people move around frequently throughout the sequences, meaning that this situation is often happening. It leads to high registration errors. Figure 6 shows some frames which present difficult situations.

To demonstrate the robustness of our proposed framework, we set PAWCS to generate more noise in foreground segmentation results. Fundamental matrices are thus calculated based on this noisy segmentation. We apply H_1 and H_2 from the new matrices to rectify two videos. From these two rectified videos, we apply our disparity calculating step to complete registration. We compare our results with the same two algorithms.

Results are shown in Table 3. In comparison with Table 2, Table 3 has higher registration error than the other. However, for the proposed framework, the error increase is small. For instance, our framework increase 3.13% error (from 62.77% to 65.90%) in Video 1 while Sonn *et al*. and St-Charles *et al*. works increased up to 9.25% (from 82.72% to 91.97%) and 9.32% (from 78.19% to 87.51%) respectively. Since current state-of-the-art techniques only estimate homographies, they are not flexible to various types of videos. The proposed method, on the other hand, relies on many components so that it is well adapted to diverse input videos.

4.4. Disparity evaluation for non-planar scenes

The main aim of this experiment is to evaluate our rectification step. The other purpose is to verify the disparity calculation step. Groundtruth for comparison is extracted

Figure 6. Some "easy" and difficult registration situations. Left: easy scene since blobs are separated. Middle and Right: people are occluding each other, leading to wrong disparity calculations.

Input video pair	R	D	
		Mean	**Standard deviation**
Video 1	7.68	7.34	3.19
Video 2	13.89	3.30	1.03
Video 3	6.23	2.81	0.92
Video 4	5.09	3.53	1.77

Table 4. Error in rectification R (%) and registration D (pixel) in comparison with rectified groundtruth of Bilodeau *et al*. [4]

from Bilodeau *et al*. [4] dataset. In order to verify the accuracy of our rectification and registration, we deal with raw input videos instead of rectified ones. First, from the diagram described in Figure 4, we segment and rectify videos using our rectification strategy. Then, disparity calculation is applied to get disparities for each object in every frame. We conducted experiments with error evaluation as follows.

First, we calculate the number of groundtruth points of the rectified videos that fall inside our registered blobs after segmentation to decide whether the rectification step is good or not. If rectification is good, many groundtruth points should fall within the blob regions. This is expressed in percentage of the number of groundtruth points. More specifically, given $A_{i,j}$ and $A'_{i,j}$ two blobs returned by the proposed framework and groundtruth points in j^{th} object of i^{th} frame respectively, this is calculated as

$$R = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(1 - \frac{\theta(A_{i,j})}{\theta(A'_{i,j})}) \tag{5}$$

where $\theta(A)$ returns the pixel count of blob A that exists in groundtruth dataset, n and m_i are number of frames and number of objects in the i^{th} frame respectively, R is thus the proportion of points that do not fall inside the registered blobs.

Second, mean of disparity is calculated as in (6).

$$D_{mean} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_n}(|D^{'}_{i,j} - D_{i,j}|) \tag{6}$$

where $D_{i,j}$ and $D^{'}_{i,j}$ are respectively the disparity of proposed framework and groundtruth.

Since the annotations for groundtruth points and disparities are sparse and do not cover all blobs, we extend the annotations by adding several points. Table 4 shows R, mean disparity error and their standard deviations with new annotations in comparison to groundtruth. Because Video 2 only has two people moving around, there is not enough blob pairs reference to have a good fundamental matrix for rectification. Thus, R is still high (13.89%). However, since the others have up to four or five people in non-planar scenes, the proposed framework quickly acquires good correspondences to calculate the fundamental matrix. As a result, R values in these videos are significant lower in comparison with Video 2. In general, not only our R but also our disparity error is low. In Video 2 and 3, since these videos have small number of people, people do not often occlude each other so our proposed framework obtains small mean disparity errors (3.30 and 2.81) and standard deviation errors (1.03 and 0.92). In contrast, the disparity errors are higher in Video 1 because there are many occlusions.

Furthermore, this increasing error in disparity evaluation also reveals the fact that our extension of groundtruth annotation is precise and worthwhile. By doing this experiments, we can note that for TIR-Vis video registration, the more people in a video, the more accurate the rectification is but this also introduces more errors in the registration. Thus, there is a tradeoff between rectification and registration quality.

5. Conclusion

We have presented an end-to-end framework for TIR-Vis uncalibrated video registration for non-planar scenes. This paper demonstrates that foreground segmentation and input videos rectification can complement each other to simplify multimodal video registration. The key ideas of the proposed framework are removing spurious blobs in the beginning of videos and preserving matches for frame to frame. A new dataset is also provided with an evaluation protocol. This framework and new dataset are significant contributions in non-planar scenes TIR-Vis video registration field.

As far as we know, our proposed framework is the first work to deal with problem of TIR-Vis video registration for non-planar scene by sparse correspondences. Thus, there is room for improvements and for exploiting characteristics of both infrared and visible videos. Although the proposed framework gets lower errors in planar scenes, there is still one existing obstacle when dealing with non-planar scenes, which is occluding people at different depths. This problem could be overcome by using blob position and information from previous frames to split blobs.

6. Acknowledgements

This work was conducted while Dinh-Luan Nguyen was doing a MITACS Globalink internship at Polytechnique Montreal. We thank Minh-Triet Tran for his helpful dicussions.

References

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

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002. 4

[3] G.-A. Bilodeau, P.-L. St-Onge, and R. Garnier. Silhouette-based features for visible-infrared registration. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 68–73, 2011. 2

[4] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi. Thermal–visible registration of human silhouettes: A similarity measure performance evaluation. *Infrared Physics & Technology*, 64:79–86, 2014. 1, 2, 5, 6, 8

[5] T. Chakravorty, G.-A. Bilodeau, and E. Granger. Automatic image registration in infrared-visible videos using polygon vertices. *arXiv preprint:1403.4232*, 2014. 1

[6] E. Coiras, J. Santamarı, C. Miravet, et al. Segment-based registration technique for visual-infrared images. *Optical Engineering*, 39(1):282–289, 2000. 2

[7] J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977. 4

[8] M. Elbakary and M. K. Sundareshan. Multi-modal image registration using local frequency representation and computer-aided design (cad) models. *Image and Vision Computing*, 25(5):663–670, 2007. 2, 5

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[10] J. Han and B. Bhanu. Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40(6):1771–1784, 2007. 2

[11] T. Hrkać, Z. Kalafatić, and J. Krapac. Infrared-visual image registration based on corners and hausdorff distance. In *Image Analysis*, pages 383–392. Springer, 2007. 5

[12] S. J. Krotosky and M. M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106(2):270–287, 2007. 1

[13] T. Mouats and N. Aouf. Multimodal stereo correspondence based on phase congruency and edge histogram descriptor. In *International Conference on Information Fusion (FUSION)*, pages 1981–1987. IEEE, 2013. 1

[14] F. P. Oliveira and J. M. R. Tavares. Medical image registration: a review. *Computer methods in biomechanics and biomedical engineering*, 17(2):73–93, 2014. 5

[15] J. P. Pluim, J. A. Maintz, and M. A. Viergever. Image registration by maximization of combined mutual information and gradient information. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 452–461. Springer, 2000. 1

[16] S. Sonn, G.-A. Bilodeau, and P. Galinier. Fast and accurate registration of visible and infrared videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 308–313, 2013. 1, 2, 6, 7

[17] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin. Online multimodal video registration based on shape matching. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015. 1, 2, 3, 4, 6, 7

[18] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin. A self-adjusting approach to change detection based on background word consensus. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 990–997, 2015. 2, 3, 6

[19] 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. 6

[20] J. Zhao and S.-c. S. Cheung. Human segmentation by fusing visible-light and thermal imaginary. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1185–1192, 2009. 2

[21] J. Zhao and S. C. Sen-ching. Human segmentation by geometrically fusing visible-light and thermal imageries. *Multimedia Tools and Applications*, 73(1):61–89, 2014. 2