Camera Motion Compensation from T-junctions in Distance Map Skeleton

Charles Beumier [1], Xavier Neyt [1]
[1] Signal & Image Centre, Royal Military Academy, 30 Avenue de la Renaissance 1000 Brussels, Belgium
beumier@elec.rma.ac.be

Abstract – **In the field of aerial surveillance, tracking targets in images is complicated by the possible motion of the camera, especially if frame differencing is used to detect moving objects. We propose in this paper to exploit the high similarity in sequences acquired from a nearly static camera. In this case distance maps grown from image edge points share many similarities and T-junctions of distance map skeletons appear to offer precisely located reference points. Each T-junction is attributed seven features: the value in the distance map, three orientations of the junction branches and R, G, B image intensities. Registering images is carried out on a division of the images into tiles, looking for the dominant translation per tile of matching T-junction points. The obtained displacement field allow for the compensation of small camera motion. This was tested on image sequences captured by a smartphone held in hand while targeting a given static scene with a few moving vehicles and pedestrians.**

Keywords – **Camera motion compensation; Image registration; Distance map; Skeleton; T-junction**

I. Introduction

In the context of security, aerial surveillance brings a clear advantage when targets have to be detected and tracked. The recent development of UAV (Unmanned Aerial Vehicle) opens new possibilities thanks to its reduced cost, rapid deployment and agility for dynamic situations. Before considering the difficult case of the UAV general motion, we are interested in videos captured by nearly static cameras attached to a semi rigid pole or partly stabilized for permanent surveillance.

In the case of a static camera, the most direct way for target detection consists in highlighting moving objects thanks to image temporal difference. In its simplest form, the difference concerns image pairs of the video sequence. In more advanced techniques, one image is a frame of the sequence and the second one is a combination of previous frames, trying to reconstruct an image of the background with high fidelity. Many background reconstruction methods have been reported in the literature as attested by [1, 2]. Our concern lies here in the image stability of the sequence to ensure the proper functioning of temporal differencing since, in the presence of camera motion, static objects of the background become false moving targets due to their apparent motion.

Because the amplitude of the motion is limited, our approach is based on image registration, as opposed to techniques relying on additional hardware to sense camera motion. Image registration methods may be classified as area-based or feature-based [3]. Many recent developments favoured the detection of interest feature points (e.g. SIFT keypoints in [4]), what is computationally more attractive. The points are attributed descriptors to ease pairing between images. To be successful, interest points should be precisely located, stable after image transformations (such as rotation, scale, image intensity and point of view changes) and provide descriptors robust to these transformations.

In the image sequences considered in this paper, most of the aforementioned image transformations are not present. The scene contains few moving objects of limited size compared to many static (background) areas, and undergoes limited translation, rotation or scale changes. In the event of a spurious larger motion, the camera is back approximately to its original position, thanks to the objective of focusing on a given scene. As a consequence our sequential images share many similar areas from which to extract simple yet useful anchor points. We observed that edge points are stable so that they can be used to derive a distance map [5] from which region features can be captured.

A skeleton provides a compact representation of an image object and is the basis for many image processing applications [6] among which image registration. Processing a distance map is a traditional way to obtain skeletons [6, 7, 8]. In our development, the full skeleton is not necessary: only the branching points called here T-junctions are exploited.

Section II outlines the methodology followed by our approach while the implementation details are described in section III. Section IV presents the results for image registration and section V discusses the sensitivity of parameters, the computational time and the limitations of the approach. Section VI concludes the paper and paves the way for improvement.

II. Methodology

To compensate for the image transformation due to camera motion in a video sequence, we need to register images to some reference frame and thus find the local displacement over each image.

From the observation that in our application most of each image consists of static parts since there are relatively few moving targets (Fig. 1), we designed a method for registering images based on simple yet efficient region features. Regions are identified in a distance map grown from edge points. Like this, the influence of gaps in region contours is not an issue.

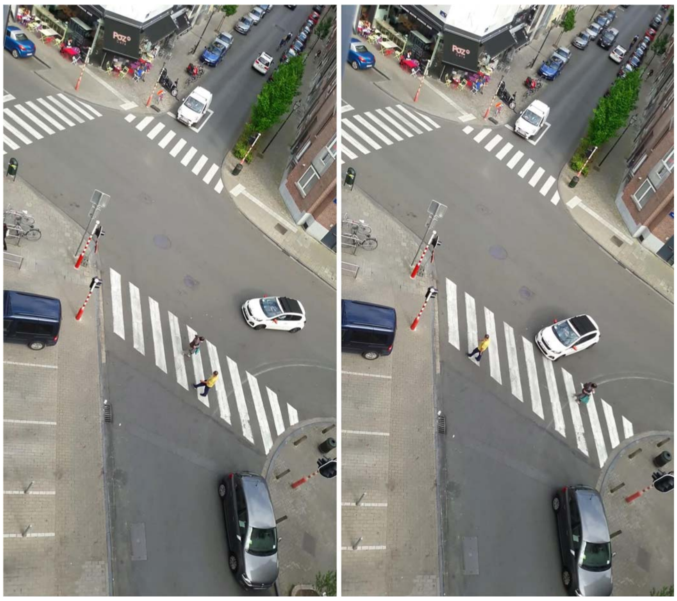

Figure 1. a) Image from sequence at t0; b) Image at t0 + 3s.

Due to the sensitivity of distance maps with spurious edge points, only the top gradient pixels are retained. These high gradient pixels are the roots (0-distance) of a distance transform in which local maxima correspond to region centres. The skeleton of the distance map, an effective and compact representation for object shape characterization [6], contains our interest points. These are the junction points of skeleton branches. They are precisely located, quite stable in similar images and they possess simple yet discriminative features for matching such as their distance value and branches orientation. Image intensity levels at T-junctions are also appropriate features since they are situated at the centre of regions, far from edges, where image intensities are likely to be uniform and stable.

The proposed image registration scheme consists in associating T-junction feature points between images. The attributes attached to each T-junction are compared to highlight possible matching pairs. These matches help finding the most likely local translation on a fix grid, arbitrary division of the image into tiles. The obtained displacement vector field defines the desired transformation used to warp the image for camera motion compensation.

III. Implementation

In this work we register images from an image sequence by matching feature points detected as skeleton T-junctions. These are derived from a distance map grown from strong gradient points.

This section describes the image processing steps followed to detect skeleton T-junctions and the procedure implemented to match them for image registration. Some words are given on how to realise the geometrical transformation (warping) needed to superimpose the image on the reference.

A. Distance transform

An image can be interpreted through its edges (pixels with high local variation of the image intensity) and/or regions (areas with low local variation). Disposing of highly correlated images from video sequences, we aim at extracting feature points for matching thanks to the medial axis of regions identified in a distance map.

Because a distance map is sensitive to spurious points, we first apply a low-pass filter with 3x3 uniform weight. We then retain edge points that are in the top 5 % of the gradient magnitude. This magnitude is computed as the norm of the vector consisting of the horizontal and vertical differences of intensity values at offsets +1 and –1 pixel.

Our distance transform computes the distance of each image pixel to the closest retained edge point. A fast implementation considers integer distance approximations and distance propagation by two successive full image scans in opposite directions [9]. The 3x3 mask used for distance propagation is the one recommended by Borgefors [9], approximating pixel distances by 4 for diagonal neighbours and by 3 for non-diagonal ones. In this operation, each pixel is finally visited 10 times for simple comparison to keep the minimal distance value.

The distance maps for the two images of Fig. 1 (which are separated by 3 seconds in the sequence) are shown in Fig. 2. The same false colour look-up table was adopted to highlight local distance differences while showing similarities between the two maps. The lowest distance values, close to 0, are displayed in black and contain the retained high gradient points (distance 0).

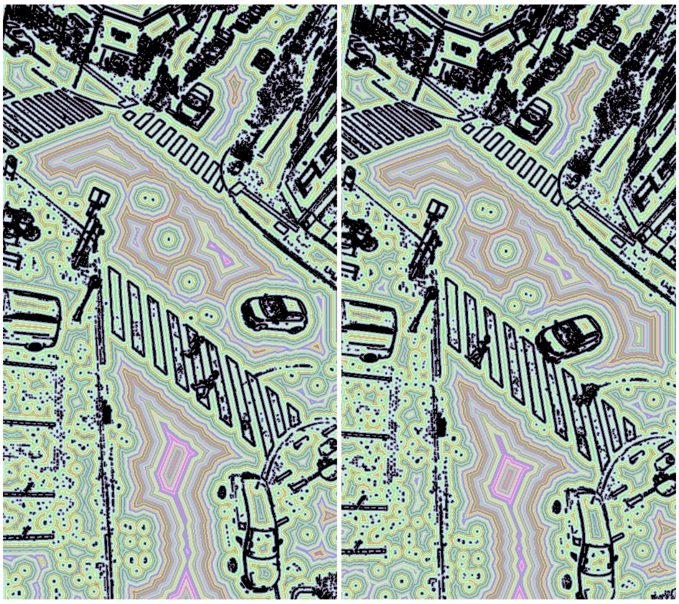

Figure 2. a) Distance map from edges of Fig. 1 (a); b) Distance map from edges of Fig. 1 (b).

B. Skeleton

A distance map contains information about regions without the need to manipulate them or maintain values into region specific tables. In particular, distance local maxima and the so-called skeleton are valuable for region size and shape characterization although they are subject to a high sensitivity to the region border details. Our focus on highly correlated images allowed for an approach based on the skeleton.

From a distance map, skeletons have been extracted in many different ways [6]. In iterative approaches, a removal operator is repeatedly applied to literally peel the objects while preserving their topology. In sequential approaches, particular points are first extracted and later connected to form the skeleton. In [7] centres of maximal disks (inscribed in objects) and saddle pixels are found thanks to distance neighbourhood comparisons. They are then connected into the skeleton. Chang [8] uses the sign pattern of local distance difference in the horizontal and vertical directions to detect ridge points that are later linked by a two-pass procedure.

We extracted skeletons for the purpose of T-junction analysis thanks to a rank filter. For each disk neighbourhood of diameter 9, the central pixel receives its rank order, after comparison with its 48 neighbours. This quite demanding processing (48 accesses and comparisons per pixel) has the advantage to easily allow for a top-N selection since many skeleton points are not first rank due to their skeleton neighbours. Rather continuous skeletons were obtained when keeping the 18 highest ranks in the rank image, as in Fig. 3. When comparing both skeletons in the figure, one can observe the high similarity between static parts but also the large sensitivity in modified areas due to moving objects.

The skeleton was useful for analysis during development and tests. However T-junctions could be extracted directly from the distance map as explained in the next subsection.

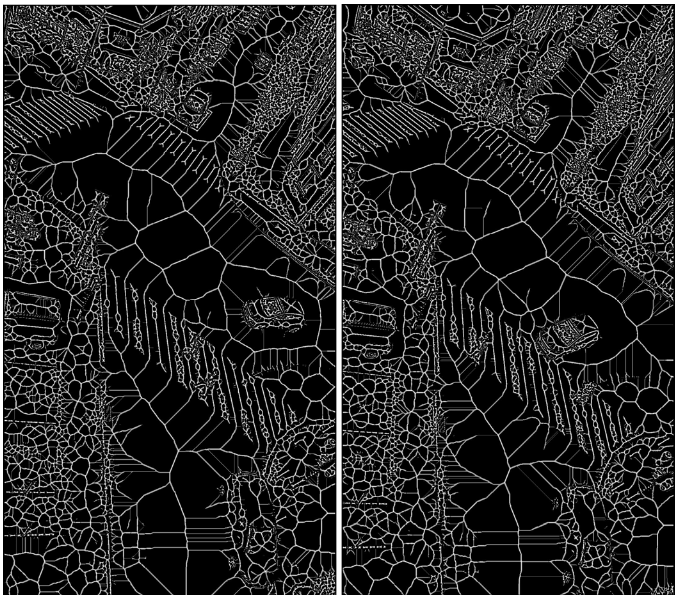

Figure 3. a) Skeleton from distance map of Fig. 2 (a); b) Skeleton from distance map of Fig. 2 (b).

C. T-junctions

T-junctions are used in this paper to name points of the skeleton having at least 3 branches. These are precisely located and stable in highly similar image regions. We detect them directly from the distance map, by keeping local 2D distance maxima in a 5x5 window that count at least 3 local distance maxima along the border of the centred 9x9 window.

Each detected T-junction is attributed a set of seven feature values. The first one is its value in the distance map. The three next features are the orientations of the skeleton branches of the T-junction (see Fig. 4). These come from the local distance maxima on the 9x9 square border. Each orientation is coded as an integer in the [0..31] range, leading to an angle granularity of about 11 degrees. The last three features are the image R, G, B intensities at the T-junction position. Additional orientations (X-junction) or spectral features are of course possible.

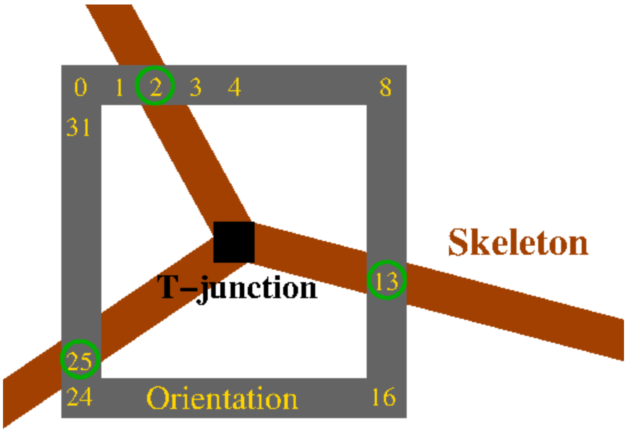

Figure 4. A T-junction and its branch orientation codes.

D. Displacement Vector Field

To limit the amount of feature point comparisons, the image is first divided into tiles of NxN pixels. Each tile keeps a list of the T-junctions located within its limits. N is typically chosen as one tenth of the largest image size to restraint the number of T-junction comparisons while allowing for a sufficient number of candidate matches.

In order to register two images, the T-junctions of one tile are compared with the T-junctions of the 3x3 neighbouring tiles of the other image. In this way, local translations of up to about (3*N)/2 pixels in each image direction are considered.

To find correspondences between two sets of T-junctions, a rough feature rejection is applied. Distance differences higher than 4 pixels, orientation differences above 2 units (22°) and intensity level differences larger than 10 are filtered out. For each acceptable correspondence, the x and y offsets, respectively dx and dy, are counted. The most important peak in (dx,dy) gives the translation candidate for the tile if its count is above a threshold C.

An image tile may miss a clear dx, dy consensus if it lacks matching T-junctions or if correspondences are polluted by the incoherent motion of moving objects. The chance to get an erroneous translation with large consensus is unlikely, except if a large moving object with apparent translation covers the tile.

The sought displacement field is rather smooth with our constraint of small camera motion. The change in displacement vector from one tile to another is limited so that the coherence can be used to check vectors and guess incorrect ones. We applied a 3x3 median filter for the inner tiles of the grid. If the estimated translation vector is far from the 3x3 median value for the considered tile, the median value is preferred.

Evaluating the displacement field on a grid not only facilitates its coherence but also simplifies the subsequent warping operation to align the image on the reference. To the grid of the reference image corresponds the grid of the image to be warped. Each square tile defines two triangles so that a direct tessellation into triangles is achieved. A triangle indeed defines an easy 2D coordinate system to localise a pixel. To avoid holes in the warped image, this image is scanned and the corresponding points are localised in the corresponding triangle of the reference image. Bilinear interpolation has been used since corresponding reference points do not necessarily fall on integer coordinates.

IV. Results

We applied the proposed approach to image sequences acquired from a SAMSUNG smartphone captured at 25 images per second, at a resolution of 1920x1080 pixels. The phone was held in hand, targeting to capture a given scene of a road intersection with pedestrian crossings.

Fig. 5 (a) shows in green the resulting displacement vector field extracted for the pair of images of Fig. 1. The translation is smooth in amplitude and varying in direction, depicting a rotational movement of the camera. We mention that the image quality is affected by the MPEG compression of stored videos. The so-called blocking artefact is strong in the grey shades of the roads but had no negative influence on the distance maps. On the contrary, some border displacement vectors have wrong direction and amplitude. They suffered from a lack of T-junctions from static parts, either because a moving object is present or since the image texture is rather poor in the area. The corrective effect of the 3x3 median filter is limited on the grid border.

The ultimate way to assess the quality of the registration process is to subtract the reference image from the image warped by the displacement field. The image difference is shown in Fig. 5 (b). The so-called frame differencing is the simplest approach for moving object detection in image sequences when camera motion is absent or compensated for. If correctly estimated, the transformation from image registration maps the static objects on top of each other. Moving objects are highlighted by the difference in image intensities, as observed in Fig. 5 (b). The visible edge-like artefacts correspond to a small compensation error of about 1 or 2 pixels. This residual error is not relevant for our application since objects of interest are at least 20 pixels in each dimension.

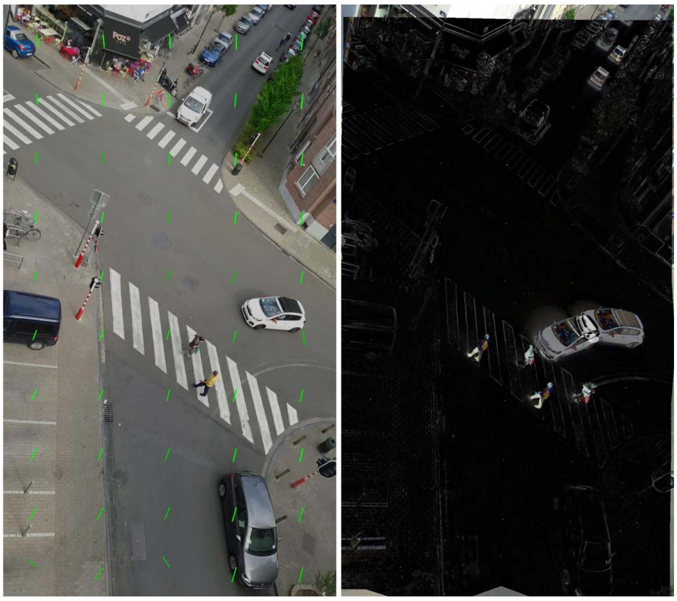

Figure 5. a) Displacement field vector for the image pair of Fig. 1; b) Image difference after registration.

V. Discussion

The parameters of the method were determined experimentally. So far no attempt was undertaken to weight the T-junction attributes (distance, orientation, intensity). They seemed to have a comparable discriminative power. No matching score was derived but, as presented in section III, matching was based on the rejection of bad T-junction associations. The thresholds for T-junction distance, orientation and intensity filtering appeared to have reasonable values. If these thresholds are too loose, the peaks in dx and dy tables become hidden by noisy values. If too strict, the peaks may be too weak to emerge. The minimum count C for a peak to be accepted was experimentally set to 5. The possibly erroneous displacement vectors for low counts are likely to be filtered by the median filter for global coherence.

The parameters concerning edge selection, T-junction detection and tile size also resulted from compromises. The edge selection criterion is not strict as long as weak edge points are discarded. The size for branch orientation (9x9) seems to be a good compromise between precision and independence from neighbouring junctions. The tile size of about 200x200 pixels appears appropriate for 1920x1080 images to have enough T-junctions for a consensus about (dx, dy) and to have a sufficient number of tiles for a correct discretization of the displacement vector.

Although we paid attention to avoid computationally prohibitive methods, the implementation is not real-time for 1920x1080 images at 25 frames a second. The time to extract T-junctions and features is less than 0.4 s with the code written in C in the Ubuntu Linux environment and running on a Intel i5-4590 at 3.3 GHz with 27 Gb of RAM, using one CPU. T-junction comparison is negligible in time as around 3000 points were extracted per image and distributed into tiles. Translated into traditional 640x480 video size, the computation time would amount to 60 ms, close to real-time.

Multi-resolution is a traditional approach for speeding up processing. In our application, a lower resolution would help finding quickly the numerous frames with tiny motion for which the compensation has not changed or is not necessary. The analysis of multiple resolutions also delivers more T-junctions since regions get reorganised through levels. This may be useful in large regions because they usually have few T-junctions. Another possible time optimisation consists in reducing the number of pixels where the gradient is evaluated. The distance map would not be much affected as edge points are mostly grouped.

Our approach for camera motion compensation was designed for nearly static camera capture and has several limitations. Based on distances to edge points, the local translation estimation will fail for scaled objects or for objects imaged from a different point of view. A limited rotation of about 20° is acceptable thanks to the tolerance of the T-junction branch orientation codes. Large orientation resilience can be achieved by sacrificing one of the three orientation codes to normalize the other two values. Finally, a sufficient number of matching T-junctions must be present all over the image, although a few isolated tiles without estimation can receive a value from neighbours. Images must thus contain a limited amount of moving (foreground) parts compared to static (background) objects.

VI. Conclusions and future work

We presented an approach for camera motion compensation in images acquired from a nearly static camera. The method exploits the image high similarity in static regions through the distance transform grown from edge points. Skeleton T-junctions are extracted directly from the distance map and are attributed features of distance, branch orientation and image intensities. The displacement field between two images is estimated from the consensus in translation of matching T-junctions in neighbouring image tiles. This displacement field helps registering the images of a sequence to highlight moving objects by frame differencing or background subtraction.

The camera motion compensation is up to 1 or 2 pixels, what hardly disturbs target detection since objects of interest usually have more than 20 pixels in both directions. As discussed in section V, the computation time on a conventional computer is close to real-time for 640x480 videos. The large proportion of very small motion in videos of nearly static cameras could be detected and handled with some trivial method to boost the speed. Other optimisations were suggested in section V.

Finally, since we are dealing with sequences, individual field vectors obtained from each image independently could be analysed and filtered temporally to increase the robustness and precision of the motion compensation.

Acknowledgment

We would like to thank the Belgian MoD and in particular the Royal Higher Institute for Defence for supporting this research.

References

[1] T. Bouwmans, “Recent Advanced Statistical Background Modeling for Foreground Detection – A Systematic Survey”, Recent Patents on Computer Science, 4 (3): 147-171, Sep 2011.

[2] M. Cristani, M. Farenzena, D. Bloisi, and V. Murino, “Background Subtraction for automated Multisensor Surveillance: A Comprehensive Review”, EURASIP Journal on Advances in Signal Processing, Vol. 2010, pp. 1-24, Feb 2010.

[3] B. Zitova and J. Flusser, “Image registration methods: a survey”, Image and Vision Computing, Vol. 21, Issue 11, pp. 977-1000, Oct 2003.

[4] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, Vol. 60, 2, pp. 91-110, 2004.

[5] R. Fabbri, L. Da F. Costa, J. Torelli, and O. Bruno, “2D Euclidean Distance Transform Algorithms: A Comparative Study”, ACM Computing Surveys, Vol. 40, No. 1, pp. 1-44, Feb 2008.

[6] P. Saha, G. Borgefors, and G. Sanniti di Baja, “A survey on skeletonization algorithms and their applications”, Pattern Recognition Letters, Vol. 76, pp. 1-12, June 2016.

[7] G. Sanniti di Baja, “Well-shaped, Stable, and Reversible Skeletons from the (3,4)-Distance Transform”, Journal of Visual Communication and Image Representation, Vol. 5, No. 1, pp. 107-115, March 1994.

[8] S. Chang, “Extracting Skeletons from Distance Maps”, Int. Journal of Computer Science and Network Security, Vol. 7, No. 7, pp. 213-219, July 2007.

[9] G. Borgefors, “Distance Transformations in Digital Images”, Computer