UAV-VIDEO REGISTRATION USING BLOCK-BASED FEATURES

Adel Hafiane, Kannappan Palaniappan[†]

Department of Computer Science
University of Missouri-Columbia
Columbia, MO 65211, USA

Guna Seetharaman

Dept. of Electrical & Computer Engineering
Air Force Institute of Technology
Dayton, OH 45433-7765, USA

ABSTRACT

We present a new approach aimed at fast multiframe registration of airborne video collected by moving platforms such as unmanned aerial vehicles. Registration is used to enable separating the moving objects from the stationary background, which is similar to estimating the egomotion of the sensor. The proposed registration algorithm is used to match the moving background and remap video frames into a common coordinate system in order to stabilize that segment of video. The major modules include dense feature detection, sparse prominent edge and corner-based feature block identification, feature block region matching to extract control points, confidence weighted robust projective transformation estimation and image warping to register a segment of frames within a given temporal window. The proposed method is shown to produce good results with small image registration error.

1. INTRODUCTION

Applications of unmanned aerial vehicles (UAVs) for surveillance, monitoring, situation awareness and resource management have steadily increased in recent years. Most UAVs have an onboard vision sub-system designed to acquire, preprocess, and transmit video images as they fly over an area of interest. The large amount of video data and effects due to camera motion make it less suitable for direct analysis by human operators. There are a number interesting challenges and opportunities to automate certain subtasks and assist the human analysts to improve the overall performance of such systems. One of the key challenges is due to the inherent camera motion - inevtiable since the UAV is a moving platform. It is often impractical to assume that geo-location and orientation of the camera will be available at a resolution and robustness required by other algorithms. Effects due to occlusion and atmospheric conditions such as illumination, cloud motion, rain, etc, compound the challenge. As a result, detecting, locating, monitoring and tracking of objects in videos becomes severely hampered, and at times impossible, shifting the mission into an offline postprocessing mode. A number of interesting papers have appeared recently designed to address specific bottlenecks in the video analysis chain. The scope of these papers vary vastly with regard to assumptions on the range of motion and complexity of objects and their relative geospatial manifests.

Video registration is an essential task designed to deal with the effects caused by camera motion [1, 2]; egomotion estimation is an alternative approach which we do not discuss in this paper. In this context, registration refers to the process of determining corresponding points or regions between two frames of a potentially dynamic scene taken at different times from different view points by a mobile camera or other sensors. Both the vantage and time vary between the two images. Although image registration has been extensively addressed in the literature over the last three decades [3, 4], a typical image pair that forms the basis of analysis has changed significantly over the years, including sensor configuration and scene dynamics that constitute the cental disparity between frames. The pervasive and persistent imaging offered by ubiquitous mobile sensors have expanded the basic model of image registration far from the canonical model used in registering a pair of satellite images in the 1970's. Modern registration is a complex instance of a broad spectrum of fusion and exploitation tasks, often involving topological and model-driven techniques suitable for domain dependent applications. Although many methods have been proposed, a technique with assured performance over a wide range of scenes still remains challenging and offers a rich ground for newer solutions.

In this paper we describe a methodology for registering video segments based on detecting prominent feature (PF) blocks and matching these blocks using SAD or NCC. PF regions are determined by dividing the image into non-overlapping blocks, extracting features in each block using 2D structure tensors sensitive to edges and corners. The center of each selected PF block represents a point of interest and block matching is used to find region correspondences between two successive images. The set of center pixel coordinates for each matching pair of PF blocks is used to compute a cumulative projective transformation matrix. The transformation matrix is then applied to warp each subse-

[†] This work was partially supported by the Air Force Research Laboratory Visiting Faculty Research Program (2007) and AFOSR/ASEE Summer Faculty Fellowship Program (2008) at Rome, NY and AFIT Award FA8604-06. The conclusions are those of the authors and do not represent the focus or policies of the Air Force Institute of Technology, USAF or US DoD.

quent image frame into the base frame for that video segmen to obtain registered images.

2. REGION-BASED FEATURE SELECTION

The first step in registration is to detect salient or prominent features (PF) that are preserved under geometric image transformations. Spatial manifests such as corners, edges, contours, and regions prove to be effective in grouping pixels within an image, and establishing a basis for registration across a pair of images. These features are generally represented by points (corner, center of gravity and line intersection), lines (Hough transform) or areas (window), and facilitate registration when the correspondence between such features (drawn from two images) is established by some means. A set of corresponding PF block regions is the most desired baseline since it can be directly used to determine the parameters of the transformation function [5]. Complex techniques do exist, which could be suitably adopted to exploit oriented properties of the spatial features, e.g. slope of line-segments, or inclusive angle associated to a corner etc. Such techniques would invariably involve fusing limited knowledge about the scene in terms of the parametric models used relating the oriented attributes of a selected feature across two instances. Depending on the context, the analysis may seek a balance between constancy and saliency of such attributes. We look at the distribution of feature point attributes within a PF block and use the centroids for displacement vectors. The tensor representation of edges and corners provide consistent characteristics since they are related to the image structure.

Many first and second derivative feature detectors and descriptors are available in the literature [5]. We use the 2D color structure tensor defined in terms of the outer product of spatial gradients in each channel with C_i representing image channels ($i = 3$ for RGB color), and further described in [6, 7]. The 2D grayscale structure tensor matrix is also referred to as the second moment or autocorrelation matrix [5]. Local descriptors based on the two eigenvalues of the structure tensor provide information about the signal in orthogonal directions. Small eigenvalues are indicative of noise so the trace of $\mathbf{J_C}$ can filter these locations.

The eigenvalues of $\mathbf{J}_C$ are correlated with the local image properties of edgeness and cornerness, defined as $\lambda_1 >> 0$, $\lambda_2 \approx 0$ and $\lambda_1 \approx \lambda_2 >> 0$ respectively. For a 2D multispectral image, the Beltrami operator defines a metric on a two-dimensional manifold $\{x, y, C_1(x, y), C_2(x, y), C_3(x, y)\}$ in the five-dimensional spatial-color space $\{x, y, C_1, C_2, C_3\}$:

$$\begin{aligned} Beltrami(\mathbf{I_{RGB}}) &= \mathbf{det}(\mathcal{I} + \mathbf{J_C}) \\ &= 1 + \mathbf{trace}(\mathbf{J_C}) + \mathbf{det}(\mathbf{J_C}) \qquad (1) \\ &= 1 + (\lambda_1 + \lambda_2) + \lambda_1\lambda_2 \end{aligned}$$

The determinant is the appropriate generalization of the gradient magnitude of intensity images to multispectral image gradients. In order to evaluate the color tensor matrix J_C two (convolution) scale factors are required – one for the spatial derivative (gradient) filters and one for integration (summation) filters. Figure 1 shows an example of the 2D color structure (or Beltrami) tensor response. Each block containing high values of tensor magnitude responses is considered to be a PF macroblock and selected for the matching process described next. Reducing the total number of PF macroblocks reduces computational cost.

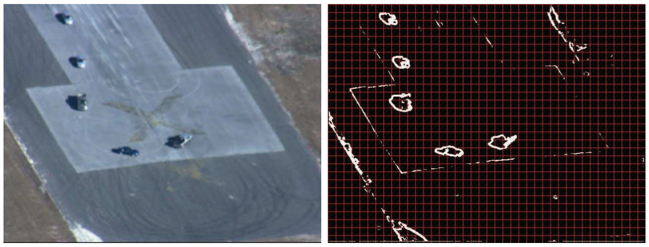
(a) Original frame (b) Beltrami tensor response

Fig. 1. Thresholded output of the 2D Beltrami tensor applied to the image shown. The gird shows correspondence between location of salient features and non-overlapping macroblocks.

3. PF BLOCK REGION-CORRESPONDENCES

Once the prominent feature (PF) macroblocks are selected based on an efficient evaluation and thresholding of Eq. 1, the next step is displacement or region correspondence matching. The matching process uses a PF block in the source image and searches for the best matching or most similar *overlapping* block contained with a search zone/window in the target image; each PF block is compared to all shifted overlapped areas by sliding the source PF block by one pixel. Two standard measures of similarity are the Euclidean distance and Normalized Cross Correlation (NCC). The L_1 approximation (referred to as sum of absolute differences (SAD)) of the L_2 Euclidean metric is used to reduce computation complexity. The NCC measure is less sensitive to absolute intensity changes between the source and target images due to the normalization terms in the denominator but is much more expensive to compute than SAD. Both were considered. The minimum of the SAD measure can be defined as,

$$\Delta X_{opt} = \arg\min_{\Delta X} \sum_{X \in \Omega} |\mathbf{I}(X + \Delta X, t - k) - \mathbf{I}(X, t)| \qquad (2)$$

The NCC between target (or reference) image $\mathbf{I}(X, t-k)$ and source (or template) image $\mathbf{I}(X, t)$ is defined as,

$$\gamma = \frac{\sum_{X\in\Omega}[\mathbf{I}(X + \Delta X, t - k) - \mu_{t-k}][\mathbf{I}(X, t) - \mu_t]}{\sqrt{\sum_{X\in\Omega}[\mathbf{I}(X + \Delta X, t - k) - \mu_{t-k}]^2 \sum_{X\in\Omega}[\mathbf{I}(X, t) - \mu_t]^2}} \qquad (3)$$

where $\mu_{t-k} = \langle \mathbf{I}(X + \Delta X, t - k) \rangle$ and $\mu_t = \langle \mathbf{I}(X, t) \rangle$ are the local intensity means (averages) in the target and template image regions respectively and the denominator is the product of the local variances. The NCC for vector images (RGB color) can be appropriately extended. We want to find the translation or displacement ΔX that maximizes the NCC measure, $\Delta X_{opt} = \arg\max_{\Delta X} |\gamma(\Delta X)|$ The NCC can also be interpreted as the cosine of the angle between the two mean corrected region blocks. If we represent the mean subtracted pixels in the target and source windows as the vectors $\overrightarrow{W_T}$, $\overrightarrow{W_S}$, respectively then,$NCC \equiv \gamma(\Delta X) = (\overrightarrow{W_T} \bullet \overrightarrow{W_S})/(\|\overrightarrow{W_T}\|\|\overrightarrow{W_S}\|)$.

4. PROJECTIVE TRANSFORMATION ESTIMATION

Once region-based block correspondences are established, we need to compute the homography relating the the two coordinate systems. This enables image $\mathbf{I}(X, t)$ to be mapped into the coordinate system of the base frame for a given video segment $\mathbf{I}(X, t-k)$. Note that we are interested in finding a good solution for the homography, and *not* on finding the unique solution for the true 3D camera motion, as our goal is mainly to compensate for and remove the effects of the background or (dominant) ground plane motion. Since UAV imagery can have significant perspective effects an projective mapping is more accurate than a single global affine transformation. Other approaches include multiple local affine projections [8] and non-rigid transformations [9]. The projective mapping function or homography uses the coordinates of the corresponding PF block centroids (control points) to find a weighted least squares solution for the transformation matrix coefficients. The homography is used to warp the image at time t into the coordinate system of the base frame at time $(t-k)$. The two images, $\mathbf{I}(x, y, t)$ and $\mathbf{I}(x, y, t-k)$ can be related by a projective transformation (or homography) when the scene points are approximately planar. Let the image coordinates of the same scene point lying on the plane π be $P(x, y)$ and $P'(x', y')$, in the view at time t and $(t-k)$ respectively. The two views can be related by the following homogeneous transformation or homography in matrix form:

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & w \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4}$$

$$P' = \mathbf{A}_{(t-k,t)} P \tag{5}$$

This transforms position P observed at time t, to position P' in the coordinate system at time $(t-k)$ via the projective transformation matrix (a backward transformation from time t to time $(t-k)$). Usually we assume $w = 1$ in matrix $\mathbf{A}$.

Suppose we are given three images, $\mathbf{I}(x, y, t-2)$, $\mathbf{I}(x, y, t-1)$, $\mathbf{I}(x, y, t)$ with corresponding planar points, P'', P', P and homography transformation matrices $\mathbf{A}_{(t-1,t)}$ and $\mathbf{A}_{(t-2,t-1)}$ that projectively maps t to $(t-1)$ (i.e. Frame 2 to Frame 1) and $(t-1)$ to $(t-2)$ (i.e. Frame 1 to Frame 0), respectively. Without loss of generality we assume for simplicity of notation that the images are sequentially sampled at one unit time intervals, t, $(t-1)$, and $(t-2)$. We can then write the two respective projective transformations as,

$$P' = \mathbf{A}_{(t-1,t)} P \quad and \quad P'' = \mathbf{A}_{(t-2,t-1)} P' \tag{6}$$

and the composite or cumulative projective transformation relating pixels in frame t to pixels in frame $(t-2)$ (i.e. pixels in Frame 2 to pixels in Frame 0), as the product of two homographies or projective maps/transformations: $P'' = \mathbf{A}_{(t-2,t-1)} \mathbf{A}_{(t-1,t)} P$ In the general case, mapping pixel positions from frame t to corresponding pixel positions in the coordinate system of frame $(t-k)$, we have

$$P(t-k, t) = \mathbf{A}_{(t-k,t)} P(t, t) \tag{7}$$

$$\mathbf{A}_{(t-k,t)} = \mathbf{A}_{(t-k,t-k+1)} \mathbf{A}_{(t-k+1,t-k+2)} \ldots . \mathbf{A}_{(t-2,t-1)} A_{(t-1,t)} \tag{8}$$

We also need to specify the coordinate system in which we reference or measure a pixel's position. Since the prime notation is limited, $P(t-k, t)$ denotes pixel position/geometry from image $\mathbf{I}(x, y, t)$ mapped to the coordinate system of image frame $\mathbf{I}(x, y, t-k)$ and $P(t, t)$ is the pixel position measured in its original coordinate system $\mathbf{I}(x, y, t)$. The elements of matrix $\mathbf{A}$ in Eq. 4 and 5 can be solved using weighted least squares, robust statistics such as LMedS or combinatorial methods such as RANSAC. Each pair of corresponding points provides three linear constraints that can be written in matrix form as shown below, where $\mathbf{a}_i^{\mathrm{T}}$ is the i^{th} row of $\mathbf{A}$ in Eq. 4, [10] ,

$$\begin{bmatrix} \mathbf{0}^{\mathrm{T}} & -w_i' \mathbf{x}_i^{\mathrm{T}} & -y_i' \mathbf{x}_i^{\mathrm{T}} \\ w_i' \mathbf{x}_i^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} & -x_i' \mathbf{x}_i^{\mathrm{T}} \\ y_i' \mathbf{x}_i^{\mathrm{T}} & x_i' \mathbf{x}_i^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \\ \mathbf{a}_3^{\mathrm{T}} \end{bmatrix} = \mathbf{0} \tag{9}$$

5. UAV-VIDEO REGISTRATION ALGORITHM

A segment of video whose length is adaptively determined based on the amount of scene change as well as parallax effects, and is typically two to a few seconds in length or about 50 to 100 frames, is registered using the following procedure:

1: Compute image spatial gradients I_x, I_y and trace of the color structure tensor matrix $\mathbf{J}_C$ at every pixel (Eq.**??**)

2: Threshold on $trace(J_C(X)) > th$ to remove noise pixels, low confidence pixels and homogeneous regions.

3: For each remaining (non-zero) feature pixel, compute the Beltrami color metric tensor (*i.e.* determinant term) at each potential feature point X_i using Eq. 1.

4: Establish highly confident PF regions for registration or tracking. Divide the image into 16 × 16 non-overlapping macroblocks, B_k, and classify prominent feature blocks based on a high PF value, which measures the percent cornerness that can be used as a block confidence measure for weighted least squares.

5: For each prominent feature block in current frame, t, find the best match in the previous frame, $(t-1)$, by maximizing NCC or minimizing SAD search (in intensity).

6: Compute the direction histogram of the motion vectors obtained in Step 5, and apply motion filtering.

7: Compute the weighted least squares (backward) homography between *adjacent* frames t and $(t-1)$, $A_{(t-1,t)}$.

```
8: Compute the homography to warp from frame $t$ to frame $(t-k)$, $A_{(t-k,t)}$, and associate it with frame $t$.
9: Warp current image $I(x,y,t)$ into the coordinate system of the active video segment base (or reference) frame $I(x,y,t-k)$ using the homography from Step 8.
```

Once a video segment has been registered then it can flow into a processing chain for object detection, tracking and verification [11]. Chunks of video segments can be analyzed in a similar fashion, interconnected and summarized.

6. RESULTS AND CONCLUSIONS

We have tested the proposed UAV-video registration algorithm using datasets from the DARPA VIVID program [11] and available online [1]. Sample results from the EgTest01 video collected at Eglin that tracks vehicles on a runway are presented. We use RMSE to measure the residual differences between the homography warped image of $\mathbf{I}(x,y,t)$ and the base frame $\mathbf{I}(x,y,t-k)$ as,

$$\text{RMSE} = \sqrt{\frac{1}{\#\mathcal{B}_{(t-k)}} \sum_{P \in \mathcal{B}_{(t-k)}} \left[\mathbf{I}(P(t-k)) - \mathbf{I}(\mathbf{A}_{(t-k,t)}P(t))\right]^2} \tag{10}$$

where $P(t) \equiv P(t,t)$ is used to simplify the notation and $\#\mathcal{B}_{(t-k)}$ is the size (in pixels) of the background set. Qualitative results showed that the registration accuracy is very good with very little error in the background except in areas of motion. The difficulty with measuring registration accuracy using a warping RMS error measure is that it depends on the relative amounts of background/camera motion and foreground/object motion. In order to avoid penalizing both background and foreground errors equally we restrict the domain to a *manually* selected set of background pixels in Eq. 10 and avoid contamination by moving object pixels to accurately assess the registration performance over a long video segment. A graph of RMSE shows that the (image intensity) RMS error compared to the base frame of a video segment increases linearly with a shift in slope around 50 frames (for SAD). So the cumulative warping error which will be scene and sensor motion dependent can be used in an adaptive framework to determine the point at which the base frame for warping needs to be reset and a new video segment started.

In terms of performance, the SAD computation by itself runs at about 8 fr/s (cumulative sums with Minkowski inequality for early termination) and NCC matching at 5 fr/s (cumulative sums for denominator terms) for 16×16 non-overlapping macroblocks and full block search within a 32×32 target window to produce $1,064$ displacement or motion vectors for a 640×480 image. The unoptimized performance is 5 and 0.5 fr/s for SAD and NCC respectively. Testing was done using a dual Intel Xeon 2.66GHz PC with 4MB cache and 12GB memory. The actual frame rate will be scene dependent but faster since feature selection will reduce the number of PF macroblocks by typically 75%. The SAD can be further sped up by using a suboptimal partial search like 4-step that evaluates on average 35 SADs instead of 289 (17 × 17) steps. The NCC uses optimized array access, register variables, cumulative sums, and loop unrolling. Adding cache coherency should make NCC faster. Parallel GPU and Cell processor implementations will enable scaling to HD video.

In addition to speed optimization, a variety of other image structures and techniques can be used in the first few steps of the algorithm. The performance of tensors based on (first derivative) gradient information is being compared to second derivative features such as the Hessian [12] and coupling between registration and other vision tasks [13]. Non-intensity descriptors, such as signed directional wavelet responses within a local neighborhood can also be used to improve matching.

[1] http://www.vividevaluation.ri.cmu.edu/datasets/datasets/datasets.html#pets2005

7. REFERENCES

[1] Y. Sheikh, S. Khan, M. Shah, and R.W. Cannata, *Geodetic Alignment of Aerial Video Frames*, p. Chapter 7, 2003.

[2] G. Zhou, C. Li, and P. Cheng, “Unmanned aerial vehicle (UAV) real-time video registration for forest fire monitoring,” in *IEEE Geoscience and Remote Sensing Symposium*, 2005, vol. 3, pp. 1803 – 1806.

[3] L. G. Brown, “A survey of image registration techniques,” *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, 1992.

[4] B. Zitová and Jan Flusser, “Image registration methods: A Survey,” *Image and Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[5] K. Mikolajczyk and *et. al* , “A comparison of affine region detectors,” *Int. J. Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005.

[6] F. Bunyak, K. Palaniappan, S. Nath, and G. Seetharaman, “Geodesic active contour based fusion of visible and infrared video for persistent object tracking,” in *8th IEEE Workshop on Applications of Computer Vision (WACV 2007)*, Austin, TX, Feb. 2007, p. Online.

[7] F. Bunyak, K. Palaniappan, S. Nath, and G. Seetharaman, “Fux tensor constrained geodesic active contours with sensor fusion for persistent object tracking,” *J. Multimedia*, vol. 2, no. 4, pp. 20–33, August 2007.

[8] G. Seetharaman, G. Gasperas, and K. Palaniappan, “A piecewise affine model for image registration in 3-D motion analysis,” in *IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, Sep. 2000, pp. 561–564.

[9] L. Zhou, C. Kambhamettu, D. Goldgof, K. Palaniappan, and A. F. Hasler, “Tracking non-rigid motion and structure from 2D satellite cloud images without correspondences,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1330–1336, Nov. 2001.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[11] Z. Yue, D. Guarino, and R. Chellappa, “Moving object verification from airborne video,” in *IEEE Int. Conf. Computer Vision Systems (ICVS)*, 2006, p. Online.

[12] H. Bay, T. Tuytelaars, and L. J. Van Gool, “SURF: Speeded up robust features,” in *Lecture Notes in Computer Science (ECCV)*, 2006, vol. 3951, pp. 404–417.

[13] C. Kambhamettu, K. Palaniappan, and A. F. Hasler, “Coupled, multi-resolution stereo and motion analysis,” in *IEEE Int. Symp. Computer Vision*, 1995, pp. 43 – 48.