A Flow-Based Approach to Vehicle Detection and Background Mosaicking in Airborne Video

Hulya Yalcin
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
hulya@cs.cmu.edu

Robert Collins
CSE Department
Penn State University
University Park, PA 16802
rcollins@cse.psu.edu

Michael J. Black
Computer Science
Brown University
Providence, RI 02912
black@cs.brown.edu

Martial Hebert
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
hebert@cs.cmu.edu

Abstract

We address the detection of vehicles in a video stream obtained from a moving airborne platform. Our approach is based on robust optical flow algorithm applied on stabilized frames. Stabilization of the frames compensates for gross affine background motion prior to running robust optical flow to compute dense residual flow. Based on the flow and the previous background appearance model, the new frame is separated into background and foreground occlusion layers using an EM-based motion segmentation. The proposed framework shows that ground vehicles can be detected and segmented from airborne video sequences while building a mosaic of the background layer.

Keywords: motion estimation, background mosaicking, tracking, optical flow, airborne imagery.

1 Introduction

Detecting moving ground vehicles from airborne video is a difficult problem because all pixels in the image are moving due to the self motion of the camera. In this paper we present a technique to detect moving vehicles by segmenting dense optic flow fields into background and occlusion layers. Figure 1 illustrates the method. The robust dense optic flow algorithm is run on motion-compensated image pairs, yielding flow fields representing background residual flow and foreground object motion. Since the residual flow of the stabilized background should be smaller in magnitude than the foreground object motion, estimated variance of the residual flow magnitude can be used in a statistical test to determine the likelihood that each pixel is from the background or foreground, providing an ownership weight to the layer segmentation process. The result of layer segmentation is a background mosaic plus an ownership weight representing the probability of each pixel belonging to either the background or moving object layer.

We propose a Bayesian framework for estimating dense optical flow over time that explicitly estimates a persistent model of background appearance. The approach assumes that the scene can be described by background and occlusion layers, estimated within an Expectation-Maximization framework. The mathematical formulation of the paper is an extension of the work in [13] where motion and appearance models for foreground and background layers are estimated simultaneously in a Bayesian framework.

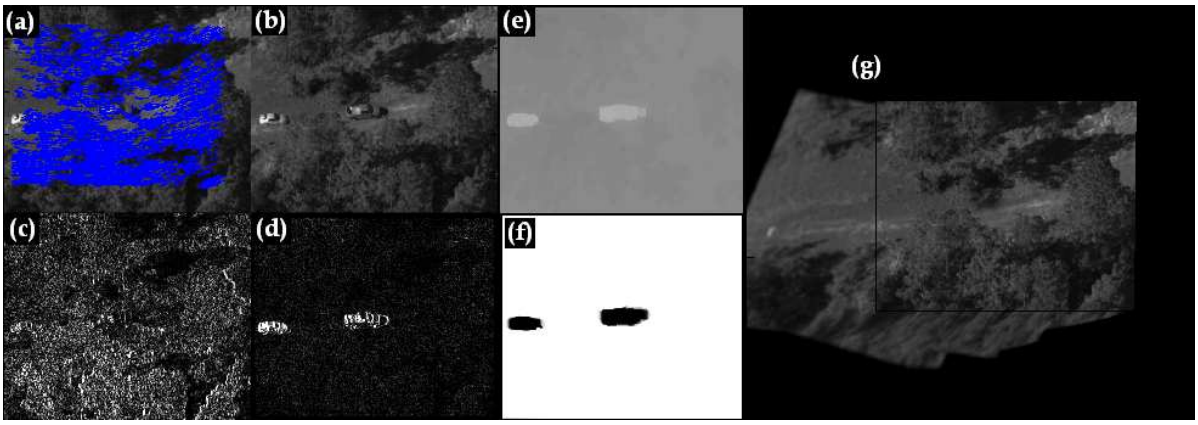

Figure 1: (a) I_{t-1} and KLT sparse flow, (b) I_t, (c) $|I_t - I_{t-1}|$, (d) $|I_t - I^S_{t-1}|$, (I^S_{t-1} : stabilized frame) (e) horizontal optic flow and background (f) ownership weight and (g) mosaic.

There are several contributions of this paper. First, we formulate a Bayesian framework for detecting and segmenting moving objects from the background on the basis of dense residual flow. The flow estimation contributes to background appearance estimation by providing a cue for differentiating between motion of background and foreground regions.

Second, We combine sparse flow results with dense optical flow results to separate video sequences into background and occlusion layers. Sparse flow features are used in two stages of the

approach. First, they provide an initial coarse stabilizaton of successive video frames. After stabilization, residual sparse flow motion statistics are used in computing motion priors when updating layer ownership weights.

Finally, computing optical flow between stabilized frames helps to cope with large background motion and accelerate the convergence of the robust optical flow. Even when the optical flow algorithm incorporates a coarse-to-fine resolution pyramid, we see improved optical flow results on stabilized frames. This is so because stabilized sequences conform better to the implicit motion prior of zero flow in typical optical flow computations.

Section 2 reviews related work on background stabilization and moving object detection from aerial video. In Section 3 we describe our approach to two frame stabilization, based on robust estimation of an affine transformation from sparse corner point correspondences. The heart of the paper is Section 4, which details the Bayesian background / foreground layer estimation procedure. Section 5 presents experimental results on two aerial tracking sequences, showing both estimated background mosaics and detected foreground object regions.

2 Related Work

This section reviews related work where background stabilization is used in detect multiple moving ground vehicles from airborne video. The recent "MTT" tracker by Alphatech [1] stabilizes video subsequences by aligning N frames to one reference frame using an 8-parameter planar perspective warp. As frames are aligned, background color statistics are recursively estimated at each pixel to form a statistical background model. Moving objects are detected by thresholding against the background Mahalanobis color distance and forming connected component blobs from flagged foreground pixels.

Bell et.al. [3] perform affine motion stabilization between adjacent pairs of frames based on a sparse set of corner matches. Intensity-based image segmentation of individual frames is performed, and the residual translation of each segmented patch is computed using an iterative Lucas-Kanade approach. Candidate moving regions are detected using thresholding and connected components on the magnitude of residual translation. Medioni et.al. [10] also stabilize adjacent frames using affine transformations computed over sparse corner features. Motion regions are detected by thresholding the magnitude of residual normal flow and applying connected components.

Tao et.al. [11] develop a practical, layer-based algorithm within a rigorous Bayesian framework that specifies data terms and priors for object appearance, motion and shape cues. The novelty that distinguishes this work from other layer-based approaches is imposition of shape constraints on the foreground object regions (the vehicles) by specifying elliptical shape priors that guide the layer segmentation to produce compact foreground regions.

Layered models of optical flow have been one of the key paradigms for simultaneously segmenting the scene and estimating its motion [2, 8, 12]. In particular, mixture model frameworks

make a soft assignment of pixels to layers. Unfortunately, this segmentation does not typically enforce spatial coherence between neighboring pixels and may, hence, be quite sparse. Additionally, these methods are typically limited to parametric motion models. An exception is [13], where a Bayesian framework is presented for estimating motion and appearance models of layers based on dense optical flow.

3 Stabilization

Two-frame stabilization is achieved by establishing correspondences between adjacent video frames and estimating an affine or higher order transformation that warps the images into alignment. We estimate image alignment by fitting a global parametric motion model to sparse optic flow. The Kanade-Lucas-Tomasi (KLT) feature tracker [4] is used to match corner features between adjacent pairs of video frames to obtain a sparse estimate of the optic flow field. For each corner feature, the method solves for a subpixel translational displacement vector that minimizes the sum of squared intensity differences between an image patch centered at the corner and a patch in the next frame centered at the estimated translated position.

A six parameter affine motion model is fit to the observed displacement vectors between two frames to approximate the global flow field induced by camera motion and a rigid ground plane. Higher order motion models such as planar projective could be used, however the affine model has been adequate in our experiments due to the large sensor standoff distance, narrow field of view, and nearly planar ground structure in aerial sequences. We use a Random Sample Consensus (RANSAC) procedure [7] to robustly estimate affine parameters from the observed displacement vectors. The benefit of using a robust procedure such as RANSAC is that the final least squares estimate is not contaminated by erroneous displacement vectors, points on moving vehicles in the scene, and scene points with large parallax.

4 Dense Motion Estimation and Background Mosaicking

In this section, we model dense motion estimation and background mosaicking in a Bayesian framework. To model the complexity of natural images where objects move and occlude each other, we introduce the notion of occlusion layer into the dense flow formulation. In particular, we introduce a background layer with appearance model and an occlusion layer and estimate these from an image sequence. The background appearance model adapts over time, and the probabilistic formulation can be used to provide a segmentation of the secene into background/occlusion regions.

4.1 Bayesian Framework

Simultaneous background mosaicking and motion estimation can be formulated as the maximization of the posterior probability

$$\arg\max_{B_t,\mathbf{u}_t} P(B_t, \mathbf{u}_t | B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_{t-1}) \tag{1}$$

where $\mathbf{u}_t = (u_t(\mathbf{x}), v_t(\mathbf{x}))$ is the horizontal and vertical flow at a pixel $\mathbf{x}$, and B_t is the *background appearance model* (intensity-based model) at time t, which serves as a "memory" of the observed region.

Using Bayes' rule, we rewrite the posterior probability as

$$\begin{aligned} P(B_t, \mathbf{u}_t | B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_{t-1}) \quad \propto \quad & P(B_t | B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t) \\ & P(\mathbf{u}_t | \mathbf{u}_{t-1}) \\ & P(\mathbf{u}_t | \mathbf{u}_t(\mathcal{G}_\mathbf{x})). \end{aligned}$$

where $\mathcal{G}_\mathbf{x}$ is the set of four neighbors for pixel $\mathbf{x}$, $P(B_t | B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t)$ is the likelihood term, and $P(\mathbf{u}_t | \mathbf{u}_{t-1})$ and $P(\mathbf{u}_t | \mathbf{u}_t(\mathcal{G}_\mathbf{x}))$ are the temporal and spatial coherence of motion respectively. The posterior holds at every pixel $\mathbf{x}$, but we omit $\mathbf{x}$ in the rest of the text for the sake of simplicity.

The goal here is to incrementally estimate the appearance model B_t and the dense motion $\mathbf{u}_t$ by taking into account the observed image, the past appearance, and the motion. As for the previous estimations and observations, we use stabilized background appearance model (B^S_{t-1}) and image observation (I^S_{t-1}) from the previous time instant.

Assuming that each image in the sequence can be separated into background and occlusion layers, the likelihood of observing image I_t can be represented as a mixture model

$$\begin{aligned} P(B_t | B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t) \quad = \quad & m^b \, P_b(B_t | I_t, I^S_{t-1}, \mathbf{u}_t) \\ & + m^{occ} \, P_{occ}(B_t | B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t) \\ & + m^o. \end{aligned} \tag{2}$$

The probability of each pixel belonging to different layers is given by the mixture probabilities m^b, m^{occ} and m^o. These mixing probabilities sum to 1, where m^o is a fixed outlier probability. In our experiments, m^o=0.

For any pixel in the current image, the likelihood for the background layer is

$$P_b(B_t | I_t, I^S_{t-1}, \mathbf{u}_t) = P(B_t | I_t) \cdot P(I_t | I^S_{t-1}, \mathbf{u}_t). \tag{3}$$

This likelihood simply enforces that the successive images in the sequence look similar when aligned using the motion, and that background appearance model be similar to the current image in regions with high background mixing probability.

The likelihood for the occlusion layer is

$$P_{occ}(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t) \;=\; P(B_t|B^S_{t-1}, \mathbf{u}_t) \cdot P(I_t|I^S_{t-1}, \mathbf{u}_t). \tag{4}$$

This likelihood enforces that successive background appearance layers look similar, even when they are occluded.

The temporal term $P(\mathbf{u}_t|\mathbf{u}_{t-1})$ simply enforces that the flow at the current instant is similar to the flow at the previous instant. The spatial term $P(\mathbf{u}_t|\mathbf{u}_t(\mathcal{G}_\mathbf{x}))$ is a standard one based on the difference between neighboring horizontal and vertical flow values. All these terms are represented with a robust likelihood function [5]. For optimization, we minimize the negative log of the posterior and these terms become robust error terms. Details are provided below.

4.2 Optimization

Given images in a sequence as well as a flow field and background appearance model at time $t-1$, we seek the appearance model B_t, optical flow field $\mathbf{u}_t$, and the mixture probabilities m^b and m^{occ} that provide a maximum likelihood fit to the data set. This problem can be considered by maximizing the posterior probability. At every new time instant, we need to estimate the background appearance model and the motion. We use the Expectation-Maximization (EM) algorithm [6] to solve for the $(B_t, \mathbf{u}_t)$ pairs.

According to the generalized EM algorithm, a local optimal solution can be achieved by iteratively optimizing the following function with respect to B_t and $\mathbf{u}_t$

$$\begin{aligned} L(B_t, \mathbf{u}_t) \;=\; & log\ P(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t) \\ & +log\ P(\mathbf{u}_t|\mathbf{u}_{t-1}) \\ & +log\ P(\mathbf{u}_t|\mathbf{u}_t(\mathcal{G}_\mathbf{x})) \\ & +\lambda(1 - m^o - m^b - m^{occ}) \end{aligned} \tag{5}$$

Note that the constraint that the mixing probabilities sum to one is imposed with a Lagrange multiplier.

At a local extremum it can be shown that the parameters m^b, m^{occ}, m^o, B_t and $\mathbf{u}_t$ must satisfy

$$q_b \cdot \frac{\partial}{\partial B_t}\left(log\ P(B_t|I_t)\right) + q_{occ} \cdot \frac{\partial}{\partial B_t}\Big(log\ P(B_t|B^S_{t-1}, \mathbf{u}_t)\Big) = 0 \tag{6}$$

and

$$(q_b + q_{occ}) \cdot \frac{\partial}{\partial \mathbf{u}_t}\Big(log\ P(I_t|I^S_{t-1}, \mathbf{u}_t)\Big) + \frac{\partial}{\partial \mathbf{u}_t}\left(log\ P(\mathbf{u}_t|\mathbf{u}_{t-1})\right) + \frac{\partial}{\partial \mathbf{u}_t}\left(log\ P(\mathbf{u}_t|\mathbf{u}_t(\mathcal{G}_\mathbf{x}))\right) = 0. \tag{7}$$

Here q_b represents the *background ownership probability*, that is the probability that the observed image value I_t belongs to the background layer. Similarly, q_{occ} represents the *occlusion layer*

ownership probability. The ownership weights are defined by

$$q_b = \frac{m^b \cdot P(B_t|I_t) \cdot P(I_t|I^S_{t-1}, \mathbf{u}_t)}{P(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t)} \tag{8}$$

and

$$q_{occ} = \frac{m^{occ} \cdot P(B_t|B^S_{t-1}, \mathbf{u}_t) \cdot P(I_t|I^S_{t-1}, \mathbf{u}_t)}{P(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t)}. \tag{9}$$

These equations for a maximum likelihood fit have been previously derived by requiring that the partial derivative of $L(B_t, \mathbf{u}_t)$ with respect to the parameters $B_t, \mathbf{u}_t$ must vanish [8, 9]. Derivation details are in Appendix.

We first estimate the ownership probabilities q_b and q_{occ} for each layer. This is the expectation step. Given these ownership probabilities, we compute the background appearance model and the motion that optimizes Eq. 6 and 7 in the maximization step.

The likelihoods and priors are modeled by a *t*- distribution of degree 3. The robust error function is given by the negative log:

$$\rho(x, \sigma, \alpha) = -log\left[\left(\frac{2\sigma^3}{\pi(\sigma^2 + x^2)^2}\right)^{\alpha}\right] \tag{10}$$

where α is a parameter that determines the relative importance of each of the likelihood and prior terms. We define the derivative of this function as $\psi(x, \sigma, \alpha)$

$$\psi(x, \sigma, \alpha) = \frac{d}{dx}\rho(x, \sigma, \alpha) = \alpha\, \frac{-4x}{\sigma^2 + x^2}.$$

After the derivations, the actual equations in the M-step are found to be

$$\begin{aligned} u(x)^{n+1} \;=\; & u(x)^n - \; (q_b + q_{occ}) \cdot \psi(I_t(x) - I^S_{t-1}(x - u_t), \sigma_{II}, \alpha_{II}) \\ & - \psi(u_t(x) - u_{t-1}(x), \sigma_{temp}, \alpha_{temp}) \\ & - \sum_{\mu \epsilon \mathcal{G}_x} \psi(u_t(x) - u_t(\mu), \sigma_{sp}, \alpha_{sp}) \end{aligned}$$

and

$$\begin{aligned} B_t(x)^{n+1} \;=\; & B_t(x)^n - \; q_b(x) \cdot \psi(B_t(x) - I_t(x), \sigma_{IB}, \alpha_{IB}) \\ & - q_{occ}(x) \cdot \psi(B_t(x) - B^S_{t-1}(x - u_t), \sigma_{BB}, \alpha_{BB}) \end{aligned}$$

where α_{II}, α_{BB}, α_{temp}, α_{sp} are the α parameters for the image likelihood, appearance prior, and temporal and spatial motion priors respectively. We used following parameters: α_{II} = 10 , α_{IB} = 5 , α_{BB} = 5 , α_{sp} = 2.5 , α_{temp} = 2 , σ_{II} = 20 , σ_{IB} = 30 , σ_{BB} = 20 , σ_{sp} = 20 , σ_{temp} = 10 .

Intuitively, the above equations (derived from Eq. 6 and 7) can be interpreted as follows: there are two terms that contribute to the background appearance model in the M-step for appearance optimization. The first term indicates that the background appearance model should adapt to the new information in the current image, change appearance if necessary, and that regions with high background layer ownership weights are more likely to be adapted to the current image since the whole term is multiplied by q_b. For regions of high occlusion layer ownership weight, the second term dominates and associates successive background appearance models using the dense motion. Simply, this term suggests that, in regions of occlusion, the background appearance model from the previous time instant should be maintained after being warped by the layer motion.

The M-step for motion optimization is identical to standard optical flow formulation since the brightness constancy term (first term) is valid in both background and occlusion regions. The second and the third terms suggest that the motion at a pixel should be similar to that of neighboring pixels in space and time.

4.3 Updating mixing probabilities

In our formulation, the mixing probabilities are simply the ownership weights. Yet, we expect these mixing probabilities, which represent the assignment of the pixels to layers, to be stable over time. To model this, we gradually update them as the ownership probabilities change.

We initially set $m_0^b = m_0^{occ} = 0.5$ and then the mixing probabilities for the next time instant are updated by a linear combination of ownership weights and motion priors as follows

$$m_{t+1}^b = \alpha_1 \cdot m_t^b + \alpha_2 \cdot q_b + \alpha_3 \cdot (1 - p(u_t)) \tag{11}$$

$$m_{t+1}^{occ} = \alpha_1 \cdot m_t^{occ} + \alpha_2 \cdot q_{occ} + \alpha_3 \cdot p(u_t) \tag{12}$$

where $p(u_t) = exp(||u_t||, \sigma_{motion\_prior})$. The motion prior variance σ_{motion_prior} is obtained using EM on sparse flow computations after stabilization. In this way, we use the motion statistics computed through stabilization to distinguish background and occlusion layers. We expect the residual background motion to be slower than the motion of occluding foreground regions, and adding a prior that models this assumption helps separate background and occlusion layers. The mixing probabilities of each layer act as a prior on every pixel representing the probability of each pixel belonging to that layer. In our experiments, $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$.

5 Experimental Results

The flow diagram of our approach is illustrated in Figure 2. Sparse flow features obtained via KLT are computed at two different stages of the algorithm. First, sparse flow features are used for obtaining the affine transformation matrix for stabilization. Once the previous frame is stabilized towards the current frame, a second set of sparse flow features are computed, and a two-component

Gaussian mixture model is fit to them using the EM algorithm. It is expected that component with the higher mixing weight represents the residual motion of the scene background, and the motions statistics of this Gaussian component are used for computing the motion priors when updating background and occlusion layer mixing probabilities.

After stabilization, dense optical flow and background appearance model are estimated. Computing optical flow between stabilized frames as opposed to current and previous frames helps to cope with large background motion and accelerate the convergence of the robust optical flow, resulting in improved optical flow results.

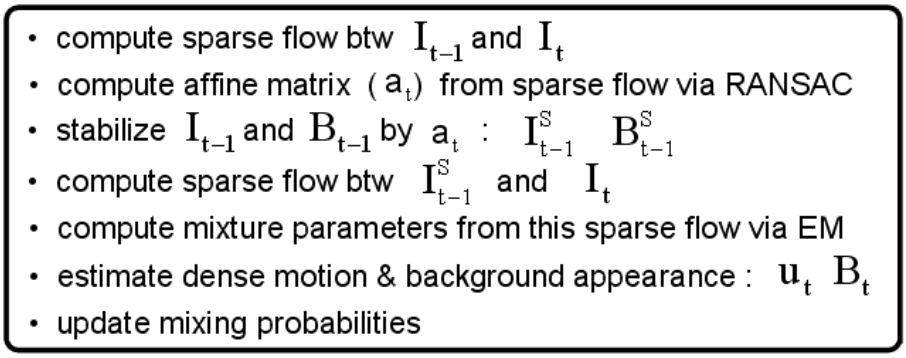

Figure 2: Flow diagram of the approach.

We experimented with our approach on two video sequences. Figures 3-4 illustrate the sparse flow features, stabilized frames, dense optical flow fields, background ownership weight and the appearance of the occlusion layer and the background mosaic.

Initially, we assume that the steady regions after stabilization belong to the background layer. Hence, we set the regions with no motion as the initial background appearance model and some regions in the background appearance model appear as blacked out. As further frames are processed, these regions are gradually recovered since the occluded regions (regions with low background ownership weight) are disoccluded and filled in with the warped appearance model from the previous time instant. Regions with high background ownership weight are updated from the current image.

From the figures, we see the ownership weights clearly delineate the moving ground vehicles. In particular, note how the shape of vehicles partially occluded by trees is extracted. This is possible because dense optical flow yields a pixel-level labeling that is more faithful to the image data than simple frame differencing on stabilized image frames.

6 Conclusions

In this paper, we presented an approach for detecting vehicles in airborne video imagery while estimating a background mosaic from the video stream. Our approach is based on robust optical

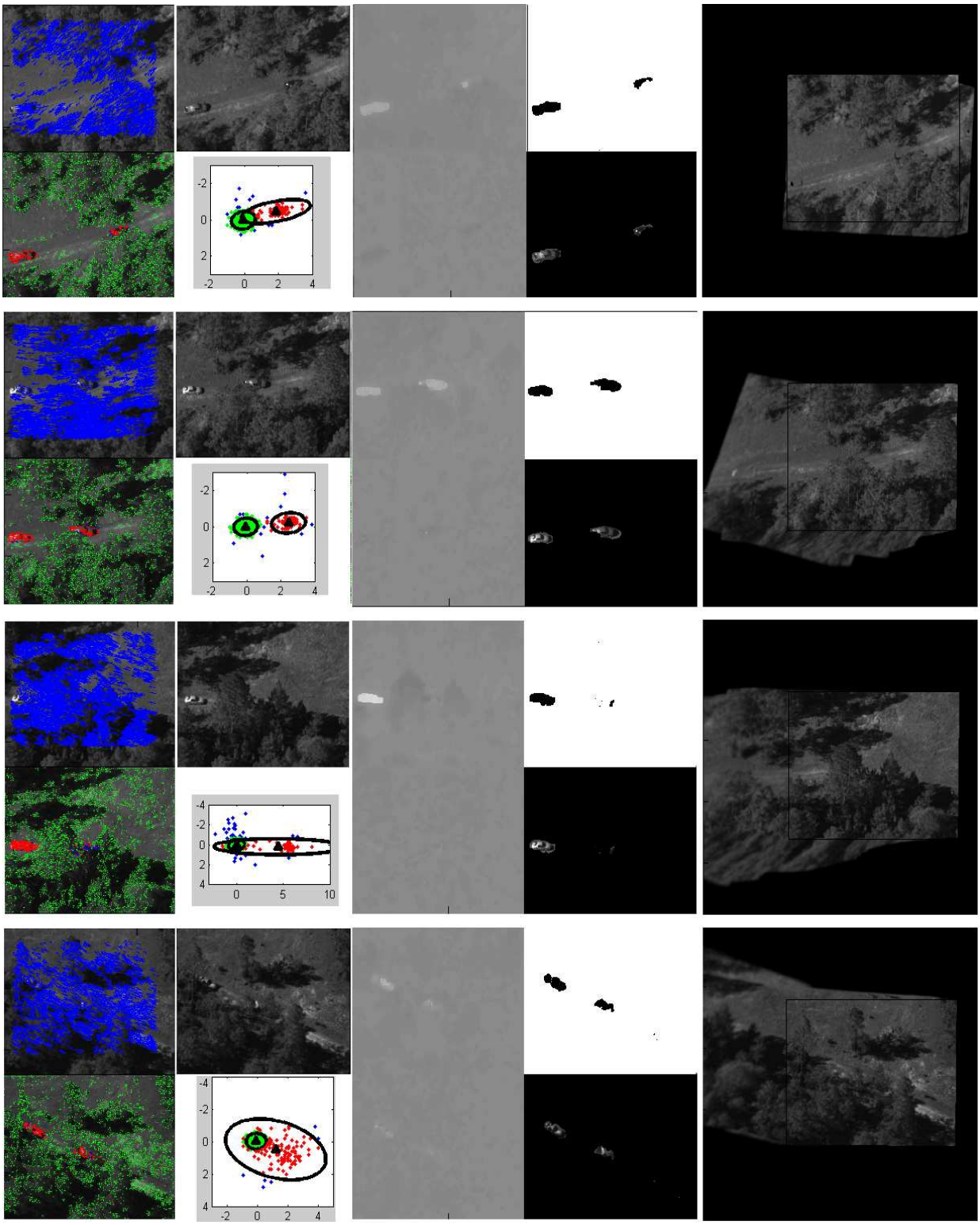

Figure 3: Results of our approach for the first sequence. KLT sparse flow before and after stabilization (column 1). Stabilized frames and EM results (column 2). Horizontal and vertical components of the robust optical flow (column 3). Background layer ownership weight and the occlusion layer appearance (column 4). Background mosaic (column 5).

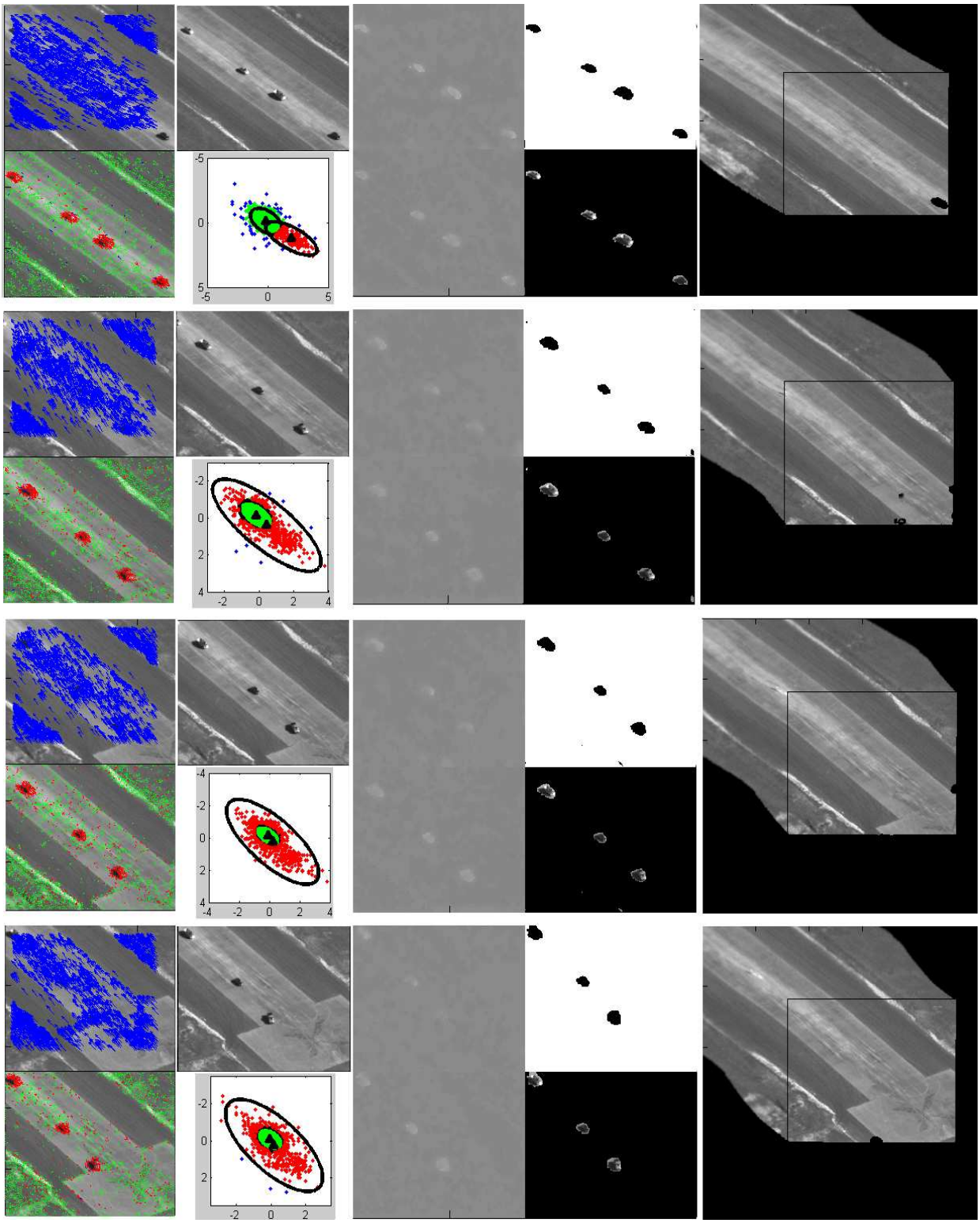

Figure 4: Results of our approach for the second sequence. KLT sparse flow before and after stabilization (column 1). Stabilized frames and EM results (column 2). Horizontal and vertical components of the robust optical flow (column 3). Background layer ownership weight and the occlusion layer appearance (column 4). Background mosaic (column 5).

flow algorithm applied on stabilized frames. Stabilization of the frames compensates for gross affine background motion prior to running robust optical flow to compute dense residual flow. Based on the flow and the previous background appearance model, the new frame is separated into background and foreground occlusion layers using an EM-based motion segmentation. The preliminary results presented here show that ground vehicles can be detected and segmented from airborne video sequences while building a mosaic of the background layer.

References

[1] P. Arambel, J. Silver, J. Krant, M. Antone, and T. Strat. Multiple-hypothesis tracking of multiple ground targets from aerial video with dynamic sensor control. In *Proc. of SPIE 5429, (Signal Processing, Sensor Fusion, and Target Recognition XIII)*, pages 23–32, Aug. 2004.

[2] S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum likelihood estimation of mixture models and mdl encoding. In *ICCV95*, pages 777–784, 1995.

[3] W. Bell, P. Felzenszwalb, and D. Huttenlocher. Detection and long term tracking of moving objects in aerial video. Technical report, Computer Science, Cornell, March 1999.

[4] S. Birchfield. KLT: an implementation of the Kanade-Lucas-Tomasi feature tracker, 1997.

[5] M.J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *ICCV93*, pages 231–236, 1993.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

[7] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24:381–395, 1981.

[8] A. Jepson and M.J. Black. Mixture models for optical flow computation. In *CVPR93*, pages 760–761, 1993.

[9] G.J. McLachlan and K.E. Basford. Mixture models: inference and applications to clustering. *Marcel Dekker Inc.*, 1988.

[10] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *PAMI*, 23(8):873–889, August 2001.

[11] H. Tao, H. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *PAMI*, 24(1):75–89, January 2002.

[12] Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *CVPR97*, pages 520–526, 1997.

[13] H. Yalcin, M. Black, and R. Fablet. The dense estimation of motion and appearance in layers. *Second IEEE Workshop on Image and Video Registration (IVR'04)*, 2004.

7 Appendix

According to the generalized EM algorithm, a locally optimal solution can be achieved by iteratively optimizing Eq. 5 wrt to parameters B_t and $\mathbf{u}_t$. Taking derivative of $L(B_t, \mathbf{u}_t)$ wrt B_t, we get

$$\begin{aligned}\frac{\partial L(B_t,\mathbf{u}_t)}{\partial B_t} &= \frac{\partial P(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)/\partial B_t}{P(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)} \\ &= \frac{m^b \cdot \frac{\partial}{\partial B_t}\left(P_b(B_t|I_t, I_{t-1}^S, \mathbf{u}_t)\right)}{P(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)} + \\ &\frac{m^{occ} \cdot \frac{\partial}{\partial B_t}\left(P_{occ}(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)\right)}{P(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)}\end{aligned}$$

Replacing $\partial P_b(B_t|I_t, I_{t-1}^S, \mathbf{u}_t)/\partial B_t$ by

$$P_b(B_t|I_t, I_{t-1}^S, \mathbf{u}_t) \cdot \frac{\partial}{\partial B_t}\left(log\ P_b(B_t|I_t, I_{t-1}^S, \mathbf{u}_t)\right)$$

and $\partial P_{occ}(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)/\partial B_t$ by

$$P_{occ}(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)\ \cdot \frac{\partial}{\partial B_t}\left(log\ P_{occ}(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)\right)$$

Since the likelihoods are defined as in equation 3 and 4, $\partial\left(log\ P_b(B_t|I_t, I_{t-1}^S, \mathbf{u}_t)\right)/\partial B_t$ is simplified as $\partial\left(log\ P(B_t|I_t)\right)/\partial B_t$ and $\partial\left(log\ P_{occ}(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)\right)/\partial B_t$ is simplified as

$$\partial\left(log\ P(B_t|B_{t-1}^S, \mathbf{u}_t)\right)/\partial B_t$$

and rewriting the equation with these changes, we get

$$\begin{aligned}\frac{\partial L(B_t,\mathbf{u}_t)}{\partial B_t} &= \frac{m^b \cdot P_b(B_t|I_t, I_{t-1}^S, \mathbf{u}_t) \cdot \frac{\partial}{\partial B_t}\left(log\ P(B_t|I_t)\right)}{P(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)} + \\ &\frac{m^{occ} \cdot P_{occ}(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t) \cdot \frac{\partial}{\partial B_t}\left(log\ P(B_t|I_{t-1}^S, \mathbf{u}_t)\right)}{P(B_t|B_{t-1}^S, I_t, I_{t-1}^S, \mathbf{u}_t)}\end{aligned}$$

The replacement trick above and further simplifications lets us define

$$\begin{aligned} q_b &= \frac{m^b \cdot P_b(B_t|I_t, I^S_{t-1}, \mathbf{u}_t)}{P(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t)} \\ &= \frac{m^b \cdot P(B_t|I_t) \cdot P(I_t|I^S_{t-1}, \mathbf{u}_t)}{P(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t)} \end{aligned}$$

and

$$\begin{aligned} q_{occ} &= \frac{m^{occ} \cdot P_{occ}(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t)}{P(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t)} \\ &= \frac{m^{occ} \cdot P(B_t|B^S_{t-1}, \mathbf{u}_t) \cdot P(I_t|I^S_{t-1}, \mathbf{u}_t)}{P(B_t|B^S_{t-1}, I_t, I^S_{t-1}, \mathbf{u}_t)} \end{aligned}$$

Here q_b (q_{occ}) represents the *ownership probability*, that is the probability that the observed image I_t belongs to background (occlusion) layer. Given values for motion and the background appearance model, these ownership weights are computed as the expectation, or E-step.

Then, the M-step is formulated in compact form as

$$\frac{\partial L(B_t, \mathbf{u}_t)}{\partial A^i_t} = q_b \cdot \frac{\partial}{\partial B_t}\left(log\ P(B_t|I_t)\right) + q_{occ} \cdot \frac{\partial}{\partial B_t}\left(log\ P(B_t|B^S_{t-1}, \mathbf{u}_t)\right).$$

At a local extremum, the right hand side of the above equation will be equal to zero

$$q_b \cdot \frac{\partial}{\partial B_t}\left(log\ P(B_t|I_t)\right) + q_{occ} \cdot \frac{\partial}{\partial B_t}\left(log\ P(B_t|B^S_{t-1}, \mathbf{u}_t)\right) = 0.$$

Similarly if we take derivative of $L(B_t, \mathbf{u}_t)$ wrt $\mathbf{u}_t$, the M-step for motion optimization will be

$$\begin{aligned} &(q_b + q_{occ}) \cdot \frac{\partial}{\partial \mathbf{u}_t}\left(log\ P(I_t|I^S_{t-1}, \mathbf{u}_t)\right) \\ &+\frac{\partial}{\partial \mathbf{u}_t}\left(log\ P(\mathbf{u}_t|\mathbf{u}_{t-1})\right) \\ &+\frac{\partial}{\partial \mathbf{u}_t}\left(log\ P(\mathbf{u}_t|\mathbf{u}_t(\mathcal{G}_\mathbf{x}))\right) = 0. \end{aligned}$$