Independent Motion: The Importance of History

Robert Pless, Tomáš Brodský, and Yiannis Aloimonos

Center for Automation Research, University of Maryland
College Park, MD, 20742-3275

Abstract

We consider a problem central in aerial visual surveillance applications – detection and tracking of small, independently moving objects in long and noisy video sequences. We directly use spatiotemporal image intensity gradient measurements to compute an exact model of background motion. This allows the creation of accurate mosaics over many frames and the definition of a constraint violation function which acts as an indicator of independent motion. A novel temporal integration method maintains confidence measures over long subsequences without computing the optic flow, requiring object models, or using a Kalman filter. The mosaic acts as a stable feature frame, allowing precise localization of the independently moving objects. We present a statistical analysis of the effects of image noise on the constraint violation measure and find a good match between the predicted probability distribution function and the measured sample frequencies in a test sequence.

1 Introduction

Independent motion detection is the first step in many visual surveillance tasks. This involves creating a model of the background image or background motion and searching for regions that violate this model. In the case of aerial imagery, the creation of this background model can be simplified by the common assumption that the scene in view is well approximated by a plane.

Past approaches calculate the background motion model by matching and tracking a set of detected features in successive frames [1], using local image correlation to approximate optic flow [3], or with multi-scale gradient based methods [5]. This background motion model serves to stabilize the image of the background plane. Then, independent motion is detected as either residual flow [1], background subtraction, or temporal differencing of intensity [7].

The approach presented here directly uses image intensity derivatives to solve for the background plane motion and changes in the camera intrinsic parameters. The independence measure is the residual normal flow – the flow in the direction of the image gradient that is not predicted by the background plane motion. Using only image derivative measurements allows us to avoid computing optical flow or finding suitable feature points to track. Optic flow methods are particularly unstable in regions with flow discontinuities or independent motion.

Poor quality video requires temporal integration over many frames to increase detection accuracy. Section 2.3 details an "optimistic" historical integration scheme, which combines information along lines of flow consistent with the normal flow.

This paper presents a method for the computation of the instantaneous planar video homography directly from image intensity derivatives, a novel scheme for optimistic temporal integration along the normal flow constraints, and an analytic and empirical study of noise effects on the magnitude of the expected residual motion vectors.

2 Implementation

We briefly synopsize the execution of the algorithm for each frame. First, compute the spatiotemporal image intensity gradients and calculate the motion of the (assumed to be planar) background. Use the spatiotemporal gradient measurements to compute normal flow, "subtract" the just computed background optic flow, and define the squared magnitude of the residual as the constraint violation measure. Then, propagate the previous independent motion confidence values along the normal flow directions and update this belief with the current constraint violation measure. Finally, use an adaptive threshold to choose what confidence level is appropriate to mark as independently moving.

Explicit details of each step are presented below, but here we point out several important implementation choices. There is intentionally no spatial integration because this limits the minimum size of object that can be detected as independent. The temporal inte-

gration does not require *any* consistency in the motions between subsequent frames; because it propagates belief along directions consistent with the normal flow in each frame, there is no assumption that either the background motion or the independent motion relative to the background is constant or slowly varying. Many aerial videos suffer from camera vibrations that manifest themselves as camera rotations.

2.1 Estimation of background motion

Assuming only that the terrain can be approximated by a plane, a general model allows arbitrary rigid motions of the camera as well as continuous changes to the internal calibration parameters. The camera model is a general pinhole with up to 5 unknown intrinsic parameters. The mapping between a scene point $\vec{M}$ and the corresponding image point $\vec{m} = [x, y, F]^{\mathrm{T}}$ can be concisely written as [2]

$$\vec{m} = \frac{\mathbf{K}\vec{M}}{\vec{M} \cdot \vec{z}}$$

where $\vec{z} = [0, 0, 1]^{\mathrm{T}}$, F is a known constant and matrix $\mathbf{K}$ represents the intrinsic camera parameters. The camera rotational parameters are represented by $[\boldsymbol{\omega}]_\times$, which is the skew symmetric matrix corresponding to the cross product with $\boldsymbol{\omega}$:

$$\mathbf{K} = \begin{pmatrix} f_x & s & \Delta_x \\ 0 & f_y & \Delta_y \\ 0 & 0 & F \end{pmatrix}, \quad [\boldsymbol{\omega}]_\times = \begin{pmatrix} 0 & -\gamma & \beta \\ \gamma & 0 & -\alpha \\ -\beta & \alpha & 0 \end{pmatrix}$$

Using the differential motion model, the general image motion field can be decomposed into a rotational component, a translational component, and a component due to the changing intrinsic parameters [8]. If the scene in view is a plane, (so that $1/Z = \mathbf{q} \cdot \vec{m}$) the motion field equation can be simplified into

$$\dot{\vec{m}} = \frac{1}{F}(\vec{z} \times (\vec{m} \times (\mathbf{A}\vec{m}))) \qquad (1)$$

where $\mathbf{A} = \mathbf{K}\mathbf{t}\mathbf{q}^{\mathrm{T}} + \mathbf{K}[\boldsymbol{\omega}]_\times\mathbf{K}^{-1} + \dot{\mathbf{K}}\mathbf{K}^{-1}$. Here, $\mathbf{t}$ is the translational velocity and $\mathbf{q}$ defines the background plane. For any planar scene, $\mathbf{A}$ completely defines the motion field that arises from any differential rigid camera motion and any continuous change in intrinsic parameters. There is not enough information in the motion field of a planar surface to decompose A to find explicitly the structure or motion parameters. In practice, the change of intrinsic parameters can only be due to focusing or zooming. For real zoom lens the pinhole camera model is not sufficient and a thick lens model has to be used [6]. As a consequence, a change of focal length induces a shift of the camera nodal point, producing a motion field that depends on the scene in view and is in fact a special case of translational field. In practice, motion parallax due to zooming is extremely small compared with other factors, especially for noisy input sequences, and can be ignored.

Estimation of background motion based on (1) is straightforward. Define $\vec{n}$ as the unit image gradient direction. Normal projection of image flow yields

$$\dot{\vec{m}} \cdot \vec{n} = \frac{1}{F}(\vec{z} \times (\vec{m} \times (\mathbf{A}\vec{m}))) \cdot \vec{n} = u_n.$$

an equation linear in the unknown elements of $\mathbf{A}$. We solve for $\mathbf{A}$ with least squares, using all the measured values u_n at image points where the spatial image gradient is sufficiently large (as usual in normal flow computation).

For most aerial sequences, simple least squares solution is sufficient, to improve the precision we perform a two step estimation. After an initial least squares estimate is computed, measurements with large residuals are considered to be outliers and discarded, and the estimation is then repeated.

2.2 Constraint Violation Function

Interpreting the spatiotemporal gradient measurements as normal flow allows a local measure of agreement between the background motion and the local motion. The solution for the background motion defines an optical flow $\vec{u}$ for each pixel. From measured image derivatives, we can calculate normal flow $\vec{u_n}$ at each pixel. The residual motion is the difference between the measured normal flow, and the projection of the computed flow onto the computed unit normal direction $\vec{n}$:

$$\vec{r} = (\vec{u} \cdot \vec{n})\,\vec{n} - \vec{u_n}. \qquad (2)$$

If $\vec{u}$ is the correct optic flow in this region of the image, and the normal flow $\vec{u}_n$ is uncorrupted by error, $\|\vec{r}\| = 0$. In regions of independent motion, $\vec{r}$ will be the vector difference between the independent object motion and the background motion projected along the image gradient direction. Section 3 expands this equation in terms of the spatiotemporal derivatives and the noise terms assumed to corrupt the measurements of these derivatives in order to find the distribution of $\vec{r}\cdot\vec{r}$, both in regions where the $\vec{u}$ is correct, and in regions of independent motion. Defining the constraint violation function on the basis of normal flow vectors has two advantages. The normal flow is required for the temporal integration used in Section 2.3, and is more independent of the magnitude of the gradient or other

local image characteristics than image intensity differences. Small errors in the background motion always lead to small residual vectors, even where the image intensity difference is large.

2.3 Temporal Integration

At this point the algorithm has computed the background motion and found points where the background motion does not account for the measured normal flow. This constraint failure is caused either by noise or by independent motion in the scene. Video with low image quality has the property there will be both background regions and independently moving regions for which the constraint violation measure will exceed any chosen threshold for $\|\vec{r}\|$. This forces accurate detection to use temporal integration to find scene points that are consistently moving independently. This temporal integration will track an "independence energy" for every pixel, propagate this energy from frame to frame, and incorporate the constraint violation measure in the current frame. Intuitively, if $\|\vec{r}\|$ is very small, we set to *zero* the independence energy – the assumption is that noise can perhaps make the background look independent, but it will almost never make the independent motion appear to be background motion.

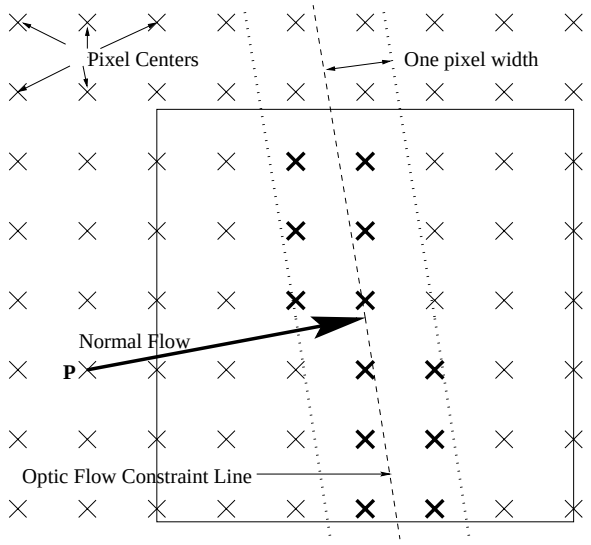

Figure 1: Energy is pushed to all of a set of pixels wherever such flow is consistent with the measured normal flow. Every pixel combines the current residual energy with the maximum energy pushed to it – All of the dark X's can use the energy from P if no other pixel sends more energy.

We thus need to find how to propagate the independence energy from frame to frame. There is a good model for pixel correspondences in the background, but the independently moving scene points do not (by definition) follow this model. We avoid computing the optical flow because optic flow algorithms are notoriously bad at motion discontinuities and small independently moving objects are essentially all motion discontinuities. To avoid the unstable optic flow computation, we use the fact that the normal flow – available directly from local spatiotemporal derivatives – defines a line of possible optic flow directions. The energy is pushed along a set of these directions as shown in Figure 1. This leads, of course, to an inexact mapping of confidence in one frame to confidence in the next frame – in effect it spreads the energy along lines of possible flow. This inexact mapping sends energy both to regions on the independently moving object (which is good), and onto background regions (which would seem to be bad). The probabilistic analysis given in Section 3, (and empirical data) finds that very soon these background pixels with independence energy are set to zero because their residual vector falls below the threshold.

Explicitly, the updated energy $\mathcal{E}'_{(i,j)}$ for a pixel at position (i,j) is:

$$\mathcal{E}'_{(i,j)} = \vec{r}_{i,j} \cdot \vec{r}_{i,j} + \max_{(x,y)} \left\{ \mathcal{C} \cdot \mathcal{E}_{x,y} \middle| (i,j) \in R(x,y) \right\}$$

In Figure 1 if P is at position (x,y), $R(x,y)$ is the set of dark X's – the set of pixels consistent with the measured normal flow. This update function sums the current squared residual flow magnitude with the maximum of $\mathcal{C}$ times independence energy that could have flowed toward this pixel, consistent with the measured normal flow. $\mathcal{C}$ is a linear interpolation function which is 1.1 if (i,j) lies exactly on the normal flow constraint line and decreases linearly to 0.5 on the boundary of region R. Essentially, the propagation is optimistic – and pragmatic. A very small residual flow magnitude is assumed to be background, and completely zeroes any independence energy. Otherwise, we combine the current residual with with the energy of any local pixel which could have an appropriate flow. This leads to an exponential growth in energy for pixels with long continuous chains of flow – consistent with the normal flow – that never have a residual flow below the threshold. This avoids any optic flow computation, but ensures that energy is passed along the correct optical flow.

2.4 Mosaics: A Stable Feature Frame

The set of successive frame to frame homographies solved for in Section 2.1 can be concatenated to find

the warping for any pair of images in the sequence. One can create a sequence mosaic by warping every frame onto the coordinate system of a chosen frame. A mosaic of strictly the background is created by removing the regions that contain the detected independently moving object from each warped frame. The redundancy inherent in the video stream allows the background mosaic to fill in areas occluded by independently moving objects.

Spatiotemporal image smoothing, and the temporal integration of independence energy results in an imprecise localization of independent motion. More accuracy can be obtained by projecting the image texture from our independently moving region onto the background mosaic and directly comparing image intensities. The differencing of pixel intensity can better localize the independent motion. This technique will fail to mark objects whose intensity is identical to the background, but this is unavoidable in the absence of a shape model of the independently moving object.

3 Probabilistic Analysis

In the absence of noise, the constraint violation function will be exactly zero for all image measurements of the background. A given noise model defines a distribution of the constraint violation measure for both background and independently moving regions. Consider noise to corrupt the image so that the measured intensity derivatives are the sum of the true derivative and a noise term: $\tilde{E_x} = E_x + n_x$, $\tilde{E_y} = E_y + n_y$, $\tilde{E_t} = E_t + n_t$. Rewriting the residual motion vector by using these derivative measurements and solving for the normal flow gives:

$$\vec{r} = \left\langle \frac{\tilde{E_x}\left(\tilde{E_t} - \tilde{E_x}u - \tilde{E_y}v\right)}{\tilde{E_x}^2 + \tilde{E_y}^2}, \frac{\tilde{E_y}\left(\tilde{E_t} - \tilde{E_x}u - \tilde{E_y}v\right)}{\tilde{E_x}^2 + \tilde{E_y}^2} \right\rangle. \tag{3}$$

In the case where the optical flow constraint equation holds, [4]

$$E_x u + E_y v + E_t = 0, \tag{4}$$

the squared magnitude of the residual vector given in (3) becomes:

$$\frac{(n_t - (un_x + vn_y))^2}{\tilde{E_x}^2 + \tilde{E_y}^2} \tag{5}$$

Because we ignore regions with small spatial gradients, $\langle \tilde{E_x}, \tilde{E_y} \rangle \gg \langle n_x, n_y \rangle$. Thus, for pieces of the image sequence where the motion $\langle u, v \rangle$ is approximately constant, if the derivative measurements are corrupted by independent Gaussian noise, then (5) is well approximated (through time) as a square of a Gaussian random variable.

When (4) is not satisfied – when the flow is measured on independently moving objects, for example – the squared magnitude of the residual vector is:

$$\frac{(E_x u_e + E_y v_e + n_t - (un_x + vn_y))^2}{\tilde{E_x}^2 + \tilde{E_y}^2} \tag{6}$$

where (u_e, v_e) is the difference vector between the background flow estimate and the actual optic flow at the point on the image. Assuming Gaussian error in the derivative measurements, this is equivalent to the distribution of the random variable $X = (a + N)^2$ where N is a zero mean Gaussian random variable and a is the magnitude of the residual flow projecting onto the gradient direction. This defines a probability density function dependent upon the relative velocity of the independently moving object. Figure 3 shows how this expected distribution of constraint violation measures changes as the relative motion increases.

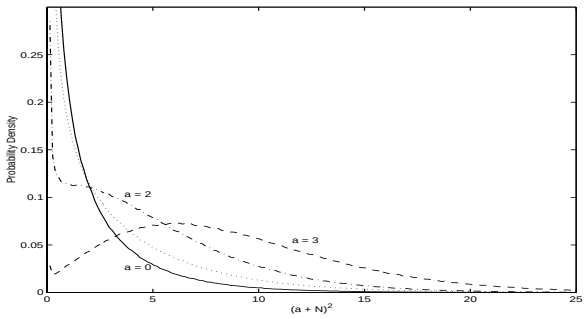

Figure 2: Probability distribution of $(a + N)^2$. Background measurements (solid line), have distribution shown with $a = 0$, also shown are a = 1,2,3; different magnitudes of relative motion.

4 Results

The test data is a video sequence taken from a camera mounted on a helicopter flying over a scene in which several people are running [1]. The video input was noisy, with a maximum frame to frame image flow of about 10 pixels and occasional large brightness changes. The independently moving people are typically 12 by 5 pixels in size. The hgih quality of the detection results is most visible as a video; the input sequence, the solution with temporal integration, and a dynamically created mosaic with enhanced object localization are available at (http://www.cfar.umd.edu/users/brodsky/detect.html).

[1] The video was kindly provided by the DARPA Airborne Visual Surveillance project

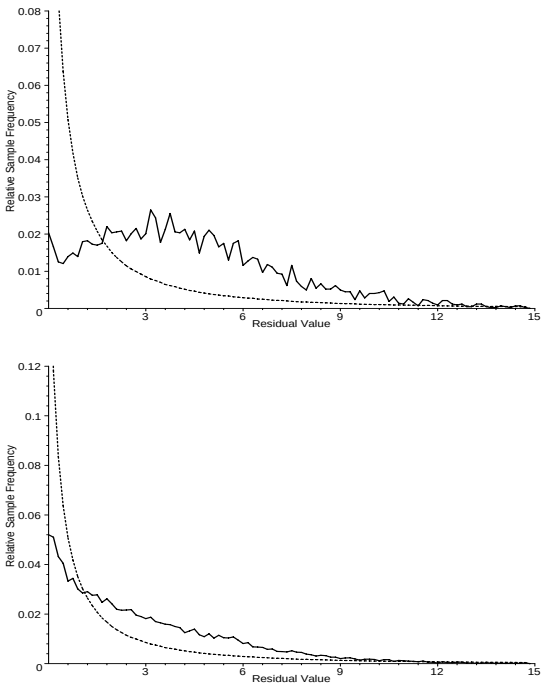

Figure 3: The distribution of constraint violation measures on the background (dotted) and the foreground for a scene segment with (top) large relative motion, (bottom) small relative motion. The small motion leads to bad localization, thus, the independent region includes some measurements of background flow.

In Figure 5, we show the tracks of the detected independently moving objects overlaid onto the background mosaic. Figure 4 shows the enhanced localization of the independently moving regions. This pixel accuracy is not possible with strictly flow based measurements because of the smoothing that is required to compute image derivatives. Empirical data shows that the sample distribution of the constraint violation measure is similar in form to the probability density function predicted in the analysis in Section 3. We show the distribution of the constraint violation measures in the background region, and the region marked as independent by the algorithm. Figure 3a is a piece of the sequence where we have very exact object localization and the relative motion is fairly large, Figure 3b is the distribution of measures from part of the sequence with much smaller relative independent motion. The importance of history is shown in Figure 6, which shows the distribution of independence energy that is built up over many frames. The background energy comes from the noise in the measurements, but does not build up over time, while the energy of the objects grows exponentially as their motion remains independent, leading to nearly disjoint energy distributions.

5 Conclusion

But for noise, accurate models of background motion would allow for precise detection of independent motion. An understanding of the noise distribution and its effects on image analysis allows algorithms to efficiently exploit the redundancy inherent in a video sequence. By approximately but efficiently computing consistent flow tracks through time, the overlapping distributions of a constraint violation function (figure 3b) can be transformed into a independence energy function with a clearly seperable distribution (figure 6).

References

[1] I. Cohen and G. Medioni. Detecting and tracking moving objects in video from and airborne observer. In *Proc. IEEE Image Understanding Workshop*, pages 217–222, 1998.

[2] O. D. Faugeras, Q.-T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In *Proc. European Conference on Computer Vision*, pages 321–334, Santa Margherita Ligure, Italy, 1992.

[3] M. Hansen, P. Anandan, K. Dana, G. van der Wal, and P. Burt. Real-time scene stabilization and mosaic construction. In *Proc. IEEE Image Understanding Workshop*, pages 457–463, 1994.

[4] B. K. P. Horn. *Robot Vision*. McGraw Hill, New York, 1986.

[5] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In *Proc. European Conference on Computer Vision*, pages 282–287, 1992.

[6] J.-M. Lavest, G. Rives, and M. Dhome. Three-dimensional reconstruction by zooming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(2):196–207, 1993.

[7] A. J. Lipton, H. Fujiyoshi, and R. S. Patil. Moving target classification and tracking from real-time video. In *Proc. IEEE Image Understanding Workshop*, pages 129–136, 1998.

[8] T. Viéville and O. D. Faugeras. The first-order expansion of motion equations in the uncalibrated case. *Computer Vision and Image Understanding*, 64:128–146, 1996.

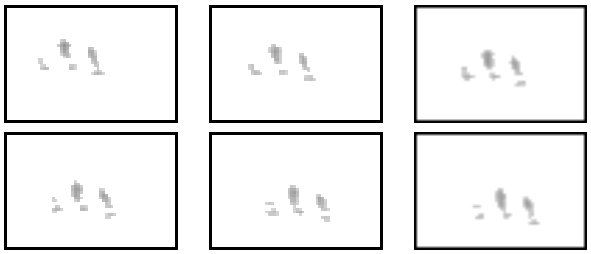

Figure 4: An example of the independent motion isolation using the image intensity differences between one frame and the stable background mosaic. Only a 60x40 pixel region from 6 consecutive frames is shown.

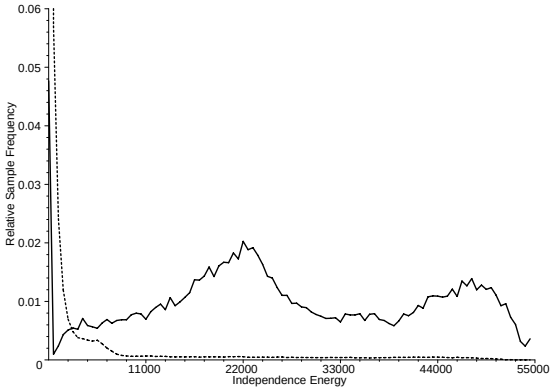

Figure 6: The importance of history: The distribution of the accumulated (over 80 frames) measure of independent motion on the background (dotted) and on a moving object. Due to the exponential growth for consistent independent motion, the two distributions are clearly separated.

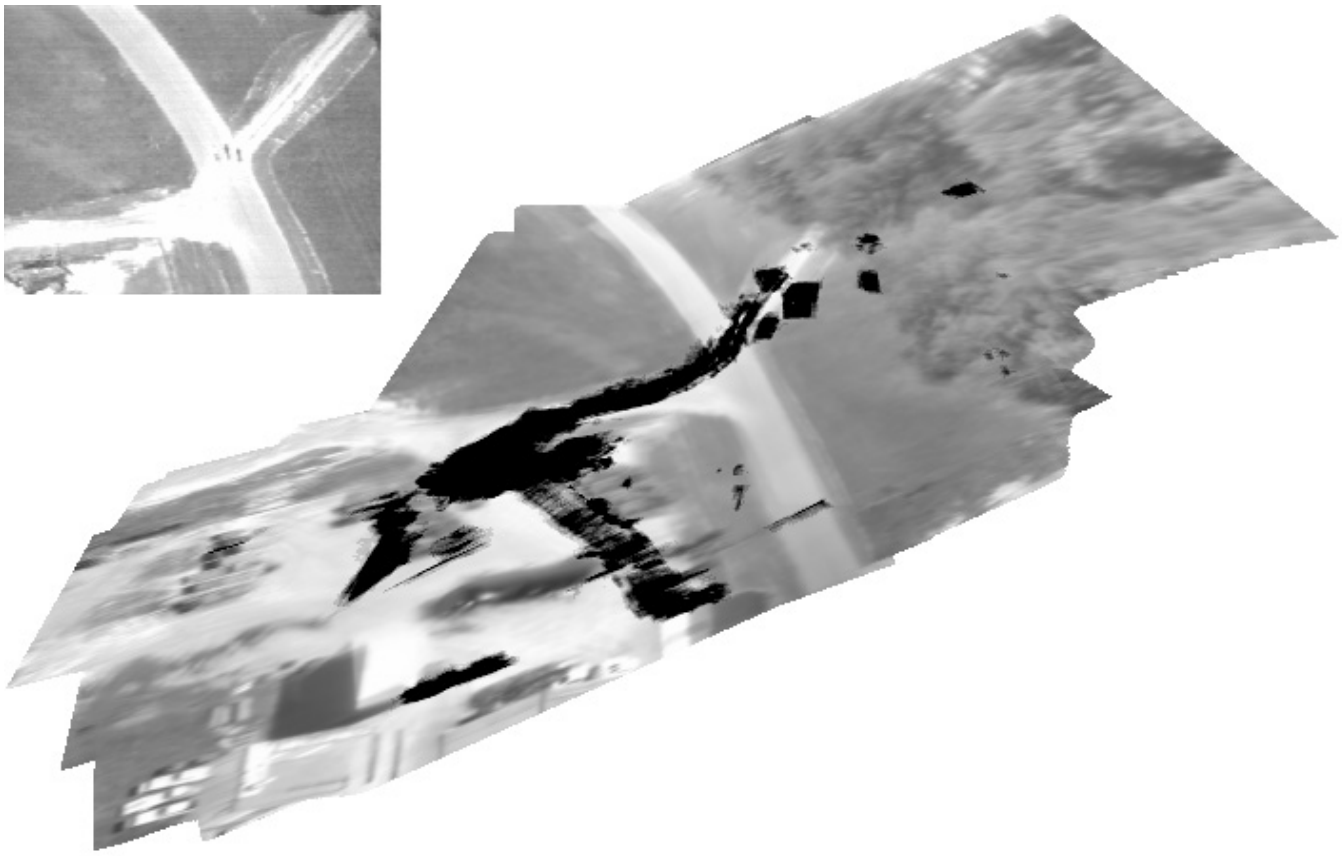

Figure 5: One input video frame and independent motion tracks overlaid onto the background mosaic. Also marked are three poles and regions of the trees which violate the planarity assumption.