Person Tracking in UAV Video

Andrew Miller, Pavel Babenko, Min Hu, and Mubarak Shah

Computer Vision Lab at University of Central Florida

1 Introduction

The UAV person tracking task for this evaluation was particularly difficult because of large, complicated, and low-quality videos, with only small images of people. We found that our best results were obtained using a combination of intensity thresholding (for IR imagery), motion compensation, interest-point detection and correspondence, and pattern classification. This can be considered a preliminary exploration into an extremely challenging problem.

2 Previous Approaches

Our previous efforts into object detection and tracking in aerial video led to the development of the COCOA system[1], which combines a multi-stage process of detection, tracking, and classification algorithms[2], with a graphical user interface for adjusting parameters.

In the first stage of processing, COCOA performs batch registration of every adjacent pair of frames, using a combination of telemetry information from the aerial vehicle, quick feature correspondence, or dense gradient information. An iterative process is used to maximize the stability of the registration, so that the aligned images can form a seamless mosaic [3].

The second stage of processing involves foreground segmentation, either with Mixture-of-Gaussian (MoG) background modeling (Figure 1), or consecutive frame differencing over a short sliding. In both cases, the information from the registration step is used to align pairs of adjacent frames so the motion of the camera does not affect the segmentation.

Finally objects are tracked from one frame to the next by establishing a correspondence based on similar location and color histogram appearance.

Since the registration stage assumes a planar projective transformation between frames, the algorithm works best when the altitude of the camera is very high, or when the scene is nearly planar. If the scene has lots of three-dimensional artifacts like trees, water towers, or tall buildings, these will show up as false foreground objects because of parallax motion, even if the registration is accurate on the ground plane. Since many of the scenes in CLEAR's dataset had several such artifacts, foreground segmentation produces lots of noisy patches.

An additional problem is the small images of people in most scenes. COCOA has proven to be fairly robust to this sort of noise in past evaluations involving vehicles because vehicles are very fast moving and large compared to

R. Stiefelhagen et al. (Eds.): CLEAR 2007 and RT 2007, LNCS 4625, pp. 215–220, 2008.

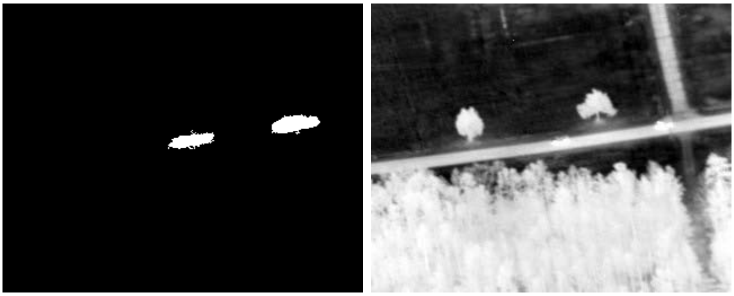

Fig. 1. Background modeling is an effective method of detecting vehicles Since the vehicles are large and fast-moving, they appear in the subtraction image as large clearly separated segments.

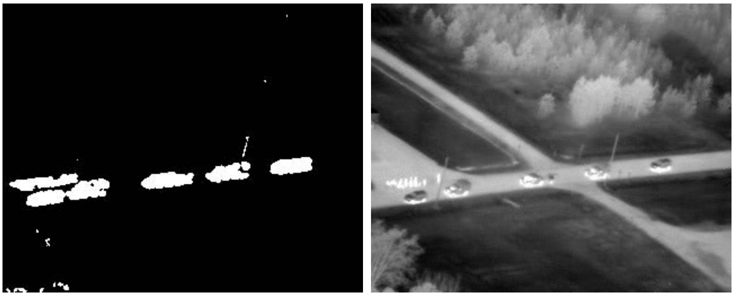

Fig. 2. Background modeling does not work as well for detecting people because the people appear as very small regions that are indistinguishable from clutter noise

the background noise. People are small and slow-moving enough that they are indistinguishable from clutter in the segmentation image.

3 Improvements in Static Detection

Rather than rely on foreground segmentation, which proved ineffective, we used a combination feature-point detection and correspondence in each static image with target recognition techniques.

3.1 Harris Corner Tracking

Harris corners provide fairly reliable features for tracking. We used greedy correspondence matching, based only on spatial information, similar to the more general approach used in [4]. The X and Y coordinates of each corner in one frame are compared to the coordinates in the next frame. The pair of points with the smallest distance between them are linked to each other by assigning the Object ID of the point in the previous frame to the point in the new frame.

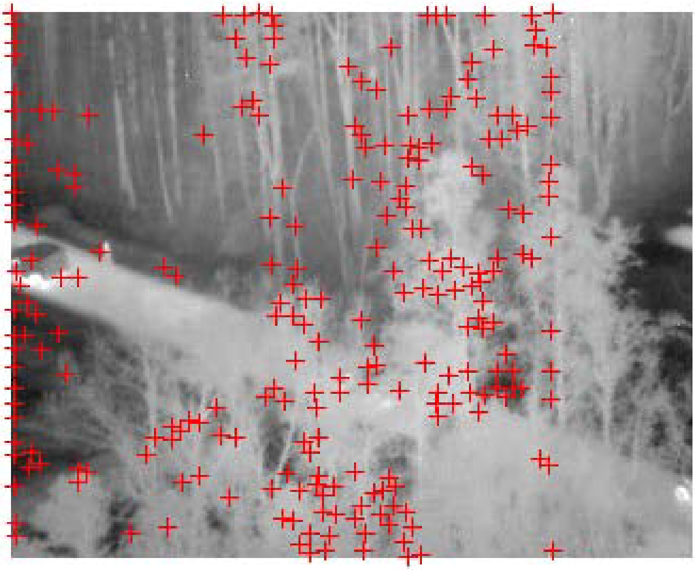

Fig. 3. Around 200 Harris corners are typically detected per frame. If a person is present in the frame, then at there will be at least one corresponding Harris corner. In this example, a person is running along the road in the left of the frame.

Then this process is repeated for the remaining unmatched points. When there are no pairs of points with a distance between them less than some threshold, the points in the new frame are given a new ID, and the points in the previous frame are forgotten.

Since the homography of the ground plane is available from the registration step in COCOA, we can project all of the interest points in the previous frame (which are assumed to be on the ground plane) onto the ground plane of the next frame. This allows for correspondence matching to be accurate even when the camera undergoes significant motion between frames.

In general, there are between 100 and 150 detected Harris corners in each frame, as in Figure 3. For each visible person, at least one corner corresponds to a spot on the edge of the person, usually the head or shoulder. Thus the remaining challenge is to suppress the false positives in a later step.

3.2 OT-MACH Filter Classification

After obtaining trajectories from Harris corner correspondence, we use an OT-MACH filter (Optimal Trade-Off Maximum Average Correlation Height) to determine whether a track is a person or something else - like a car, or just noise, either of which we should discard. An OT-MACH filter essentially generates a single image template using several training examples of the target, which can be correlated with the image to detect the target.

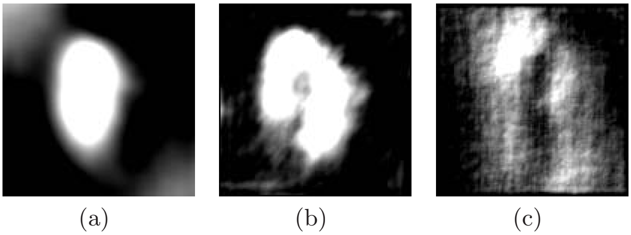

Fig. 4. OT-MACH Filters generated for: (a) zoomed out IR Intensity, (b) zoomed out IR gradient magnitude, and (c) close up EO gradient magnitude

One filter was generated from a training set of chips of people in all sizes by resizing the images to a common scale. Additional filters were generated by rescaling the original filter to 80% and 120% of the original size. We also used an Edge-Enhanced OT-MACH filter [5] to improve performance in color imagery.

In each frame, the object is compared to the filter to determine whether or not it is a person or not. For each frame when it is classified as a person, a counter associated with the object is incremented. If the counter is greater than a percentage of at least 20% of the total number of frames, then the object is considered a person for its entire duration.

To better localize the tracked person, the filter is correlated with the image in a small spatial window around the feature point, and the location with the highest correlation value is taken as the object's final estimated position.

3.3 Intensity and Gradient Histogram

Another way of removing false positives is to compare the the distribution of intensity values and gradient magnitudes of a chip to a generated model. Rather than compare the entire distributions, we can compare only the highest value in the chip as a token for the whole. The histograms for the maximum intensity and gradient magnitude are shown in Figure 5. The intensity histogram is only used for IR imagery, since people in IR will appear very bright, but can be of nearly any color in EO.

These probabilities are thresholded for each frame, and the number of 'passing' frames over an entire trajectory is used to discard more false positives.

4 Results

Although there have been difficulties in obtaining a quantitative metric for the performance of our algorithm, our own subjective analysis of our output indicates that there are still very many false positives and some missed detections.

The detection and tracking works fairly consistently, but the classification is still quite poor. The views of people are too small for filters to be meaningful when they are scaled down to an appropriate size.

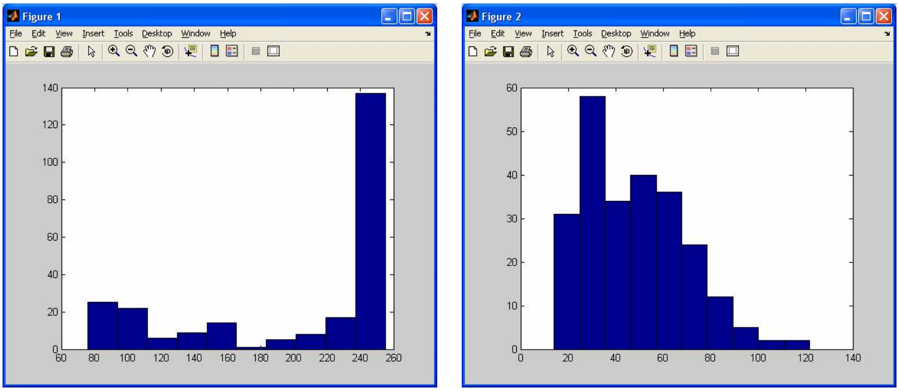

Fig. 5. Intensity (left) and gradient magnitude(right) histograms. People appear very bright in IR imagery, thus the intensity histogram is dominated by the upper bin. The gradient magnitude histogram indicates that most chips of people have a maximum gradient magnitude between 20 and 70.

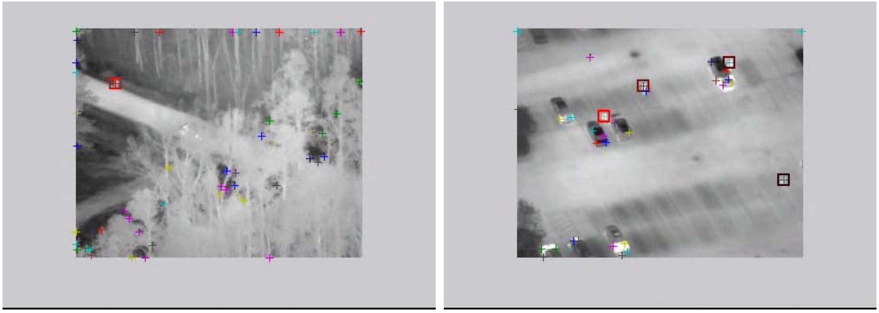

Fig. 6. Sample results of combining Harris corner detection and tracking with OT-MACH filter classification. A box is drawn around objects with a track-length of more than 10 frames, and at least 20% classification score. The brightness of the box indicates the classification score beyond 20%.

5 Conclusion

Our contribution to this task is the two-part framework of performing correspondence tracking on feature points with lots of positives and then using a combination of classification techniques to suppress the false positives. This approach may prove to be robust in ambiguous scenarios.

Ultimately it seems that the dataset is too difficult to make significant progress in tracking. There were a relatively small number of videos compared to the number of varying conditions. In some cases the camera changed zooms drastically, out of focus, fully occluded, or covered in rain droplets.

As a testament to the difficulty of the dataset, we sat a human volunteer down to watch all of the videos once and asked them to simply point out any people they saw. Later, a larger group watched the videos several times and debated whether or not certain artifacts looked enough like people. The human volunteer only noticed people in 8 of the 30 videos, while the larger group agreed on people in 25 of the videos.

If the dataset is improved and extended in future evaluations, we think it will greatly improve researcher's abilities to make meaningful advances in tracking technology.

References

1. Ali, S., Shah, M.: Cocoa - tracking in aerial imagery. In: SPIE Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications, Orlando (2006)
2. Javed, O., Shah, M.: Tracking and object classification for automated surveillance. In: The Seventh European Conference on Computer Vision, Denmark (2002)
3. Sheikh, Y., Zhai, Y., Shah, M.: An accumulative framework for alignment of an image sequence. In: Proceedings of Asian Conference on Computer Vision (2004)
4. Shafique, K., Shah, M.: A noniterative greedy algorithm for multiframe point correspondence. In: IEEE Trans. Pattern Anal. Mach. Intell. (2005)
5. Ahmed, J., Jafri, M.N., Shah, M., Akbar, M.: Real-time edge-enhanced dynamic correlation and predictive open-loop car-following control for robust tracking. Machine Vision and Applications Journal, Manuscript submission ID MVA-May-06-0110 (accepted, 2007)