Recognizing Humans in Motion: Trajectory-based Aerial Video Analysis

Yumi Iwashita[1]
yumi@ieee.org
M. S. Ryoo[1]
mryoo@jpl.nasa.gov
Thomas J. Fuchs
Thomas.Fuchs@jpl.nasa.gov
Curtis Padgett
curtis.w.padgett@jpl.nasa.gov

Jet Propulsion Laboratory,
California Institute of Technology
Pasadena, CA, USA

Abstract

We propose a novel method for recognizing people in aerial surveillance videos. Aerial surveillance images cover a wide area at low resolution. In order to detect objects (e.g., pedestrians) from such videos, conventional methods either utilize appearance information from raw videos or extract blob information from background subtraction results. However, people seen in low resolution images have less appearance information, and hence are very difficulty to classify based on their appearance or blob size. In addition, due to heavy camera movements caused by aerial vehicle ego-motion and wind, the system is expected to generate many noisy false detections including parallax. The idea presented in this paper is to detect and classify objects from aerial videos based on their motion: we analyze a trajectory of each object candidate, deciding whether it is a person-of-interest or simple noise based on how it moved. After objects are tracked by a Kalman filter-based tracking, we represent their motion as multi-scale histograms of 'orientation changes', which efficiently captures movements displayed by objects. Random forest classifiers are applied to our new representation to make the decision. The experimental results illustrate that our approach recognizes objects-of-interest (i.e., humans) even when there exist a large number of false detection/tracking, and it does it more reliably compared to the approaches with previous paradigm.

1 Introduction

Aerial video surveillance has been providing new opportunities to monitor activities in a large area. As the volume of aerial surveillance data grows, automatic scene analysis is becoming increasingly critical. An initial step for such a system is the detection and tracking of moving objects such as people and vehicles.

[1] These authors contributed equally to the paper.

[2] Yumi Iwashita is currently at Kyushu University in Japan. This research was conducted while she was a visiting researcher at JPL.

© 2013. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

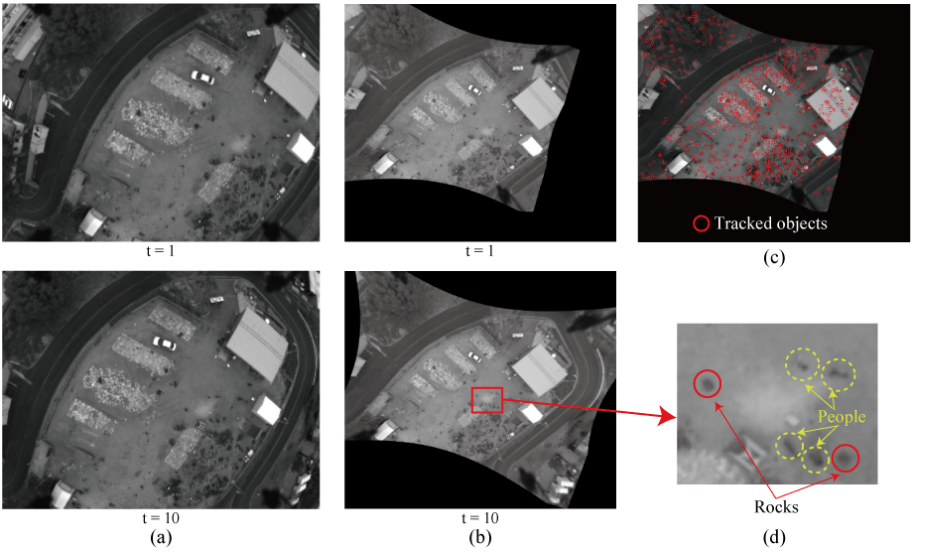


Figure 1: (a) Examples of captured aerial images, (b) rectified and stabilized images, (c) tracked objects (red circles), (d) an enlarged area including rocks and people whose appearances are similar.

In general, the technique of tracking objects consists of the following three steps: (i) video images, which are captured by an airplane / unmanned aerial vehicle (UAV), are stabilized at the ground plane, (ii) object areas are extracted by background subtraction techniques or appearance-based detectors, and (iii) detected objects are tracked in sequential images by several approaches, such as Kalman filters, particle filters, a graph-based model etc [10, 11, 12, 18]. Due to the ego-motion of the camera on the platform as shown in Fig. 1 (a), which is taken by a UAV, a global image motion is induced. Thus, the image stabilization is necessary in the first step (Fig. 1 (b)). The second step seeks to determine changes induced by motions on the stabilized scene. These changes stem from not only moving objects in the scene, but also parallax induced motion and stabilization error. Thus tracking results generally include a lot of false detections as shown in Fig. 1 (c).

To increase the accuracy of the object tracking in aerial images, several methods utilized object classification techniques, such as observed object size, histogram of gradient (HOG) and Haar-like features [9, 16]. Objects classified by these conventional methods were assumed to have enough image resolution. However, in case that the number of pixels on the object is small due to low image resolution, it is hard to obtain appearance information of the object. Figure 1 (d) shows an example magnified area of an aerial image, and yellow dotted circles show people and the others are rocks, whose appearance is almost the same with the person. Thus, a new algorithm is required to address this challenge to classify objects.

1.1 Approach overview

This paper proposes a novel method to detect/recognize people by classifying object candidates in low resolution aerial images. The idea of the proposed method to classify moving objects is based on their motion, not their appearance. After objects are tracked by a

Kalman filter-based tracking, we represent their motion as new features, named multi-scale histograms of orientation changes, which efficiently summarizes movements of the object candidates. The contributions of our approach are (i) the introduction of a new concept that objects can be better recognized using their motion information particularly in aerial videos and (ii) our novel feature representation to capture object motion effectively and efficiently.

We evaluate our method with aerial images such as Fig. 1 (a), which include buildings, cars and people. Objects with enough number of pixels, such as cars, can be distinguished by appearance information, however it is hard to classify people due to the small number of pixels and lack of appearance information. We will show our proposed method can classify people.

2 Previous works

Extensive work has been done on object tracking with potential application to aerial data. Reilly *et al*. detected moving objects by median background model, and objects are tracked using bipartite graph matching with the combination of road orientation and object context [12]. Xiao *et al*. proposed a joint probabilistic relation graph approach to track a large number of vehicles [17]. This method utilized vehicle behavior model from road structure to detect and track in wide area. Keck *et al*. proposed a real-time system for detecting and tracking moving objects from aerial images [7]. These papers focused on vehicle tracking. Vehicles are detected from tracked objects using road structure model, tracking assumption where distance traveled by targets are relatively long, etc. However, there are tracking errors due to tracks which do not follows the assumption and false detections due to parallax and registration errors.

To increase the accuracy of the detection of objects in aerial images, several methods utilized object classification techniques have been proposed. Xiao *et al*. proposed a car and people classifier based on image histogram of gradient (HOG) [16]. Leitloff *et al*. proposed a method to detect cars by adapting boosting in combination with Haar-like features [8], and Schmidt *et al*. proposed a slightly similar method with [8] to detect people based on Haar-like features [13]. Teutsch *et al*. extracted appearance features, such as moments and local binary pattern (LBP), with a 9-NN classifier [14]. In these methods the image resolution of object area is large enough to obtain appearance information, so objects were detected successfully.

Aerial surveillance images cover a wide area at low resolution. Thus it is hard to track small objects, such as pedestrians, due to low contrast and small number of pixels on the subject. Mattyus *et al*. proposed a method to classify tracks using the size of the objects [9]. This method doe not use the appearance information, thus it can be applied to low resolution images. However, this method cannot distinguish objects and people whose shape is close.

3 Tracking and feature extraction

This section describes a method for object tracking in aerial images and feature extraction from tracks. In our method, we assume that the processing for image stabilization was done in advance. First, we explain a method to extract objects by a background subtraction and track objects by the Kalman filter. Then, we introduce a method to extract features from their tracks.

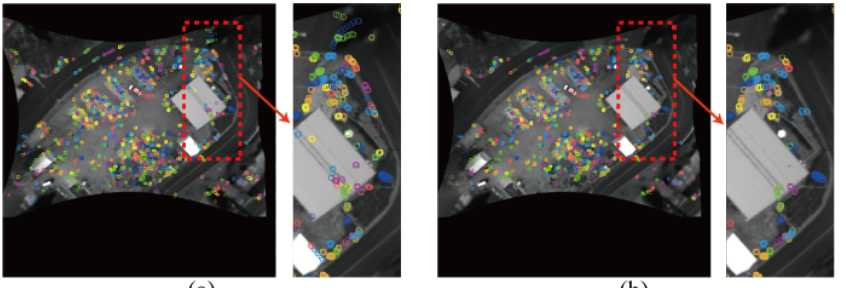


Figure 2: (a) An example of results of object tracking, (b) results after filtering tracks based on their length. Most tracks are originated by noise, but the filtering based on track durations was insufficient to remove them.

3.1 Object extraction and tracking

We extract object areas by a background subtraction where the background at each pixel is modeled using a mixture of Gaussian.

Tracking objects is performed using the Kalman filter (KF) [4]. Each time a new observation is received, it is associated to the correct track among the set of the existing tracks. If it is a new object, a new track is created. To associate with the new observation and the correct track, a distance between the position of the new observation and a predicted position of each existing track is computed. If the calculated distance is less than a velocity covariance of the track, these two are considered to be correlated. When a track is not correlated with new observations, the track is deactivated. Figure 2 (a) shows results of object tracking, and we are able to observe that there exist a large number of detections: these detections not only include objects such as humans, but also include falsely detected areas due to parallax and stabilization error. Figure 2 (b) shows results after filtering tracks based on their duration (i.e., discarding tracks whose lengths are too short), but there still exist many false tracks.

3.2 Feature extraction

After tracking object candidates, we represent their motion information using our histogram-based features. This subsection presents our new motion feature representation named histogram of orientation changes, which essentially is a set of multi-temporal-scale histograms concatenating trajectory orientation changes. The idea is that tracks originated by humans participating in activities will contain movements completely different from tracks generated by noisy false object detections which is likely to move with random orientation changes. In addition, we present an extended version of our histogram representation, which captures movement magnitudes as well as their orientation changes. This version constructs our multi-scale orientation histograms for each magnitude range, representing orientation changes when the object is showing fast motion vs. when showing slow motion.

3.2.1 Histogram of orientation changes

The first histogram is based on relative orientation changes between every pair of vectors, which are defined using three consecutive observed points $p_{\Delta f}(n-\Delta f)$, $p_{\Delta f}(n)$, $p_{\Delta f}(n+\Delta f)$ with a constant frame difference Δf. Δf essentially is a time-scale describing how much

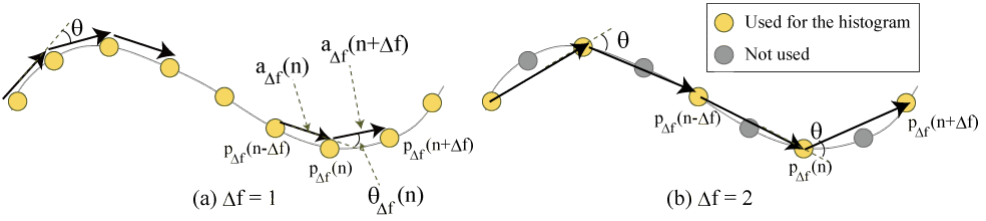


Figure 3: Examples of orientation changes (Δ f=1 and 2).

details (i.e., detailed motion) from the trajectory the representation will consider. N is the total number of observed points of a track and $1 \leq n < N$. We define vectors $a_{\Delta f}(n+\Delta f)$ and $a_{\Delta f}(n)$ as $a_{\Delta f}(n+\Delta f) = p_{\Delta f}(n+\Delta f) - p_{\Delta f}(n)$ and $a_{\Delta f}(n) = p_{\Delta f}(n) - p_{\Delta f}(n-\Delta f)$ (Fig. 3). The relative orientation change $\theta_{\Delta f}(n)$ between $a_{\Delta f}(n+\Delta f)$ and $a_{\Delta f}(n)$ is defined as

$$\theta_{\Delta f}(n) = atan\frac{(a_{\Delta f}(n+\Delta f))_y}{(a_{\Delta f}(n+\Delta f))_x} - atan\frac{(a_{\Delta f}(n))_y}{(a_{\Delta f}(n))_x} \tag{1}$$

If $\theta_{\Delta f}(n) < 0$, we add 2π to $\theta_{\Delta f}(n)$. Δf has different variations of frame difference so that the orientation can show not only local information but also semiglobal information of the track. In this paper we set Δf as multiples of 2 $(1, 2, 4, ..., F)$. The histograms of a track T is $H_1(T) = [h_1(T), ..., h_{\Delta f}(T), ..., h_F(T)]$. A value of the wth histogram bin of each histogram $h_{\Delta f}(T)$ is

$$h_{\Delta f}(T)[w] = \sum_n \eta_{\Delta f}(n), \tag{2}$$

where

$$\eta_{\Delta f}(n) = \begin{cases} 1 & (w-1)\Delta\theta \leq \theta_{\Delta f}(n) < w\Delta\theta \\ 0 & otherwise \end{cases} \tag{3}$$

The total number of bins is W $(1 \leq w \leq W)$, and $\Delta\theta = 360/W$.

Figure 4 shows examples of calculated histograms for a person, a car, and stabilization error $(\Delta f = 1, 2, 4)$. In this case $W = 12$ $(\Delta\theta = 30$ [degree]), and the magnitude of each bin shows $h_{\Delta f}[w]$. Since a person tends to walk straight, the histogram value is bigger in bins close to 0 degree compared with the rest of the angles. Moreover, histograms of the person at different frame difference show similar tendency since he / she does not move randomly. Histograms of the car also have the similar tendency but slightly different. In case of the stabilization error, the histogram value is distributed and histograms at different frame are different due to its random motion.

3.2.2 Histogram of orientation changes for multiple magnitudes

The above histogram does not include magnitude information of each vector $a_{\Delta f}(n)$, thus we introduce one more parameter, magnitude $g(m)$, to the histogram H_1. $g(m)$ shows the number of pixels, and it is used to categorize the histogram H_1 based on the magnitude information. m is $1 \leq m < M$, and M is the total number of magnitude categories. The new histograms of the track T is $H_2(T) = [H_1^1(T), H_1^2(T), H_1^m(T), ..., H_1^M(T)]$, where $H_1^m(T)$ is a histogram at a magnitude range m. $H_1^m(T)$ is defined as $[h_1^m(T), ..., h_{\Delta f}^m(T), ..., h_F^m(T)]$. A value of the wth histogram bin of each histogram $h_{\Delta f}^m(T)$ is

$$h_{\Delta f}^m(T)[w] = \sum_n \eta_{\Delta f}^m(n), \tag{4}$$

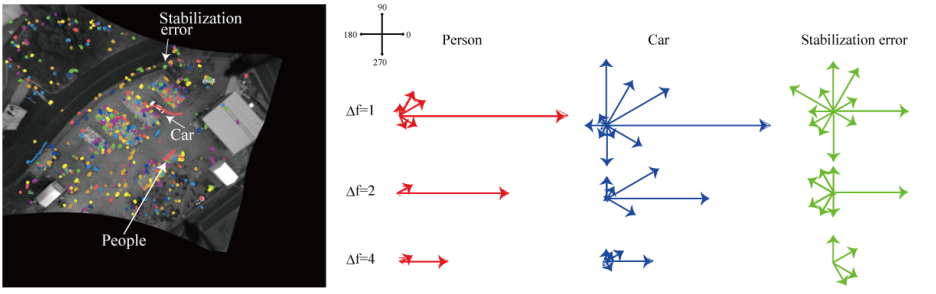


Figure 4: Examples of histogram of orientation changes (Δ f=1, 2, 4).

where

$$\eta^m_{\Delta f}(n) = \begin{cases} 1 & (w-1)\Delta\theta \le \theta(n,\Delta f) < w\Delta\theta \\ & \wedge\ g(m-1) \le |a_{\Delta f}(n+\Delta f, n)| < g(m) \\ 0 & otherwise \end{cases} \tag{5}$$

3.2.3 Segmenting a trajectory into tracklets

Instead of constructing our histogram representation for each trajectory, we segment each trajectory into a set of multiple tracklets (with a fixed temporal duration such as 4 seconds) and build their representations. The idea is to capture salient moment the trajectory is showing human-like motion, and take advantage of it to make the overall recognition of what that trajectory is. That is, trajectories tend to be long and may contain multiple behaviors since aerial images cover wide area, and normalizing all such behaviors will make the system lose a great amount of information. For example a person, who keeps standing for a while, may start walking. Features (orientation changes) obtained while the person is standing will be similar to features of different objects such as false object detection due to stabilization error, and features captured during walking will have unique human-like characteristics. Therefore, in order to capture multiple aspects of the observed trajectories, we separate the track into tracklets (each of them has K frames). Both histograms H_1 and H_2 are calculated from all tracklets.

4 Classification

We use a decision forest classifier [1, 2, 3, 5] based on the features describe in Section 3 to distinguish between trajectories of pedestrians and the ones introduced by noise, camera parallax, and/or other moving objects (e.g., cars).

The setting is challenging for classification due to the fact that (i) the distribution of class labels is very unbalanced comprising 100-fold more tracklets from noise than from pedestrians, (ii) the classes themselves are extremely heterogeneous due to the fact that pedestrians behave in various ways and noise is induced by a plethora of sources and (iii) the histogram-based feature representation is very sparse comprising a large number of non-informative covariates. To overcome these challenges we employ a decision forest, which is an ensemble classifier consisting of randomized decision trees that are induced from bootstraps of the training data. In the last decade decision forest have proven their performance for computer vision in a vast array of applications ranging from medical imaging [6] to space exploration

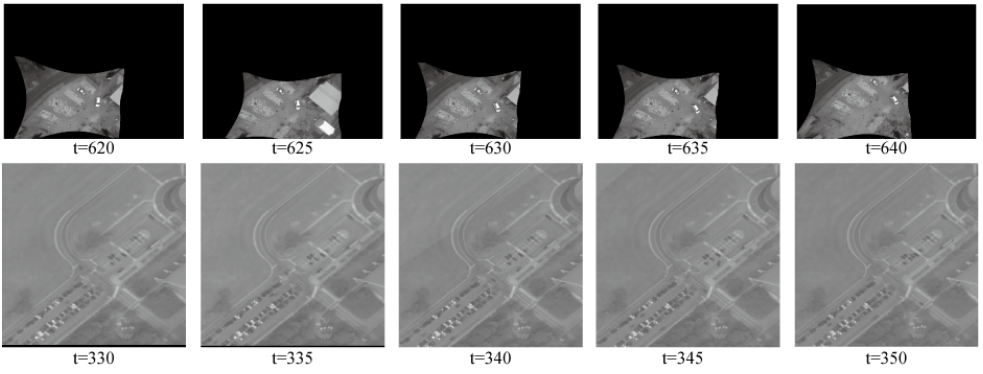


Figure 5: Example snapshots from the two aerial video datasets we are using for experiments. The upper figures are from the quadrotor dataset, and the lower figures are from the AngelFire dataset.

[15]. In this application, the non-linearity of the classifier allows us to handle the intra-class variability while the randomization allows for robust handling of the large number of noise dimensions. Overall these properties make decision trees a good choice for this task.

For every feature set reported in Section 5 we learned a decision forest model [2] from training data, each consisting of 200 trees. In addition to out-of-bag data, evaluation was performed on an external test set and reported in Figure 6. Classification was performed per tracklet, which were then fused into trajectories by taking the maximum confidence estimate of the forest for the tracklets comprising a trajectory. Finally the trajectories were weighted by the number of frames they span to prevent punishing long continuous trajectories compared to short ones.

5 Experiments

In this section, we implement and evaluate our trajectory-based human recognition methodology, while comparing it with other classification works.

5.1 Dataset

In order to evaluate the accuracy of our trajectory classification using motion representation, we constructed a new video dataset composed of videos captured using an aerial vehicle. We mounted a ground-looking camera on a small aerial vehicle (a quadrotor [AscTec Pelican]), and obtained its videos while flying it $30 \sim 40$ meters above ground. Their image resolution was 640×480, 7.5 fps, and the size of each human was smaller than 10 pixels by 10 pixels. A total number of frames was approximately 10000 (i.e., $\sim$22 minutes). Due to heavy wind, the quadrotor was shaking very frequently. A homography-based stabilization algorithm was applied to the raw videos.

In this dataset, more than 10 actors were asked to move on a ground while the quadrotor was recording the videos, by performing typical activities observed in aerial videos including human-vehicle interactions such as 'a group of people gathering at one location, waits for the vehicle, and gets into the vehicle as it arrives' and human object interactions such

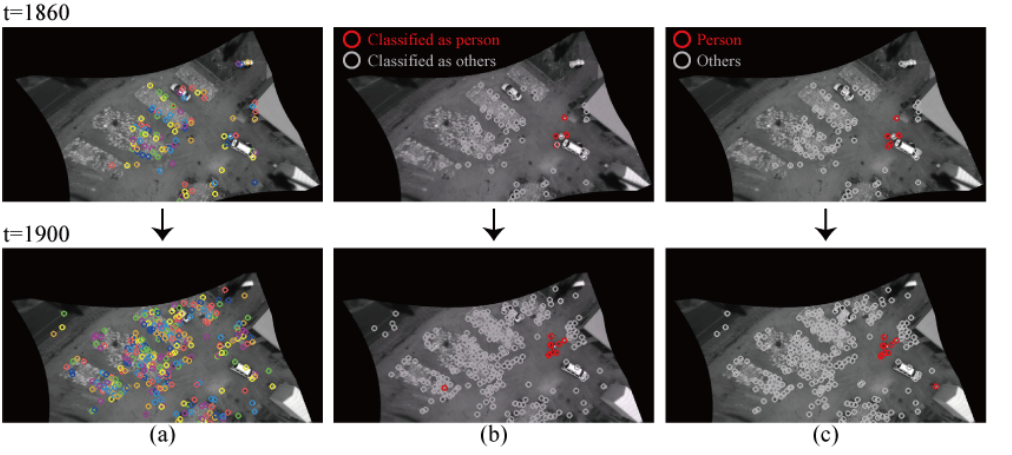


Figure 6: (a) Example results of object tracking (time t=1860 and 1900), (b) classification results by the proposed method, and (c) ground truth. The video contains the activity of 'a group of person unloading an object from the vehicle', and we are able to observe that our approach correctly recognizes human tracks as opposed to other noise.

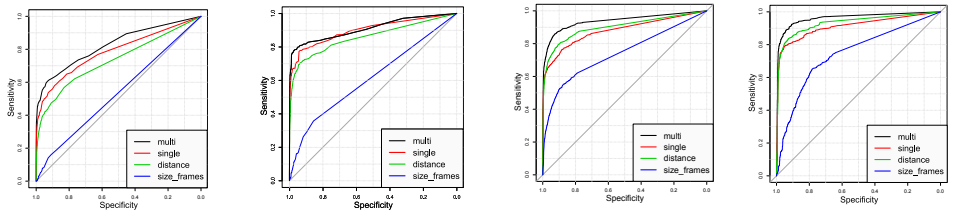
as 'people carrying boxes'. Trajectories of humans were estimated using our standard detection/tracking algorithm, and their ground truth labels were provided (i.e., whether the trajectory is originated by the actors-of-interest).

In addition, we used another dataset composed of videos captured using an airplane: the AngelFire dataset. Their image resolution was 2070×1556. and the size of each human was smaller than 10 pixels by 10 pixels. A total number of frames was approximately 900 with very low frequency (e.g., 2 fps). A homography-based stabilization algorithm was applied to the raw videos. In this dataset, more than 80 people walked in a parking lot while the airplane was recording the videos. Trajectories of humans were estimated using our standard detection/tracking algorithm, and their ground truth labels were provided. Figure 5 shows examples.

5.2 Implementation

We implemented four types of trajectory classification approaches. First, we implemented (1) our approach using our basic motion histogram ignoring movement magnitudes (Section 3.1) and (2) the multi-scale version of our approach considering different magnitudes (Section 3.2).

In addition to our approaches, we implemented two baselines: (3) a classifier filtering out each noisy track based on 'the blob size of the object and the number of frames the object was being tracked', and (4) an approach recognizing noise tracks based on the amount of per-frame movement (i.e., movement magnitude). The first baseline can be viewed as a filtering approach which has been commonly used in conventional systems [2]. The second baseline construct a histogram representation counting the number of frames the tracked object was showing large (or low) motion. That is, it counts the number of frames the track was showing unrealistic jumps. Both these baselines use an identical classifier to our approach (i.e., random forest).



(a) Quadrotor + tracklet (b) Quadrotor + per-frame (c) AngelFire + tracklet (d) AngelFire + per-frame

Figure 7: ROC curves of our approaches and baselines, tested with two different datasets and two different settings. Multi-magnitude version of our approach (black) performed superior to all the others, including our approach without multi-magnitude (red), the baseline only using movement magnitude (green), and the baseline using blob size and track duration (blue).

5.3 Evaluation

We evaluated performance of the implemented approaches in terms of true positive rates (i.e., $tp/(tp+fn)$, also called 'Sensitivity') and false positive rates (i.e., $fp/(fp+tn)$, also described as 1 - 'Specificity'). More specifically, we plotted a ROC curve of each trajectory classification approach, which describes how true positive rate and false positive rate changes as the decision boundary (i.e., probability threshold) changes. Figures 6 (a) $\sim$ (c) show examples of tracking results, classification results by the proposed method, and their ground truths.

In order to train/test the classifiers, we divided an entire dataset into two subsets: one for training and the other for testing. That is, half of our videos were used for the training and the other half was used for the testing, without any overlap between them. The mean accuracy was obtained by repeating this trianing-testing splits multiple rounds.

The approaches were evaluated using two different settings. First, we measured ROC curves by evaluating tracklet-level decisions, without integrating their decisions. This setting shows how well our tracklet-based classifications are made, which serve as basis for the final trajectory-level decisions. The other setting measures per-frame decision accuracy after integrating tracklet decisions into trajectory-level decisions. That is, this setting tests how many humans at each frame were correctly labeled as 'humans', once tracklet-level decisions are integrated into trajectory-level decisions.

Figure 7 shows ROC curves of the approaches. We are able to clearly observe that the classification using our methods (i.e., our histogram-based motion representation) performs superior to the other approaches. Particularly, our approaches performed very superior to the previous conventional approach of using object blob size and the number of frames. This confirms the advantages of our approach considering detailed trajectory motion by constructing their histogram-based representations over the previous simple filtering. Compared to baseline approaches, our approach obtained much higher true positive rates at a low false-positive range, which is particularly important for applications in practice.

The approaches showed better performance for the AngelFire dataset compared to our quadrotor dataset in general. This is due to the fact that the AngelFire dataset was captured using a larger aerial vehicle flying at much higher altitude with less ego-motion. The quadrotor dataset was more challenging since its heavy motion due to wind caused many false tracks, and our approach was able to successfully overcome such problems.

6 Conclusion

We proposed a new approach of recognizing persons based on their motion from aerial videos. We introduced our new feature representation to capture motion information of detected object candidates, and illustrated that our feature representation in conjunction with random forest classifiers enables much more reliable detection of moving persons. In contrast to appearance-based detection and classification, our approach was able to correctly estimate human locations from very low-resolution aerial videos.

Acknowledgment: The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

References

[1] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3):81–227, 2011.

[4] M. Djalalov, H. Nisar, Y. Salih, and A. Malik. An algorithm for vehicle detection and tracking. *International Conference on Intelligent and Advanced Systems*, pages 1–5, 2010.

[5] Thomas J. Fuchs and Joachim M. Buhmann. Inter-active learning of randomized tree ensembles for object detection. In *Computer Vision Workshops, IEEE 12th International Conference on Computer Vision*, pages 1370–1377, 2009.

[6] Thomas J. Fuchs and Joachim M. Buhmann. Computational pathology: Challenges and promises for tissue analysis. *Journal of Computerized Medical Imaging and Graphics*, 35(7):515–530, April 2011.

[7] M. Keck, L. Galup, and C. Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, 2013.

[8] J. Leitloff, S. Hinz, and U. Stilla. Vehicle detection in very high resolution. satellite images of city areas. *IEEE Transactions on Geoscience and Remote Sensing*, 48(7): 2795–2806, 2010.

[9] M. Mattyus, C. Benedek, and T. Sziranyi. Multi target tracking on aerial videos. *ISPRS Istanbul Workshop 2010 on Modeling of optical airborne and spaceborne Sensors*, 2010.

[10] E. Pollard, A. Plyer, B. Pannetier, F. Champagnat, and G. Besnerais. Gm-phd filters for multi-object tracking in uncalibrated aerial videos. *International Conference on Information Fusion*, pages 1171–1178, 2009.

[11] T. Pollard and M. Antone. Detecting and tracking all moving objects in wide-area aerial video. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 15–22, 2012.

[12] V. Reilly, H. Idrees, and M. Shah. Detection and tracking of large number of targets in wide area surveillance. *11th European Conference on Computer Vision*, pages 186–199, 2010.

[13] F. Schmidt and S. Hintz. A scheme for the detection and tracking of people tuned for aerial image sequences. *ISPRS conference on Photogrammetric image analysis*, 2011.

[14] M. Teutsch, W. Kruger, and N. Heinze. Detection and classification of moving objects from uavs with optical sensors. *SPIE, Signal Processing, Sensor Fusion, and Target Recognition XX*, 2011.

[15] David R. Thompson, William Abbey, Abigail Allwood, Dmitriy Bekker, Benjamin Bornstein, Nathalie A. Cabrol, Rebecca Castano, Tara Estlin, Thomas J. Fuchs, and Kiri L. Wagstaff. Smart cameras for remote science survey. In *Proceedings of the 10th International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)*, 2012.

[16] J. Xiao, H. Cheng, H. Feng, and C. Yang. Object tracking and classification in aerial videos. *SPIE 6967, Automatic Target Recognition XVIII*, 2008.

[17] J. Xiao, H. Cheng, H. Sawhney, and H. Feng. Vehicle detection and tracking in wide field-of-view aerial video. *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 679–684, 2010.

[18] Q. Yu, I. Cohen, G. Medioni, and B. Wu. Boosted markov chain monte carlo data association for multiple target detection and tracking. *International Conference on Pattern Recognition*, 2:675–678, 2006.