Focus-of-Attention for Human Activity Recognition from UAVs

G.J. Burghouts, A.W.M. van Eekeren, J. Dijk

TNO Intelligent Imaging, The Netherlands

ABSTRACT

This paper presents a system to extract metadata about human activities from full-motion video recorded from a UAV. The pipeline consists of these components: tracking, motion features, representation of the tracks in terms of their motion features, and classification of each track as one of the human activities of interest. We consider these activities: walk, run, throw, dig, wave. Our contribution is that we show how a robust system can be constructed for human activity recognition from UAVs, and that focus-of-attention is needed. We find that tracking and human detection are essential for robust human activity recognition from UAVs. Without tracking, the human activity recognition deteriorates. The combination of tracking and human detection is needed to focus the attention on the relevant tracks. The best performing system includes tracking, human detection and a per-track analysis of the five human activities. This system achieves an average accuracy of 93%. A graphical user interface is proposed to aid the operator or analyst during the task of retrieving the relevant parts of video that contain particular human activities. Our demo is available on YouTube.

Keywords: human activity recognition, tracking, human detection, motion-in-motion, airborne platforms, pattern recognition.

1. INTRODUCTION

UAVs are recording large amounts of full-motion video data [1]. Typically, this data is collected to gather actionable intelligence. Extracting valuable intelligence is a time-demanding task for the analysts. The reason for this, is that the UAV usually observes a wide area for a long time. In many cases, the relevant parts of the video data concern particular human activities and events of interest [2]. For the analyst, it is hard to retrieve such events from the large amounts of video data, as typically they are rare. In populated areas it is hard to find the few relevant events in the midst of all activities, while in desolated areas it is hard to keep focus. To support the operator and analysts during their tasks, we propose a fully automated system that is able to recognize and localize particular human activities in UAV video data. Recognized activities are stored as metadata and can be exploited in an online, causal system for intelligence gathering, or for offline forensic analysis.

The system needs to fulfill several functional requirements. The camera is moving, hence the system needs to cope with this ego-motion. Extracted features need to be invariant to this disturbing factor. Stabilization is essential, but it can never remove all the spurious details that result from 3D movement through a 3D world [3]. The system has to be able to ignore such inevitable artefacts. Finally, the system needs to know where the humans are, because its goal is to recognize and localize human activities. An example is given in Figure 1, where the camera motion is apparent. Realistic, imperfect tracks have to be analyzed by the system in order to recognize the activity performed by the human.

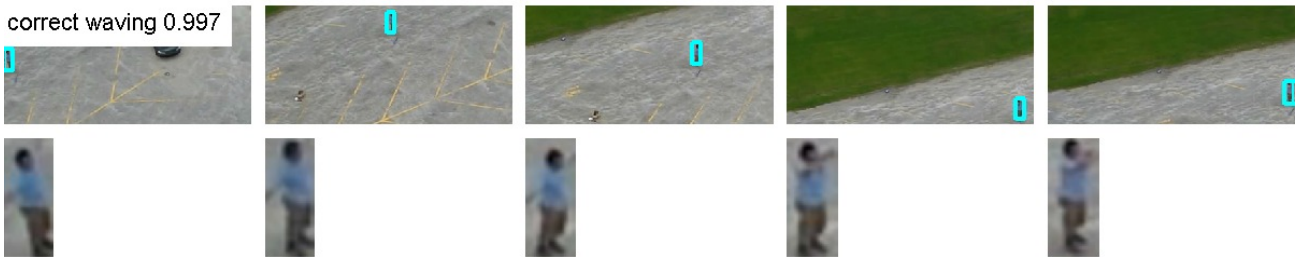

Figure 1. The capability addressed in this paper: recognition and localization of human activities in aerial video data. The tracking (blue boxes) and activity prediction (text box in the upper left) perform well, even though the camera motion is severe. The tracker keeps the person in focus and the activity is correctly interpreted as 'waving' with a high probability.

Electro-Optical and Infrared Systems: Technology and Applications XI, edited by David A. Huckridge, Reinhard Ebert, Proc. of SPIE Vol. 9249, 92490T · © 2014 SPIE · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2067569

To address these requirements, the system comprises several components. Stabilization is a pre-processing step. To capture the motion of the activities, robust and distinctive motion features are extracted from the video data. The objects in the scene are detected and tracked. The motion features are associated to the tracks, to know which object displays particular motion patterns. A human detector is deployed to distinguish humans from other objects. A model of human activities is learned from labeled examples of the activities of interest. These components are widely used for video analysis [4,5] and also for UAV video data [6]. The big unknown is how each of these components contributes to reliable human activity predictions in the challenging recording conditions. The contribution of this paper is three-fold: (A) we validate the merit of each component, (B) we add a focus-of-attention mechanism to perform a per-track analysis of the human activities, and, (C) we show how the algorithms can be combined in a robust system, resulting in a good human activity recognition capability for UAV video data. One of the interesting findings is that the focus-of-attention mechanism is key to robust predictions of human activities.

We demonstrate the robustness and discriminative power of our system on the publicly available UCF-ARG dataset [7]. Five activities have been annotated. Experiments show that the activities can be distinguished well, when the components are combined. Our demo is available on YouTube, which shows 16 UAV streams, where 15 persons are walking and 1 person is running. We show that it is hard for the human eye to recognize the running person [8]. Our method is able to instantly pinpoint the running person, by zooming in on this activity in order to highlight it for the human analyst.

The paper is organized as follows. Section 2 describes the related work. In Section 3 we summarize the system, its components and the various configurations. Section 4 describes the dataset, annotations and experimental setup. Section 5 discusses the results. Finally, in Section 6, we conclude with the main findings.

2. RELATED WORK

Systems have been proposed to analyze UAV video data in an automated manner and present relevant information to a user. Examples are military persistent wide-area aerial surveillance [1], mapping image sequences onto maps [10], tracking cars for traffic monitoring [9], long-term tracking of targets by a swarm [11], and detection of human activities in the video streams [3]. Before we discuss the novelty of this paper, we will first motivate which insights and components we will adopt from the research community.

2.1 Recent progress

One of the generic challenges for the analysis of UAV imagery, is stabilization of the video frames. Simple methods such as estimation of the homography by RANSAC [12] have proven to be effective if parallax effects are not severe. For complex scenes that generate complicating artefacts, advanced methods have been proposed [13]. The scenes that will be considered in this paper are not very complex and therefore a simple method suffices. Visual tracking has received a lot of attention by the computer vision community [14]. In this paper we are interested in the tracking of humans. These trackers have improved by exploiting better person detections by e.g. the FPDW detector [15]. We expect that a person detector will improve the tracking greatly in case of misalignments due to imperfect stabilization, especially when the detector can be run often, which is the case with a real-time detector such as [15]. Schemes to combine different type of detections and motion prediction have been proposed [14], which have increased the performance. The rationale is that each type of detector performs well on different cases, for instance standing people [15] or moving objects [16]. In this paper, some activities involve movement, where others are performed in one place. We expect that a combined tracker is needed for these activities. Representing the motion of the body has been extensively studied, ranging from motion templates [17] to localized motion features such as the spatio-temporal interest points (STIP) [18]. Motion templates are highly dependent on a precise alignment, which will not be the case if the stabilization is not perfect. Motion features such as STIP have proven to be very distinctive [19]. However, when the camera moves, a

recent alternative, Improved Trajectories [20], is to be preferred. They will be the feature of choice in this paper. Various activity recognition methods have been proposed, some based on a simple bag-of-words model [21], where others involve various sources such as tracking and group attributes [22]. Various features have been combined in activity recognition methods such as [23]. In this paper, human activities are considered which involve characteristic motion patterns (e.g. wave), shapes (e.g., dig), and movement (e.g., run). Therefore, we use motion features and adopt the approach as taken by [22] to associate motion features to tracks, and performing a track-based prediction as in [5,24] of the human activity that is being displayed.

2.2 Contributions of this paper

The novelty of this paper is that we show how the abovementioned algorithms can be combined in a robust system, resulting in a good human activity recognition capability for UAV video data. The common datasets for human activity recognition are about movies, sport, YouTube, surveillance [25]. Experiments on UAV videos are currently limited, to setting a baseline by the bag-of-words approach [26] and evaluating the performance of an advanced method to remove camera motion from the motion field [3]. To our knowledge, there is no earlier work in the area of UAV video analysis that analyzes a human activity recognition system by validating the merit of each component. Given the relevance of such a system for UAV operators and analysts, and given the increase of UAVs as the means for monitoring and gathering intelligence, such a validation is a very useful contribution to the research community and industry. In this paper, the goal is to establish how the recognition system should be designed and which components are needed. In Section 2.1 we learn that many components are available, but their merit is unclear. Their merit will be assessed in this paper. Further, there is no clear view yet on how a system should decide where and when a relevant activity happens. To provide such information, some form of spatio-temporal segmentation or localization is needed. In this paper, we contribute a focus-of-attention mechanism to perform a per-track analysis of the human activities. Finally, it is unclear how a recognition system may be deployed by a UAV operator or analyst. As a first step towards deployment, we have implemented a demo of our recognition system together with a mockup graphical user interface. In summary, there are four novelties in this paper: we validate the merit of each component; we add a focus-of-attention mechanism to perform a per-track analysis of the human activities; we show how the algorithms can be combined in a robust system; we provide a short demo showing how a human activity recognition capability may aid the analysis of UAV video data.

3. SYSTEM FOR HUMAN ACTIVITY RECOGNITION

This section describes our method to recognize human activities in UAV video data. A flowchart of this method is depicted in Figure . As a generic first step, motion features are computed. We distinguish various setups of the system. Motion features can be interpreted based on the whole image sequence (system #1). As an alternative, features are associated to each track. For systems #2 and #3, all features from all tracks in the image sequence are accumulated, respectively for all tracks from all objects (system #2) or only for tracks which are classified as human (system #3). Section 3.2 describes how tracks are classified as human vs. other objects. The alternative is to interpret each track individually, which is done by system #4 (all tracks from all objects) and system #5 (only tracks which are classified as human). Figure 2 shows the components of the system and how they are combined in the various setups #1 to #5, which are ordered by increasingly complexity. The system components are described in Sections 3.1 to 3.4.

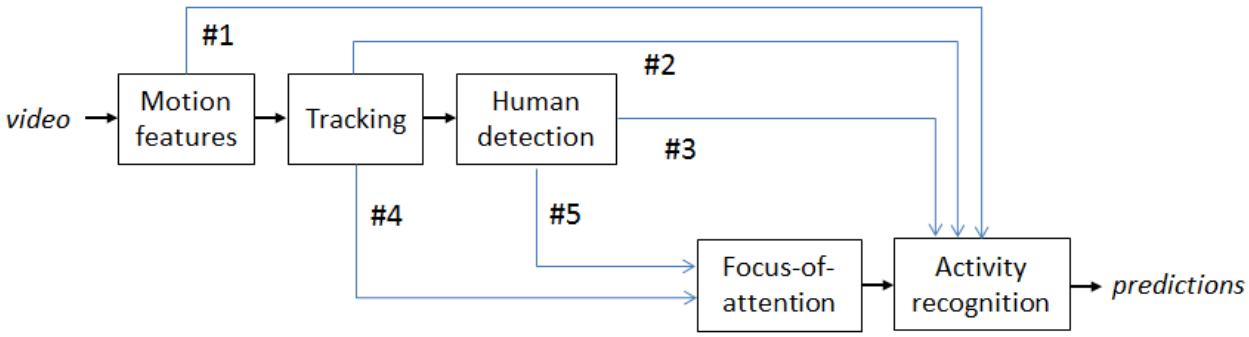

Figure 2. Outline of the system with its components and how they are combined to obtain five different system variations.

3.1 Motion features

As motion features, we adopt the state-of-the-art improved trajectories [20]. The camera moves severely. Improved trajectories are robust to this effect, because the camera motion is compensated for by means of frame-to-frame stabilization and removal of features that are due to ego-motion. The improved trajectories are represented by motion boundaries (on the contours of objects) and histograms of gradients (HOG) and optical flow (HOF) [18], which have proven very distinctive for human activity recognition, even under severe camera motion [20].

3.2 Tracking and human detection

For tracking we exploit a dedicated human detector, i.e., FPDW [15], and a generic moving object detector [16], to generate candidates. Where the human detector is very suitable to detect people who are standing straight up, the moving object detector is able to detect people when they are moving. Tracks are acquired by frame-to-frame matching of detections based on their appearance [26]. Details of our tracker can be found in [12]. Tracks that involve human detections, are designated as humans.

3.3 Focus-of-attention

We distinguish various setups of the method. Features can be interpreted based on the whole image sequence (system #1). Alternatively, features are associated to each track, after which every track is interpreted (systems #2 to #5). The procedure to associate features to a track is detailed in [22]. In system #2, all tracks including their features are accumulated as one feature representation for a given image sequence. In system #3, the same operation is performed yet only for tracks that are designated as human. In systems #1, #2 and #3 all features for a given image sequence are fed at once to the human activity recognition algorithm. Systems #4 and #5 interpret each track individually, see e.g., [5,24], based on its features. System #4 does so for each track, whereas system #5 only interprets tracks that are designated as human. To make the final interpretation of which human activity is observed for a given image sequence, the maximum posterior probability is taken from all tracks.

3.4 Activity recognition

The method for the recognition of human activities from motion features has been described in detail in [21]. We summarize it here. The model for each activity is obtained by a bag-of-features approach. The motion features are transformed into visual words. A set of features is represented as a frequency count of the words. The quantizer of choice is a random-forest [27]. It has a high discriminative power because it exploits the labels for each activity during the training phase. This way of quantization led to good performance in our recent experiments [23]. The final step is the SVM classifier with a $\chi 2$ kernel. The classifier serves as the detector for each activity. For each set of features (either from the complete image sequence or only for one track), we obtain a posterior probability for each of the activities. It is assigned the activity with the maximum posterior probability.

4. EXPERIMENTS

4.1 Videos and Human Activities

For the experiments, the UCF-ARG dataset [1] is selected, which has been recorded from an airborne camera, undergoing severe ego-motion. From this data, we have selected five human activities: ‘walk’, ‘run’, ‘dig’, ‘wave’, ‘throw’. This selection is interesting, because some activities have a characteristic movement and speed (walk and run). The remaining three activities do not involve movement; they occur in place. Their differences are subtle, as some observables are shared. Throwing involves the picking up of an item, which is similar to the motion patterns of digging. Throwing also involves arm motion, which has similarities to waving. For each of the five activities, there are 48 videos. The videos have a resolution of 960 x 540 pixels, recorded at 29 fps. The 48 videos for each activity are captured by recording 12 persons where each person performs the activity four times. The scene is not complex; it is a parking lot

with some cars on it, and it is surrounded by grass. Each video has been annotated manually, where the annotation indicates the person performing the activity. An example of such an annotation is shown in Figure 3.

Figure 3. An example of our annotation of the human activity 'digging' from the UCF-ARG dataset. Upper row: original frames showing the annotation (green boxes), lower row: zoom in on the manual annotation.

4.2 Tracks and Labels for Learning

The experiments are all performed with the automated tracker from Section 3.2. An example of the automated tracker is shown in Figure 4, for the activity 'digging' from the UCF-ARG dataset. The imperfect coverage of the tracker is a very common error. Our tracker is evaluated extensively in [12]. The average track duration is 3 seconds. The topic of this paper is human activity recognition in a fully automated system, so we use the tracks as-is. A track is designated to be human if at least one of its bounding boxes was classified as human by the human detector. To obtain the label for each track for the training of the model, i.e., which human activity is performed during the track, the track is intersected with the manual annotations from Section 4.1. In case of overlap, the track is labeled as the annotated activity and used as such for the training of the per-track interpretation of systems #4 and #5.

Figure 4. An example of the automated tracker, showing the boxes on the person for the activity 'digging' from the UCF-ARG dataset. Upper row: the track is displayed in the original frames (blue boxes), lower row: zoom in on the track.

4.3 Recognition and Cross-validation

There are 12 persons involved in the dataset, each performing all five activities four times, yielding 240 videos in total (see also Section 4.1). One prediction is produced per video, by assigning the human activity that has the maximum posterior probability (see also Section 3.4), either from interpreting all features in the video (system #1), or for all tracks combined (systems #2 and #3) or the two per-track interpretation systems (#4 and #5). For the per-track interpretation, the prediction per video is obtained by keeping the one maximum posterior probability from all interpreted tracks in the

video. Hence, for each video one prediction is obtained, which is compared to the activity label of the video. This enables us to determine the classification accuracy and to analyze the confusions between the five human activities. The performance of each system is measured by the average classification accuracy across the five activities. For cross-validation, a leave-one-person-out scheme is deployed, which is common for evaluation of human activity recognition [25]. This yields 12 folds, because there are 12 persons in the dataset. For each fold, all videos of one person are in the test set, and the videos of the remaining eleven persons are in the training set. This per-fold training and testing is repeated 12 times and the results are accumulated. Finally, the classification accuracy for each activity is established and an average accuracy is computed.

5. RESULTS

5.1 Performance evaluation

The five system configurations from Section 3 have been assessed and compared, using the experimental setup from Section 4. All results have been summarized in Table 1. The performance of system #1, without tracking, is not very good. There are many errors for the activities 'wave' and 'throw' (33% accuracy), and 'dig' does not perform well either (50% accuracy). However, 'walk' and 'run' can be reasonably distinguished with this simple system. A large gain in classification accuracy is observed when the analysis is performed on features from tracks (systems #2 – #5). System #2 is the most simple setup of how to use tracks: all features from all tracks are accumulated. This gives a significantly better performance than system #1: the accuracy increases from 57% to 79% on accuracy. When only the tracks that are designated as human are used, the accuracy increases further to 88%. For the systems which are based on an interpretation per track, i.e., systems #4 and #5, Table 1 shows that the accuracy can be further improved to 93%. As expected, system #5 performs better than its counterpart without focus-of-attention, i.e., system #3. Apparently the human detection has a significant positive impact on the best result that can be achieved with the considered components. Interestingly, this does not hold for system #4 in comparison to its counterpart without focus-of-attention, i.e., system #2. System #4 performs less than system #2, while one may expect that it would work better because it involves more advanced analysis. This counterintuitive finding can be explained from the following insight. When there is no human detection involved, there are many partial tracks, which are fragmented both in space and time. This gives partial coverage of the human activities. With the human detection, the partial tracks are discarded, as there are no human detections on such tracks. This is the reason why system #5 performs well. However, without human detection, the partial tracks give rise to errors (system #4). It causes so many errors that it is better to consider all features of all tracks in the video at once (system #2). All in all, this experiment shows that the optimal configuration is system #5, including tracking, human detection and the focus-of-attention.

System	Tracking	Human detection	Focus-of-attention	Walk	Run	Dig	Wave	Throw	Classification accuracy
#1	-	-	-	75%	91%	50%	33%	33%	57%
#2	Yes	-	-	85%	94%	60%	71%	83%	79%
#3	Yes	Yes	-	90%	94%	79%	90%	88%	88%
#4	Yes	-	Yes	79%	91%	58%	83%	65%	75%
#5	Yes	Yes	Yes	94%	94%	94%	94%	91%	93%

Table 1. Human activity recognition results for five activities from the UCF-ARG dataset. System #5 involves all components (tracking, human detection, per-track analysis) and it performs best.

We are interested in the errors by system #5, i.e., which human activities are confused. The confusion matrix is shown in Figure 5. As expected, 'walk' and 'run' are confused (6%). 'Wave' and 'dig' are also confused. Although the activities look different in the video, partial tracks cause the 'dig' activity to be only partially observed. When only the arms of the 'dig' movement are observed, it may look like 'wave'. This happens during training time and causes confusion in the learned model. For specific examples of correctly recognized activities, errors, and illustrations of the typical confusions, we refer to Section 5.3.

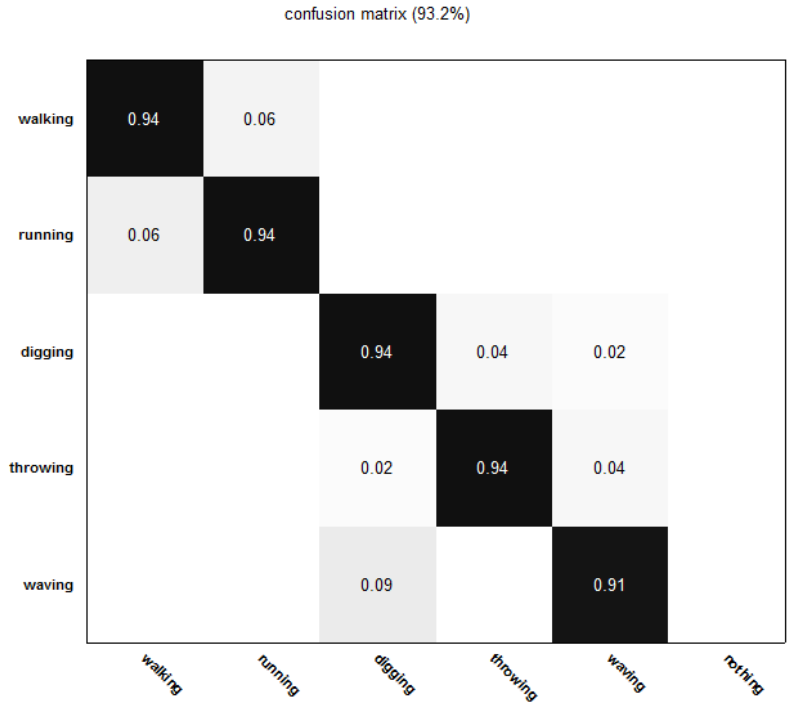

Figure 5. Human activity recognition results for five activities from the UCF-ARG dataset, obtained by system #5, which involves all components (tracking, human detection, per-track analysis) and it performs best..

5.2 Robustness to distracting human activities

In a second experiment, the robustness of the systems #1 – #5 is validated for the realistic condition where another activity happens simultaneously in the video stream. The rationale of this experiment is that an adequate system should be robust to another distracting human activity at the same time in the video stream. To simulate this condition for the UCF-ARG dataset, the track, feature, human detection and focus-of-attention data of each video has been augmented with the data of another randomly selected video. The performance of systems #4 and #5 will remain the same, because it involves a per-track analysis. The performance of systems without such focus-of-attention, i.e., systems #1 – #3, are tested. Clearly, performance deteriorates when there are multiple activities and the system has no ability to interpret them separately. The average classification accuracy drops by significantly for system #1, from 57% to 39%. The performance drops more than 50% for the systems with tracking (systems #2 and #3). The focus-of-attention is critical to scenes with multiple human activities present.

5.3 Recognition examples

For the best system (#5) we provide some classification examples, to provide some qualitative insights in the performance of our human activity recognition system. Running can still be recognized, even if the legs are not fully

covered by the track, see Figure 6 (top row). When the running person is observed from behind and from above, the fast and pronounced leg motion is not very clear and it may look like walking (bottom row).

Figure 6. 'Run': correct (top) and wrong (bottom) classification.

Throwing can still be recognized, even if the bounding boxes are not well aligned, see Figure 7 (top row). This misalignment is very common for this activity, because at each time step a different part of the body moves. First the torso moves, to pick the item up from the ground. Then the legs move, to get upright, after which the item is thrown by an arm movement. When only the arm movement is observed, the throwing is confused with waving, even though there is only one arm involved which moves fast (bottom row).

Figure 7. ‘Throw’: correct (top) and wrong (bottom) classification.

5.4 YouTube demo

The challenge for the human operator is to identify particular human activities in the large amounts of video data, as typically the interesting activities are rare, e.g., running. In populated areas it is hard to find the few relevant events in the midst of all activities. The rationale of this paper is to support the operator and analysts during their tasks, by a fully automated system that is able to recognize and localize particular human activities in UAV video data, as shown in the experiments and screenshots in the preceding subsections. Those recognized activities are stored as metadata and can be exploited in an online, causal system for intelligence gathering, or for offline forensic analysis. Yet, it is unclear how a recognition system may be deployed by the operator or analyst.

As a first step towards deployment, we have implemented a demo of our recognition system together with a mockup graphical user interface. In this demo, we show how hard it is for the human eye to find the one running person in the midst of 15 other persons who are all walking, in a setup with 16 camera feeds which are displayed in a 4x4 display matrix. The demo is available on our YouTube channel (www.youtube.com/intelligentimaging), where the reader can assess the difficulty of finding the runner. Figure 8 shows the overlays produced by our mockup interface, where it is immediately clear where the runner is. The operator or analyst is pointed to the runner by an orange bounding box, which is shown in more detail in the zoomed-in inset video frame at the bottom of this graphical user interface.

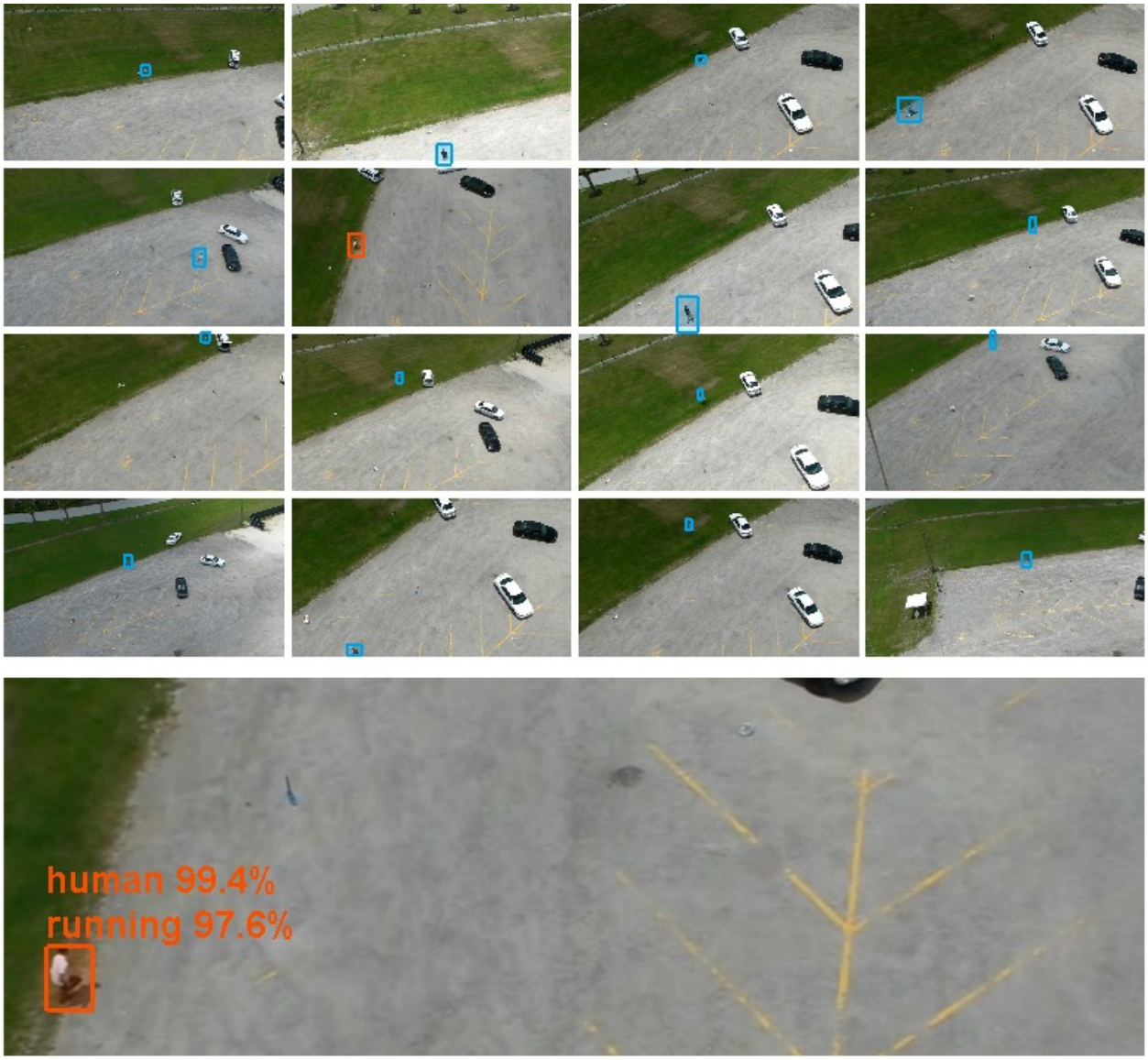

Figure 8. Our system can find the runner (orange box) in the midst of 15 other people who are walking (blue boxes). A mockup graphical user interface highlights the runner to the operator or analyst, so that proper actions can be taken or relevant intelligence can be gathered.

6. CONCLUSIONS

We have presented a system to extract metadata about human activities from full-motion video recorded from a UAV. The goal is to aid the human operator or analyst during his or her task to retrieve the few relevant occurrences of human activities in the midst of many other activities and/or long-duration video streams. The human activities under consideration are walk, run, throw, dig, wave. Walk and run have a characteristic movement and speed, while the remaining three activities do not involve movement; they occur in place. Their differences are subtle, as all activities involve arm and leg movements. General components that are needed for human activity recognition from UAV imagery are stabilization (to compensate for severe ego-motion of the camera), tracking, motion features, representation of the tracks, and classification of each track as one (or none) of the human activities.

One central question that we answered in this paper is the merit of each component for activity recognition. We have learned that the combination of tracking and human detection is needed to focus the attention on the relevant tracks that are involved in the human activities of interest. The best performing system includes tracking, human detection and a per-track analysis of the five human activities. This system achieves an average accuracy of 93%. Running and walking are very well recognized. The other activities involve detailed body movements in one place and these are more difficult to recognize. The errors are mainly related to tracking. They are typically due to a too short tracking of the activity. This especially holds for the compound and/or repetitive activities. Another error source is an insufficient spatial coverage of the activity, which happens in some cases of outward arm and leg movements. Improved tracking or post-track fixes are expected to increase the performance of the human activity recognition technology.

Another question that we explored in this paper is how the operator or analyst can be aided by human activity recognition technology. We have shared a conceptual graphical user interface to support the task of retrieving the relevant parts of video that contain particular human activities. A demo of this interface is available on YouTube, where the viewer is challenged to find the one runner in the midst of 15 other people who are walking. As shown in the online demo, and also in the experiments in this paper, our system can find the runner and highlight it to the user, thereby aiding him or her to focus on the relevant parts of the video.

REFERENCES

[1] Nagendran, A., Harper, D., Shah, M., "New system performs persistent wide-area aerial surveillance", SPIE Newsroom, http://spie.org/x41092.xml?ArticleID=x41092 (2010).

[2] Higgins, R.P., "Automatic event recognition for enhanced situational awareness in UAV video", Military Communications Conference (2005).

[3] Wu, S., Oreifej, O., Shah, M., "Action Recognition in Videos Acquired by a Moving Camera Using Motion Decomposition of Lagrangian Particle Trajectories", ICCV (2011).

[4] J. Aggarwal, M. Ryoo, Human activity analysis: A review, ACM Comput. Surv. 43 (3), 1-43 (2011).

[5] H. Bouma, G.J. Burghouts, L. de Penning, et al., Recognition and localization of relevant human behavior in videos, Proc. SPIE 8711 (2013).

[6] Wang, J., Zhang, Y., Lu, J., Xu, W., "A Framework for Moving Target Detection, Recognition and Tracking in UAV Videos", Affective Computing and Intelligent Interaction (2012).

[7] Nagendran, A., Harper, D., Shah, M., UCF-ARG dataset, University of Central Florida, "Aerial, Rooftop and Ground camera", http://crcv.ucf.edu/data/UCF-ARG.php (2010).

[8] Intelligent Imaging video channel on YouTube, "From UAV images to actionable intelligence", https://www.youtube.com/watch?v=IRBl7lbqqBs (2014).

[9] Heintz, F., Rudol, P., Doherty, P., “From images to traffic behavior - A UAV tracking and monitoring application”, International Conference on Information Fusion (2007)..

[10] Lin, Y. Medioni, G., "Map-Enhanced UAV Image Sequence Registration and Synchronization of Multiple Image Sequences", CVPR (2007).

[11] Dasgupta, P., “A Multiagent Swarming System for Distributed Automatic Target Recognition Using Unmanned Aerial Vehicles”, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 38, no. 3 (2008).

[12] Eekeren, A.W.M., Burghouts, G.J., Dijk, J., "Detection of humans and moving objects from an airborne platform", SPIE (2014).

[13] Oreifej, O., Li, X., Shah, M., “Simultaneous Video Stabilization and Moving Object Detection in Turbulence”, IEEE PAMI (2013).

[14] Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M., "Visual Tracking: An Experimental Survey", IEEE PAMI (2014).

[15] Dollár, P., Belongie, S., Perona, P., "The Fastest Pedestrian Detector in the West", BMVC (2010).

[16] Stauffer, C., Grimson, W., "Adaptive background mixture models for real-time tracking", CVPR (1999).

[17] Bobick, A., Davis, J., "The recognition of human movement using temporal templates", IEEE PAMI, 23(3):257–267 (2001).

[18] Laptev, I., "On space-time interest points", IJCV, 64 (2/3) (2005).

[19] Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., "Learning Realistic Human Actions from Movies", CVPR (2008).
[20] Wang, H., Schmid, C., "Action Recognition with Improved Trajectories", IEEE International Conference on Computer Vision (2013).
[21] Burghouts, G.J., Schutte, K., "Spatio-Temporal Layout of Human Actions for Improved Bag-of-Words Action Detection", Pattern Recognition Letters (2013).
[22] Andersson, M., Patino, L., Burghouts, G.J., Flizikowski, "Activity Recognition and Localization on a Truck Parking Lot", IEEE AVSS (2013).
[23] Burghouts, G., Schutte, K., Bouma, H., den Hollander, R., "Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos," Machine Vision and Applications 25(1), 85–98 (2014).
[24] Burghouts, G.J., Schutte, K., ten Hove, R.J-M., van den Broek, S.P., et al., "Instantaneous Threat Detection based on a Semantic Representation of Activities, Zones and Trajectories", Signal, Image and Video Processing, pending revision (2014).
[25] Liu, H., Feris, R., Sun, M.T., "Benchmarking human activity recognition", CVPR Tutorial, CVPR (2012).
[26] Oh, S., Hoogs, A., Perera, A., et al., "A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video", CVPR (2011).
[26] Bouma, H., Borsboom, S., den Hollander, R., Landsmeer, S., Worring, M., "Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination", Proc. SPIE, vol. 8359 (2012).
[27] Breiman, L., "Random forests", Machine Learning, 45(1), 5-32 (2001).