Human Detection and Motion Analysis from a Quadrotor UAV

IOP Conf. Series: Materials Science and Engineering **405** (2018) 012003 doi:10.1088/1757-899X/405/1/012003

Human Detection and Motion Analysis from a Quadrotor UAV

Asanka G Perera[1], Ali Al-Naji[1,2], Yee Wei Law[1], and Javaan Chahl[1,3]

[1] School of Engineering, University of South Australia, Mawson Lakes, SA 5095, Australia.
[2] Electrical Engineering Technical College, Middle Technical University, Al Doura 10022, Baghdad, Iraq.
[3] Joint and Operations Analysis Division, Defence Science and Technology Group, Melbourne, Victoria 3207, Australia.

E-mail: asanka.perera@mymail.unisa.edu.au

Abstract. This work focuses on detecting humans and estimating their pose and trajectory from an umnanned aerial vehicle (UAV). In our framework, a human detection model is trained using a Region-based Convolutional Neural Network (R-CNN). Each video frame is corrected for perspective using projective transformation. Using Histogram Oriented Gradients (HOG) of the silhouettes as features, the detected human figures are then classified for their pose. A dynamic classifier is developed to estimate forward walking and a turning gait sequence. The estimated poses are used to estimate the shape of the trajectory traversed by the human subject. An average precision of 98% has been achieved for the detector. Experiments conducted on aerial videos confirm our solution can achieve accurate pose and trajectory estimation for different kinds of perspective-distorted videos. For example, for a video recorded at 40m above ground, the perspective correction improves accuracy by 37.1% and 17.8% in pose and viewpoint estimation respectively.

1. Introduction

Unmanned aerial vehicles (UAVs) can be deployed in a variety of search and rescue, and surveillance applications by leveraging its mobility and operational simplicity. In some situations, a UAV's ability to recognize the actions of a human subject is desirable, then take responsive actions. Recognizing human actions from videos captured from a static platform is a challenging task owing to the articulated structure and range of possible poses of the human body. Recognition is further challenged by the quality of videos which include perspective distortion, occlusion, and motion blur.

The study presented in this paper is focused on using a UAV to recognize human subjects from an aerial video and to estimate the gait sequence and movement trajectory. Our solution consists of the following steps: (i) The human detector is trained using the method of Region-based Convolutional Neural Network (R-CNN) [1] with aerial images selected from publicly available aerial image datasets and our field images. (ii) The perspective correction step compensates for perspective distortion in aerial images.

Multiple pre-annotated homography matrices are used for different levels of distortion caused by different camera elevation angles. The experimental results show that this technique enhances performance in gait and trajectory estimation for aerial videos. (iii) The segmentation step

generates the silhouettes and uses Histograms of Oriented Gradients (HOG) [2] as feature descriptors. (iv) The pose estimation uses a dynamic classifier inspired by [3, 4]. (v) The trajectory estimation step estimates the shape of the human subject's path using 3-D skeletons and localizing them with respect to the initial pose and viewpoint.

The key contribution of this paper is a preliminary solution that a vision-capable quadrotor will be able to use for human detection, pose estimation and trajectory estimation. This study proposes to use an R-CNN detector and a perspective correction module in combination with a novel dynamic classifier architecture. Unlike other designs, our classifier uses temporal relationships between poses to achieve efficient pose and trajectory estimations.

2. Related work

Aerial videos are always subject to some level of perspective distortion due to their aerial viewpoint. It is necessary for videos to be perspective-corrected before classification. Projective transformation, or homography, is a standard technique for correcting perspective distortion [5], but this traditional approach requires the *vanishing point* to be manually specified. Rogez et al. [6] used manually determined *vertical scene lines* to estimate the vanishing point and localize the reconstructed poses based on the vanishing point. Our homography step is similar to theirs, the difference being that we determine the vanishing point based on the altitude and angle of the camera.

Projective transformation has been used to achieve improved results for videos from overhead cameras [7, 8]. In [7], affine transformation has been applied using the 3D scene information for perspective correction. Their experiments show that perspective correction has a noticeable impact on recognition performance. Li et al. [8] reported a human detection accuracy of 87.2% for CAVIAR dataset [9] when the images are corrected for perspective using 3D scene information. When using the 2D search and in-plane rotation the accuracy was only 38.3%. This 48.9% improvement was achieved for the static videos while our 37.1% and 17.8% accuracy improvements in pose and viewpoint estimation were achieved from dynamic videos (40m above ground).

Dynamic classifier selection (DCS) [10] is based on the *local accuracy estimation* of each individual classifier. The main idea is selecting an individual classifier which is most likely correct for a given sample. The final classification decision is made only by the selected classifier. A relatively similar classifier was developed by Ko et al. [4] by integrating a majority voting system. Our classifier follows the DCS principles, but it selects the best individual classifier without executing the entire ensemble of classifiers.

Using UAVs in human detection and activity recognition missions is a relatively new topic. Some studies focused on human detection methods from aerial videos in relation to search and rescue missions [11, 12]. These studies aimed at identifying humans lying or sitting on the ground. Some notable approaches related to human identity recognition in low-resolution aerial videos are weighted voter-candidate formulation by Oreifej et al. [13] and blob matching using an adaptive reference set by Yeh et al. [14]. Monajjemi et al. [15] developed a UAV onboard gesture recognition system to identify periodic movements of waving hands from other periodic movements like walking and running in an outdoor environment. Our experimental set-up is most similar to Monajjemi et al.'s.

3. Methodology

This section provides details on human detection, perspective correction, segmentation and feature extraction, pose estimation and trajectory estimation.

3.1. Human detection

A human detection model is trained using the R-CNN method originally presented in [1]. R-CNN combines region proposals with Convolutional Neural Networks (CNN). In the pre-processing stage, it uses a region proposal algorithm [16] before running the CNN. R-CNN is considered to be a state-of-the-art visual object detection system that combines bottom-up region proposals with rich features computed by a CNN [1].

For experimentation, the CNN features of 510 selected images were used to train the detector. 510 images were selected to represent different human subjects from a range of viewpoints. The images were selected from publicly available MoBo Aligned dataset [17], VIRAT video dataset [18], mini-drone video dataset [19], UCF aerial action dataset [20], PETS 2006 dataset [21] and our aerial field videos (see Figure 1). All the human instances in the images were labeled. The images were randomly indexed in order to mix them properly. Then, the dataset was randomly stratified as 0.8:0.2 for training and testing data respectively and achieved a 98% accuracy for human detection.

Transfer learning [22] was applied to retraining the AlexNet [23] neural network with the new CNN features. For this task, AlexNet pre-trained network was selected because it has been pre-trained on 1.2 million ImageNet [24] images of 1000 classes, some of which were trained on images of humans in different settings, and showed the best performance in the ImageNet Large Scale Visual Recognition Challenge in 2012 [23].

3.2. Perspective correction

The relative orientation between the human subject and the camera can be represented in a *horizontal coordinate system* (see Figure 2(a)). In the horizontal coordinate system, $\phi \in [0, \pi/2]$ is the elevation/tilt angle, whereas $\theta \in [0, 2\pi)$ is the azimuth/pan angle. The azimuth angle is calculated in the radial direction between the heading direction of the human subject and the camera center axis on the horizontal plane. The vertical perspective distortion occurs when $\phi > 0$, and worsens as ϕ gets larger. When $\phi = 90°$, perspective distortion cannot be corrected. For $60° \leq \phi < 90°$, the captured images have a severely distorted perspective that is very difficult to compensate. Therefore in this study, the maximum ϕ is limited to $60°$.

Perspective correction is done by mapping the distorted image plane (see Figure 2(b)) to the undistorted vertical plane through *homography*. Segments on the undistorted vertical plane then enable the matching of test and training images. Given an image, for every homogeneous point on the image plane, $\boldsymbol{x}$, there exists a homography matrix $\mathbf{H}$ that maps it to a homogeneous point, $\boldsymbol{x}'$, on the undistorted vertical plane, i.e.,

$$\boldsymbol{x}' = \mathbf{H}\boldsymbol{x}. \tag{1}$$

The matrix $\mathbf{H}$ depends on the elevation angle ϕ. Instead of calculating $\mathbf{H}$ for each video, It was calculated offline for each of the following ϕ values: $\arctan(10/30) = 18.4°, \arctan(20/30) = 33.7°, \arctan(30/30) = 45.0°$ and $\arctan(40/30) = 53.1°$. To calculate $\mathbf{H}$, four points were manually selected in a sample video frame to (i) delineate the area of interest and (ii) generate the vertical scene lines, as shown in Figure 2(b). The vertical scene lines define the homography matrix $\mathbf{H}$.

3.3. Segmentation and feature extraction

After perspective correction, the human silhouette was segmented. The size of the silhouette in the image plane varies depending on the direct distance between the camera and the human subject. Perspective correction alone cannot address this scaling issue. Thus, the test silhouette is scaled up or down to match the scale of the training images. Prior to feature extraction, the test videos are annotated for pose and viewpoint.

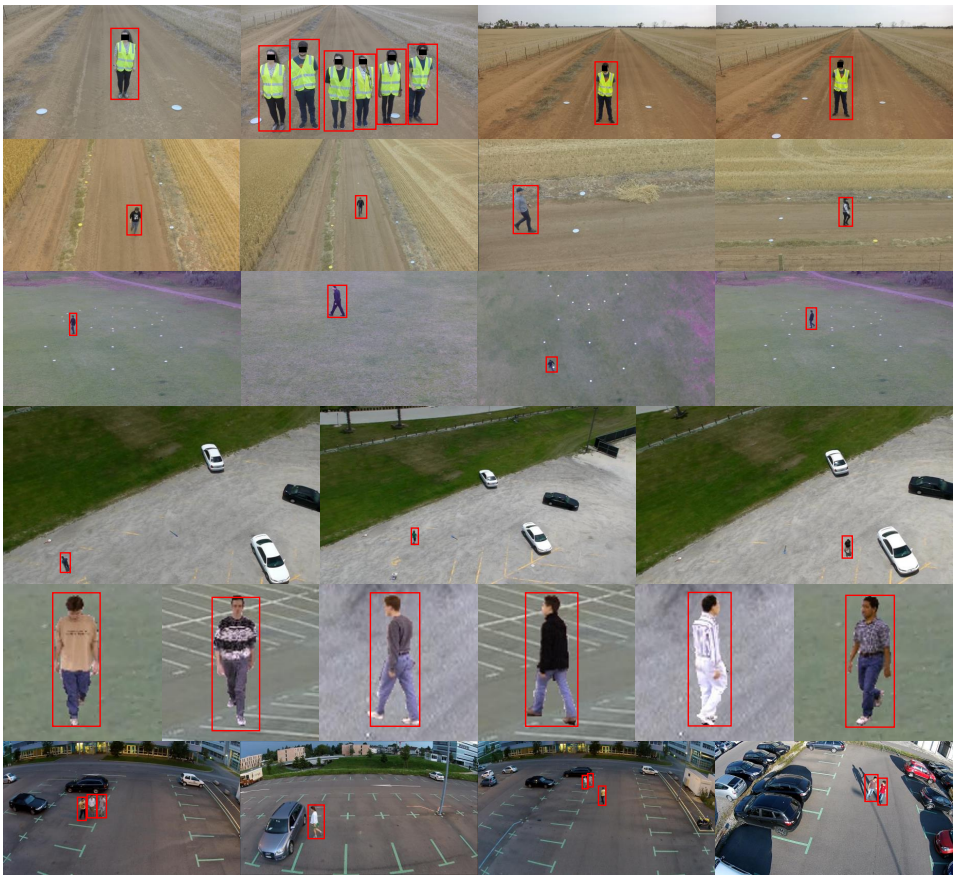

Figure 1. Sample images with annotated bounding boxes. First three rows represent some images from our field videos. Third to sixth rows correspond to some selected images from UCF aerial action [20], MoBo Aligned [17] and mini-drone [19] datasets respectively.

For each frame, the RGB image was converted into a binary image and its bounding box area was segmented. Noise was removed using a Gaussian filter and small objects containing fewer than a threshold number of pixels were also removed. The remaining blob or blobs were considered to represent the human silhouette. Currently, the denoising parameters and segmentation parameters were customized for each video clip to obtain the best possible silhouette, so they are subject to improvements.

For feature extraction, the image window was divided into small spatial regions called "HOG cells" [2]. The weighted gradients in a HOG cell form a 1-D histogram which represented the orientation of the edge lines. The feature vector was formed from the HOG blocks, each of which represents a group of HOG cells.

3.4. Pose estimation

A training dataset was created from 1017 silhouette images to identify the eight sub-steps of the human gait cycle (P_1 to P_8 in Figure 3 (b)). This training dataset should not be confused with the 510 images used to train the R-CNN detector. Each sub-step (or pose) had viewpoints from eight radial directions (azimuth angles that are 45° apart), giving rise to $8 \times 8 = 64$ pose-viewpoint pairs. The finite number of elevation-azimuth angle pairs are equivalent to the

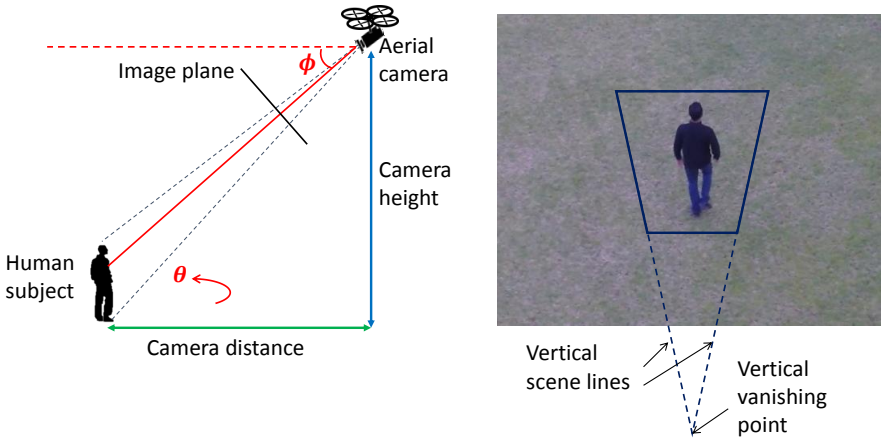

Figure 2. (a) The UAV hovers at a known camera height and angle. (b) An input image with vertical scene lines. The scene lines are manually constructed according to the elevation angle. The blue color box in the middle is the area of interest for homography in the vertical plane. The vanishing point is the point where parallel scene lines would meet each other on the image plane.

discretized viewing hemisphere described in [6].

The collected training data consisted of 64 labels, representing eight sub-steps of the gait cycle and eight viewpoints (see Figure 3 (b)). A training dataset of n observations is denoted by

$$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}, \tag{2}$$

where $\boldsymbol{x}_i$ is the ith feature vector, and y_i the ith label.

It was assumed that the human subject walked forward at a constant speed, does not take sharp turns and does not twist the body while turning. This assumption does not preclude left, right, or backward turns, as long as the turning is not abrupt, as exemplified by the yellow-border windows in Figure 3(b).

The classification output was considered to be a state. The admissible state transitions restrict the next classifier prediction to be one of the states the current state can transition to (yellow-border windows in Figure 3(b)). Given the current pose and viewpoint, when a new image is available, the associated pose was predicted to be either the current pose, or the pose in the next sub-step of the gait cycle. When the pose changes from the current state to the next state, the viewpoint of the next pose has to be one of the following: the same viewpoint (moving straight), 45° clockwise from the current viewpoint (turning left), or 45° anticlockwise from the current viewpoint (turning right).

Our dynamic classifier selection (DCS) architecture consists of 64 4-class SVM classifiers denoted $C_4(P_i, V_j)$, $i, j \in \{1, \ldots, 8\}$. The classifier $C_4(P_i, V_j)$ is associated with pose P_i and viewpoint V_j, and is trained to recognize the set of four classes:

$$\{(P_i, V_j), (P_{i\boxplus 1}, V_j), (P_{i\boxplus 1}, V_{j\boxminus 1}), (P_{i\boxplus 1}, V_{j\boxplus 1})\}, \tag{3}$$

where $i, j \in \{1, \ldots, 8\}$ and the operators $\boxplus, \boxminus$ are defined as follows:

$$i \boxplus j = (i + j + 1) \mod 8 - 1, \tag{4}$$
$$i \boxminus j = (i - j - 1) \mod 8 + 1. \tag{5}$$

For example, the classifier $C_4(P_4, V_5)$ is trained to recognize the four classes labeled a, b, c and d in Figure 3 (b).

As depicted in Figure 3 (a), our classification process works in two stages: (i) the initialization stage and (ii) the DCS stage. In the initialization stage, the first q video frames are classified using classifier C_{64}. C_{64} is a single 64-class SVM classifier trained with the complete dataset (64 classes). The DCS stage starts with the $(q+1)$th video frame. In this stage, each frame is classified with a classifier chosen based on the class label predicted by the previous iteration.

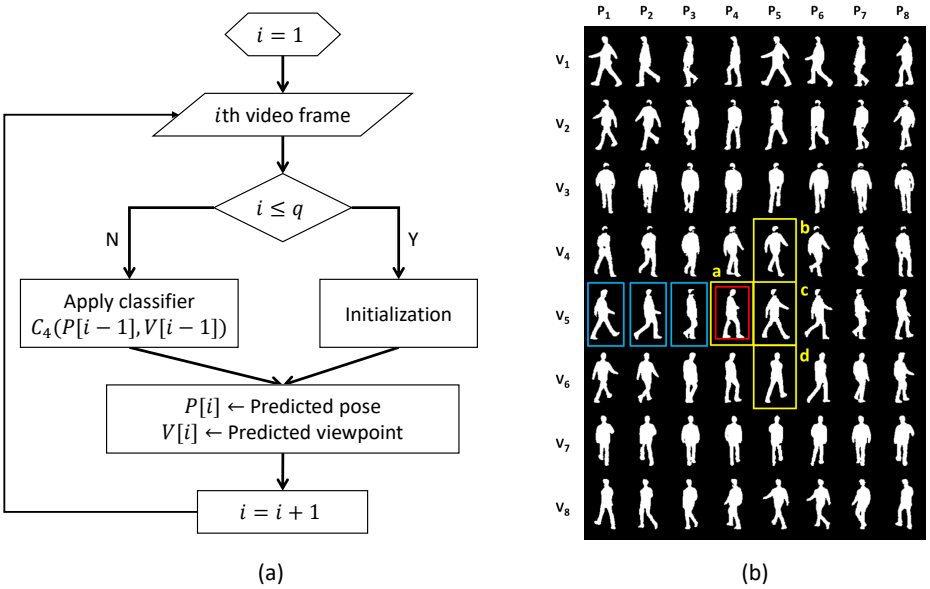

Figure 3. (a) Flowchart for pose-viewpoint classification by dynamic classifier selection. Prior to the workflow, all classifiers should have been trained. (b) The training dataset consists of 64 classes of pose-viewpoint pairs. The rows and columns represent the poses and the viewpoints respectively. Each silhouette in the figure is a random image from the pose-viewpoint subset it belongs. Blue- and red-border windows show four consecutive initialized frames (from left to right). Once initialized, the pose and the viewpoint of the most recently initialized image (red-border window) are used to select the next classifier. In this example, the training image subsets of the next classifier are shown in yellow-border windows.

For elaboration, consider the example in Figure 3(b). Suppose $q = 4$, and the blue- and red-border windows are sample classes predicated by the classifier C_{64}. The red-border window highlights the class predicted for the qth frame. Since this class is (P_4, V_5), the classifier $C_4(P_4, V_5)$ is chosen to classify the $(q+1)$th frame. The training subsets for $C_4(P_4, V_5)$ are highlighted with the yellow-border windows a, b, c and d.

The most significant difference between our classifier architecture and existing architectures in the recent literature [3, 25, 26] is that ours does not execute all the classifiers to make a decision. Instead, only the relevant classifier is selected for every next image. The relevance of the classifier is determined by its training subsets, and the training subsets are selected based on the state transition graphs.

3.5. Trajectory estimation

Trajectory estimation refers to the estimation of the shape of the path traversed by the human subject. Each estimated viewpoint serves as an estimation of the walker's orientation. For each

estimated orientation, a 3-D pose is reconstructed from the estimated pose. The algorithm can be described as follows:

- Whenever an estimated pose is the same as the previous, he subject is assumed to remain at the same location. Such predictions occur due to the camera's high frame rate and/or the subject's slow movements.
- Whenever an estimated pose differs from the previous, the subject is assumed to have moved a fixed distance from the location of the previous pose. When the orientation changes, the next pose is positioned at a fixed distance from the location of the previous pose at an angle of $\pm 45°$ (+ve for right turns, −ve for left turns).

4. Experimental results

The experiments were conducted at different heights with original aerial videos and perspective-corrected videos. For trajectory estimation, each estimated trajectory is plotted on a 2-D plane with *unitless* axes, and the starting location mapped to the origin. Along a trajectory, the estimated poses were reconstructed using a 3-D, 13-jointed skeletal models. The proximty of the estimated trajectories to the actual trajectories was assessed visually.

All the videos were captured from a rotorcraft UAV (see Figure 4(b)) in a slow and low-altitude flight mode. For recording videos, a GoPro Hero 4 black camera with an anti-fish eye replacement lens (5.4mm, 10MP, IR CUT) and a 3-axis Solo gimbal was used. The images were sampled at a rate of 30fps. In order to ease the segmentation process, the videos were recorded with an uncluttered background and with the human subject wearing dark clothes. The UAV-captured videos were segmented as described in Section 3.3. These experiments were conducted using HOG features.

Certain assumptions were made to ease the coordinate transformation between the camera and the human subject. The human subject was assumed to be upright on a flat ground. The camera roll angle was considered to be zero. The roll, pitch and yaw angles of the UAV were assumed to be zero during the slow flying. Hence, the flight dynamics of the UAV has negligible effects on the true camera elevation angle. The camera elevation angle and the height were directly recorded via the UAV control interface. The UAV was operated at a known ground distance (camera distance) from the human subject.

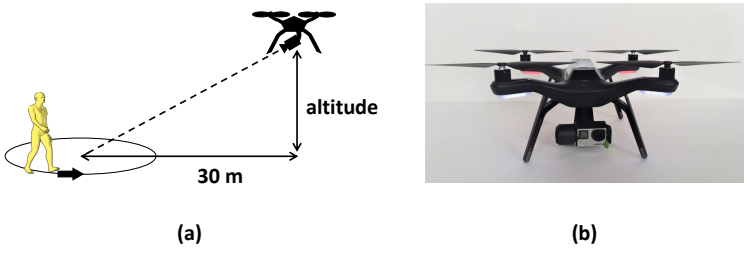

Figure 4. (a) A human subject was filmed walking on a circle, while the UAV stays pointed at the center of the circle. (b) The rotorcraft UAV, namely a 3DR Solo, used for experimentation.

As depicted in Figure 4(a), a human subject is filmed walking on a marked circle by a UAV pointing at the center of the circle. The experiment was conducted to analyze the effect of perspective distortion in detail. The UAV was flown at heights of 10m, 20m, 30m and 40m (see Figure 6). The lowest height of 10m caused negligible perspective distortion, but at $h = 40$m ($\phi = 53.1°$), the video suffers from severe perspective distortion. The main observations are:

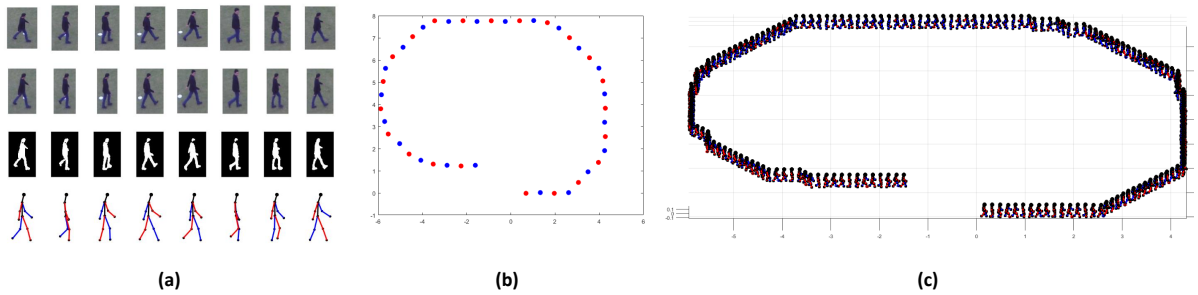

Figure 5. Results for the altitude of 20m: (a) The top row shows a series of cropped video frames. The second row is the perspective-corrected version of the top row. The third row shows the segmented silhouettes and the bottom row shows the estimated poses. (b) The estimated trajectory, where each dot marks where both feet touch the ground, with red representing right foot in front and blue representing left foot in front. (c) The estimated trajectory and 3-D reconstruction of the estimated poses.

Table 1. Estimation errors of the dynamic classifier for perspective-distorted (PD) and perspective-corrected (PC) videos. e_{pose} and $e_{\text{viewpoint}}$ are estimation errors for pose and viewpoint respectively.

Altitude	**Perspective distortion**	**#frames**	e_{pose}	$e_{\text{viewpoint}}$
$h = 10$m	No distortion	787	**23.5%**	**13%**
$h = 20$m	PD	784	**22.1%**	**17.2%**
	PC		39.9%	20.4%
$h = 30$m	PD	810	56.7%	44.8%
	PC		**40.6%**	**37.2%**
$h = 40$m	PD	817	74.4%	42.6%
	PC		**37.3%**	**24.8%**

- In terms of pose and viewpoint estimation accuracies, perspective correction helps the dynamic classifier.
- The advantage of perspective correction is more pronounced on more distorted videos.

Table 1 shows an overall reduction in estimation errors when perspective correction is applied. This conclusion is further indicated in improved trajectory estimations in Figure 6.

5. Discussion

The human detector trained using R-CNN successfully identifies the humans in aerial videos. However, it is trained to focus human detection from relatively clutter-free backgrounds. The error rate might be high for detecting humans in a complex background. The robustness of the detector can be improved by using a range of images from different settings as training set images.

A drawback of the dynamic classifier is its dependency on the initialization. Like all classifiers, C_{64} sometimes makes mistakes, throwing the $C_4(\cdot,\cdot)$ classifiers off-course. A potential improvement is to re-initialize the dynamic classifier (see Figure 3(a)) periodically.

HOG features are traditionally considered to be handcrafted features, and in some areas they have been replaced by CNN features. However, for the classification stage we are interested in

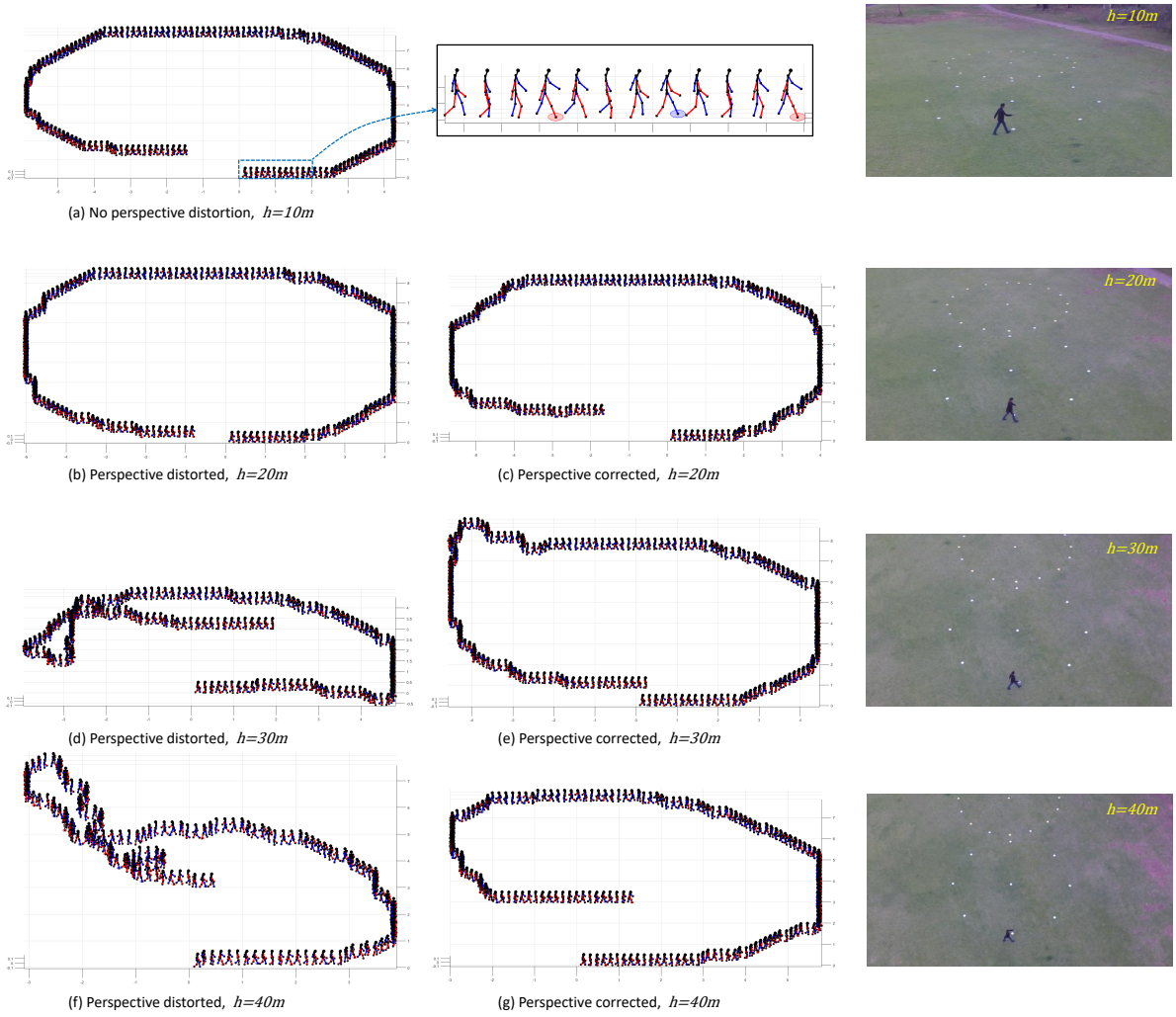

Figure 6. Reconstructed poses and their trajectories correspond to four different heights.

the shapes of the human subject. Our observation for the overall robustness of HOG features is dependent on silhouettes and hence significantly depends on the edges. However, segmentation of aerial images (for HOG) is very challenging due to varying resolution and background, and can benefit from the latest advances in semantic segmentation.

The results confirm the intuition that perspective correction is imperative for severely perspective-distorted videos. Our solution has problems with purely frontal or back views, because frontal and back silhouettes do not provide sufficient details for differentiating poses. A potential solution is provided by the mobility of the aerial platform itself. The UAV can be programmed to seek a good elevation angle and azimuth angle, before it starts analyzing the human subject's action. This will require control algorithms and machine intelligence that go beyond the scope of this work.

6. Conclusion

This paper discusses a solution for human detection from perspective distorted aerial videos and estimates their pose and trajectory. The approach consists of R-CNN-based detection,

perspective correction by homography, HOG feature extraction and dynamic classifier selection. The detector trained on different aerial and fronto-parallel images achieved nearly ideal accuracy in the experiments. The dynamic classifier consists of 64 4-class classifiers and enables robust classification results. Trajectory estimation provides the shape of the path traversed by the human subject, and is dependent on viewpoint estimation. The study discussed in this article is limited to estimating walking gaits. Our future work includes equipping UAVs with the ability to recognize human activities.

Acknowledgment
This project was partly supported by Project Tyche, the Trusted Autonomy Initiative of the Defence Science and Technology Group.

References

[1] Girshick R, Donahue J, Darrell T and Malik J 2014 Rich feature hierarchies for accurate object detection and semantic segmentation *2014 IEEE Conference on Computer Vision and Pattern Recognition* pp 580–587 ISSN 1063-6919
[2] Dalal N and Triggs B 2005 Histograms of oriented gradients for human detection *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* vol 1 pp 886–893 vol. 1 ISSN 1063-6919
[3] Woods K, Kegelmeyer W P and Bowyer K 1997 Combination of multiple classifiers using local accuracy estimates *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19** 405–410 ISSN 0162-8828
[4] Ko A H, Sabourin R, Britto A S and Jr 2008 From dynamic classifier selection to dynamic ensemble selection *Pattern Recognition* **41** 1718 – 1731 ISSN 0031-3203
[5] Hartley R and Zisserman A 2003 *Multiple view geometry in computer vision* (Cambridge university press)
[6] Rogez G, Orrite C, Guerrero J and Torr P H 2014 Exploiting projective geometry for view-invariant monocular human motion analysis in man-made environments *Computer Vision and Image Understanding* **120** 126 – 140 ISSN 1077-3142
[7] Bak S, Zaidenberg S, Boulay B and Brmond F 2014 Improving person re-identification by viewpoint cues *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* pp 175–180
[8] Li Y, Wu B and Nevatia R 2008 Human detection by searching in 3d space using camera and scene knowledge *2008 19th International Conference on Pattern Recognition* pp 1–5 ISSN 1051-4651
[9] 2004 URL `http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/`
[10] Perera A G, Law Y W, Al-Naji A and Chahl J 0 2018 Human motion analysis from UAV video *International Journal of Intelligent Unmanned Systems* **0** 00–00 (*Preprint* `https://doi.org/10.1108/IJIUS-10-2017-0012`) URL `https://doi.org/10.1108/IJIUS-10-2017-0012`
[11] Rudol P and Doherty P 2008 Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery *Aerospace Conference, 2008 IEEE* pp 1–8 ISSN 1095-323X
[12] Andriluka M, Schnitzspan P, Meyer J, Kohlbrecher S, Petersen K, von Stryk O, Roth S and Schiele B 2010 Vision based victim detection from unmanned aerial vehicles *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* pp 1740–1747 ISSN 2153-0858
[13] Oreifej O, Mehran R and Shah M 2010 Human identity recognition in aerial images *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* pp 709–716 ISSN 1063-6919
[14] Yeh M C, Chiu H K and Wang J S 2016 Fast medium-scale multiperson identification in aerial videos *Multimedia Tools and Applications* **75** 16117–16133 ISSN 1573-7721
[15] Monajjemi M, Bruce J, Sadat S A, Wawerla J and Vaughan R 2015 UAV, do you see me? Establishing mutual attention between an uninstrumented human and an outdoor UAV in flight *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* pp 3614–3620
[16] Zitnick C L and Dollár P 2014 Edge boxes: Locating object proposals from edges *Computer Vision – ECCV 2014* ed Fleet D, Pajdla T, Schiele B and Tuytelaars T (Cham: Springer International Publishing) pp 391–405
[17] Rogez G, Rihan J, Orrite-Uruñuela C and Torr P H S 2012 Fast Human Pose Detection Using Randomized Hierarchical Cascades of Rejectors *International Journal of Computer Vision* **99** 25–52 ISSN 1573-1405 URL `https://doi.org/10.1007/s11263-012-0516-9`
[18] Oh S, Hoogs A, Perera A, Cuntoor N, Chen C C, Lee J T, Mukherjee S, Aggarwal J K, Lee H, Davis L, Swears E, Wang X, Ji Q, Reddy K, Shah M, Vondrick C, Pirsiavash H, Ramanan D, Yuen J, Torralba

A, Song B, Fong A, Roy-Chowdhury A and Desai M 2011 A large-scale benchmark dataset for event recognition in surveillance video *CVPR 2011* pp 3153–3160 ISSN 1063-6919

[19] Bonetto M, Korshunov P, Ramponi G and Ebrahimi T 2015 Privacy in mini-drone based video surveillance *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* vol 04 pp 1–6

[20] 2009 UCF Aerial Action Data Set `http://crcv.ucf.edu/data/UCF_Aerial_Action.php` [Online; accessed 19-Jan-2018]

[21] 2006 PETS 2006 Benchmark Data `http://www.cvg.reading.ac.uk/PETS2006/data.html` [Online; accessed 19-Jan-2018]

[22] Yosinski J, Clune J, Bengio Y and Lipson H 2014 How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems 27* ed Ghahramani Z, Welling M, Cortes C, Lawrence N D and Weinberger K Q (Curran Associates, Inc.) pp 3320–3328

[23] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems 25* ed Pereira F, Burges C J C, Bottou L and Weinberger K Q (Curran Associates, Inc.) pp 1097–1105 URL `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`

[24] Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L 2009 Imagenet: A large-scale hierarchical image database *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp 248–255 ISSN 1063-6919

[25] Kuncheva L I, Bezdek J C and Duin R P 2001 Decision templates for multiple classifier fusion: an experimental comparison *Pattern Recognition* **34** 299 – 314 ISSN 0031-3203

[26] Tulyakov S, Jaeger S, Govindaraju V and Doermann D 2008 *Review of Classifier Combination Methods* (Berlin, Heidelberg: Springer Berlin Heidelberg) pp 361–386 ISBN 978-3-540-76280-5