Geometric Constraints for Human Detection in Aerial Imagery

Vladimir Reilly, Berkan Solmaz, and Mubarak Shah

Computer Vision Lab, University of Central Florida, Orlando, USA
{vsreilly,bsolmaz,shah}@eecs.ucf.edu

Abstract. In this paper, we propose a method for detecting humans in imagery taken from a UAV. This is a challenging problem due to small number of pixels on target, which makes it more difficult to distinguish people from background clutter, and results in much larger searchspace. We propose a method for human detection based on a number of geometric constraints obtained from the metadata. Specifically, we obtain the orientation of groundplane normal, the orientation of shadows cast by humans in the scene, and the relationship between human heights and the size of their corresponding shadows. In cases when metadata is not available we propose a method for automatically estimating shadow orientation from image data. We utilize the above information in a geometry based shadow, and human blob detector, which provides an initial estimation for locations of humans in the scene. These candidate locations are then classified as either human or clutter using a combination of wavelet features, and a Support Vector Machine. Our method works on a single frame, and unlike motion detection based methods, it bypasses the global motion compensation process, and allows for detection of stationary and slow moving humans, while avoiding the search across the entire image, which makes it more accurate and very fast. We show impressive results on sequences from the VIVID dataset and our own data, and provide comparative analysis.

1 Introduction

In recent years improvements in electronics and sensors have allowed for development and deployment of Unmanned Aerial Vehicles (UAVs) on greater and greater scale, in a wide variety of applications, including surveillance, military, security, and distaster relief operations. The large amount of video data obtained from these platforms, requires automated video analysis tools, whose capabilities must include object detection, tracking, classification and finally scene and event analysis. While a number of methods and systems exist for detecting and tracking vehicles in UAV video (e.g. [1] [2]), the same cannot be said about human detection.

State of the art human detection methods such as [3] [4] [5] [6] [7], are designed to deal with datasets containing imagery taken from the ground, either in surveillance or consumer imagery scenario. People in that type of imagery are

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part VI, LNCS 6316, pp. 252–265, 2010.

fairly large (e.g. 128x64 in the case of INRIA dataset). Also the camera in such scenarios is generally oriented with the ground plane. In our case, the humans are much smaller as seen in Figure 1. On average they are about 24x14 pixels in size, and have no visible parts, this makes part detection methods such as [4] and [6] inapplicable. Bag of feature models such as [5] also have great difficulty due to a very small number of interest points that can be found. Another issue is that since the camera is mounted on a moving aerial platform, the imaged size and visible distinguishing features of a person can be reduced even further when the camera is at a high elevation angle. Also, the moving aerial platform introduces a large number of possible orientations at which a human can appear in the scene. Due to lack of good distinguishing features of the human body in aerial imagery, a brute force image search generates many false detections, and is also quite slow. Hence, previous two works that specifically deal with aerial imagery ([8] and [9]), opt to constrain the search with preliminary processing.

A very popular approach is to constrain the search using motion as in [10], or Xiao et. al. [8]. They assume that only moving objects are of interest, and adopt a standard aerial surveillance pipeline. First, they compensate for global camera motion, then they detect moving objects, and finally classify each moving object as either a person or vehicle using the combination of histograms of oriented gradients (HOG) and a support vector machine proposed in [3]. The problem with the motion constraint, is that since people are viewed from far away, their motion is very subtle and difficult for the system to pick up. Of course, if people are stationary, then the system cannot detect them at all. If there are shadows present in the scene, then a number of additional problems arise. It is difficult to localize the human, since its shadow is part of the moving blob, which also makes the blobs more similar to each other making it more difficult to track them. See Figure 8 for examples of these failures.

Miller et. al. avoid the moving object assumption [9], by assuming that at least one Harris corner feature point will be detected on the human in each frame. This generates a large number of candidates which are then suppressed through tracking of the Harris corners in global reference frame. Each corner is then classified using a OT-MACH filter. If a track contains more human classifications than 20% of total track length, all points within track are labelled as human. The problem with the above approach is the large number of potential human candidates; they report 200 for a 320x240 image, and the need for a sophisticated tracker to filter them out.

We propose a very different approach. In particular we constrain the search by assuming that humans are upright shadow casting objects. We utilize directed low level computer vision techniques based on a set of geometric scene constraints derived from the metadata of the UAV platform. Specifically, we utilize the projection of the ground plane normal to find blobs normal to the ground plane, these give us an initial set of potential human candidates. Similarly we utilize the projection of shadow orientation to obtain a set of potential shadow candidates. We then obtain a refined set of human candidates, which are pairs of shadow and normal blobs that are of correct geometric configuration, and relative size.

Fig. 1. On the left, are frames from some of the sequences, also examples of humans. The humans are only around 24x14 pixels in size, and are difficult to distinguish from the background. On the right, still image shadow detection using techniques from [11], pixels belonging to humans, and large parts of background were incorrectly labelled as gradient that belongs to shadow.

This is once again done based on projected directions, as well as the ratio of assumed projected human height and projected shadow length.

Once the refined set of candidates has been obtained, we extract wavelet features from each human candidate, and classify it as either human or clutter using a Support Vector Machine (SVM). Note that the main idea behind our geometric constraints is to improve the performance of any detection method by avoiding full frame search. Hence other models, features, and classification schemes suitable for aerial imagery can be used. Additionally, our method can be used to alleviate object localization problems associated with motion detection in presence of strong shadow.

The advantage of our constraints is that they do not require motion detection, registration, and tracking, which are time consuming, and can have their own problems. Additionally our method does not suffer degraded performance in presence of strong shadows. A slight disadvantage is that to get the full benefit, a strong shadow is necessary. However the initial set of candidates which we generate without using the shadow still performs better than brute force full-frame search (see section 4).

In absence of metadata, a static image shadow detector can be used to find the shadows in the image. For this purpose we extend the geometry detection method to work as a novel shadow detection method described in section 3.3. We found that standard shadow detection methods such as [11] and [12] perform poorly on real data (see Figure 1). The methods are based on obtaining illumination invariant (shadow-less) images, and comparing edges between these and original images. Since the humans and their shadows look similar in our data, the illumination invariant images would remove parts of shadows, humans and strong background gradients.

The main contribution of this paper is a novel method constraining human detection in aerial video, as well as a shadow detection method. In future work we will extend it to other object types. Our use of shadow is somewhat counterintuitive, since instead of treating it as a nuisance, we actually use it to help with the detection.

2 Ground-Plane Normal and Shadow Constraints

2.1 Metadata

The imagery obtained from the UAV has the following metadata associated with most of the frames. It has a set of aircraft parameters *latitude*, *longitude*, *altitude*, which define the position of the aircraft in the world, as well as *pitch*, *yaw*, *roll* which define the orientation of the aircraft within the world. Metadata also contains a set of camera parameters *scan*, *elevation*, *twist* which define the rotation of the camera with respect to the aircraft, as well as *focal length*, and *time*. We use this information to derive a set of world constraints, and then project them into the original image.

2.2 World Constraints

The Shadow is generally considered to be a nuisance in object detection, and surveillance scenarios. However, in the case of aerial human detection, the shadow information augments the lack of visual information from the object itself, especially in the cases where the aerial camera is close to being directly overhead. We employ three world constraints.

- The person is standing upright perpendicular to the ground plane.
- The person is casting a shadow.
- There is a geometric relationship between person's height and the length of their shadow. See Figure 2.

Given *latitude*, *longitude*, and *time*, we use the algorithm described in [13], to obtain the position of the sun relative to the observer on the ground. It is defined by the azimuth angle α (from the north direction), and the zenith angle γ (from the vertical direction). Assuming that the height of the person in the world is k we find the length of the shadow as $l = \frac{k}{\tan(\gamma - 90)}$, where γ is the zenith angle of the sun. Using the azimuth angle α we find the groundplane projection of the vector pointing to the sun, and scale it with the length of the shadow $\mathbf{S} = \langle l\cos(\alpha), l\sin(\alpha), 0\rangle$.

2.3 Image Constraints

Before we can use our world constraints for human detection, we have to transform them from the world coordinates to the image coordinates. To do this we use the metadata to obtain the projective homography transformation that relates image coordinates to the ground plane coordinates. For an excellent review of the concepts used in this section see [14].

We start by converting the spherical *latitude* and *longitude* coordinates of the aircraft to the planar Universal Transverse Mercator coordinates of our world $X_w = east$, and $Y_w = north$. Next, we construct a sensor model that transforms any image point $\mathbf{p'} = (x_i, y_i)$ to the corresponding world point $\mathbf{p} = (X_w, Y_w, Z_w)$. We do this by constructing the following sensor transform.

$$\Pi_1 = T^a_{Zw} T^e_{Xw} T^n_{Yw} R^y_{Zw} R^p_{Xw} R^r_{Yw} R^s_{Za} R^e_{Xa} R^t_{Ya}, \tag{1}$$

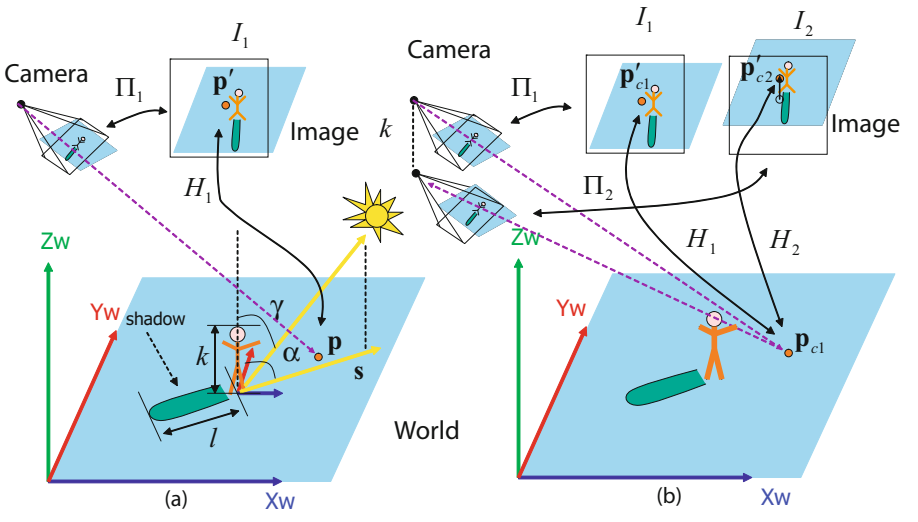


Fig. 2. Left, the sensor model Π_1 maps points in camera coordinates into world coordinates (since the transformation between image and camera coordinates is trivial we do not show it in the image).**X** corresponds to East direction, **Y** to North, **Z** to vertical direction. Vector **S** is pointing from an observer towards the sun along the ground. It is defined in terms of α - azimuth angle between northern direction and the sun. Zenith angle γ is between vertical direction and the sun. The height of a human is k, and the length of the shadow is l. We place the image plane into the world, and raytrace through it to find the world coordinates of the image points (we project from the image plane to the ground plane). We compute a homography H_1 between image points and their corresponding world coordinates on groundplane. Right, illustrates how we obtain the projection of the groundplane normal in the original image. Using a lowered sensor model Π_2 we obtain another homography H_2, which maps points in camera coordinates to a plane above the ground plane. Mapping a world point $\mathbf{p}_{c1}$ using H_1, and H_2, gives two image points $\mathbf{p}'_{c1}$, and $\mathbf{p}'_{c2}$. Vector from $\mathbf{p}'_{c1}$ to $\mathbf{p}'_{c2}$ is the projection of the normal vector.

where T^a_{Zw}, T^e_{Xw}, and T^n_{Yw} are translations for aircraft position in the world - *altitude*, *east*, and *north* respectively. R^y_{Zw}, R^p_{Xw}, and R^r_{Yw} are rotations for the aircraft - yaw, pitch and roll respectively. R^s_{Za}, R^e_{Xa} and R^t_{Ya} are rotation transforms for camera - scan, elevation, and tilt, respectively.

We transform 2D image coordinates $\mathbf{p}' = (x_i, y_i)$ into 3D camera coordinates $\hat{\mathbf{p}}' = (x_i, y_i, -f)$, where f is the *focal length* of the camera. Next, we apply the sensor transform from equation 1, and raytrace to the ground plane (see Figure 2 **(a)**).

$$\mathbf{p} = RayTrace(\Pi_1 \hat{\mathbf{p}}'). \tag{2}$$

Ray tracing requires geometric information about the environment, such as the world height at each point, this can be obtained from the digital elevation map

of the area - DEM. In our case, we assume the scene to be planar, and project the points to the ground plane at zero altitude $Z_w = 0$.

For any set of image points $\mathbf{p'} = (x_i, y_i)$, raytraycing gives a corresponding set of ground plane point $\mathbf{p} = (X_w, Y_w, 0)$. Since we are assuming only one plane in the scene we only need correspondences of four image corners. We then compute a homography, H_1, between the two sets of points, such that $\mathbf{p} = H_1\mathbf{p'}$. Homography, H_1, will orthorectify the original frame, and align it with the North Direction. Orthorectification removes perspective distortion from the image and allows the measurement of world angles in the image. We use the inverse of the homography H_1^{-1} to project the shadow vector defined in world coordinates into the image coordinates. (see Figure 4 **(a)**).

$$\mathbf{S'} = \mathbf{S}H_1^{-1}. \tag{3}$$

Now, we obtain the projected ground plane normal (refer to Figure 2 **(b)**). We generate a second sensor model, where we lower the camera along the normal direction Z_w, by k, which is the assumed to be a person's height.

$$\mathit{\Pi}_2 = (T^a_{Zw} - [I|k])T^e_{Xw}T^n_{Yw}R^y_{Zw}R^p_{Xw}R^r_{Yw}R^s_{Za}R^e_{Xa}R^t_{Ya}. \tag{4}$$

Using the above sensor model $\mathit{\Pi}_2$ we obtain a second homography H_2 using the same process that was used for obtaining H_1. We now have two homographies, H_1 maps the points from the image to the ground plane, and H_2 maps the points from the image to a virtual plane parallel to the ground plane that is exactly k units above the ground plane. We select the center point of the image $\mathbf{p'}_{c1} = (x_c, y_c)$, and obtain its ground plane coordinates $\mathbf{p}_{c1} = H_1\mathbf{p'}_c$. Then we map it back to the original image using H_2, $\mathbf{p'}_{c2} = H_2^{-1}\mathbf{p}_c$. The projected normal is then given by

$$\mathbf{Z'} = p'_{c2} - p'_{c1}. \tag{5}$$

We compute the ratio between the projected shadow length and the projected person height as

$$\eta = \frac{|\mathbf{S'}|}{|\mathbf{Z'}|}. \tag{6}$$

3 Human Detection

3.1 Constraining the Search

In order to avoid the search over the entire frame, the first step in our human detection process is to constrain the search space of potential human candidates. We define the search space as a set of blobs oriented in direction of shadow, and direction of normal. To do so we utilize the image projection of the world constraints derived in the previous section - the projected orientation of the normal to the ground plane $\mathbf{Z'}$, the projected orientation of the shadow $\mathbf{S'}$, and the ratio between the projected person height, and projected shadow length η. See Figure 3.

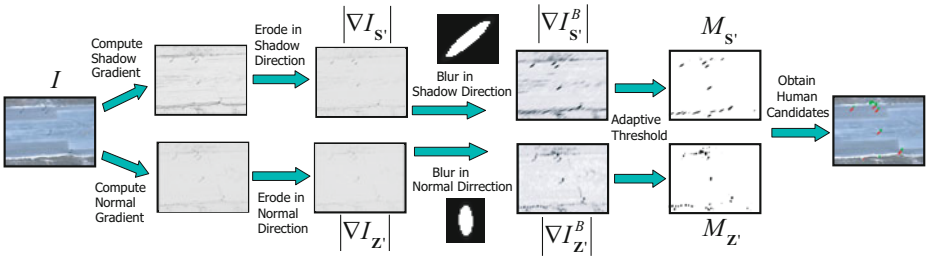


Fig. 3. This figure illustrates the pipeline of applying image constraints to obtain an initial set of human candidates

Given a frame I, we compute gradient oriented in the direction of the shadow by applying a 2D Gaussian derivative filter,

$$G(x,y) = \cos(\theta)2xe^{-\frac{x^2+y^2}{\sigma^2}} + \sin(\theta)2ye^{-\frac{x^2+y^2}{\sigma^2}}, \tag{7}$$

θ is the angle between the vector of interest and the x axis,and take its absolute value. To further suppress gradient not oriented in the direction of the shadow vector we perform structural erosion along a line in the direction of the shadow orientation:

$$|\nabla I_{\mathbf{S'}}| = erode(\nabla I, \mathbf{S'}). \tag{8}$$

We obtain $|\nabla I_{\mathbf{Z'}}|$ using the same process. Next, we smooth the resulting gradient images with an elliptical averaging filter whose major axis is oriented along the direction of interest:

$$I^B_{\mathbf{S'}} = |\nabla I_{\mathbf{S'}}| * G_{\mathbf{S'}}, \tag{9}$$

where $B_{\mathbf{S'}}$ is an elliptical averaging filter, whose major axis is oriented along the shadow vector direction, this fills in the blobs. We obtain $I^B_{\mathbf{Z'}}$ using $G_{\mathbf{Z'}}$. Next, we apply an adaptive threshold to each pixel to obtain shadow and normal blob maps.

$$M_{\mathbf{S'}} = \begin{cases} 1 \text{ if } I^B_{\mathbf{S'}} > t \cdot mean(I^G_{\mathbf{S'}}) \\ 0 \text{ otherwise,} \end{cases} \tag{10}$$

See Figure 4 for resulting blob maps overlayed on the original image. We obtain $M_{\mathbf{Z'}}$ using the same method. From the binary blob maps we obtain a set of shadow and object candidate blobs using connected components. Notice that a number of false shadow and object blobs were initially detected, and later removed.

3.2 Exploiting Object Shadow Relationship

The initial application of the constraints does not take into account the relationship between the object candidates and their shadows, and hence generates many false positives. Our next step is to relate the shadow and human blob maps, and

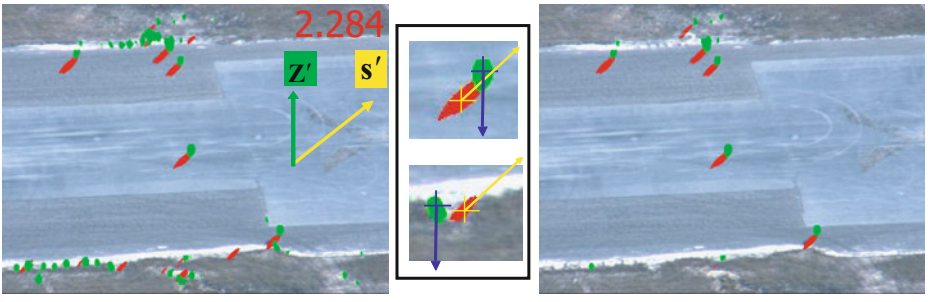


Fig. 4. (a) shows shadow blob map $M_{\mathbf{S}'}$ (shown in red), and normal blob map $M_{\mathbf{Z}'}$ (shown in green), overlayed on the original image. Notice there are false detections at the bottom of the image. Yellow arrow is the projected sun vector $\mathbf{S}'$, the projected normal vector $\mathbf{z}'$ is shown in green, and the ratio between the projected normal and shadow lengths is 2.284 (b) shows example candidates being refined. A valid configuration of human and shadow blobs (top) results in an intersection of the rays, and is kept as a human candidate. An invalid configuration of blobs (bottom) results in the divergence of the rays, and is removed from the set of human candidates. (c) shows refined blob maps after each normal blob was related to its corresponding shadow blob.

to remove shadow-human configurations that do not satisfy the image geometry which we derived from the metadata. We search every shadow blob, and try to pair it up with a potential object blob, if the shadow blob fails to match any object blobs, it is removed. If an object blob never gets assigned to a shadow blob it is also removed.

Given a shadow blob, $M^i_{\mathbf{S}'}$, we search in an area around the blob for a potential object blob $M^j_{\mathbf{Z}'}$. We allow one shadow blob to match to multiple normal blobs, but not vice versa,since the second case is not very likely to be observed. The search area is determined by major axis lengths of $M^i_{\mathbf{S}'}$ and $M^j_{\mathbf{Z}'}$. For any object candidate blob, $M^j_{\mathbf{Z}'}$ that falls within the search area, we ensure that it is in the proper geometric configuration relative to the shadow blob (see Figure 4 **(b)**) as follows. We make two line segments l^i, and l^j, each defined by two points as follows $l^i = \{c_i, c_i + Q\mathbf{S}'\}$, and $l^j = \{c_j, c_j - Q\mathbf{Z}'\}$. Where c_i, and c_j are centroids of shadow and object candidate blobs respectively, and Q is a large number. If the two line segments intersect, then the two blobs exhibit correct object shadow configuration.

We also check to see if the lengths of the major axes of $M^i_{\mathbf{S}'}$ and $M^j_{\mathbf{Z}'}$ conform to the projected ratio constraint η. If they do then we accept the configuration.

Depending on the orientation of the camera in the scene, it is possible for the person and shadow gradients to have the same orientation. In that case the shadow and object candidate blobs will merge, the amount of merging depends on the similarity of orientations $\mathbf{S}'$ and $\mathbf{Z}'$. Hence, we accept the shadow object pair if

$$\frac{M^i_{\mathbf{S}'} \cap M^j_{\mathbf{Z}'}}{M^i_{\mathbf{S}'} \cup M^j_{\mathbf{Z}'}} > q(1 - abs(\mathbf{S}' \cdot \mathbf{Z}')), \tag{11}$$

where q was determined empirically. For these cases the centroid of the person candidate blob is not on the person. Therefore for these cases we perform localization, where we obtain a new centroid by moving along the shadow vector $\mathbf{S}'$, as follows

$$\tilde{c} = c + \frac{m}{2}(1 - \frac{1}{\eta})\frac{\mathbf{S}'}{\|\mathbf{S}'\|}, \tag{12}$$

where m is the length of the major axis of shadow blob $M^i_{\mathbf{S}'}$.

3.3 Constraints without Metadata

Having all of the metadata, quickly provides a set of strict constraints for a variety of camera angles, and time of day. However, there may be cases when the metadata is either unavailable, or worse, is incorrect. In such cases it is acceptable to sacrifice some of the generality, and computation time to obtain a looser set of constraints that still perform well. Assuming that humans are vertical in the image, and ignoring the ratio between the size of humans and their shadows, we can still exploit the orientation of the shadow in the image, as well as the relationship between humans and their shadows, as described below.

We find the orientation of the shadow in the image in the following manner. We quantize the search space of shadow angle θ between 0° and 360°, in increments of d (we used 5 in our experiments). Keeping the normal orientation fixed, and ignoring shadow to normal ratio, we find all human candidates in image I for every orientation θ using technique described in sections 3.1 & 3.2 (see Figure 5). We track the candidates across different θ. Similar angles θ will detect the same human candidates. Therefore, each human candidate C_i has a set Θ_i for which it was detected, and a set O_i which is a binary vector, where each element corresponds to whether the shadow and human blobs overlapped. Then, the set of orientations for which it was detected due to overlap is Θ^o_i, and the set of orientations for which it was detected without overlap is $\Theta^{\bar{o}}_i$ (see Figure 5). We remove any candidate which has been detected over less than p orientations, since a human is always detected as a candidate if shadow and normal orientations are similar, and the resulting blobs overlap according to equation 11 (as in 5 (b) & (f)). Here p depends on quantization, we found that it should encompass at least 70°.

If there are two or more humans casting shadows on planes parallel to the ground plane (poles will work for the task as well), their orientations will be consistent. We find the optimal shadow orientation $\hat{\theta}$ by treating each $\Theta^{\bar{o}}_i$ as a sequence and then finding the longest common consecutive subsequence β among all $\Theta^{\bar{o}}$. Subsequence β must span at least 20° but no more than 40°. Finally, the optimal orientation $\hat{\theta} = mean(\beta)$. If we cannot find such a subsequence then there are either no shadows, or the orientation of the shadow is the same as the orientation of the normal, so we set $\hat{\theta}$ to our assumed normal. Figure 5

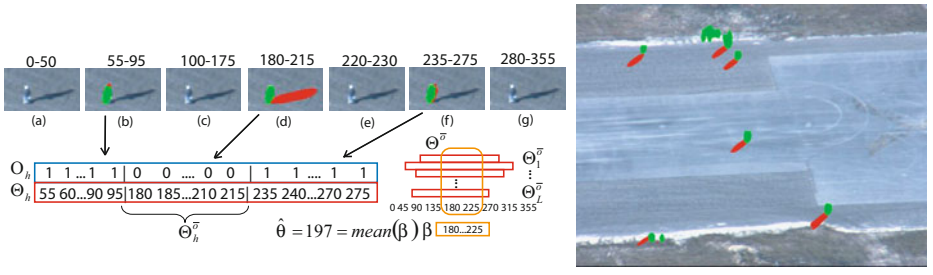


Fig. 5. (The flow chart shows our method for finding optimal shadow orientation for a given image in the absence of metadata. Top row shows human candidate responses obtained for different shadow orientations. A human candidate is then described by a vector of orientations for which it was detected, and a binary overlap vector. Optimal orientation $\hat{\theta}$ is the average of longest common consecutive non-overlapping subsequence of orientations among all human candidates. The image on the rights shows refined human candidate blobs for an automatically estimated shadow orientation of 35°, without metadata. Corresponding metadata derived value of θ for this frame is 46.7°. Blobs that were detected using metadata can be seen in fig. 4.

shows an example frame for which human candidates, were detected using the automatically estimated shadow orientation. There is a 10° difference between estimated orientation, and orientation derived from the metadata. This is the same frame as in Figure 4, qualitative examination of the shadow blobs, seems to indicate that the estimated orientation is more accurate than the one derived from the metadata, however the computation time of obtaining it is much larger. In practice this issue can be dealt with in the following manner. The angle can be estimated in the initial frame, and in subsequent frames it can be predicted and updated using a Kalman filter.

3.4 Object Candidate Classification

Wavelets have been shown to be useful in extracting distinguishing features from imagery. So in the final step of our method, we classify each object candidate as either a human or non-human using a combination of wavelet features and SVM (Figure 6). We chose wavelet features over HOG because we obtained higher classification rate on a validation set. We suspect that this is due to the fact that in the case of HOG, the small size of chips does not allow for the use of optimal overlapping grid parameters reported in [3], giving too coarse sampling. We apply Daubechies 2 wavelet filter to each chip, where the low-pass, and high-pass filters for a 1-D signal are defined as

$$\phi_1(x) = \sqrt{2}\sum_{k=0}^{3} c_k \phi_0(2x-k),\ \psi_1(x) = \sqrt{2}\sum_{k=0}^{3}(-1)^{k+1} c_{3-k}\phi_0(2x-k), \quad (13)$$

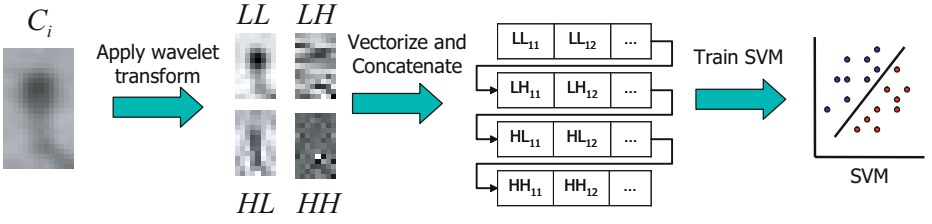


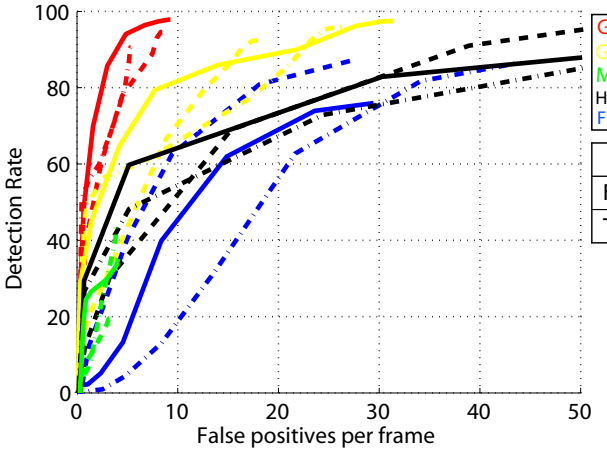
Fig. 6. Object candidate classification pipeline. Four wavelet filters (LL, LH, HL, HH) produce scaled version of original image, as well as gradient like features in horizontal vertical and diagonal directions. The resulting outputs are vectorized, normalized, and concatenated to form a feature vector. These feature vectors are classified using SVM.

here $c = (\frac{(1+\sqrt{(3)})}{4\sqrt{(2)}}, \frac{(3+\sqrt{(3)})}{4\sqrt{(2)}}, \frac{(3-\sqrt{(3)})}{4\sqrt{(2)}}, \frac{(1-\sqrt{(3)})}{4\sqrt{(2)}})$, are the Daubechies 2 wavelet coefficients, and ϕ_0 is either row or column of original image, and . In the case of the 2D image, the 1D filters are first applied along x, and then y directions. This gives to four outputs LL, LH, HL, HH. Where LL is a scaled version of the original image, and LH, HL, and HH, correspond to gradient like features along horizontal, vertical and diagonal directions. We used only one level, since adding more did not improve the performance. We vectorize the resulting outputs, normalize their values to be in the $[0, 1]$ range, and concatenate them into a single feature vector. We train a Support Vector Machine [15] on the resulting feature set using the RBF kernel. We use 2099 positive and 2217 negative examples $w \times h$: 14×24 pixels in size.

During the detection stage, we compute the centroid of the remaining object candidate blobs $M^i_{\mathbf{Z}'}$, extract a $w \times h$ chip around each centroid, extract wavelet features, and classify the resulting vector using SVM. If focal length data is available then the chip size could be selected automatically based on the magnitude, and orientation of the projected normal $|\mathbf{Z}'|$. Note, that this would amount to the use of absolute scale information, which would require a minor change in the geometry portion of the method to account for the effect of perspective distortion. The change amounts to computation of multiple shadow, and normal vector *magnitudes* for different regions of the image. However, since the sequences in the VIVID 3 dataset do not have correct focal length information, the size of the people in the images is approximately the same, and there is generally little perspective distortion in aerial video, we selected the $w \times h$ to be equal to the size of chips in the training set.

4 Results

We performed both qualitative and quantitative evaluation of the algorithm. Qualitative evaluation is shown on sequences from VIVID3 and 4 as well as some of our own data. The data contains both stationary and moving vehicles and people, as well as various clutter in the case of VIVID4. Vehicles cast a shadow, and



	Seq1	Seq2	Seq3
Frames	1191	1006	823
Total People	4892	2000	3098

Fig. 7. SVM confidence ROC curves for sequences 1 (dashed-dotted), 2 (dashed), and 3 (solid). Our Geometry based method with shadow, object-shadow relationship refinement, and centroid localization is shown in red. Yellow curves are for our geometry based method without the use of object-shadow relationship refinement, or centroid localization. A standard full frame detector (HOG) is shown in blue. Green shows results obtained from classifying blobs obtained through registration, motion, detection, and tracking, similar to [8]. Black curves are for our modified implementation of [9], which uses Harris corner tracks.

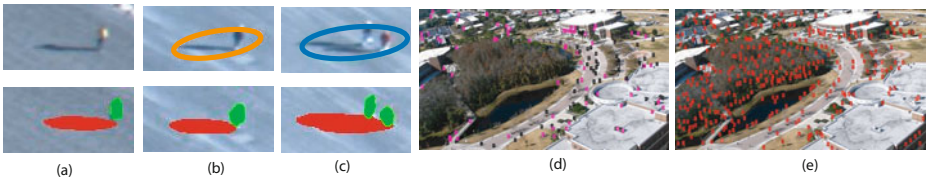


Fig. 8. (a) (b) and (c) compare motion detection (top row), and our geometry based method (bottom row). (a) Human is stationary and was not detected by the motion detector. (b) Moving blob includes shadow, the centroid of blob is not on the person. (c) Two moving blobs were merged by the tracker because of shadow overlap, centroid is not on either person. By contrast our method correctly detected and localized the human candidate (green). (d) and (e) compare geometry constrained human detection, and full frame HOG detection. Human candidates that were discarded by the wavelet classifier as clutter are shown in magenta, candidates that were classified as human are shown in **black**. Unconstrained full frame detection (e) generates many false positives.

are usually detected as candidates, these are currently filtered out in the classification stage, however we plan to extend the geometry method for vehicle detection as well. For quantitative evaluation we evaluated our detection methods on three sequences from the DARPA VIVID3 dataset of 640x480 resolution, and compared the detection against manually obtained groundtruth. We removed the frames where people congregated into groups. We used the following evaluation criteria *Recall* vs False Positives Per Frame (FPPF). Recall is defined as $\frac{TP}{TP+FN}$, where FN is number of false negatives, TP is the number of true positives in the frame. To evaluate the accuracy of the geometry based human candidate detector method, we require the centroid of the object candidate blob to be within w pixels of the centroid blob, where w is 15. We did not use the PASCAL measure of 50% bounding box overlap, since in our dataset the humans are much smaller, and make up a smaller percentage of the scene. In INRIA set inroduced in [3], an individual human makes up 6% of the image, in our case the human makes up about 0.1%. Under these circumstances small localization errors, result in large area overlap difference, hence we feel that the centroid distance measure is more meaningful for aerial data. Figure 7 compares ROC curves for our geometry based method with and without the use of object-shadow relationship refinement, and centroid localization, conventional full frame detection method (we used HOG detection binaries provided by the authors), and standard motion detection pipeline of registration, detection, and tracking. Figure 8 shows qualitative detection results. Conventional full frame detection is not only time consuming, (our MATLAB implementation takes several hours per 640x480 frame), but it also generates many false positives. By contrast preprocessing the image using geometric constraints to obtain human candidates, is not only much faster (6 seconds per frame), but gives far better results. Geometric constraints with the use of shadow based refinement, and centroid localization provide the best performance. However even without these additional steps, the geometric constraint based only on the projection of the normal still give superior results to full frame, as well as motion constrained detection. Motion based detection suffers from problems discussed in section 1, and shown in Figure 8. Which is why the green ROC curves in Figure 7 are very short. We implemented a part of [9] method, where instead of using the OT-Mach filter, we used our wavelet SVM combination for classification. These ROC curves are shown in black. We suspect that the poor performance is caused by poor tracking results. They simply used a greedy approach based on euclidian distance between the corners without any motion model. Therefore if a track contains corners belonging to both people and background, the 20% track length classification heuristic would introduce many false positives.

5 Conclusions

We proposed a novel method for detecting pedestrians in UAV surveillance imagery. This is a difficult problem due to very small size of humans in the image, and a large number of possible orientations. Our method takes advantage of the metadata information provided by the UAV platform to derive a series of geometric constraints, and to project them into the imagery. In cases when metadata is

not available we proposed a method for estimating the constraints directly form image data. The constraints are then used to obtain candidate out of plane objects which are then classified as either human or non-human. We evaluated the method on challenging data from the VIVID 3 dataset, and obtained results superior to both full frame search, motion constrained detection, and Harris tracks constrained detection [9].

Acknowledgement

This research was funded in parts by Harris Corporation and US government.

References

1. Cheng, H., Butler, D., Basu, C.: ViTex: Video to tex and its application in aerial video surveillance. In: CVPR (2006)
2. Xiao, J., Cheng, H., Han, F., Sawhney, H.: Geo-spatial aerial video processing for scene understanding and object tracking. In: CVPR (2008)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1 (2005)
4. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
5. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR (2005)
6. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 69–82. Springer, Heidelberg (2004)
7. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: CVPR (2007)
8. Xiao, J., Yang, C., Han, F., Cheng, H.: Vehicle and person tracking in UAV videos. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) RT 2007 and CLEAR 2007. LNCS, vol. 4625, pp. 203–214. Springer, Heidelberg (2008)
9. Miller, A., Babenko, P., Hu, M., Shah, M.: Person tracking in UAV video. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) RT 2007 and CLEAR 2007. LNCS, vol. 4625, pp. 215–220. Springer, Heidelberg (2008)
10. Bose, B., Grimson, E.: Improving object classification in far-field video. In: CVPR (2004)
11. Xu, L., Qi, F., Jiang, R.: Shadow removal from a single image. Intelligent Systems Design and Applications 2 (2006)
12. Finlayson, G., Hordley, S., Lu, C., Drew, M.: On the removal of shadows from images. IEEE PAMI 28 (2006)
13. Reda, I., Anreas, A.: Solar position algorithm for solar radiation applications. NREL Report No. TP-560-34302 (2003)
14. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004), ISBN: 0521540518
15. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`