Noname manuscript No.
(will be inserted by the editor)

Shadow Casting Out Of Plane (SCOOP) Candidates for Human and Vehicle Detection in Aerial Imagery

Vladimir Reilly · Berkan Solmaz · Mubarak Shah

Received: date / Accepted: date

Abstract In this paper, we propose a method for detecting humans and vehicles in imagery taken from a UAV. This is a challenging problem due to a limited number of pixels on target, which makes it more difficult to distinguish objects from background clutter, and results in much larger search space. We propose a method for constraining the search based on a number of geometric constraints obtained from the metadata. Specifically, we obtain the orientation of ground plane normal, the orientation of shadows cast by out of plane objects in the scene, and the relationship between object heights and the size of their corresponding shadows. We use the aforementioned information in a geometry-based shadow, and ground-plane normal blob detector, which provides an initial estimation for locations of shadow casting out of plane (SCOOP) objects in the scene. These SCOOP candidate locations are then classified as either human or clutter using a combination of wavelet features and a Support Vector Machine. To detect vehicles, we similarly find potential vehicle candidates by combining SCOOP and inverted-SCOOP candidates and then classify them using wavelet features and SVM. Our method works on a single frame, and unlike motion detection based methods, it bypasses the entire pipeline of registration, motion detection, and tracking. This method allows for detection of stationary and slowly moving humans and vehicles while avoiding the search across the entire image, allowing accurate and fast localization. We show impressive results on sequences from VIVID and CLIF datasets and provide comparative analysis.

Keywords human detection · vehicle detection · aerial surveillance · UAV · shadow · metadata

1 Introduction

Every year Unmanned Aerial Vehicles, or UAVs, are becoming more widespread in both military and civilian applications, including surveillance, rescue, and reconnaissance [36] [37] [26]. In the course of these operations video data containing useful information is collected. This information may be useful during the mission itself or may become useful at a later date. The ever-increasing number of UAV missions equates to a backlog of data which becomes quite large; thus requiring too many man-hours to analyze manually. This calls for automated video analysis tools whose capabilities include registration [40], object detection [16], tracking [37] [41], classification [38], and scene and event analysis [8] [17]. In this paper, we will focus on detecting pedestrians and vehicles in UAV imagery. This,however, is a challenge. Problems include smaller object sizes, varying orientations, motion blur, and camera motion.

One straightforward approach to this problem, is to apply a state-of-the-art static frame detection algorithm such as [9] [10] [18] [21] [29]. This approach, however, runs into the problem of small object size, which may make it impossible to construct a meaningful model. Methods that perform parts detection explicitly, such as [10] [21] [7] [34] [31], will not be able to construct meaningful models for individual parts at resolutions of just 24x14. Bag-of-feature methods, such

Vladimir Reilly
University of Central Florida, Orlando
E-mail: vsreilly@eecs.ucf.edu

Berkan Solmaz
University of Central Florida, Orlando
E-mail: bsolmaz@eecs.ucf.edu

Mubarak Shah
University of Central Florida, Orlando
E-mail: shah@eecs.ucf.edu

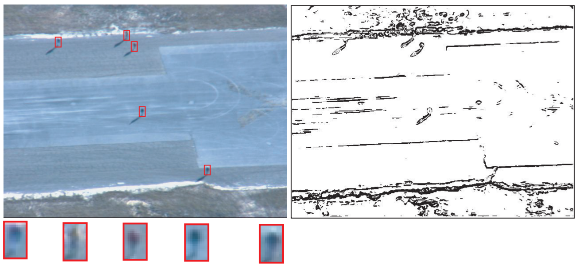

Fig. 1 On the left, a frame from one of the sequences, below it are examples of humans present in the frame. The humans are only around 24x14 pixels in size and are difficult to distinguish from the background. On the right, gradients belonging to shadow labelled using techniques from [39]. Gradients belonging to humans and large parts of background were incorrectly labelled as gradient belonging to shadow.

as [18], also have difficulty constructing models because only a small number of interest points can be found.

The objects are so small that even holistic methods such as [32] [9] [33] [2] will have issues with extracting sufficient discriminative information. Another issue introduced by the small object size is the need to process a very large number of windows across an entire image. This obviously increases processing time and generates many false positives, especially if the object model is not sufficiently discriminative. The above problems can be further compounded by motion blur and varied orientation of objects within the scene.

These issues of size, performance, and speed have prompted researchers working specifically on object detection in aerial video to limit their search space using the following constraints: motion [3] [38] [37], feature points [22], multi-modal techniques [28] [12], semantic constraints [8] [37], metadata derived scale constraints [4], and 3D reconstruction [36].

The authors of [38] assume that only moving objects are of interest and adopt a standard aerial surveillance pipeline. The authors compensate for global camera motion before detecting moving objects. They then classify each moving object as either a person or a vehicle using a combination of histograms of oriented gradients (HOG) and SVM proposed in [9]. Motion constraint can be a problem in the case of human detection; since people are viewed from far away their motion is subtle and difficult for the system to pick up. Of course if people are stationary, then the system cannot detect them at all. For vehicles, motion constraint may not be a reliable cue either since targeted vehicles can remain stationary, for example at an intersection, for a long time.

Rudol and Doherty use an infrared camera in [28] to constrain the search for humans that are sitting or prone. They first threshold the infrared imagery to isolate areas of human body heat then projected those areas to a color camera. Then they applied a cascaded ADABOOST classifier to those areas to detect humans. Unfortunately, this process requires mounting two cameras on the UAV;in this case,reliable detection of body heat requires expensive high-quality infra red cameras.

Gaszczak et. al. also use an infrared camera in [12] both for human detection and vehicle heat-signature confirmation. Initial human detections were found in the infrared camera using cascaded ADABOOST and then refined using a generative shape model. Initial vehicle detections on the other hand, were obtained in EO imagery using cascaded ADABOOST and then verified by region-growing of hot spots in the IR camera.

Cascaded ADABOOST is also applied to EO imagery by Breckon et. al. [4] to detect vehicles. The initial detections obtained were then refined by using the height and field of view of the UAV to filter out detections which did not conform to proper vehicle sizes.

Skokalski, and Breckon suggest a framework for detecting salient objects in color aerial imagery [30]. Given an input frame they extracted nine feature images including a contrast map of the meanshift image and various normalized color channels. Next, they applied edge-detection and gradient operators to each then combined them using AND and OR operators. The contrast map is weighted by the inverse of the probability of belonging to the global color distribution of the image. While the method shows impressive results without relying on metadata, it requires color information for most of its features. Conversely, in grey-scale imagery with non-uniform background it would lose most of it's discriminative ability.

We propose a very different approach. In this paper, we explore the idea of constraining the search by using shadows cast by humans and vehicles as a form of saliency. Objects themselves are small and lack distinguishing features making them difficult to detect. However, if an object casts a shadow, it provides an additional source of gradient information that can be exploited. In addition, the relationship between the object and its shadow serves as an additional que as to the location of the object. This allows us to construct a detection method that works on a single grayscale frame obtained from an EO sensor, avoiding motion detection, registration, and tracking.

Relying on shadows may seem like an overly stringent requirement, however in the case of motion detection and tracking, the shadows would have to be explicitly processed anyways since their presence causes a number of problems. It is difficult to localize humans within moving blobs since their shadows are part of the moving blobs. Shadows also make the blobs more simi-

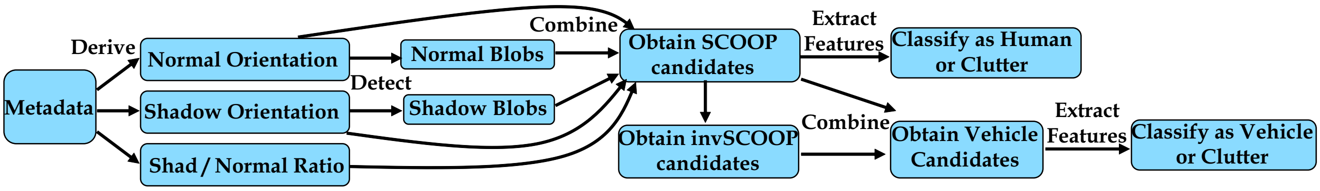

Fig. 2 Overall pipeline of our system. First, we use metadata to derive geometric constraints. Second, we find normal and shadow blobs in the image. Third, we combine blobs using geometric constraints into SCOOP candidates. Fourth, we combine blobs using geometric constraints into inverted SCOOP candidates. Fifth, we combine SCOOP and inverted SCOOP into vehicle candidates. Finally, each SCOOP candidate we extract wavelet features and classify it as human or clutter; from every vehicle candidate we extract wavelet features and classify it as human or clutter.

lar to each other making it more difficult for the tracker to differentiate them when they overlap. See Figure 18 for examples of these failures. Hence, there is a lot of work on detecting, and removing shadows from moving objects in a surveillance scenario for both people [20] [6] [1] [24], and vehicles [14] [19] [35] [15] [42] (See [25] for a survey of these approaches). Our primary focus is not to remove shadows from moving blobs, though our method can be applicable, rather to use shadow for human and vehicle detection in static imagery.

Our approach is to constrain the search by assuming that humans are upright shadow casting objects and vehicles are box-like shadow casting objects. We use low level computer vision techniques based on a set of geometric scene constraints derived from the metadata of the UAV platform. Specifically, we use the projection of the ground-plane normal to find blobs normal to the ground-plane. These blobs give us an initial set of potential out-of-plane object candidates. Similarly, we use the projection of shadow orientation to obtain a set of potential shadow candidates. We then obtain a refined set of what we call Shadow Casting Out Of Plane (SCOOP) candidates. SCOOP candidates are pairs of shadow and normal blobs that are of correct geometric configuration and relative size. Once the refined set of candidates has been obtained, we extract wavelet features from each SCOOP candidate and classify it as either human or clutter using a Support Vector Machine (SVM). In addition to obtaining regular SCOOP candidates we can also obtain inverted SCOOP candidates, where we reverse assumed directions of normal and shadow. We then combine SCOOP and inverted SCOOP candidates into vehicle candidates, which we then classify as vehicle or clutter.

In absence of metadata, a single image shadow detector (such as [11] [39] [23]) 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 6. We found that standard shadow detection methods such as [39] and [11] perform poorly on real data (see Figure 1). These methods are based on obtaining illumination invariant, or shadow-less images, and comparing edges between these and original images. Since humans and their shadows look similar in our data, the shadow-less images would remove parts of shadows, humans, and strong background gradients.

The main contribution of this paper is a novel method for detecting shadows in UAV imagery, and using those shadows to obtain candidate SCOOP objects to aid the detection of humans and vehicles. Our method has the following assumptions and operational requirements.

First, we assume that objects in the scene are casting a shadow. If the weather prevents the casting of shadow, then human detection can still proceed, where only the blobs related to the ground-plane normal are used as a constraint at a cost of degraded performance. In the case of vehicles, the detection can still proceed as if the shadow was there, but will also result in degraded performance.

Second, we assume that humans are sufficiently upright for their shadow to be visible. The human can be standing, or sitting down, but not lying on the ground.

Third, we use metadata associated with the UAV imagery in order to determine the shadow orientation in the world. We also use it to find the orientation of ground-plane normal and shadow in the imagery, as well as their relative size. Errors in the metadata result in incorrect orientations in the image. However, the SCOOP detection method has approximately $+10° - 10°$ robustness to errors in orientation. If metadata is not available we provide a method to determine the shadow orientation automatically assuming that the ground-plane normal is fixed.

Fourth, we make a planar scene assumption when we determine the orientation of the shadow in the world and when we project it into the image. When the scene is not planar, this assumption has an effect equivalent to incorrect metadata, where the shadow's orientation and/or length in the image is not correct. However, since our method is resistant to orientation errors, its performance will degrade gradually as the surface

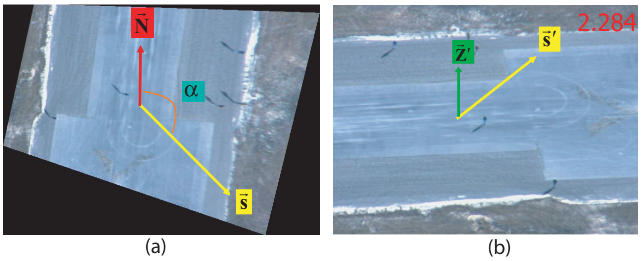

Fig. 3 **(a)** Orthorectified frame from one of the sequences. The vertical direction is aligned with the world north direction **N**. The sun vector **S** is defined by azimuth angle α between the north vector and the vector pointing to the sun. **(b)** is the original frame showing the projected sun vector $\mathbf{S'}$, the projected normal vector $\mathbf{z'}$, and the ratio between the projected normal and shadow lengths 2.284.

roughness increases. Note that if a Digital Elevation Map (DEM) is available, we do not have to make this assumption. Rather, we can explicitly compute different shadow orientations for different surfaces in the scene and regions of the image. The automated orientation detection method requires two shadow casting objects to be present on a plane in order to find the correct shadow orientation for that plane.

2 Ground-Plane Normal and Shadow Constraints

In this section we describe how metadata of the UAV can be used to define a set of constraints in the world, as well as how we can project those constraints into the image space.

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*, and *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 and vehicle detection, the shadow information augments the lack of visual information from the object, especially in the cases when the aerial camera is almost directly overhead. For human detection we define three world constraints:

- The person is standing upright.
- The person is casting a shadow.
- There is a geometric relationship between person's height and the length of their shadow (see Figure 4).

For vehicle detection, the constraints are:

- The vehicle is a box-like object.
- There is a geometric relationship between the boundaries of the vehicle and the boundaries of its shadow.

Given the *latitude*, *longitude*, and *time* of day, we use the algorithm described in [27] 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

$$l = \frac{k}{\tan(\gamma - \pi/2)}, \quad (1)$$

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 [13].

We start by converting the spherical *latitude* and *longitude* coordinates of the aircraft to the planar Universal Transverse Mercator coordinates of our world $X_w = \mathit{east}$ and and $Y_w = \mathit{north}$. 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

$$\Pi_1 = T^a_{Zw} T^e_{Xw} T^n_{Yw} R^y_{Zw} R^p_{Xw} R^r_{Yw} R^s_{Za} R^e_{Xa} R^t_{Ya}. \quad (2)$$

Matrices T^a_{Zw}, T^e_{Xw}, and T^n_{Yw} are translations for aircraft position in the world: *altitude*, *east*, and *north* respectively. Matrices R^y_{Zw}, R^p_{Xw}, and R^r_{Yw} are rotations

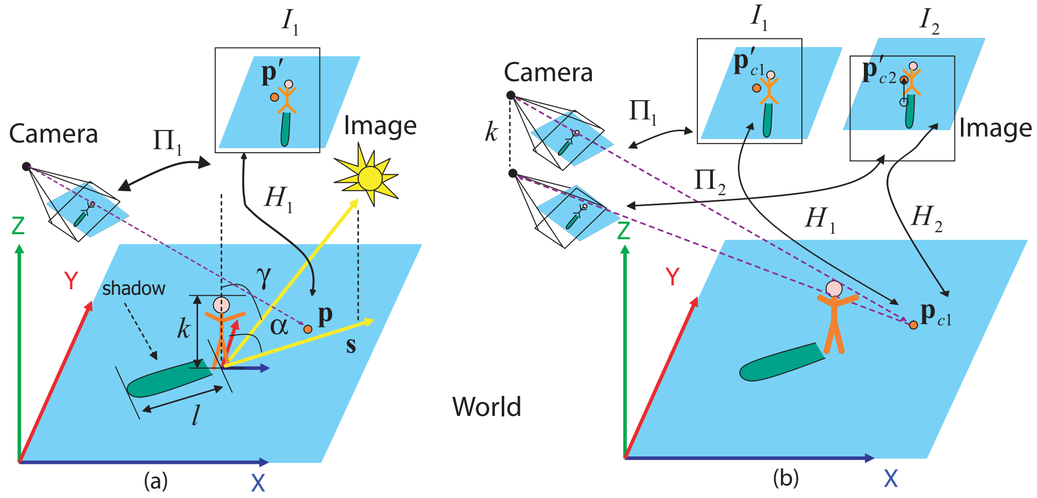

Fig. 4 On the left, the sensor model Π_1 maps points in camera coordinates into world coordinates (since the transformation between image and camera coordinates is trivial we do not show it in the image).$\mathbf{X}$ corresponds to East direction, $\mathbf{Y}$ to North, $\mathbf{Z}$ to vertical direction. Vector $\mathbf{S}$ 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.

for the aircraft: yaw, pitch and roll respectively. Matrices 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 2 and raytrace to the ground plane (see Figure 4 **(a)**)

$$\mathbf{p} = RayTrace(\Pi_1 \hat{\mathbf{p}}'). \tag{3}$$

Ray tracing requires geometric information about the environment, such as the world height at each point. This information 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 $\mathbf{p}' = (x_i, y_i)$, ray tracing gives a corresponding set of ground plane points $\mathbf{p} = (X_w, Y_w, 0)$. Since we are assuming that only one plane exists 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 (see Figure 3 **(a)**). Orthorectification removes perspective distortion from the image and allows for 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 3 **(b)**).

$$\mathbf{S'} = \mathbf{S}H_1^{-1}. \tag{4}$$

Next, we obtain the projected ground plane normal (refer to Figure 4 **(b)**). We generate a second sensor model,

$$\Pi_2 = (T^a_{Zw} - [I|k])T^e_{Xw}T^n_{Yw}R^y_{Zw}R^p_{Xw}R^r_{Yw}R^s_{Za}R^e_{Xa}R^t_{Ya}, \tag{5}$$

where we lower the camera along the normal direction Z_w, by k, which is the assumed height of the person.

Using the above sensor model Π_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

$$\mathbf{Z'} = p'_{c2} - p'_{c1}. \tag{6}$$

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

$$\eta = \frac{|\mathbf{S'}|}{|\mathbf{Z'}|}. \tag{7}$$

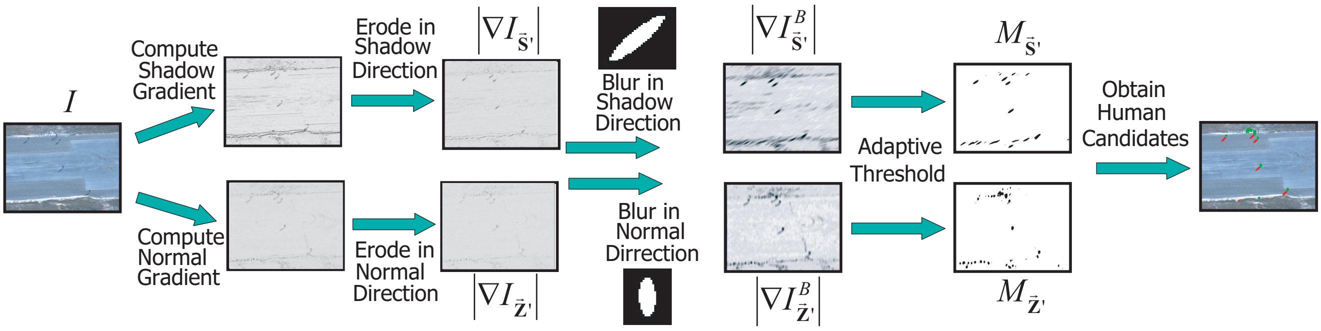
I
Compute Shadow Gradient
Erode in Shadow Direction
$|\nabla I_{\vec{\mathbf{S}}'}|$
Blur in Shadow Direction
$|\nabla I^B_{\vec{\mathbf{S}}'}|$
$M_{\vec{\mathbf{S}}'}$
Adaptive Threshold
Obtain Human Candidates
Compute Normal Gradient
Erode in Normal Direction
$|\nabla I_{\vec{\mathbf{Z}}'}|$
Blur in Normal Dirrection
$|\nabla I^B_{\vec{\mathbf{Z}}'}|$
$M_{\vec{\mathbf{Z}}'}$

Fig. 5 The pipeline of utilizing image constraints to obtain an initial set of human and normal blobs, by applying a series of oriented filters to the original image.

3 Human Detection

Now that the world constraints have been projected into the image, we can avoid searching over the entire frame and instead search the space of potential object candidates. We define the search space as a set of pairs of blobs oriented in the direction of shadow and direction of normal. These combinations of normal and shadow blobs represent a set of shadow casting out of plane (SCOOP) candidates which can belong to humans, vehicles, or background. In the case of human detection, a single SCOOP candidate makes for a sufficient human candidate. Since vehicles are more complex objects, we have to use the combination of a SCOOP and an inverted SCOOP in order to obtain a vehicle candidate.

3.1 Detecting Shadow and Normal Blobs

The first step of human detection is to detect blobs that potentially belong to out of plane objects, and blobs belonging to shadows. To do so, we use 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 height of the person, and projected shadow length η (see Figure 5). This image contains gradients oriented in many different directions, therefore we employ directed filters to enhance gradients oriented in the directions of interest while suppressing gradients oriented in other directions.

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

$$G(x,y) = \cos(\theta) 2x e^{-\frac{x^2+y^2}{\sigma^2}} + \sin(\theta) 2y e^{-\frac{x^2+y^2}{\sigma^2}}, \qquad (8)$$

and take the absolute values of its responses. In the above equation θ is the angle between the vector of interest and the x axis. To further suppress gradients not oriented in the direction of the shadow vector we perform structural erosion along a line in the direction of the shadow orientation

$$|\nabla I_{\mathbf{S}'}| = erode(\nabla I, \mathbf{S}'). \qquad (9)$$

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:

$$I^B_{\mathbf{S}'} = |\nabla I_{\mathbf{S}'}| * G_{\mathbf{S}'}, \qquad (10)$$

where $B_{\mathbf{S}'}$ is an elliptical averaging filter, whose major axis is oriented along the shadow vector direction. This process 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

$$M_{\mathbf{S}'} = \begin{cases} 1 \text{ if } I^B_{\mathbf{S}'} > t \cdot mean(I^G_{\mathbf{S}'}) \\ 0 \text{ otherwise.} \end{cases} \qquad (11)$$

See Figure 6 for resulting blob maps overlaid 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 blobs using connected components. Notice from Figure 6 that a number of false shadow and object blobs were initially detected. In the next section, we describe how to remove those false positives by combining normal and shadow blobs into SCOOP candidates.

3.2 Exploiting Object Shadow Relationship to generate SCOOP candidates

The initial application of the constraints does not take into account the relationship between the normal blobs and their shadows, hence generating many false positives. Our next step is to relate the shadow and normal blob maps. We obtain a set of SCOOP candidates and remove shadow-normal configurations that do not

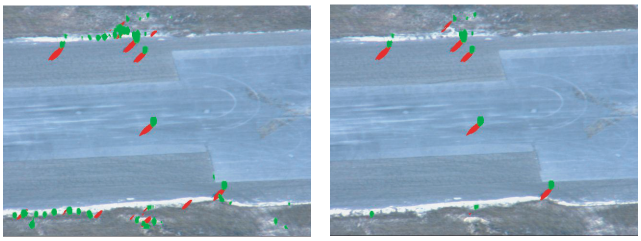

Fig. 6 On the left, a shadow blob map $M_{\mathbf{S}'}$ (shown in red) and normal blob map $M_{\mathbf{Z}'}$ (shown in green) overlayed on the original image. There are false detections at the bottom of the image. On the right, refined blob maps after each normal blob was related to its corresponding shadow blob. The false detection at the bottom are now gone.

satisfy the image geometry which we derived from the metadata. We search every shadow blob, trying 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 for a single shadow blob to be assigned to multiple normal blobs, but not vice versa since the second case is rarely 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 7) 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 normal 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

$$\frac{M^i_{\mathbf{S}'} \cap M^j_{\mathbf{Z}'}}{M^i_{\mathbf{S}'} \cup M^j_{\mathbf{Z}'}} > q(1 - abs(\mathbf{S}' \cdot \mathbf{Z}')), \tag{12}$$

where q was determined empirically. For these cases, the centroid of the object 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

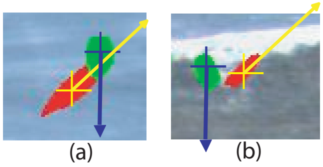

Fig. 7 **(a)** A valid configuration of normal and shadow blobs results in an intersection of the rays and is kept as a SCOOP candidate. **(b)** An invalid configuration of blobs results in the divergence of the rays and is removed from the set of SCOOP candidates.

$$\tilde{c} = c + \frac{m}{2}(1 - \frac{1}{\eta})\frac{\mathbf{S}'}{\|\mathbf{S}'\|}, \tag{13}$$

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

3.3 Obtaining Human Candidates

The outlined procedures generate the final set of SCOOP candidates $\mathbf{K}^{\mathbf{Z}'}_{\mathbf{S}'} = \{\mathbf{k}^{\mathbf{Z}'}_{\mathbf{S}'1}, ..., \mathbf{k}^{\mathbf{Z}'}_{\mathbf{S}'n}\}$, where each candidate $\mathbf{k}^{\mathbf{Z}'}_{\mathbf{S}'} = \{M^i_{\mathbf{S}'}, M^j_{\mathbf{Z}'}\}$ is a pair of normal and shadow blobs that were detected at normal orientation $\mathbf{Z}'$ and shadow orientation $\mathbf{S}'$. Since a single SCOOP candidate is sufficient to capture a human in low resolution aerial video, $\mathbf{K}^{\mathbf{Z}'}_{\mathbf{S}'}$ is the set of human candidates where we classify each normal blob $M_{\mathbf{Z}'}$ as human or clutter.

3.4 Classifying Human Candidates

The final step of human detection is to classify each SCOOP candidate $\mathbf{k}^{\mathbf{Z}'}_{\mathbf{S}'}$ as either human or clutter. For this purpose, we compute the centroid of the normal blob $M^i_{\mathbf{Z}'}$ of each remaining SCOOP candidate, and extract a $w \times h$ chip around that centroid. We then extract wavelet features from each chip and apply a Support Vector Machine (SVM) classifier (see Figure 8). We use the Daubechies 2 wavelet filter, where the low-pass (L) and high-pass (H) filters for a 1-D signal are defined as

$$\phi_1(x) = \sqrt{2}\sum_{k=0}^{3} c_k \phi_0(2x - k), \tag{14}$$

$$\psi_1(x) = \sqrt{2}\sum_{k=0}^{3} (-1)^{k+1} c_{3-k} \phi_0(2x - k), \tag{15}$$

where ϕ_0 is either row or column of the original image and $c = (\frac{(1+\sqrt{(3)})}{4\sqrt{(2)}}, \frac{(3+\sqrt{(3)})}{4\sqrt{(2)}}, \frac{(3-\sqrt{(3)})}{4\sqrt{(2)}}, \frac{(1-\sqrt{(3)})}{4\sqrt{(2)}}),$

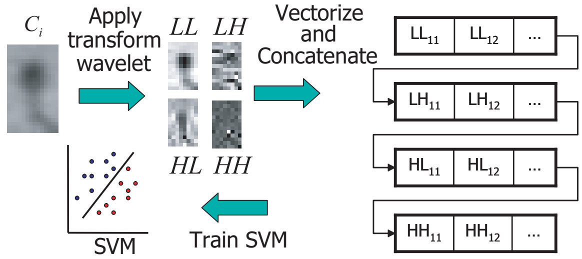

Fig. 8 Object candidate classification pipeline. Four wavelet filters (LL, LH, HL, HH) produce scaled version of original image as well as gradient like features in horizontal vertical and diagonal directions. The resulting outputs are vectorized, normalized, and concatenated to form a feature vector. These feature vectors are classified using SVM.

are the Daubechies 2 wavelet coefficients. In the case 2D signals, such as images, the 1D filters are first applied along x, and then y directions. This produces 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 [5] on the resulting feature set using the RBF kernel. We use 2099 positive and 2217 negative examples $w \times h$ pixels in size.

Note that if focal length data is available, then the chip size could be selected automatically based on the magnitude and orientation of the projected normal $|\mathbf{Z'}|$. Additionally, if perspective distortion in the imagery is fairly strong, we would have to assume different sizes of humans and shadows for different regions of the image, which would require a minor change in the geometric part of the method. The change would include computing multiple shadow and normal vector *magnitudes* for different regions of the image. Since there is little perspective distortion in VIVID, no drastic zoom changes within a video, and the focal length information provided is not correct, we kept $w \times h$ constant over the entire video. We selected $w \times h$ to be 24×14 equal to the size of images in the training set.

4 Vehicle Detection

Since vehicles are larger, and more complex objects than humans (at typical aerial surveillance resolutions), a single SCOOP candidate is too simple to serve as a vehicle candidate. While the vehicle does generate a SCOOP (see Figure 9 **(b)**), the normal blob of the SCOOP (shown in green) captures only a small part of the vehicle without giving a clear idea of the size of the vehicle or the location of its centroid. This makes it difficult to localize the vehicle for actual classification. Additionally, if our goal is to minimize the number of classifications that we want to perform, then treating every SCOOP candidate as a potential vehicle is less than optimal since SCOOPs can be generated by humans, poles, or simply background gradients. As can be seen in Figure 9 **(a)**, a vehicle will have at least two shadow and two normal blobs associated with it. Therefore, rather than using a single SCOOP (one shadow and one normal blob) in order to obtain vehicle candidates we fuse two shadow and two normal blobs in the following manner.

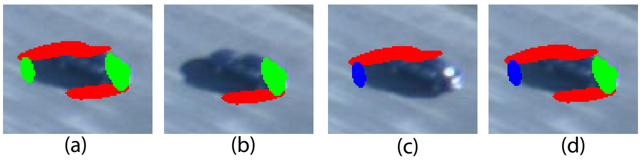

Fig. 9 **(a)** All shadow and normal blobs obtained using method described in section 3.1. **(b)** SCOOP candidate detected using the method described in section 3.2 at normal orientation $\mathbf{Z}'$ and shadow orientation $\mathbf{S}'$. **(c)** inverted SCOOP candidate that was detected at normal orientation $-\mathbf{Z}'$ and shadow orientation $-\mathbf{S}'$. **(d)** A car candidate assembled from a SCOOP and an inverted-SCOOP. candidates.

4.1 Obtaining Vehicle Candidates

In addition to obtaining a set of SCOOP candidates $\mathbf{K}^{\mathbf{Z}'}_{\mathbf{S}'}$ as described in section 3.2, we use the exact same process in obtaining a set of inverse-SCOOP candidates $\mathbf{K}^{-\mathbf{Z}'}_{-\mathbf{S}'}$, where we negate the direction of the normal $\mathbf{Z}'$ and the direction of the shadow $\mathbf{S}'$. Next, we combine the regular and inverted-SCOOP candidates.

For every scoop candidate $\mathbf{k}^{\mathbf{Z}'}_{\mathbf{S}'} = \{M^i_{\mathbf{S}'}, M^j_{\mathbf{Z}'}\}$, we search for the closest inverted SCOOP candidate $\mathbf{k}^{-\mathbf{Z}'}_{-\mathbf{S}'} = \{M^k{}_{-\mathbf{S}'}, M^k{}_{-\mathbf{Z}'}\}$. The distance between the candidates is defined as the Euclidian distance between the points e_1 and e_2 (see Figure 10 **(a)**). Where e_1 is a point on shadow blob $M^i_{\mathbf{S}'}$, of SCOOP candidate $\mathbf{k}^{\mathbf{Z}'}_{\mathbf{S}'}$ that is closest to the corresponding normal blob $M^j_{\mathbf{Z}'}$, and e_2 is a point on shadow blob $M^k{}_{-\mathbf{S}'}$ of inverse-SCOOP candidate $\mathbf{k}^{-\mathbf{Z}'}_{-\mathbf{S}'}$, that is furthest from its corresponding normal blob $M^l{}_{-\mathbf{Z}'}$. Since we want the SCOOP candidates to enclose the car, we impose an additional constraint that the orientation of the vector between the centroids of the shadow blobs $M^j_{\mathbf{Z}'}$ and $M^k{}_{-\mathbf{S}'}$ must be at least 30° different from the orientation of the shadow blobs. Since this gives better localization and excludes a large part

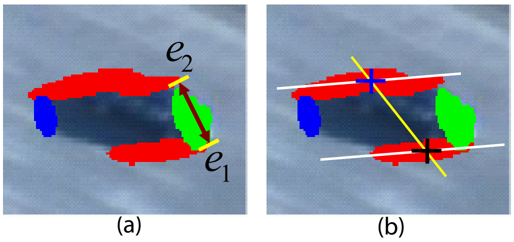

Fig. 10 **(a)** The distance computation between the SCOOP and inverted SCOOP candidates. The distance is computed between points e_1 and e_2. **(b)** The orientation constraint between the SCOOP and inverted SCOOP candidates that must be satisfied for their configuration to be considered a valid car candidate.

of the shadow, a vehicle candidate is represented by a bounding box that encompasses both the SCOOP and the centroid of the shadow blob of the inverse SCOOP.

A single vehicle may have multiple SCOOP pairs detected on it. Distractions like specular reflections or decals can create multiple SCOOP and inverse-SCOOP candidates for a single vehicle. Therefore, as a final processing step we merge bounding boxes that have more that 50% overlap given by the following formula

$$\frac{A \cap B}{min(A,B)}, \tag{16}$$

where A and B are areas of two vehicle candidate bounding boxes. Dividing by the minimum of the two areas makes it easier to merge boxes. In cases where a small bounding box is enclosed within a larger one, it will be merged. We perform the merge procedure recursively until no additional bounding boxes can be merged.

4.2 Classifying Vehicle Candidates

As with humans, the final stage of vehicle detection is classifying each vehicle candidate as either vehicle or clutter. Since vehicle candidates encompass the entire object, we simply extract the region within the candidate's bounding box and resize it to 40 × 40. Next we extract Daubechies 2 wavelet coefficients and classify them using SVM. The vehicle training set consisted of 1900 positive and 1889 negative examples. We included vehicles at different orientations in the positive set.

5 Handling Multiple Scales

Scale information is required at two stages of our method. First, it is needed at the candidate detection stage for setting the parameters of the various image processing masks from pipeline 5 as well as proximity settings in the refinement stage. Second, it is needed at the classification stage for extracting chips of correct sizes. If metadata is available, detailed, and accurate, then the scale can be trivially extracted from it. If metadata is not available, then different scales can be handled as follows. First, we apply the candidate detector at an assumed scale and obtain a set of candidates. The size of the candidates obtained determines then determines the scale for the classification stage. We perform this process for several scales and then select the scale for which the classifier gave maximum confidence. The increments in scale of the candidate detection method can be coarse, because as we show in figure 26, the candidate detector method is robust to error in scale from 0.5× to 2× of the original image.

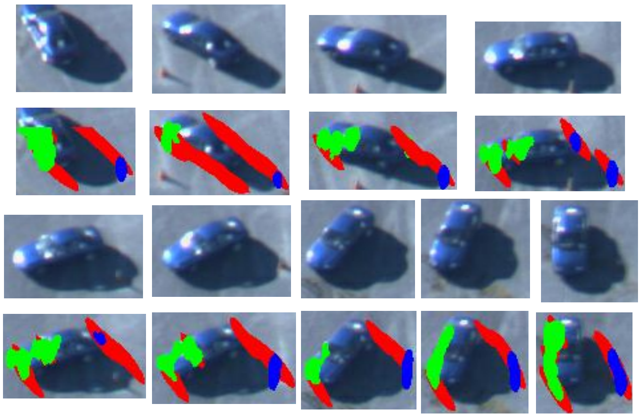

Fig. 11 Car candidates for different orientations of the vehicle. Shadow Casting Out Of Plane (SCOOP) object candidates are green, while inverted Shadow Casting Out Of Plane (invSCOOP) object candidates are blue.

6 Constraints without Metadata

Having all of the metadata provides a set of strict constraints for a variety of camera angles and times of day. However, there may be cases when the metadata is either unavailable or is incorrect. In such cases it is acceptable to sacrifice generality and computation time to obtain a looser set of constraints that still perform well. If we assume that humans are vertical in the image and ignore the ratio between the size of humans and their shadows, we can determine the orientation of the shadow in the image by exploiting the relationship between out of plane objects and their their shadows.

When SCOOP objects are present in the scene the shadows they cast will be at similar orientations. Therefore, SCOOP objects will be consistently detected in the shadow orientation range of about −10° to +10° of the actual orientation. If we assume that the scene is mostly

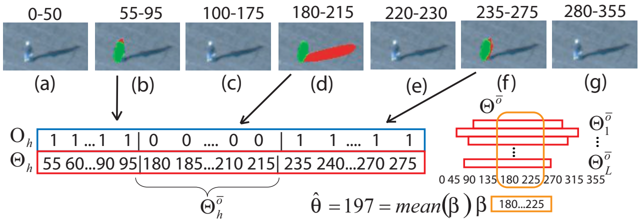

Fig. 12 Method for finding optimal shadow orientation for a given image in the absence of metadata. Top row shows human candidate responses obtained for different shadow orientations. A human candidate is then described by a vector of orientations for which it was detected and a binary overlap vector. Optimal orientation $\hat{\theta}$ is the average of longest common consecutive non-overlapping subsequence of orientations among all human candidates.

planar, then the orientation range will be consistent across all SCOOP objects in the scene. We apply the SCOOP detector at all orientations of the shadow, then search for the longest range of orientations consistent across multiple SCOOP objects. This corresponds to the longest common consecutive subsequence (LCCS) among orientation ranges of all potential SCOOP objects detected in the scene. Hence we need at least two SCOOP objects on the plane. If the scene is not planar, shadow orientation ranges for some SCOOP objects will exhibit a shift. This shift will simply shorten the length of the LCCS, though a reasonable orientation allowing for SCOOP detection will still be found. Multiple planes, each with drastically different orientations, will correspond to multiple LCCSs which can be found if two or more SCOOP objects are present on each plane. The details of this method are as follows.

We quantize the search space of shadow angle θ between 0 and 2π in increments of d (we used $\pi/36$ 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 techniques described in sections 3 and 3.2 (see Figure 12). 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. The set of orientations for which it was detected due to overlap is Θ_i^o, and the set of orientations for which it was detected without overlap is $\Theta_i^{\bar{o}}$ (see Figure 12). 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 12 (as in 12 (b) & (f)). Here p depends on quantization; we found that it should encompass at least 70°.

We now find the optimal shadow orientation $\hat{\theta}$ by treating each $\Theta_i^{\bar{o}}$ as a sequence and then finding the longest common consecutive subsequence β among all $\Theta^{\bar{o}}$. We favor subsequences that are shared among the most SCOOP candidates. 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 13 shows an example frame for which human candidates were detected using the automatically estimated shadow orientation. There is a 10° difference between the estimated orientation and the orientation derived from the metadata. This is the same frame as in Figure 6, qualitative examination of the shadow blobs indicates 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, the angle can be estimated in the initial frame and then predicted in subsequent frames using a Kalman filter.

In order for clutter to affect the method, the clutter has to adhere to two major constraints. First, the clutter has to contain structures which are consistent with shadow casting out of plane objects at orientations different from the true orientation that we are looking for. Second, the orientation of the shadow component of these structures would have to be consistent among the different structures in order to allow for the LCCS to be detected among them. If such structures do exist in the image, then rather than requiring at least 2 true shadow casting out of plane objects, we would require $N + 1$ where N is the number of *consistent* confuser structures in the image. Note that we do not require the presence of humans in the scene, inanimate shadow casting objects serve the same purpose. In figure 28 we show the result of automated shadow orientation estimation for the cluttered balloon data image. In panel **(a)** we show SCOOP candidates detected for one of the incorrect orientations, however more SCOOP candidates were detected around the correct orientation shown in panel **(b)**. The larger number of true consistent SCOOP objects dominated the inconsistent potential confusers, and allowed the method to find the correct shadow orientation in the image.

7 Results

In this section we present both quantitative and qualitative results of human and vehicle detection on the VIVID and CLIF datasets. These results are compared

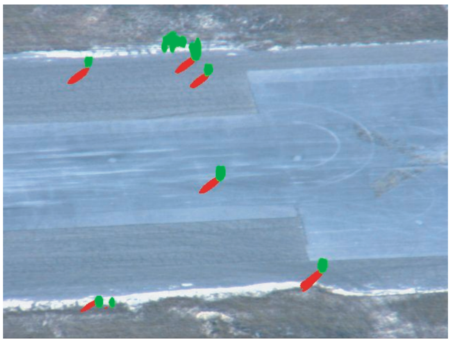

Fig. 13 Refined human candidate blobs for an automatically estimated shadow orientation of 35° without metadata. Corresponding metadata derived value of θ for this frame is 46.7°. Blobs that were detected using metadata can be seen in Fig. 6 (b).

against motion constrained detection, Harris corner constrained detection, and an unconstrained full frame search.

7.1 Human Detection

We evaluated our detection methods on three sequences from the DARPA VIVID3 and two sequences from the APHill dataset and compared the detections against manually obtained groundtruth. The image sizes for the two datasets are 640x480 and 720x480 respectively. The table in Figure 15 shows the number of frames and the number of people in each sequence for the VIVID 3 dataset. We removed the frames where people congregated into groups.

We used the following evaluation criteria *Recall* (True detection rate) vs. False Positives Per Frame (FPPF). Recall is defined as $\frac{TP}{TP+FN}$, where FN is number of false negatives in the frame and TP is the number of true positives. 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 the INRIA set introduced in [9], 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 a large area overlap difference. Hence, the centroid distance measure is more meaningful for aerial data.

Figure 14 compares ROC curves for the following methods: our geometry based method with and without the use of object-shadow relationship refinement and centroid localization, our geometry method augmented with temporal information, conventional full frame detection method (we used HOG detection binaries provided by the authors), and standard motion detection pipeline of registration, detection, and tracking.

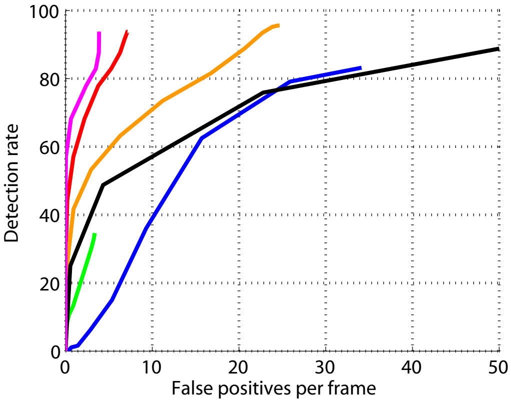

Fig. 14 SVM confidence ROC curves on VIVID3 sequences. Our Geometry method based on classifying SCOOP candidates is shown in red. Orange curves are for our geometry based method without the use of object-shadow relationship refinement or centroid localization.Magenta curves are for our geometry based method augmented with temporal information. 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 [38]. **Black** curves are for our modified implementation of [22] which uses Harris corner tracks.

	Sequence1	Sequence2	Sequence3
Frames	1191	1006	823
Total People	4892	2000	3098

Fig. 15 This table provides details on the VIVID sequences that were used for quantitative evaluation of the human detection methods.

Figure 23 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 the proposed geometric constraints to obtain human candidates is not only much faster (0.72 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 gives superior results to full frame and motion constrained detection.

Motion based detection suffers from problems discussed in Section 1 and shown in Figure 18. Which is why the green ROC curve in Figure 14 is very short. We implemented a part of the method found in [22], where instead of using the OT-Mach filter, we used our wavelet SVM combination for classification. This ROC curve is shown in black. We suspect that the poor

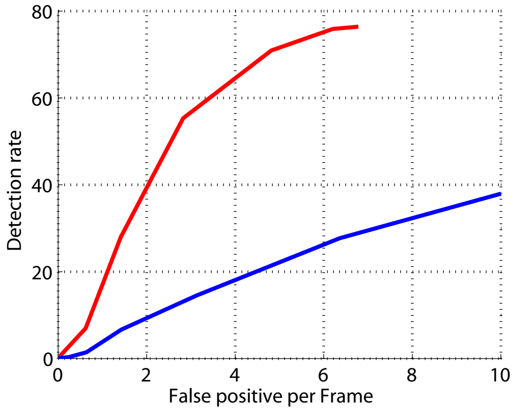

Fig. 16 SVM confidence ROC curves for human detection on APHill sequences. Our Geometry method based on classifying SCOOP candidates is shown in red. A standard full frame detector (HOG) is shown in blue.

	Sequence1	**Sequence2**
Frames	285	331
Total People	419	772

Fig. 17 This table provides details on the APHill sequences that were used for quantitative evaluation of the human detection methods.

performance is due to the large number of initial candidates, since multiple corners are likely to occur on man-made background objects, their shadows, and airfield markings.

By contrast, when proper initialization and localization is provided, temporal information is a convenient way of improving performance. This is illustrated by the magenta curve in Figure 14. In this case, detections obtained by refined geometric constraints (red curve) are tracked in a homography-constrained global coordinate system. Objects that persisted for less than 5 frames are discarded. Classification is performed using wavelets and SVM: if 20% of the track is classified as human, then the entire track is labelled as human. The above technique further suppresses false positives, however the maximum detection rate is slightly reduced since we probably removed incidents of short human tracks.

Similar conclusions can be drawn from the APHill data, the ROC for which are shown in figure 16. The geometry based method achieves far fewer false positives then than full frame detection.

7.2 Vehicle Detection

We performed quantitative evaluation of vehicle detection on the entire frame range of the same VIVID3

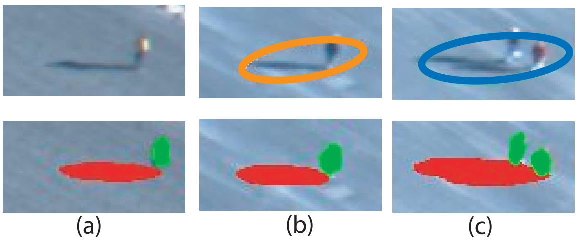

Fig. 18 Qualitative comparison of motion detection (top row) and our geometry based method (bottom row). (a) Human is stationary and was not detected by the motion detector. (b) Moving blob includes shadow, the centroid of blob is not on the person. (c) Two moving blobs were merged by the tracker because of shadow overlap, centroid is not on either person. By contrast our method correctly detected and localized the human candidate (green).

sequences used in the quantitative evaluation of human detection, as well as the same sequences used in the APHill evaluation. Figure 20 shows the number of frames and vehicles contained within each sequence in VIVID, while figure 22 shows the same for the APHill dataset. In order to determine true detections, we used the 33% bounding box overlap criteria from [16]. Figure 19 shows the Recall vs FPPF ROC curves of the same 6 detection methods described in 7.1, with a few slight differences specific to vehicle detection. In the geometry method enhanced with temporal information (the magenta curve), we removed tracks of lengths less than ten. In the case of geometry method without using the SCOOP concept (the orange curve), we represented vehicle candidates as pairs of normal blobs instead of classifying each blob individually. In the Harris corner constrained method (the black curve), we did not track and classify individual Harris corners. Because a single vehicle will have multiple corners belonging to it, we clustered the locations of the corners using MeanShift then tracked and classified the resulting clusters.

At this resolution, the vehicles are much larger than humans and have a sufficient amount of interesting visible features. Therefore, as can be seen in Figure 19, the performance of all detection methods is much better than in the human case. Another thing to note, is that the relative performance of some of the methods is now different. Full frame detection receives very large performance gain, since the classifier can construct a good object model at this resolution. The gains from constraining the search using SCOOP candidates is moderate, though the speed advantage is still there.

Another drastic difference is the performance of motion detection based method (green curve). It is actually better than the geometry method without utilizing temporal information (red curve). This is because the camera observes the scene persistently allowing for a

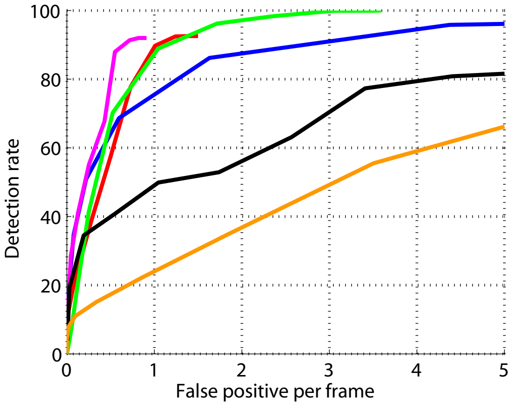

Fig. 19 SVM confidence ROC curves VIVID3 sequences. Our Geometry method based on combining multiple SCOOP candidates is shown in red. Orange curves are for our geometry based method without the use of object-shadow relationship refinement or centroid localization.Magenta curves are for our geometry based method augmented with temporal information. 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 [38]. **Black** curves are for classifying tracks of clusters of Harris corners.

	Sequence1	Sequence2	Sequence3
Frames	1812	1785	1813
Total Vehicles	1719	1742	2019

Fig. 20 This table provides the details for the 3 VIVID sequences used for quantitative vehicle detection evaluation

good background model to be constructed. Additionally, unlike humans, the vehicles are always moving fairly quickly in these sequences. This allows the background subtraction to detect them with ease. Another issue is that unlike in the case of humans, the shadows cast by the vehicle do not severely distort the motion blob creating localization problems at the classification stage and the ground-truth comparison stage.

Detection based on grouping of normal blobs without utilizing the SCOOP concept (orange curves) is not very meaningful. This is because it essentially represents the vehicle as two neighboring parallel gradients, ignoring the box-like nature of the object.

Finally, representing shadow casting vehicles as clusters of Harris corners is rather challenging, since multiple Harris corners are detected on the vehicle and its shadow. These corners have to be clustered to estimate the location of the vehicle. The relative location and quantity of these corners differs with changes in orientation of the vehicle, and corners from background and neighboring objects can become part of the cluster as the vehicle moves through the scene. This, of course, causes localization problems, not unlike those of motion detection in the human case.

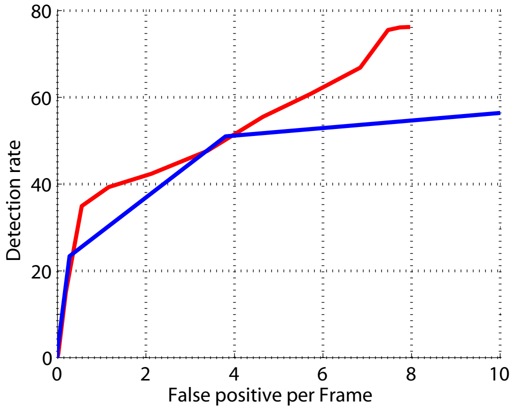

Fig. 21 SVM confidence ROC curves for vehicle detection on APHill sequences. Our Geometry method based on classifying SCOOP candidates is shown in red. A standard full frame detector (HOG) is shown in blue.

	Sequence1	Sequence2
Frames	285	331
Total Vehicles	336	849

Fig. 22 This table provides details on the APHill sequences that were used for quantitative evaluation of the vehicle detection methods.

Qualitative evaluation (see Figure 27) was performed on segments of the CLIF 2007 dataset. Here the resolution is smaller than in VIVID and there are more confusers. In this case, full frame detection once again starts to perform poorly and there is a clear gain from using SCOOP candidates to constrain the search. In the case of CLIF, the camera is very close to nadir, making the normals very short and in the case of vehicles almost non existent. However, our method can still find the candidates. Additionally, if the vehicle is oriented with the orientation of the shadow, the actual presence of the shadow is not even necessary to be able to detect it as a candidate, since the primitive blobs will be detected on the sides of the vehicle. Whether a vehicle can be correctly detected as a candidate while it is at a 45° angle relative to shadow's orientation depends on whether the resolution is high enough to be able to detect the primitive shadow blobs on the corner of the vehicle.

8 Conclusions

We proposed a novel method for detecting humans and vehicles in aerial imagery. This method works on a single image and is is based on constraining the search space of the image by detecting Shadow Casting Out Of

Plane (SCOOP) object candidates. Our method takes advantage of the metadata information provided by the UAV platform to derive a set of geometric constraints, and to project them into the imagery. In cases where metadata is not available, we proposed a method for estimating the constraints directly from image data. The constraints were then used to obtain candidate out-of-plane objects which were then classified as either human or non-human. For vehicles we combined multiple SCOOP candidates to obtain a vehicle candidate. In the case of humans, we evaluated the method on challenging data from the VIVID 3&2 as well as APHill datasets, and obtained results superior to both full frame search, motion constrained detection, and Harris track constrained detection. In the case of vehicle detection, we performed evaluation on the VIVID3 and CLIF datasets obtaining superior results. The purpose of the method is to augment the performance of any full frame classifier but could also be used for shadow detection and removal in the case of background subtraction based surveillance.

Acknowledgements

This research was partially supported by the Harris corporation and Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-10-C-0112. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the Harris corporation or DARPA.

References

1. Bi, S., Liang, D., Shen, X., Wang, Q.: Human cast shadow elimination method bad on orientation information measures. ICAL (2007)
2. Bissacco, A., Yang, M.H.: Detecting humans via their pose. NIPS (2007)
3. Bose, B., Grimson, E.: Improving object classification in far-field video. CVPR (2004)
4. Breckon, T., Barnes, S., Eichner, M., Wahren, K.: Autonomous real-time vehicle detection from a medium-level uav. UAVS (2009)
5. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001). Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
6. Chang, J.C., Hu, W.F., Hsieh, J.W., Chen, Y.S.: Shadow elimination for effective moving object detection with gaussian models. ICPR (2002)
7. Chen, Y.T., Chen, C.S., Hung, Y.P., Chang, K.Y.: Multi-class multi-instance boosting for part-based human detection. ICCV (2009)
8. Cheng, H., Butler, D., Basu, C.: ViTex: Video to tex and its application in aerial video surveillance. CVPR (2006)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. CVPR **1** (2005)
10. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. CVPR (2008)
11. Finlayson, G., Hordley, S., Lu, C., Drew, M.: On the removal of shadows from images. IEEE PAMI **28**(1) (2006)
12. Gaszczak, A., Breckon, T., Han, J.: Real-time people and vehicle detection from uav imagery. SPIE (2011)
13. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)
14. Hsieh, J.W., Yu, S.H., Y.-S., C., Hu, W.F.: A shadow elimination vethod for vehicle analysis. ICPR (2004)
15. Hu, H., Huang, Y.Q., Li, L.M.: Moving vehicle shadow elimination approach based on mark growing of multi-feature fusion. ICACIA (2010)
16. Kembhavi, A., Harwood, D., Davis, L.: Vehicle detection using partial least squares. IEEE PAMI (2011)
17. Kluckner, S., Mauthner, T., Roth, P.M., Bischof, H.: Semantic classification in aerial imagery by integrating appearance and height information. ACCV (2009)
18. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. CVPR (2005)
19. Liu, Z., Zhao, F., Yang, H.: A new method of moving shadow elimination combining texture and chrominance of moving foreground region bsed on criterion. WCICA (2010)
20. Martel-Brisson, N., Zaccarin, A.: Moving cast shadow detection from a gaussian mixtrue shadow model. CVPR (2005)
21. Mikolajczyk, K., C., S., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. ECCV (2004)
22. Miller, A., Babenko, P., Hu, M., Shah, M.: Person tracking in UAV video. CLEAR (2007)
23. Panagopoulos, A., Samaras, D., Paragios, N.: Robust shadow and illumination estimation using a mixture model. CVPR (2009)
24. Porikli, F., Thornton, J.: Shadow flow: A recursive method to learn moving cast shadows. ICCV (2005)
25. Prati, A., Mikic, I., Trivedi, M.M., Cucchiara, R.: Detecting moving shadows: Algorithms and evaluation. IEEE TPAMI **25** (2003)
26. Quaritsch, M., Kruggl, K., Wischounig-Strucl, D., Bhattacharya, S., Shah, M., Rinner, B.: Networked uavs as aerial sensor network for disaster management applications. Elektrotechnik und Informationstechnik (127) (2010)
27. Reda, I., Anreas, A.: Solar position algorithm for solar radiation applications. NREL Report No. TP-560-34302 (2003)
28. Rudol, P., Doherty, P.: Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery. IEEE Aerospace (2008)
29. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. CVPR (2007)
30. Sokalski, J., Breckon, T.: Automatic salient object detection in uav imagery. UAVS (2010)
31. Tian, T.P., Sclaroff, S.: Fast multi-aspect 2d human detection. ECCV (2010)
32. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. PAMI **30** (2008)
33. Wang, X., Han, T., Yan, S.: An hog-lbp human detector with partial occlusion handling. ICCV (2009)
34. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. ICCV (2005)

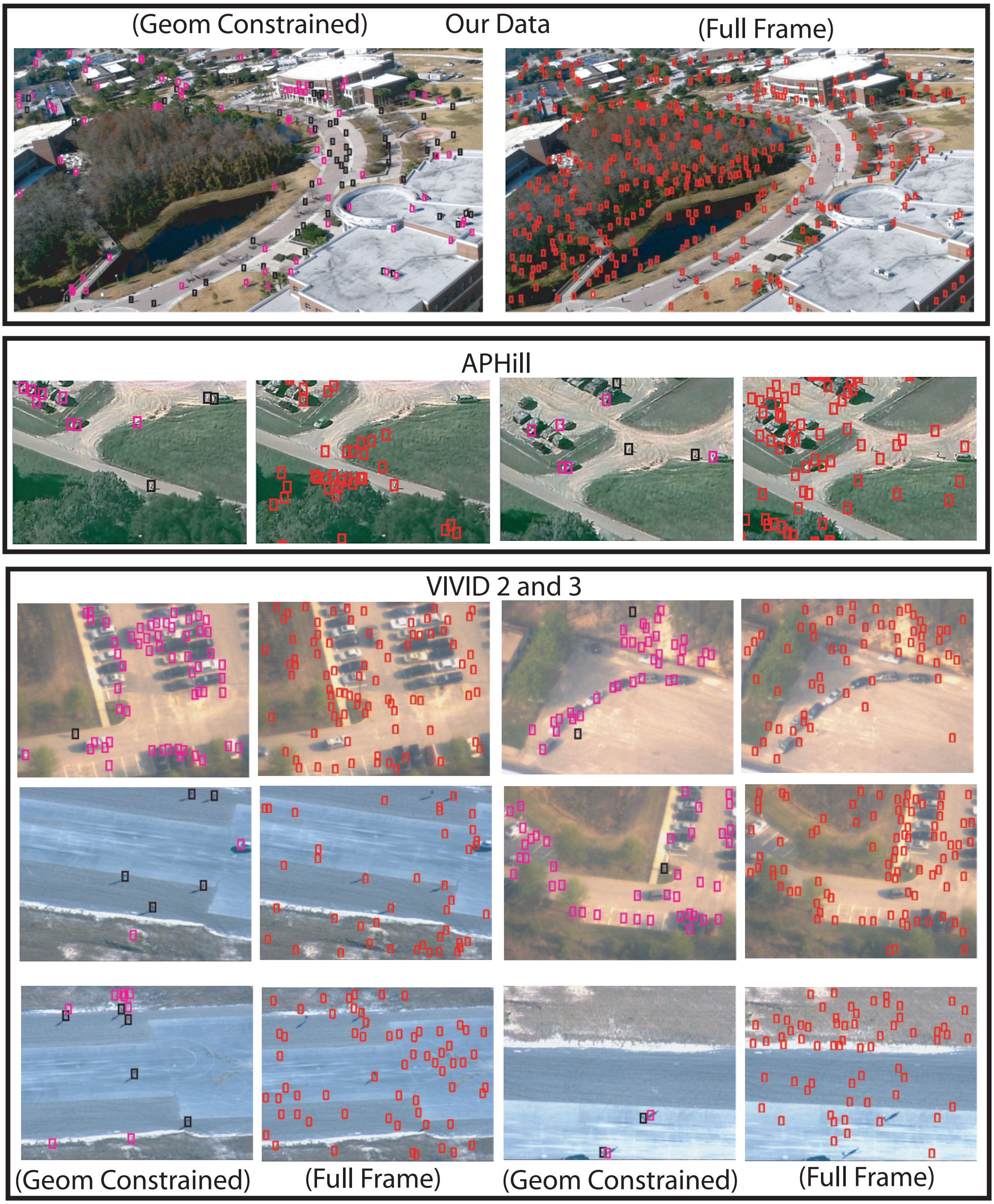

Fig. 23 Qualitative detection results on VIVID 3, VIVID 2 and some of our own data. Columns labelled (Full Frame) show the result of full frame search (HOG) applied to entire frame (human detections are shown in red). Columns labelled (Geom Constrained) show the results of our geometry constraint based method. 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**.

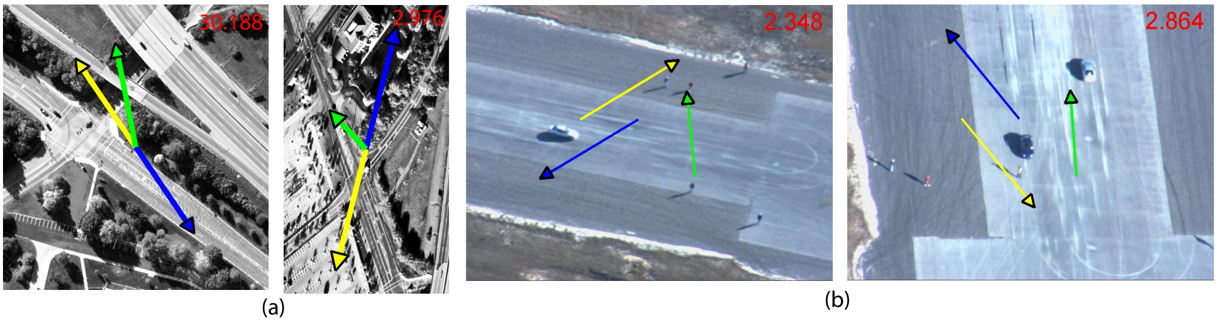

Fig. 24 Metadata derived geometric constraints for images from the CLIF 2007 dataset **(a)**, and the VIVID dataset **(b)**. Yellow arrow, is the vector $\mathbf{S'}$ pointing towards the sun. Blue arrow is a vector pointing in the direction of the shadow (reverse of sun direction). Green arrow is the vector $\mathbf{Z'}$ pointing in the direction of the normal. The shadow to normal ratio is shown as red text. Note that in the first image the camera is close to NADIR, hence the shadow to normal ratio is very high.

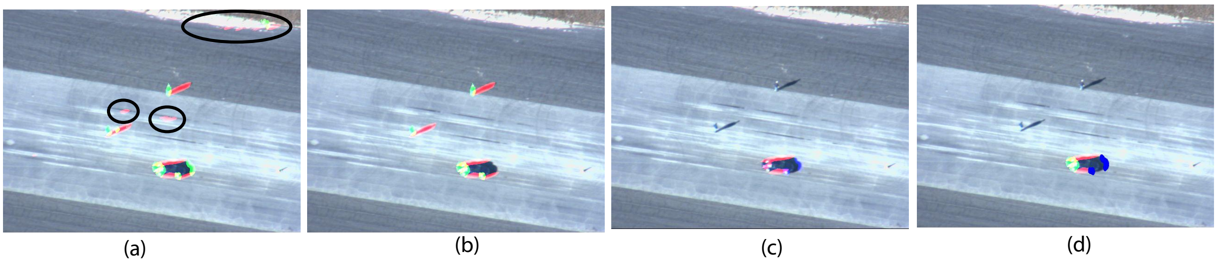

Fig. 25 The candidate refinement process. **(a)** All of the shadow and normal blobs that were obtained using the method described in Section 3.1. **(b)** The set $\mathbf{K}^{\mathbf{Z}'}_{\mathbf{S}'}$ of refined SCOOP candidates. Note that a lot nonsense shadow and normal blobs (circled in **black**) were removed from areas of strong gradient. **(c)** The set $\mathbf{K}^{-\mathbf{Z}'}_{-\mathbf{S}'}$ of refined inverse scoop blobs. Finally, **(d)** show vehicle candidates obtained by combining SCOOP and inverse SCOOP.

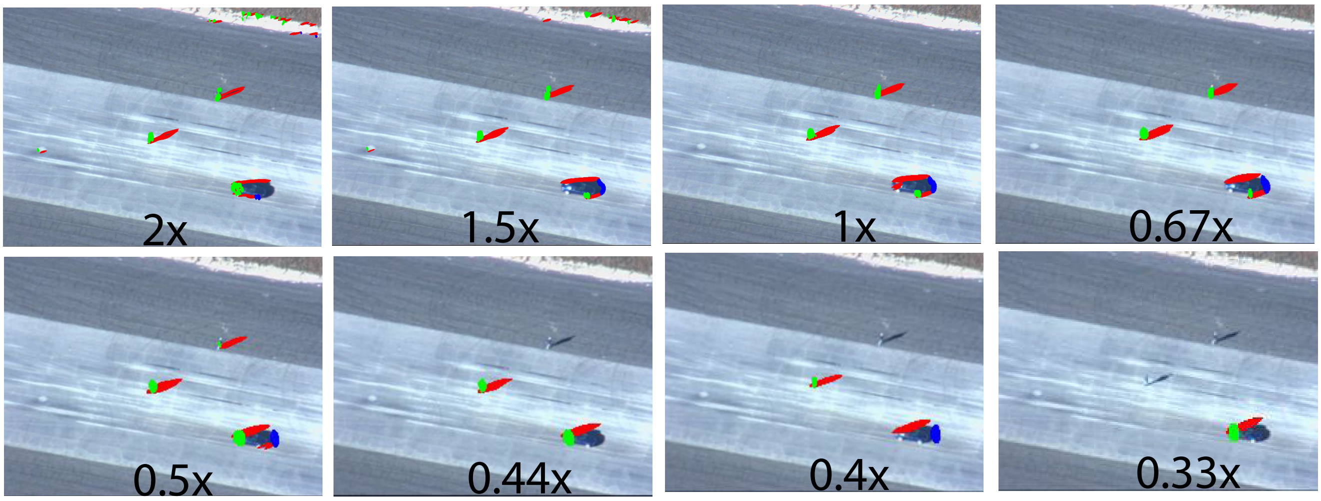

Fig. 26 Human and car candidates detected at different image scales. Green blobs are normal blobs associated with SCOOP candidates, textclorblueblue blobs are normal blobs associated with invSCOOP candidates. We used the same settings at all stages of the candidate detection method, however we resized the image prior to the application of the processing pipeline by the amount shown at the bottom of each image, and then resized the results to the same size for display purposes.

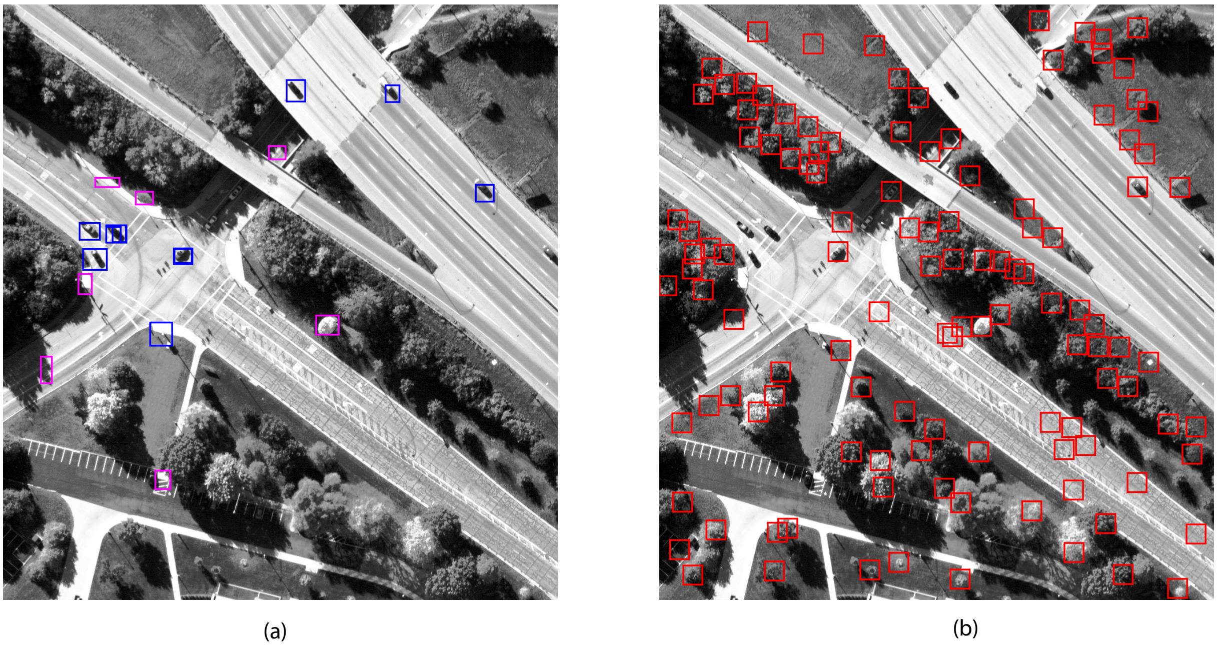

Fig. 27 **(a)** shows vehicle detection results for geometrically constrained method on the CLIF 2007 dataset. Candidates that were classified as vehicles are shown in blue, candidates that were discarded by the classifier are shown in magenta. **(b)** Show results for full frame detection in red. Full frame detection has generated a lot more false positives. Note that there are two vehicles that our method misses because their shadows are obscured by the shadow cast by the bridge full frame detector misses only one of those vehicles.

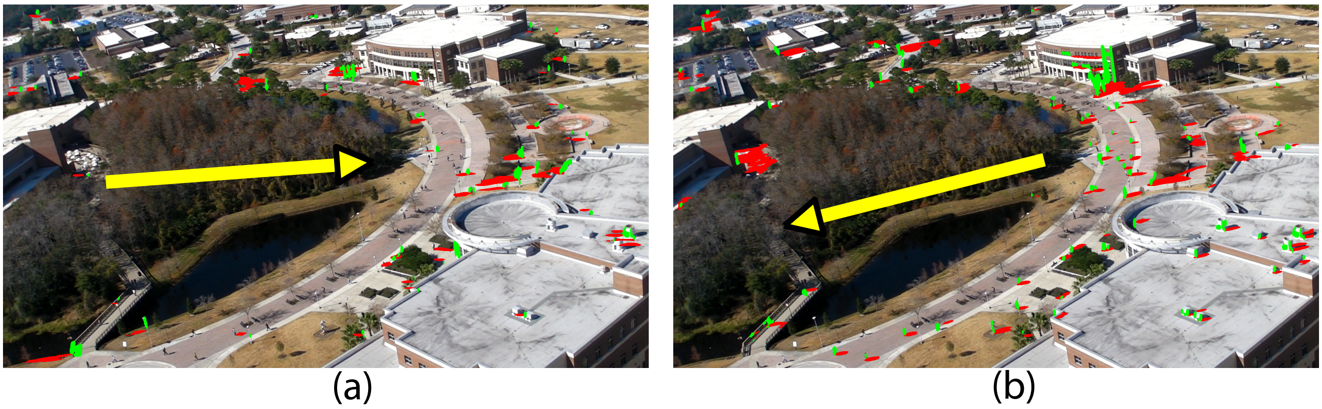

Fig. 28 Refined SCOOP detections output by automated shadow orientation estimation method for **(a)**: one of the incorrect sun directions which were discarded by our method, and **(b)**: the final sun direction which was selected by our method. Green blobs belong to the normal blobs, red blobs belong to the shadow blobs. The yellow arrow indicates the direction of the sun for which the SCOOPs were detected.

35. Wu, Q., Luo, X., Li, H., Liu, P.: An improved multi-scale retinex algorithm for vehicle shadow elimination based on variational kimmel. SWUATC (2010)
36. Xiao, J., Cheng, H., Han, F., Sawhney, H.: Geo-spatial aerial video processing for scene understanding and object tracking. CVPR (2008)
37. Xiao, J., Cheng, H., Sawhney, H., Han, F.: Vehicle detection and tracking in wide field-of-view aerial video. CVPR (2010)
38. Xiao, J., Yang, C., Han, F., Cheng, H.: Vehicle and person tracking in aerial videos. Multimodal Technologies for Perception of Humans (2008)
39. Xu, L., Qi, F., Jiang, R.: Shadow removal from a single image. Intelligent Systems Design and Applications **2** (2006)
40. Yahyanejad, S., Wischounig-Strucl, D., Quaritsch, M., B., R.: Incremental mosaicking of images from autonomous, small-scale uavs. AVSS (2010)
41. Yilmaz, A., Javed, O., Shah, M.: Object tracking a survey. ACM Comput. Surv **38** (2006)
42. Yoneyama, A., Yeh, C.H., Jay Kuo, C.C.: Moving cast shadow elimination for robust vehicle extraction based on 2d joint vehicle/shadow models. AVSS (2003)