Pedestrian Detection in Low-resolution Imagery by Learning Multi-scale Intrinsic Motion Structures (MIMS)

Jiejie Zhu, Omar Javed, Jingen Liu, Qian Yu, Hui Cheng, Harpreet Sawhney
SRI International
{jiejie.zhu,omar.javed,jingen.liu,qian.yu,hui.cheng,harpreet.sawhney}@sri.com

Abstract

Detecting pedestrians at a distance from large-format wide-area imagery is a challenging problem because of low ground sampling distance (GSD) and low frame rate of the imagery. In such a scenario, the approaches based on appearance cues alone mostly fail because pedestrians are only a few pixels in size. Frame-differencing and optical flow based approaches also give poor detection results due to noise, camera jitter and parallax in aerial videos. To overcome these challenges, we propose a novel approach to extract Multi-scale Intrinsic Motion Structure features from pedestrian's motion patterns for pedestrian detection. The MIMS feature encodes the intrinsic motion properties of an object, which are location, velocity and trajectory-shape invariant. The extracted MIMS representation is robust to noisy flow estimates. In this paper, we give a comparative evaluation of the proposed method and demonstrate that MIMS outperforms the state of the art approaches in identifying pedestrians from low resolution airborne videos.

1. Introduction

In recent years, large-format wide-area sensors are increasingly being used for persistent surveillance tasks, including border security, force protection and aerial surveillance. The wide-area sensors are usually placed on high towers, aero-states or unmanned aerial vehicles for these applications. The goal of employing such sensors is to detect targets and activities of interest, at the largest possible distance while covering the greatest possible area (preferably in tens of square kilometers). Therefore, these sensors have low ground sampling resolution, typically 0.3m-0.5m Ground Sampling Distance (GSD), in order to cover a large area and low frame-rate (2-5 Hz) due to large image size. Automated analysis tools are critical for such sensors as the size of the area monitored and the number of objects to track are beyond continuous manual inspection.

Most of existing automated analysis tools for wide-area sensors have focused on vehicle tracking/classification

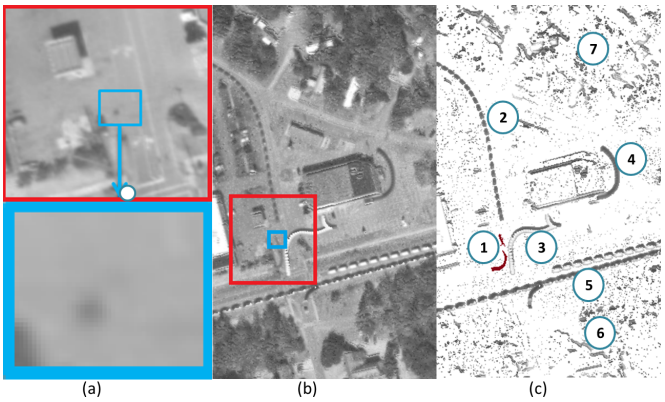


Figure 1: (a) shows portion of a single frame from a wide-area imagery. A pedestrian is a dark dot with a size of a few pixels even in a zoomed-in view. (b) shows 50 frames foregrounds overlaid on the top of background, where the trajectories of moving objects such as cars pedestrian (1), vehicles (2,3,4,5) and parallax (6,7) become visible. It is worth noting that this overlaid view is after stabilization. (c) shows the residual pixels, *i.e.* the pixels that are different from the background of the middle image. The residual pixels may come from three categories: moving objects such as vehicles, pedestrians and parallax that are static but cannot be compensated by a global motion due to the structures above the ground plane and registration errors. Among all residual pixels, only a very small portion belongs to pedestrians, which has been marked as red pixels in the right image. Our goal is to detect pedestrians by using their intrinsic motion patterns.

[19, 4, 22]. There is very little existing work focussed on detection and tracking of pedestrians from large format wide-area imagers. This is because tracking pedestrians in wide area surveillance has unique challenges (see Figure 1). The main challenge for wide-area dismount detection is that extremely low resolution, i.e., GSD is around 0.3m to 0.5m per pixel, therefore pedestrians cover only 4 - 9 pixels in the image, please see Figure 1. At this scale the typical shape

or appearance based models for object detection, such as HOG [5], deformable part models [6] and shape based models [2] no longer provide significant discrimination against the background.

One viable solution of detecting pedestrian from wide area aerial imagery is to exploit motion. Most existing motion based approaches [18, 21] use background subtraction, frame difference and Histogram of Optical Flow (HOF) as features for classification. These motion based features are sensitive to the noise in image registration, and optical flow estimation. Thus use of these features often results in high false alarm rate when pedestrians comprise only a few pixels in the imagery.

To solve the aforementioned challenges, let us first observe the residual motion image in Figure 1c, which is the motion after compensating the global camera motion. The residual motion comes from moving objects (e.g., pedestrian and vehicles), parallax, and inevitable registration errors. To distinguish the motion of pedestrian from that of others, one may think of directly using the location, or velocity, for detection, but these features are not discriminative because different types of objects or noise may have similar location, speed or direction. We believe that, however, the motion that comes from the same type of object (e.g., pedestrian) forms a manifold in a space such as (x,y,vx,vy), and from which we can select a unique pattern capturing the intrinsic properties of the object. One such intrinsic property is the local dimensionality of the topology of the motion, which is location, speed and trajectory shape invariant. Therefore, we propose a novel approach to discover the local dimensionality based on the local topology of the motion manifold for pedestrian detection.

However, without specifying a scale of the local topology, the dimensionality may not be meaningful. As an example, a pedestrian can be regarded as a point when viewing from afar, while it is a 3D object when looking at it closely. As a result, we pair the dimensionality and scale to model a pedestrian's motion pattern. This leads to two questions: 1) How to robustly estimate the dimensionality and 2) how to pick up the right scale. To answer these questions, we propose a learning-based tensor voting[16] approach. Basically, tensor voting provides local dimensionality of motion patterns and its saliency at a specific scale. However, in practice, it is hard to manually select the right scale(s). Thus, in our work, we use tensor voting to generate a whole spectrum of features at various sampled scales and employ feature selection to form a compact discriminative representation. These extracted features encode the intrinsic properties of pedestrian's motion pattern at various scales. We refer to these features as Multi-scale Intrinsic Motion Structure (MIMS) features.

In summary, our contributions include: (1) We propose a novel approach to discover MIMS features for pedestrian detection in aerial videos. The MIMS representation is robust to noise and invariant to location, velocity and shape of trajectory. (2) We introduce a learning strategy for selection of invariant features, and (3) present a thorough evaluation of the proposed approach on WAAS videos along with a comparison with the state of the art.

2. Related Work

Human detection in aerial surveillance videos has received significant attention [18, 21]. Appearance based and frame-by-frame motion based feature analysis are two main pedestrian detection approaches reported in the literature.

Reilly et al. [20] uses background subtraction for detection and employs a geometric feature that measures the ratio between a people's height and the size of its corresponding shadows to filter out non-human area. This ratio is within a range if both camera and sun's location are known. A voting method is proposed in [17] to recognize pedestrians by measuring the appearance similarity between labeled samples and candidate locations. Candidate locations are predicted by a SVM classifier using HOG descriptors. Sokalski et al. [23] developed a salient object detector based on color information. Paper [10] detects human from thermal UAV images using two methods. One is to classify human from non-human from thermal signatures with respect to orientation, thermal contrast and size. The other is to extract and match human silhouette using shape descriptors. A similar technique, i.e., matching human thermal silhouettes, is also applied in [7].

Unlike these appearance-based approaches that work mainly on low-altitude UAV platforms, we are dealing with lower resolution data with GSD around 0.3m to 0.5m per pixel, therefore pedestrians cover only 4 - 9 pixels in the image. At this scale the typical shape or appearance based models mentioned above do not perform well.

Lee [12] uses motion to detect pedestrians in a cluttered scene. The method consisted of extracting the silhouettes of moving humans and using perceptual grouping to reduce the impact of clutter and noise in the detection process. Yu and Medioni [24] use motion structure analysis approach in a 4D space for vehicle detection. They manually selected a few scales of intrinsic structure and used segmentation to extract the vehicle motion from a 4D motion space. Prokaj et al. [19] use background subtraction for vehicle detection in airborne videos and learn vehicle motion patterns from initial tracklets to improve tracking. Zhao and Medioni [25] propose another tensor voting based approach to learn directional motion patterns which can be used as a prior to improve tracking.

Unlike our approach, the above mentioned approaches neither attempt to detect very small moving objects nor analyze the pattern at multiple scales to extract a complete signature profile of the object of interest for detection.

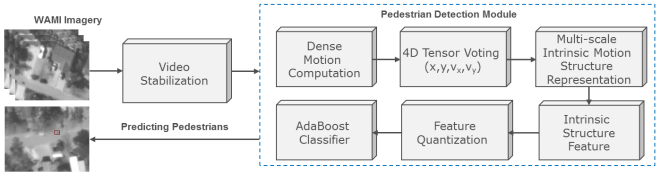


Figure 2: The overall framework of our approach. Modules within the blue dashed line are the focus of this work. We used a technology similar to [11] to stabilize the video an d used Liu's [13]'s optical flow code to generate dense optical flow to feed our pedestrian detection algorithm.

3. Algorithm Overview

The overall algorithm framework is shown in 2. Given a video clip of WAMI imagery, we first stabilize the video using an approach similar to [11]. The stabilization process compensates the camera motion and enforce the motion smoothness constraint. Then, we compute dense optical flows on the stabilized video clips for every two consecutive frames.

Given (x, y, v_x, v_y) from multiple frames, we use tensor voting to group such 4D samples by selecting different neighborhood size in both spatial and temporal domain in the voting process. We then compute the principle axes of the group and define (1) the index of the maximal eigenvalue difference as motion dimensionality d and (2) the maximal eigenvalue difference as motion saliency s. Since each group is estimated using a specific neighbor size, it is straightforward to deem that the motion dimensionality and saliency are extracted at a scale σ and (d, s, σ) is called Intrinsic Motion Structure (IMS) which captures the intrinsic property of the motion. For each pixel of the imagery, we extract multiple IMS at different scale, and concatenate to represent the pixel. Such a compact Multi-scale IMS (MIMS) representation is trained using an AdaBoost-like method to detect pedestrians which is introduced in section 5.

4. Intrinsic Motion Structure Discovery

In this section, we describe how our approach discover motion dimensionality and saliency of a pixel IMS in a 4D space (x, y, v_x, v_y) at a specific scale.

When a 2D point (x, y) traverses continuously on a 2D plane with a velocity (v_x, v_y), it actually produces a motion shape in a 4D space of (x, y, v_x, v_y). The 2D point can be referred to a pixel, a patch or an object on the image at a certain scale. In 4D space, we call the temporal shape of the motion as a fiber, which is different to the trajectory on 2D space of (x,y). See the pedestrian's motion trajectory in Figure 3 for an example the 2D trajectory corresponds to a fiber in the 4D space. If a set of such 2D points (*e.g.* a set of points on the same object) are moving under the similar motion pattern, they basically form a fiber bundle, which presents a different shape to a fiber at a different scale. As shown in 3, the trajectory of vehicle actually form a fiber bundle at the current scale. Please note that at a different scale, the fiber bundle may be a single fiber.

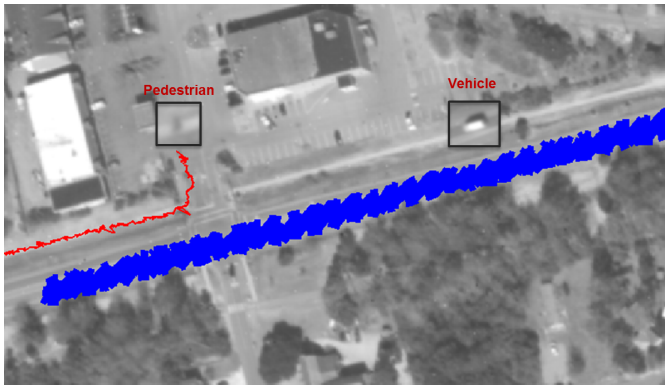


Figure 3: Example of vehicle and pedestrian motion structure. The red trajectory represents pedestrian, which corresponds a fiber in the 4D space (x,y,vx,vy). The blue one represents a vehicle, which corresponds to a fiber bundle in the 4D space.

one may notice neither the fiber nor its individual components (e.g., location, velocity) can be directly used to differentiate pedestrians from other objects, because pedestrians can be anywhere with different velocities, even their trajectory shapes may be similar to that of other objects. However, features derived from a combination of location, speed and motion trajectory from same sources or objects share intrinsic properties that can distinguish pedestrians from others. This leads us to develop a novel feature that is independent of object's location, speed and motion trajectory. Such an intrinsic representation captures the essential geometrical properties of a pedestrian's motion and therefore offers considerable advantages when the pedestrians becomes tiny and the background becomes more complex.

4.1. IMS Extraction using Tensor Voting

Given samples in the 4D space (x, y, v_x, v_y), we use tensor Voting [16] to group 4D samples because of its robustness to noise. Essentially, tensor voting can be regarded as an unsupervised computational framework to recover the intrinsic local geometric information, which is encoded in a symmetric, non-negative definite matrix. This local geometry describes a moving object's local motion structure (i.e., motion dimensionality and saliency) which can be derived

by examining its eigen system as:

$$\mathbf{T} = \sum_{i=1}^{N} \lambda_i \mathbf{e}_i \mathbf{e}_i^T = \sum_{i=1}^{N-1} (\lambda_i - \lambda_{i+1}) \sum_{k=1}^{i} \mathbf{e}_k \mathbf{e}_k^T + \lambda_N \sum_{i=1}^{N} \mathbf{e}_i \mathbf{e}_i^T \quad (1)$$

where $\{\lambda_i\}$ are the eigenvalues arranged in descending order, $\{\mathbf{e}_i\}$ are the corresponding eigenvectors, and N is the dimensionality of the input space which is 4 for our 4D space. The decomposition in Eq.1 provides a way to interpret the motion local geometry.

The motion dimensionality and saliency reveal the local motion structure of a moving object. The motion dimensionality d is the index number of the largest gap between two consecutive eigenvalues,λ_i, λ_{i+1}, *i.e.*

$$d = \arg\max_i (\lambda_i - \lambda_{i+1}) \quad (2)$$

The motion saliency is the largest difference of two consecutive eigenvalues $\lambda_d - \lambda_{d+1}$. In other words, a local motion structure, whose normal space and tangent space are d dimensional and (N-d) dimensional respectively, is the most salient interpretation to $\mathbf{T}$. The corresponding eigenvectors $\mathbf{e}_1, ..., \mathbf{e}_d$ span the normal space of the structure and $\mathbf{e}_d + 1, ..., \mathbf{e}_N$ span the tangent space.

4.2. Multi-scale IMS (MIMS) Representation

When computing IMS, the only free parameter is σ that controls the neighborhood size in voting. The right neighborhood size of σ to be used in voting is often unknown unless hand-crafted. Interestingly, previous work often neglected σ when describing motion dimensionality. In fact, the motion dimensionality is meaningful only when σ is specified. For example, the 2D sheet of the vehicle motion shown in Figure 3 may become an object with 3D volume when its corresponding spatial scale increases (*e.g.* viewed under a microscope) or it shrinks to a line when its corresponding scale decreases (*e.g.* under a telescope).

Instead of using a hand-crafted single scale to represent IMS, we pair up the scale and the dimensionality at multiple scales,

$$\{(d_i, s_i, \sigma_i), i = 1, \ldots, k\} \quad (3)$$

where d_i and s_i are the intrinsic dimensionality and its corresponding saliency at scale σ_i. $||\sigma||$ is the total number of scales. In this representation, one can avoid the scale selection in a hand crafting manner and improve the tolerance of increased noisy. In addition, the multi-scale representation is able to capture the changing of dimensionality when the scale varies, such as shown in Figure 4.

5. Learning Pedestrians from MIMS

We regard pedestrians detection from airborne video as a binary classification problem. In other words, we aim

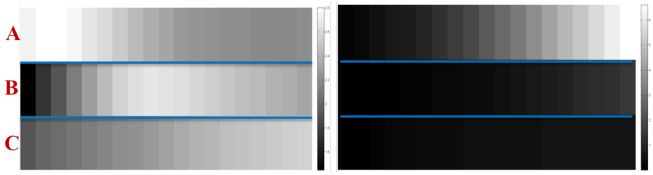


Figure 4: This example demonstrates the change of the motion dimensionality (left panel) and saliency (right panel) with respect to scale σ. We selected 20 uniformly sampled σ from 0.2 to 4 (increasing from left to right) for dataset I. Each row (A,B,C) shows the mean statistical values of all instances from the same object type. Row A, B and C represent pedestrian, vehicle, and background, respectively. Obvious patterns can be observed, such as the motion dimensionality of pedestrians (row A) keep decreasing with increased scales, while that of vehicle (row B) first increases then decreases. The dimensionality of background (Row C) keeps increasing with scale.

to differentiate the pedestrians' MIMS from other objects' MIMS. We first introduce how we can derive features from (d_i, s_i, σ_i) and then introduce feature selection using Adaboost to speed up classification process.

5.1. MIMS Distribution

Given a scale σ, the intrinsic feature is encoded into two values (d, s) as shown in Eq.3, where $d \in \mathbb{N}^+, 1 \leq d < N$ and $s \in \mathbb{R}^+$. These two values capture a 2D distribution of the intrinsic structure at a certain scale. Suppose we normalize s between $[0, 1]$ and choose k bins to quantize the normalized saliency value as $\tilde{s}$, we can build a 2D look-up table to approximate the 2D distribution of intrinsic structure at this scale. In our case, d can only be $[1, 2, 3, 4]$ and we quantize the saliency into 16 bins. This partition of the feature space corresponds to a partition of the sample space.

For the binary pedestrians classification problem, a sample is represented as a tuple $\{\mathbf{x}, y\}$, where $\mathbf{x}$ is one residual pixel and y is class label whose value can be $+1$ (pedestrian) or -1 (non-pedestrian). The feature value at a certain scale σ can be represented as

$$f(\mathbf{x}, \sigma) = (d, \tilde{s}) \quad (4)$$

where the feature value can be regarded as a 2D indices of the 2D look-up table. Suppose we use $\mathbf{f}$ to represent the 2D indices, then

$$W_{\mathbf{f}}^c = P(f(\mathbf{x}, \sigma) \in bin_{\mathbf{f}}, y = c), c = \pm 1 \quad (5)$$

encodes the 2D distribution of intrinsic features among the positive and negative training samples.

5.2. Boosting MIMS Feature

According to the real-valued version of AdaBoost algorithm [9], the weak classifier $h(\mathbf{x})$ based on MIMS can be

defined as a piece-wise function

$$h(\mathbf{x},\sigma) = \frac{1}{2}\ln\left(\frac{W_{\mathbf{f}}^{+1}+\varepsilon}{W_{\mathbf{f}}^{-1}+\varepsilon}\right), \text{when} \mathbf{f} = f(\mathbf{x},\sigma) \quad (6)$$

where $W_{\mathbf{f}}^{+1}$ and $W_{\mathbf{f}}^{-1}$ are defined in Eq.5.

Given the characteristic function,

$$B_{\mathbf{f}}(u) = \begin{cases} 1 & u \in bin_{\mathbf{f}} \\ 0 & \text{otherwise} \end{cases}, \mathbf{f} \in \text{2Dindices} \quad (7)$$

the weak classifier based on MIMS can be formulated as:

$$h(\mathbf{x},\sigma) = \frac{1}{2}\sum_{\mathbf{f}}\ln\left(\frac{W_{\mathbf{f}}^{+1}+\varepsilon}{W_{\mathbf{f}}^{-1}+\varepsilon}\right)B_{\mathbf{f}}(f(\mathbf{x},\sigma)) \quad (8)$$

For each intrinsic feature at a certain scale, one weak classifier is built. Then the real AdaBoost algorithm [8] is used to learn strong classifiers, called layers, from the weak classifier pool. The strong classifier $H(\mathbf{x})$ is a linear combination of a series of weak classifiers selected

$$H(\mathbf{x}) = \sum_{i=1}^{T}\alpha_i h(\mathbf{x},\sigma_{\mathbf{i}}) + b \quad (9)$$

where where T is the number of weak classifiers in $H(\mathbf{x})$, σ_i is the scale in the ith selected weak classifier and b is a threshold.

6. Experiments

We performed both qualitative and quantitative evaluation of the algorithm on two datasets (see section 6.1). Dataset I is in house imagery. Dataset II is a publicly available aerial video dataset [20].

6.1. Experimental Datasets

Dataset I consists of aerial imagery at a frame rate of approximately 2.0Hz. Image size is 512×512 at a GSD of 0.25m. Figure 1 (b) shows an example of the imagery captured. This dataset is very challenging since a pedestrian usually occupies only a few pixels and the shadow cast by pedestrian is not visible at all. Thus, the appearance based approaches, such as HOG human detector [5] and the approach in [20] that is relying on shadow cannot work on this dataset.

Our algorithm works not only for the scenario with pedestrians in wide area aerial videos but also for the case with regular aerial videos with high resolution of pedestrians. To verify this, we tested our approach on a public dataset [20] (dataset II in this paper), where pedestrians often occupy 20-40 pixels compared with 5-10 in dataset I. The image resolution of Dataset II is 640x480.

Both datasets consist of three sequences. We use two sequences for training and leave one sequence out for testing. We report the average performance of the three runs. Ground truth labels of pedestrians in these two datasets are manually generated.

In these two datasets, we consider following three main types of residual motion pixels as non-pedestrians:

- Pixels from parallax motion. This mainly includes high-rise buildings, trees and background noise.
- Pixels from slowly moving vehicles. When vehicles are about to stop or start, their motion magnitudes are similar to that of pedestrians.
- Pixels from shadows. Motion inside shadows are not robust because all pixels appear to be dark and it is hard to localize their corresponds in the next frame.

6.2. Evaluation Metric and Baseline Features

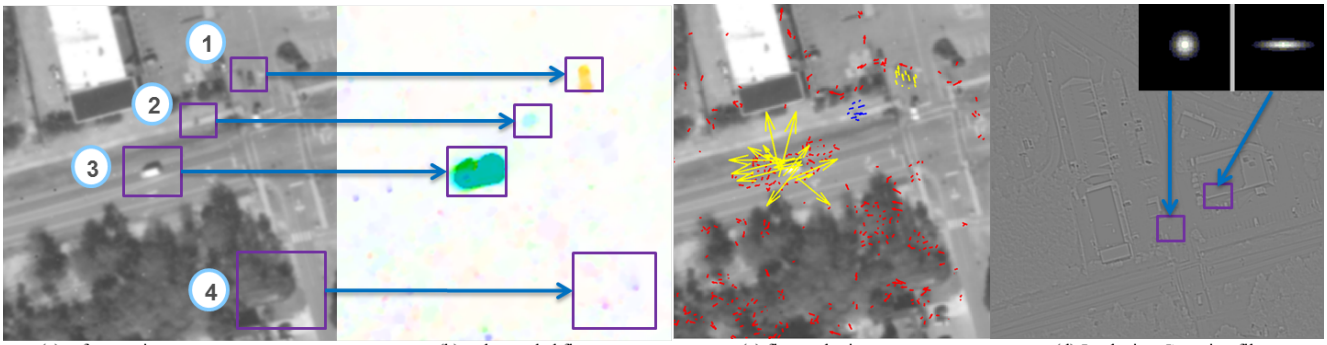
Since our ground truth labels is pixel-wise, the accuracy in ROC curve is reported pixel-wise. Following [20], we do not use the PASCAL measure of 50% bounding box overlap because the pedestrians in our datasets are very small, and make up a very small percentage ($< 0.1\%$) of the scene. Under these circumstances, pixel-wise results provides better measure than box overlap based measures. In addition, we report the false alarm per minute per square kilometer.

For comparison, we implemented two types of baseline features: appearance based and flow based features. Both are widely used in pedestrian detection. Here are our implementation details for each of the baselines:

- Laplacian of Gaussian (LoG) Filter [1]: at every single pixel, we compute its Laplacian of Gaussian response to measure how different it is as comparing to its surroundings. We sample various sizes of Gaussian standard deviation (e.g., inner filters are sized from [1,1] pixels to [6,6] and outer filters are sized from [2,2] to [7,7]). Pair of inner and outer filter combination produces a response as a feature value.
- Histogram of Oriented Gradient (HOG) [5]: HOG is computed at every 5x5 image patch sampled every 2 pixels. The empirical patch size and sample rate give better results on our datasets.
- Frame-by-Frame Flow: we compute per-pixel flow from two consecutive frames , and the features of each pixel are the concatenation of the horizontal flow, the vertical flow and the flow magnitude.

In addition to report results on MIMS, we also report results for intermediate features, such as motion dimensionality and saliency. We also compute features using alternative manifold estimation methods for comparison. Here are the implementation details:

- Motion Dimensionality and Saliency: given current frame f_t, we form a 9-frame 3D volume of flow



(a) reference image (b) color-coded flow (c) flow velocity (d) Laplacian Gaussian filter

Figure 5: This example shows the flows computed from two consecutive frames. (a) shows the original image, (b) shows the color coded flow for the same image. Hue indicates the flow direction and the saturation indicates the flow magnitude. (c) optical flows shown as vectors on the same image. Due to the occlusion and shadows, the surround flows of (3) are noisy (see red small arrows overlapped with yellow arrows on the vehicle. (d) shows an example of LoG filter response on the image. The inset images show that shape of a pedestrian's response is closer to a circle while that of the building, where parallax motion can cause false detection, is closer to a line.

from the previous and next 4 frames of f_t. To compute the motion dimensionality and saliency at a certain pixel p_i^t, we apply the tensor voting on a group of pixels, which are the neighbors of p_i^t in the 4D space (x, y, v_x, v_y), to compute the dimensionality and saliency of p_i^t as described in section 4.

- MIMS: we compute motion dimensionality and saliency at scale $\sigma = (0.4, 0.6.., 4)$ and combine them by quantizing the saliency into 16 bins and dimensionality to 4 bins (maximal 4 dimensions). This gives us a 2D histogram with 16×4 bins. For each sample at a scale, its index in this 2D histogram will be regarded as features.

- Diffusion Map [15]: we construct a graph to include multiple frame features (x, y, v_x, v_y). The intrinsic structure of co-occurrence feature is computed by clustering similar feature from multiple Gaussian kernels to define the neighborhood size. For each Gaussian kernel, a bag of word (BoW) descriptor is created and each feature is represented using this BoW. We concatenate all BoWs into feature vector and train a classifier using the method introduced in 5 .

6.3. Results and Discussion

Table 1 summarizes the detection rate and False Alarm per km per minute from Dataset I. To verify that both appearance-based and flow-based features are not discriminative enough for pedestrian detection in aerial videos with low resolution, we performed a set of experiments on detecting pedestrians using HOG, LoG, and Frame-2-Frame flow features. Both HOG and flow features are widely used in object and pedestrian detection. The results are shown

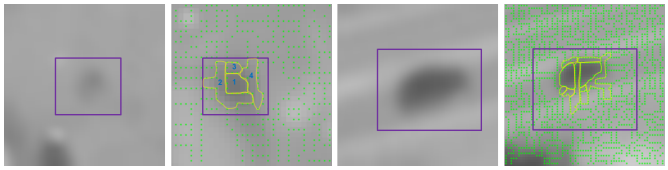


Figure 6: Examples of super-pixel from dataset I for a pedestrian (the left two panels) and a vehicle (the right two panels) from low contrast and low resolution aerial video. We used watershed segmentation with gradient magnitude at fine level to segment the objects.

in Table 1. We can see HOG and LoG obtain similar results, and both of them are much worse than that of ID and MIMS. This illustrates that pedestrians from aerial images in Dataset I have less discriminative appearance features which are unable to differentiate the pedestrians from other moving objects and noise. Although the flow feature works better than appearance based feature, it does not achieve acceptable detection rate either. We believe this is because the motion from only two frames is often noisy.

Interestingly, combining MIMS and appearance features further improves the performance, as shown in Per-pixel LoG+MIMS row of Table 1. It seems to tell us that, although the pedestrians are small, the contrast sensitive filter may still be helpful in detection. However, MIMS still provides the key contribution since the improvement is not much as compared to that of MIMS.

Other manifold learning approaches may be able to discover the intrinsic geometric structure of the data. For example, diffusion maps is able to perform multi-scale data analysis too and has shown good results [15]. However, unlike tensor voting, diffusion maps is not able to dig out the

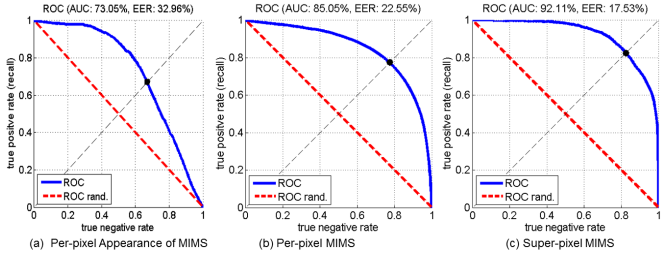


Figure 7: ROC curves from the experiment of Laplacian of Gaussian Filter, per-pixel MIMS and super-pixel MIMS. Overall, super-pixel based approach outperforms the per-pixel based approach and the Laplacian of Gaussian Filter has the lowest accuracy.

intrinsic local dimensionality and motion saliency. As we can see the results in Table 1, our approach performs much better than diffusion maps on this problem.

To further improve the system performance in terms of efficiency and accuracy, we also show results on super-pixels. The result is listed in Table 1, which outperforms all other approaches. One reason that accounts for this is that super-pixel leverages tensor voting by explicitly grouping spatial pixels in a frame-by-frame basis. This reduces the effect of noise to tensor voting at pixel level. An example of super-pixel representation of pedestrian and vehicle is shown in Figure 6. Figure 7 compares the ROC curve of pixel based and super-pixel based approach. We show two examples of our qualitative results from dataset I in Figure 9.

In order to compare with a public available aerial video pedestrian detection work in the literature, we compared our approach with the geometric approach in [20]. We achieved (in green) comparable result to their best performance sequence (see Figure 8 in red). It is worth noting that our approach does not rely on any shadow information, which is not reliable for pedestrian detection from low-contrast and/or low resolution aerial videos.

7. Conclusion

In this paper, we proposed a novel feature for detecting pedestrians in WAAS surveillance imagery. Our MIMS feature encodes the local structure of the motion pattern by computing the intrinsic dimensionality and saliency of the motion manifold at a number of scales . The discriminative scales are selected via a learning based approach. The local dimensionality and structure estimates enable us to differentiate background clutter, parallax and vehicles from pedestrians from noisy optical flow estimates. Our evaluation shows that the MIMS feature outperforms the state of art appearance and motion based features for pedestrian detection.

Table 1: Quantitative results from Dataset I. Coverage shows the detection rate and FPS shows the false positive score that is measured in a 60 seconds (120 frames) video within a square kilometer area.

Features	Coverage	FAPkmPmin
HOG	62%	1378
Laplacian of Gaussian (LoG)	63%	889
Frame-2-Frame Flow	67%	668
Intrinsic Dimensionality (ID)	69%	568
Structure Saliency (SS)	66%	1023
ID+SS (MIMS)	72%	493
Diffusion Map	70%	536
Per-pixel LoG+MIMS	78%	367
Super-pixel LoG+MIMS	80%	115

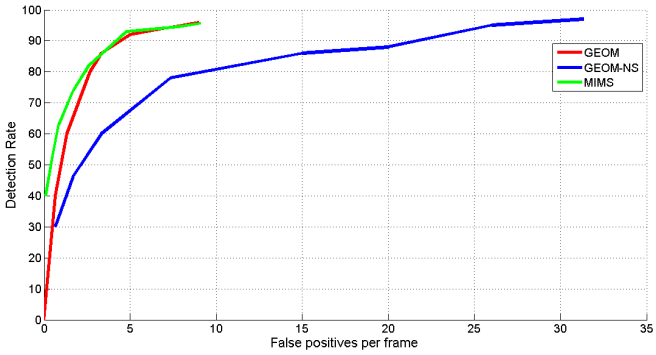


Figure 8: Comparison with Reilly's approach [20] using shadows. Note that our approach does not rely on shadow-cue which may not be available in all the aerial videos.

8. Acknowledgment

This project is funded by Office of Naval Research (ONR) under contract number: N00014-08-C-0339.

References

[1] M. Agrawal, K. Konolige, and M. R. Blas. Censure: Center surround extremas for realtime feature detection and matching. In *ECCV (4)*, pages 102–115, 2008. 5

[2] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *NIPS*, pages 831–837, 2000. 2

[3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV (4)*, pages 25–36, 2004.

[4] J. Choi, Y. Dumortier, J. Prokaj, and G. G. Medioni. Activity recognition in wide aerial video surveillance using entity relationship models. In *SIGSPATIAL/GIS*, pages 466–469, 2012. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005. 2, 5

Figure 9: Qualitative results from two video clips (best viewed by zooming in 4x). The left image is captured at a shopping mall where many pedestrians are working out/into buildings. The right image is captured at a location near a restaurant, where few pedestrians appeared. The detected pedestrians are labeled using a red circle. Their corresponding tracks within 20 frames are shown in green color. In the left example, there are altogether four pedestrians detected but one is false alarms when a dark vehicle is about to stop. In the right example, there are two pedestrians. One is walking on the pavement, and the other is crossing a road. A false detection is also showed up when the vehicle is about to stop. We believe these false alarms can be successfully removed by vehicle tracking which is relatively more mature than pedestrian detection approaches.

[6] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International Journal of Computer Vision*, 95(1):1–12, 2011. 2

[7] P. Doherty and P. Rudol. A uav search and rescue scenario with human body detection and geolocalization. In *Australian Conference on Artificial Intelligence*, pages 1–13, 2007. 2

[8] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*, pages 23–37, 1995. 5

[9] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000, 1998. 4

[10] A. Gaszczak, T. Breckon, and J. Han. Real-time people and vehicle detection from UAV imagery. In *Proc. SPIE Conference Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, volume 7878, 2011. 2

[11] K. J. LaTourette and M. D. Pritt. Dense 3d reconstruction for video stabilization and georegistration. In *IGARSS*, pages 6737–6740, 2012. 3

[12] M.-S. Lee. Detecting people in cluttered indoor scenes. In *CVPR*, pages 1804–1809, 2000. 2

[13] C. Liu. Beyond pixels: Exploring new representations and applications for motion analysis. In *PhD. Thesis. Massachusetts Institute of Technology*, 2009. 3

[14] J. Liu, Y. Yang, I. Saleemi, and M. Shah. Learning semantic features for action recognition via diffusion maps. *Computer Vision and Image Understanding*, 116(3):361–377, 2012.

[15] J. Liu, Y. Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *CVPR*, pages 461–468, 2009. 6

[16] P. Mordohai and G. G. Medioni. Dimensionality estimation, manifold learning and function approximation using tensor voting. *Journal of Machine Learning Research*, 11:411–450, 2010. 2, 3

[17] O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. In *CVPR*, pages 709–716, 2010. 2

[18] T. Pollard and M. Antone. Detecting and tracking all moving objects in wide-area aerial video. In *CVPR Workshops*, pages 15–22, 2012. 2

[19] J. Prokaj, X. Zhao, and G. G. Medioni. Tracking many vehicles in wide area aerial surveillance. In *CVPR Workshops*, pages 37–43, 2012. 1, 2

[20] V. Reilly, B. Solmaz, and M. Shah. Geometric constraints for human detection in aerial imagery. In *ECCV (6)*, pages 252–265, 2010. 2, 5, 7

[21] V. Reilly, B. Solmaz, and M. Shah. Shadow casting out of plane (scoop) candidates for human and vehicle detection in aerial imagery. *International Journal of Computer Vision*, 101(2):350–366, 2013. 2

[22] X. Shi, H. Ling, E. Blasch, and W. Hu. Context-driven moving vehicle detection in wide area motion imagery. In *ICPR*, pages 2512–2515, 2012. 1

[23] J. Sokalski, T. P. Breckon, and I. Cowling. Automatic salient object detection in uav imagery automatic salient object detection in uav imagery, 2010. 2

[24] Q. Yu and G. G. Medioni. Motion pattern interpretation and detection for tracking moving vehicles in airborne video. In *CVPR*, pages 2671–2678, 2009. 2

[25] X. Zhao and G. G. Medioni. Robust unsupervised motion pattern inference from video and applications. In *ICCV*, pages 715–722, 2011. 2