Preprints of the 2012 20th Mediterranean Conference on
Control & Automation (MED), Barcelona, Spain, July 3-6,
2012

Object Classification in Traffic Scenes using Multiple Spatio-Temporal Features

Guruprasad Somasundaram, Vassilios Morellas
and Nikolaos Papanikolopoulos
Deparment of Computer Science
University of Minnesota
200 Union Street SE, Minneapolis
MN 55455
{guru,morellas,npapas}@cs.umn.edu

Saad Bedros
Honeywell Labs
1985 Douglas Drive N
Golden Valley
MN 55422
saad.bedros@honeywell.com

***Abstract*— Object classification is a widely researched area in the field of computer vision. Lately there has been a lot of attention to appearance based models for representing objects. The most important feature of classifying objects such as pedestrians, vehicles, etc. in traffic scenes is that we have motion information available to us. The motion information presents itself in the form of temporal cues such as velocity and also as spatio-temporal cues such as optical flow, DHOG [6], etc. We propose a novel spatio-temporal feature based on covariance descriptors known as DCOV which captures complementary information to the DHOG feature. We present a combined classifier based on properties of tracked objects along with the DHOG and the DCOV features. We show based on experiments on the PETS 2001 dataset and two video sequences of bicycle and pedestrian traffic that the proposed classifier provides competent performance for distinguishing pedestrians, vehicles and bicyclists. Our method is also adaptive and benefits from the availability of more data for training. The classifier is also developed with real-time feasibility in mind.**

I. Introduction

Video surveillance is one of the prime motivators for computer vision research. Powerful computer vision algorithms have obviated the requirement for hours of painstaking manual examination of traffic footage. The amount of information that can be extracted is difficult to quantify and over the years innovations in computer vision and machine learning research have aided us in obtaining more and more information from traffic videos which would have earlier been deemed difficult. For automatic extraction of information about the traffic count of each object class, we need to develop powerful object classifiers.

The additional challenge involved in developing object classifiers for traffic data is the effect of view we have over the scene. Most traffic cameras have a bird's eye view of the scene and therefore the objects look much smaller than in many other scenarios. An example of surveillance under near and far view conditions is shown in figure 1. Near field is often favored since we obtain more resolution of the objects and therefore we obtain better classification performance. This is particularly true of feature based models such as bag-of-words [13] using interest point detectors like SIFT (Scale Invariant Feature Transform), SURF (speeded up robust features) etc. In far field situations there usually is not enough interest point features detected to build meaningful bag-of-words models. Under such circumstances we resort to low level reliable features. When the resolution of the images is low, we use a more overall description of the images using feature descriptors such as the histogram of oriented gradients (HOG) [9], region covariance descriptor [18], etc. In addition to describing the appearance using these descriptors we also provide spatio-temporal description of a sequence of objects using features like DHOG and the proposed DCOV features. These features are simple and efficient to compute. The DHOG as well as the DCOV features can be computed just from the gradient and color information of the extracted blobs. An Intel Quad Core processor (2.4GHz) machine with 4GB of RAM, can process upto 15 frames per second for a video of size 320×240 pixels.

Formulating a suitable classifier is as important as devising descriptive features. Support vector machines (SVM) [8] have been successfully employed for a variety of different classification and regression tasks. A thorough discussion of SVM based object classification using different types of features is provided in [20]. We can build multiple classifiers based on different types of features and combine their classification decisions suitably in order to enhance the overall performance. Often times the combined classifier outperforms any of the individual classifiers used in the combination. We propose a Naive Bayes classifier for combining classification using tracked object properties such as blob area, perimeter, eccentricity, solidity along with spatio-temporal features such as DHOG and DCOV. The combined classifier provides the best performance overall due to the complementary nature of the individual classifiers.

In the next section we discuss some relevant prior work in literature dealing with the problem of classifying moving objects. Then we propose the differential covariance feature (DCOV) based on the Jensen-Bregmann divergence of positive semi-definite matrices in section III. This is followed by a discussion of our method of combining different features for classification. Finally we analyze the performance of our classifier on real data obtained from the PETS 2001 dataset as well as two bicyclists and pedestrian traffic videos in section IV and provide concluding remarks in section V.

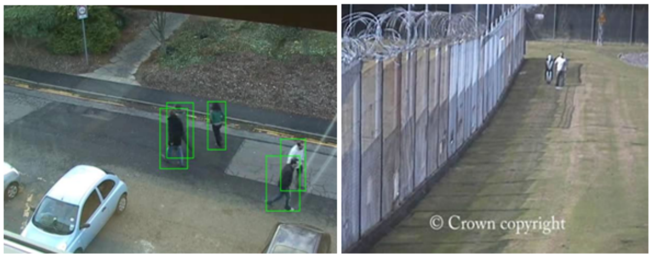


Fig. 1. The difference between near and far field-of-view surveillance. Objects are much smaller in far field thereby we can obtain much less appearance information.

II. Related Work

Object classification and detection literature is dominated by methods that perform classification using single images. These methods aim towards classifying objects under varying conditions of illumination, scale, orientation, shear, etc. The bag-of-words model [13] focuses on learning the distribution of stable features that can be reliably matched across varying conditions to recognize objects. Other models such as the parts-based discriminative object models [11], [10], [14] focus on identifying different parts of an object and understanding the spatial correspondence between them. These approaches are however more suitable to scenarios where there are more pixels in the target object. This is the near field situation.

The far field scenario suffers from fewer pixels on target thereby rendering the bag-of-words, parts-models, etc. unreliable. However for our analysis we deal with situations involving static cameras and moving objects. Under such circumstances we can perform background modeling and subtraction to obtain only the moving objects. These foreground objects are usually characterized based on simple morphological features such as blob area, aspect ratio, solidity, eccentricity, etc. [4], [3], [16]. These methods are not robust to tracking artifacts such as blob merging, splitting and inaccurate tracking resulting in wavering blob extraction. These effects cause simple measures such as area, perimeter, etc. to fail and these measures are often not view independent.

A very popular spatio-temporal measure is optical flow [5], [1] estimation. Accurate estimation of optical flow can yield discriminating information about the motion of objects. Since pedestrians move distinctly compared to vehicles, a suitable description of the motion is obtained using optical flow. In [19] segmentation of rigid bodies using dense optical flow is discussed. Despite many efforts to improve the computational burden of estimating optical flow, it is still considered demanding for real-time needs. Also pedestrians are not rigid bodies due to the amorphous nature of the silhouette of the human structure. Non-rigid body segmentation is feasible using optical flow albeit expensive [15].

Recently in [21], [6] methods that use simple appearance based features which also take advantage of the temporal information are discussed. Chen *et al.* introduced the DHOG feature based on the changes introduced in the HOG descriptor due to rigid and non rigid motion of vehicles and pedestrians. They argued that by using the weighted differences of HOG features discrimination can be obtained. Even though HOG features provide good description of objects they only capture the spatial distribution of gradients and their orientations. We propose the use of covariance feature which can measure the distribution of multiple features of different types such as color, first order and second order gradients, etc. in one representation. By observing the differences in the covariance feature (DCOV) in a manner similar to the DHOG feature we obtain additional information for discrimination between different object classes.

III. Approach

In this section we first propose the differential covariance descriptor (DCOV). Then we describe our baseline classifier based on blob properties and then finally we describe the Naive Bayes combined classifier which is used to enhance the baseline classifier using the DHOG and DCOV.

A. Differential Covariance Descriptor

The region covariance descriptor was introduced by Tuzel *et al.* for the purpose of object classification and detection. For any region of interest we can compute different kinds of features F at each pixel location in that region of interest. In the original paper of Tuzel *et al.* they propose the following set of features.

$$F(x,y) = [x,\ y,\ R(x,y),\ G(x,y),\ B(x,y) \\ |\frac{\partial I(x,y)}{\partial x}|,\ |\frac{\partial I(x,y)}{\partial y}|,\ |\frac{\partial^2 I(x,y)}{\partial x^2}|,\ |\frac{\partial^2 I(x,y)}{\partial y^2}|]^T \quad (1)$$

Here x,y are the pixel locations. R, G and B are red, green and blue color channel values at that location. I is gray scale intensity, and the partial derivatives represent the first and second order gradients along the x and y directions respectively. We can also add more feature of interest to this set providing more descriptive power. Given this feature set at each x, y location, we can compute the covariance descriptor of a particular region as

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \mu)(f_i - \mu)^T \quad (2)$$

Where f_i is the feature descriptor of the i^{th} pixel in the region of interest. μ is the mean feature of all pixels in the region. Hence given the 9 dimensional feature F, the covariance matrix C will be 9×9 dimensions. This covariance matrix measures the correlation between related features of each pixel in a region. It gives us a way of combining different types of feature into one descriptor. Just as the HOG descriptor models the distribution of oriented gradients, the covariance is related to the distribution of all the features over which it is computed. Traditionally covariance descriptors are used in nearest neighbor searches. A variety of distance metrics exist for comparing positive semi-definite matrices in general which can be applied to covariance matrices.

One such metric is proposed in [12] and described here:

$$\rho(C1, C2) = \sqrt{\sum_{i=1}^{n} ln^2 \lambda_i(C1, C2)} \quad (3)$$

Where ρ is the distance between two covariance matrices C1 and C2. Here λ_i are the generalized eigen values of C1 and C2. The log-euclidean distance metric has been studied for comparing diffusion tensors [2]. By taking the matrix logarithm of the covariance matrices and then directly comparing the upper triangular portion of the matrices with the euclidean metric gives good matching results and it is also faster for very large covariance matrices.

A more theoretically correct and fast metric for computing the distance between covariance matrices was proposed by Cherian *et al.* [7]. This metric is based on the Jensen-Bregmann logdet divergence (JBLD) and it is a dissimilarity measure between two positive semi-definite matrices. It is defined as follows:

$$J(C1, C2) = log|\frac{C1 + C2}{2}| - \frac{1}{2}log|C1.C2| \quad (4)$$

We determined that out of the three metrics, we obtained best run times with the JBLD metric. Also theoretical justification as to why this is a more suited metric is provided in [7]. Given the distance metric we can compute the differential covariance (DCOV) feature. We will be dealing with blob images obtained from tracking. These blobs are typically very small images. For example a pedestrian blob occupies around 20×40 pixels. Since these blobs are tracked across multiple frames, we have information about the distance traveled by the object between successive frames. For each frame we can compute the covariance descriptor of the blob: C_i for the i^{th} frame. Then the DCOV feature is computed as follows

$$dcov = \frac{\sum_{i=2}^{n} J(C_i, C_{i-1}) \times d_i}{D} \quad (5)$$

Where d_i is the distance traveled by the object between frame $i-1$ and i, and D is the total distance traveled by the object during the course of the tracking and n is the total number of frames. In comparison the DHOG feature is computed as follows:

$$dhog = \frac{\sum_{i=1}^{n} \Theta_i \times d_i}{D} \quad (6)$$

where

$$\Theta_i = 1 - \sum_{j} min(HOG_{i-1}^{j}, HOG_{i}^{j}) \quad (7)$$

is the intersection of the two histograms between successive frames i and $i-1$ and the superscript j represents the j^{th} component of the histogram. In both cases the features are weighted based on the distance traveled between successive frames to counter problems like objects becoming temporarily stationary.

B. Baseline Moving Object Classifier

Figure 2 describes the general flow of operation in our classifier. While processing the traffic video, we update a background model using the mixtures of gaussians approach [17]. After subtracting the background, we extract the foreground objects (blob) and use a Kalman filter on the blob position and velocity information to perform tracking. With tracking we obtain information about the blobs such as area, velocity, blob perimeter, eccentricity etc. Our baseline classifier is based on the approach in [16] which uses blob properties and velocities to classify pedestrians and bicyclists. However we modify their approach by training support vector machines on a set of blob features such as area, convex area, perimeter, eccentricity and solidity. In addition we learn a gaussian model for the velocities. Both the support vector machine and the gaussian pdf of velocities estimate a probability for each blob P_{svm} and $P_{velocity}$ as to which class they belong to. In this framework, we can also handle multiple moving objects in a single frame as they are tracked invidually. Then the classification is done as follows:

$$label = \arg\max_{class} P_{svm}(class) \times P_{velocity}(class) \quad (8)$$

The class that maximizes the product of the two estimated probabilities is the predicted label. This represents our baseline classifier. The properties used by this classifier are simple and fast to compute and can be reliably computed across varying conditions of resolution, viewpoint etc but they may not good features to use for classificatio. However for our test scenarios this classifier performs well.

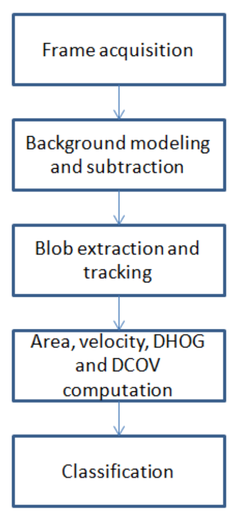


Fig. 2. A flowchart showing the sequence of processing performed on each frame of a video sequence.

C. Combined Naive Bayes Classifier

According to the Naive Bayes approach, each classifier estimates a probability of membership of an object to a particular class based on a feature X_i. Then the individual predictions can be combined in a Bayesian manner as shown in Eq. 9. In our case we have the baseline classifier based on morphological properties and velocity. Then we have classification using DHOG and DCOV values. For these two properties we again model them individually as gaussian distributions. With the three classifiers we can obtain a combined probability estimate which maximizes the product in equation 9.

$$P(class|X_{1....n}) = \arg\max_{class} \prod_{i=1}^{n} P(class|X_i) \qquad (9)$$

IV. Experiments and Results

We compare and analyze the different representation schemes and their corresponding classifiers on two datasets. The first dataset is the PETS 2001 dataset. This dataset consists of far field occurrences of pedestrians and vehicles. This dataset serves primarily as a test bench for evaluating tracking algorithms. However due to the relatively low resolution of the moving objects it presents a challenge for classifying objects using appearance based models. The distributions of area and perimeter values of pedestrians and vehicles from one sequence in this dataset are shown in figure 4. The distribution of velocities of pedestrians and vehicles from the same sequence is also shown. We do not use any calibration information. The training data is obtained from the training sequences for both cameras in this dataset (dataset 1 to 3). The test sequences are evaluated individually based on the models obtained from the corresponding training sequences. Then the overall performance is measured by averaging the accuracies for the individual sequences. The PETS 2001 is a rather small dataset in terms of the volume of the detected traffic objects. In each sequence there were only 20 to 30 objects out of which a majority were pedestrians. Table II shows the accuracies of classification of the baseline classifier, DHOG classifier, DCOV classifier and the combined classifiers separately. We notice that the combined classifier has better performance than any of the individual classifiers illustrating the fact that the information provided by each of the individual features are complementary to each other. In other words in situations where area or perimeter may not be enough to distinguish an object the DHOG value or the DCOV value might be useful. Because we pack in multiple features into the covariance descriptor, the DCOV feature often tends to be more informative.

The second dataset that we evaluated upon consists of two traffic video sequences of pedestrians and bicyclists. The reason we use this dataset is that unlike vehicles and pedestrians, bicyclist and pedestrian velocities are not far apart. Also the morphological structure of bicyclists is quite different compared to pedestrians and vehicles. The bicyclist is also a composite mixture of a person and a bicycle, making the problem of appearance based classification between pedestrians and bicyclists non-trivial. In this dataset the first sequence is obtained from a bicycle trail where pedestrians and bicyclists are more commonly found as individuals. Pedestrians tend to move quicker due to activities such as jogging. Bicyclists also tend to have more speeds than usual.

Class	DHOG	DCOV
Pedestrian	0.25	0.28
Vehicle	0.15	0.18
Bicyclist	0.4	0.32

TABLE I
Generally observed DHOG and DCOV values for different classes.

Classifier	Accuracy
Baseline	90.4%
DHOG Gaussian	89.1%
DCOV Gaussian	90.4%
Baseline + DHOG	92.1%
Baseline + DCOV	94.2%
Combined	95.7%

TABLE II
PETS 2001 Dataset Results

More intra-class variability in the appearance of bicyclists can also be found due to the presence of tandems, recliners, etc. The second sequence in this dataset is obtained from a university scenario. In this case pedestrians tend to move slower and have the tendency to form groups. Bicyclists ride more carefully thereby moving slower and situations such as walking a bicycle are not uncommon thereby making the velocities more ambiguous. Sample frames from the two sequences are shown in figure 6. The area and perimeter value distribution and the velocity distributions from the two sequences are shown in figures 7, 8, 9(a), and 9(b) respectively. By analyzing the DHOG and DCOV values for pedestrians and bicyclists we can learn about the generalizability of these two features to more classes. This dataset has a lot more instances of pedestrians and bicyclists than what the PETS 2001 dataset had. There were about 350 bicyclists and 400 pedestrians between the two sequences. We used about 50 each for obtaining training data for the classifiers. The rest were used for testing. The experimental results on the two datasets are illustrated in tables III and IV respectively. Again we notice an improvement in the performance of the baseline classifier when information from the DCOV and DHOG features are added. In one scenario the DHOG feature performs slightly better than the DCOV and in the other case the trend is reversed. This shows the complementary nature of these two features.

Another interesting observation was that the DHOG and DCOV values were somewhat consistent across different scenarios and settings. A general trend for the DHOG and DCOV values are shown in table I. From these results we observe that the DCOV and DHOG values can be extended to classify objects such as bicyclists in addition to pedestrians and vehicles. Both these features are simple and fast to compute and provide competitive performance on challenging datasets.

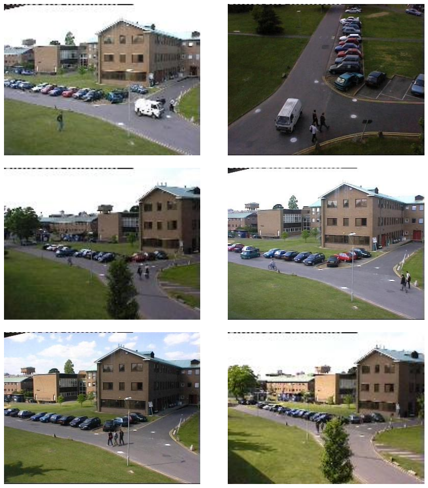
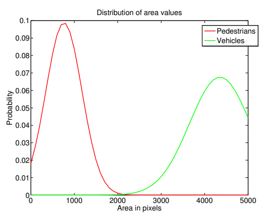
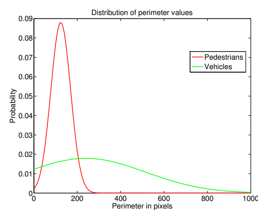


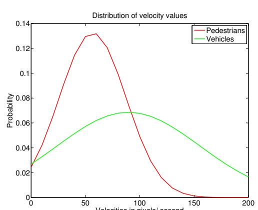
Fig. 3. Images from the PETS 2001 Dataset.



(a) Area



(b) Perimeter

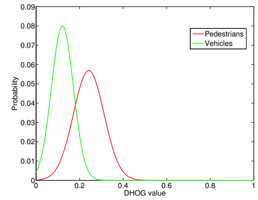


(c) Velocity

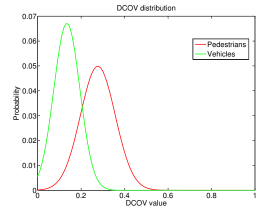
Fig. 4. Probability distribution of area,perimeter and velocity values for pedestrians and vehicles from a sample training sequence of the PETS 2001 dataset. Due to fewer samples the gaussian approximations of the distribution are shown. The velocity gaussian is used to make prediction based on velocity whereas the area and perimeter are used in a SVM classifier.

V. Conclusions

In conclusion we summarize our contributions. We propose a novel spatio-temporal feature called the differential covariance feature (DCOV). This feature models the change in appearance of objects as defined multiple types of cues including color when compared to the DHOG values which measures the change in the distribution of gradients only. We also show that the DCOV and DHOG are complementary and can be used in tandem in Naive Bayes setting to obtain improved classification performance. We validate our claim with classification experiments on the PETS 2001 (pedestrians vs. vehicles) dataset and real traffic video data from bicyclist and pedestrian traffic (pedestrians vs. bicyclists). The obtained DCOV values are consistent across different scenarios and can be used reliably as discriminant for different traffic object classes. Since the computed features are



(a) DHOG



(b) DCOV

Fig. 5. Probability distribution of DHOG & DCOV values for pedestrians and vehicles from a sample training sequence of the PETS 2001 dataset.

scalars, simple classifiers suffice to provide discrimination.

VI. Acknowledgments

This material is based upon work supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract #911NF-08-1-0463 (Proposal 55111-CI), Honeywell Labs, and the National Science Foundation, through grants #IIP-0443945, #IIP-0726109, #CNS-0708344, #CNS- 0821474, #IIP-0934327, and #IIS-1017344.

References

[1] L. Alvarez, J. Weickert, and J. Sánchez, "Reliable estimation of dense optical flow fields with large displacements," *International Journal of Computer Vision*, vol. 39, no. 1, pp. 41–56, 2000.

[2] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic resonance in medicine*, vol. 56, no. 2, pp. 411–421, 2006.

[3] B. Bose and E. Grimson, "Improving object classification in far-field video," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[4] L. Brown, "View independent vehicle/person classification," in *Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*. ACM, 2004, pp. 114–123.

Fig. 6. Top: Images from the university walkway sequence. Pedestrian crowds are more common. Bottom: Images from the bicycle trail.

Classifier	Accuracy
Baseline	89.76%
DHOG Gaussian	75.45%
DCOV Gaussian	78.16%
Baseline + DHOG	90.04%
Baseline + DCOV	92.11%
Combined	93.88%

TABLE III
Performance on Bicycle Trail Data

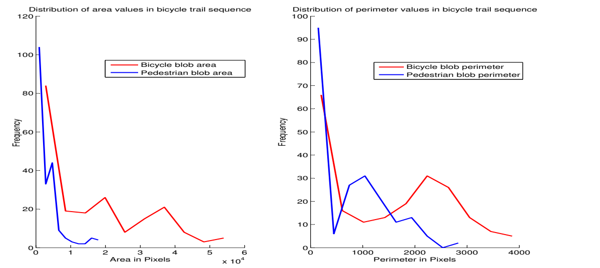


Fig. 7. Area and perimeter distribution of bicyclists and pedestrians in the bicycle trail. Area values are separated well however the perimeter values are not.

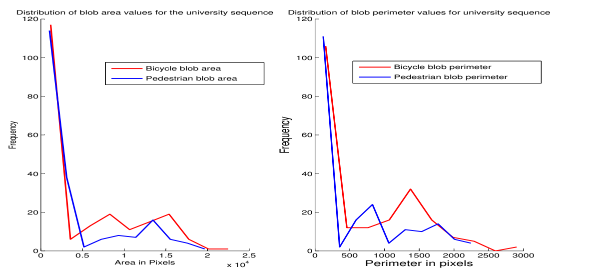
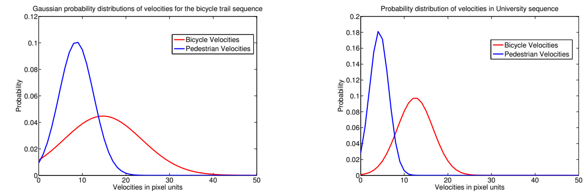
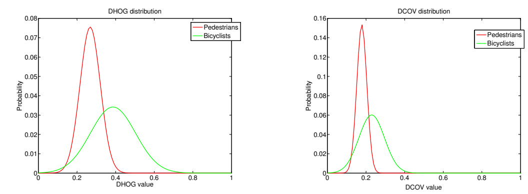


Fig. 8. Area and parameter distribution of bicyclists and pedestrians in the University walkway. Both area and perimeter distributions are not separated well.



(a) Bicycle Trail (b) University walkway

Fig. 9. Velocity distribution of bicyclists and pedestrians.



(a) DHOG (b) DCOV

Fig. 10. Probability distribution of DHOG & DCOV values for pedestrians and bicyclists from training data obtained real traffic videos.

[5] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," *Computer Vision-ECCV 2004*, pp. 25–36, 2004.
[6] L. Chen, R. Feris, Y. Zhai, L. Brown, and A. Hampapur, "An integrated system for moving object classification in surveillance videos," in *IEEE Fifth International Conference onAdvanced Video and Signal Based Surveillance, 2008. AVSS'08.* IEEE, 2008, pp. 52–59.
[7] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Efficient similarity search for covariance matrices via the jensen-bregman logdet divergence," in *International Conference on Computer Vision (ICCV)*, 2011.
[8] C. Cortes and V. Vapnik, "Support vector networks," in *Machine Learning*, 1995, pp. 273–297.
[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition*, C. Schmid, S. Soatto, and C. Tomasi, Eds., vol. 2, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005, pp. 886–893. [Online]. Available: http://lear.inrialpes.fr/pubs/2005/DT05
[10] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *2009 IEEE 12th International Conference on Computer Vision.* Ieee, 2010, pp. 229–236.
[11] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.
[12] W. 'Forstner and B. Moonen, "A metric for covariance matrices," *Technical report, Dept. of Geodesy and Geoinformatics, Stuttgart University (1999).*
[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.
[14] J. Liebelt, C. Schmid, and K. Schertler, "Viewpoint-independent object class detection using 3d feature maps," in *IEEE Conference onComputer Vision and Pattern Recognition, 2008.* IEEE, 2008, pp. 1–8.
[15] A. Lipton, *Local application of optic flow to analyse rigid versus non-rigid motion.* Carnegie Mellon University, The Robotics Institute, 1999.
[16] Y. Malinovskiy, J. Zheng, and Y. Wang, "Model-free video detection and tracking of pedestrians and bicyclists," *Computer-Aided Civil and Infrastructure Engineering*, vol. 24, no. 3, pp. 157–168, 2009.
[17] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 1999.
[18] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *Computer Vision–ECCV 2006*, pp. 589–600, 2006.
[19] J. Weber and J. Malik, "Rigid body segmentation and shape description from dense optical flow under weak perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 139–143, 1997.
[20] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Conference on Computer Vision and Pattern Recognition Workshop*, June 2006.
[21] L. Zhang, S. Li, X. Yuan, and S. Xiang, "Real-time object classification in video surveillance based on appearance learning," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on.* IEEE, 2007, pp. 1–8.

Classifier	Accuracy
Baseline	94.69%
DHOG Gaussian	88.4%
DCOV Gaussian	86.59%
Baseline + DHOG	95.12%
Baseline + DCOV	95.07%
Combined	96.1%

TABLE IV
PERFORMANCE ON UNIVERSITY WALKWAY DATA