MultiClass Object Classification in Video Surveillance System

Mohamed Elhoseiny, Amr Bakry

***Abstract*—There is growing demand for automated public safety systems for detecting unauthorized vehicle parking, intrusions, un-intended baggage, etc. One impacting factor for these applications is object detection and recognition in surveillance systems. This is chalengeable problem since the purpose of the surveillance videos is to capture wide landscape of the scene; resulting in small, low resolution and occluded images for the objects. The goal of this project is to design and implement recognition system for objects in outdoor surveillance videos. In this paper we extracted as many as 25000 features to make an inherent study of the system with many parameter settings. We managed to build up an efficient Object classification system with different configurations. The system was evaluated based on various parametrs and models mainly SVM and AdaBoost. Configuration of the systems include both domain shift features and feature selection.**

I. Introduction

Object classification is an important building block that impacts reliability of many applications of surveillance systems, including the public safety applications and video indexing/tagging to semantically search given video. Though, object classification in outdoor surveillance systems is chalengeable problem, for the following reasons: uncontrollable circumstances (e.g. fog, rain, lighting and haze); incomplete appearance details of the moving objects due to occlusions and large distance between the camera and the moving objects; and very low images resolution since we sometimes have to deal with moving object occupying small area (~50 squared pixels) in the video frames. Figure 1 shows examples of annotated images captured from surveillance cameras. These reasons altogether make the state of the art approach [[12]] for object detection and recognition barely detect object in a surveillance frames however it does good job for recognizing object in non-surveillance images. Besides the quality performance of our module, there is another factor we have to consider, time performance of the classification. This factor is crucial if we are going to use our module for further processing of the video such as activity recognition.

This paper presents a study that ends up with a robust object classification system. The cotribution of this paper could be highlighted in the following points:

1) Building Object Detection system with high recognition accuracy.
2) Showing the contrast between involving Appearance features (HOG [[8]]$\approx$ 25000 features) and Non-Visual Features ($\approx$ 150 features).
3) Tool was built up to help treating and filtering noisy features and selecting the most efficient set of features.
4) Using different classification approaches (SVM and AdaBoost), along with different parameters settings and Cross Validation methodology.

The rest of this paper is organized as follows. Section 2 gives brief survey for object detection and tracking as background for this work. Section 3 lists the related work and the existing literature for object recognition and classification. In section 4, we present the work flow and proposed approach. In section 6, we show the experimental results and discussion. Finally, section 7 presents the conclusion and the future work.

II. Background

Our system could be classified as video processing system and hence. Our work share a common front end in Video processing system which is mainly Object Classification and Tracking. The following two section presents relevant work in object detection and classification.

A. Object Detection and Segmentation

A lot of research work has been done to achieve high accuracy of object detection till the moment. The challenge has been to maintain the balance between precise detection and real time processing. Detection has different jobs to do depending on the application. This section presents the outer shell of object detection as a major area in computer vision research. Color based Object Detection(CBOD) , an approach of object detection that is applied where color of object is predefined or labeled with specific color. In these static environments detection is done by thresholding around the defined color. This approach is accurate and can be easily achieved in real time, however; it is dummy against unknown environment where colors of the objects are not predefined. In the same way, skin detection was investigated as an important application under CBOD [14], [15] . It has been used as a preprocessing to detect skin colored objects such as hand, face.

Another approach that depends on motion is Motion based Object Detection (MBOD). MBOD depends on building models to characterize areas that belong to the background and accordingly foreground objects is the complement (Figure 3). This technique is the most widely used in surveillance systems.

Massive research efforts have been done seeking robust background model of MBOD for different environments. Thanarat Horprasert et al presented an algorithm that is able to detect moving objects with static background but they handled shadow removal as a challenging problem because shadow is also moving but it is not an object [16]. Rui Tan,et al built a model specialized for extract vehicles as the main function of

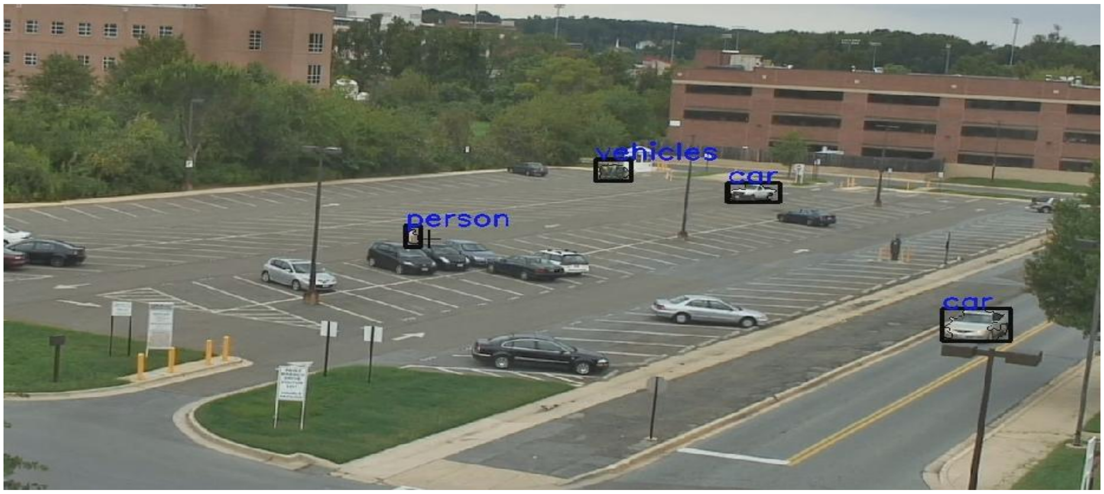


Figure 1: Online run of our system

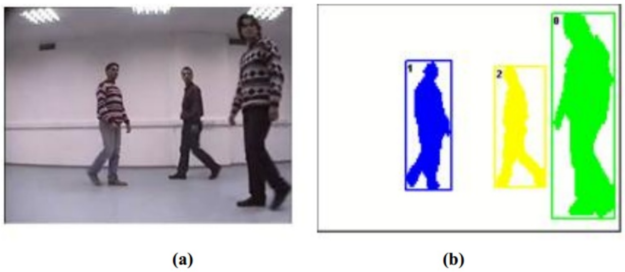


Figure 2: (a) Current Frame , (b) Labeled Detected objects[20].

traffic segmentation [17]. Rui Tan,et al used Adaptive Gaussian mixture model in which the dimension of the parameter space at each pixel can adaptively reflects the complexity of pattern at the pixel. Sheng-Yan Yang, et al combined spatial and color coherency with the pixel-wise GMM to determine the background model [18] that leads to an accurate object contours even in dynamic scenes. Actually presented work above assumes that camera is static, so some of the work has been directed to achieve discrimination of objects even when camera is in motion. Yan Zhang et al presented robust moving object detection algorithm [19]where camera is fixed on a moving vehicle, Experimentations on numerous real world driving videos have shown the effectiveness of the proposed technique. Moving objects such as pedestrians and vehicles up to 40 meters away from the camera have been reliably detected at 10 frames per second on a 1.8GHz PC. After each pixel of the frame is classified as background or foreground pixel, the foreground regions are segmented to define regions of objects (Figure 2) using Connected Component Labeling Algorithm [20].

B. *Object Tracking*

Object tracking has the function of mapping corresponding objects across frames captured from image acquisition system either by camera or stored video for the aim of maintaining their location (Fig. 2.7). Object tracking simply do a tangible job for which object detection, classification are performed.

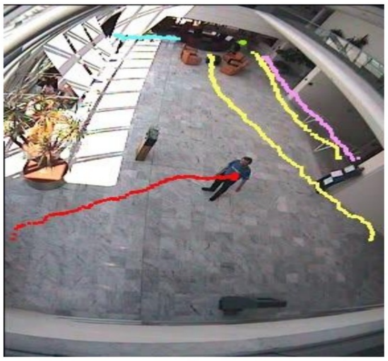


Figure 3: Object Tracking

Figure 3 Object Tracking Dedeoğlu et al [21] exploited the object features such as size, center of mass, bounding box and color histogram which are extracted in previous steps to establish a matching between objects in consecutive frames. Also their approach handles occlusion in a heuristic way by defining the state of merging the objects and the state of object split. A combination of correlogram and histogram information is used to model object and human color distributions by Mart Balcells Capellades, et al [22]to be able to detect when people merge into groups and segment them during occlusion. Identities are preserved during the sequence, even if a person enters and leaves the scene. The system is also able to detect when a person deposits or removes an object from the scene.

III. Related Work

This section presents both work related to object classification generallyu and still images and minor work present in videos.

A. *Object Classification in still Image*

There were various methods used for object classification in still images. An example of still images datasets include PASCAL dataset (Figure) .In the first part, methods for object classification typically extract features by applying some

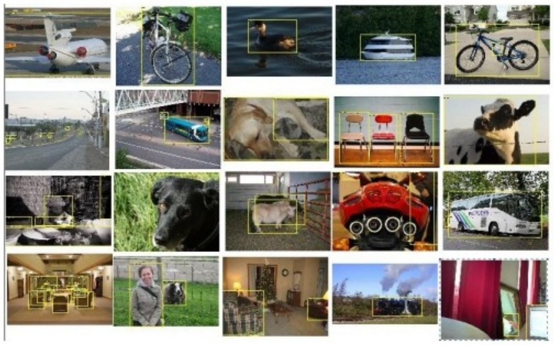


Figure 4: PASCAL dataset sample images

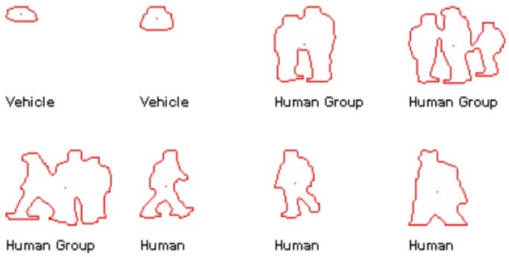


Figure 5: Contour Based Object Classification [21]

salient point detectors on the images. The survey by Schmid et al. [28] evaluated the repeatability rate and information content of various interesting point detectors. They compared contour based, intensity based and parametric model based methods. They found that Harris point detector [29] and its multi-scale variation perform better or at least equivalent to other detectors in two aspects: repeatability and information content. Matas et al. [30] proposed detection algorithm for an affinely-invariant stable subset of extremal regions, named the maximally stable extremal regions (MSER). Integrated in the SIFT descriptor [31], the difference of Gaussian (DoG) is also a good keypoint detector and widely used. One comparison shows that the salient parts in images are detected no matter it belongs to objects or noisy background. Some other methods for scene categorization [32], [33] just used regular grid on the images to extract features from rectangle patches. Random sampling is also used [34]. In these systems, salient regions are detected in the image but not all are supposed to be keypoints of object that we are looking for. Some will lie on the background or cluttered. The successful usage of these points after detection will depend on descriptors and classification.

B. Object Classification in Videos

Object classification has the task of categorizing the object according to its type (e.g. ball, car, human, etc). The raw input is the silhouette of the object to be classified. This type of classification is named shape based classification and this is the commonly used type for surveillance systems generally or action recognition specially. Dedeoğlu, Yiğithan et al presented an approach in [21] that is able to classify objects as human, human group and vehicle (Figure 2.5) based on silhouette template database. Distance function is measured between the query silhouette to be classified and the database. The query silhouette will be categorized as class C if the shortest distance is found between that silhouette and a silhouette belonging to class C within specified tolerance considering probability of silhouettes of untrained objects.

Jianpeng Zhou et al presented human classification algorithm based on codebook learning named DSCL (distortion sensitive competitive learning) [23] as a part of human tracking system. The concept of object classification is also used to categorize and classify postures of the same object.[24] Where posture of human is classified using Support Vector Machine (SVM) that have as input affine invariant Fourier descriptor of the human contour corresponding the posture. The same idea can be used to determine posture of hand or sign language alphabetic . [25], [26] are instances of using classification for the aim of sign language translation. (It is worthy to mention that in sign language application skin color is the approach of object detection, however in human postures classification is preprocessed by Motion Based Object Detection). Another direction that uses a combination of features is the work by Yehezkel and Boaz [27]. They presented an approach is classify objects based on various features.

IV. Proposed approach

In this section, we present in details how work flows in this project, and the proposed approaches. Figure 6shows the architecture of the proposed.As shown in the figure, The system takes as an input the video frames which is input to background subtraction module that detection objects from motion. Then Object detection information (i.e. bounding boxes, contour of extracted objects current frame, binary frame) is relayed to feature Extraction Module that extract the object's features. Third step is configuration phase that take the full length feature vector X and derive another feature vector X' out of it. Finally, object classification phase takes the feature vector $X'$and used it for learning or classification depending on the mode of the system. Each of the following subsections intends detail the system phases.

A. Dataset and Background subtraction

Building security and activity recognition system might be an important application/extension to our work. In such systems, the main subject is visual events. So in this work, we focus on moving objects only. Our main dataset is VIRAT dataset [6], [7], which is designated for activity recognition purposes. For extracting the moving objects, we have used background subtraction technique [2]. Yet, we found that the resulting blobs are noisy and does not help greatly for extracting the entire area for moving objects. So that, we relax our condition some how and use the provided bounding box annotation by the dataset for extracting complete objects. Even though, we use the results of background subtraction for acquiring some features for the objects, as described bellow.

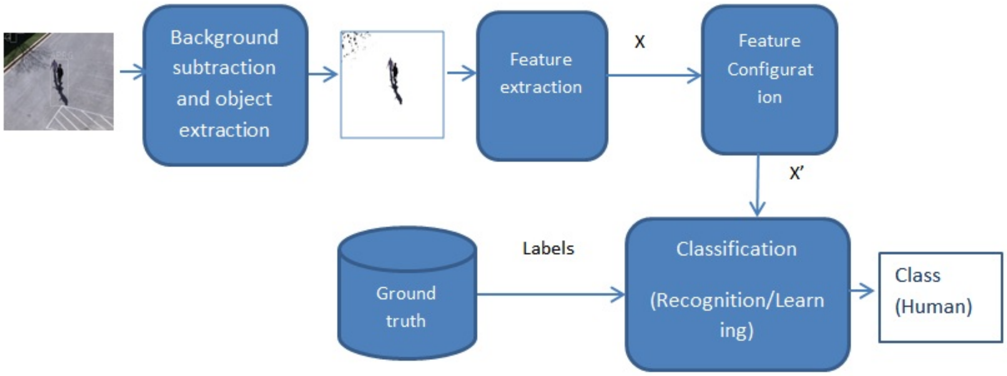


Figure 6: System Architecture

B. Feature extraction

As aforementioned, the purpose of this work is to study the effectiveness of features and evaluation different Machine Learning algorithms on the extracted features. For this purpose, we extracted many features detailed in the following sections. We have also developed a tool to extract these features from the videos in a balance and less noisy approach.

1) HOG features: Histogram of Oriented Gradients (HOG) [8] are feature descriptors used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. The approach basically performs four main operations. (1) Gradient computation, (2) Orientation binning (Discretization from the computed gradient, (3) Descriptor Blocks, The purpose of this step is to avoid changes in illumination and contrast, the gradient strengths should be locally normalized. This requires splitting pattern into grouping the cells together into larger, spatially connected blocks. We followed *CVPR2006-WorkShop* guidelines to build up a HOG descriptor for each of the K classes in our system.

2) Luminance Symmetry: Many objects appear in the image with symmetric texture and shape with respect to their symmetry line. A good assumption of the symmetry line is the major axis of the ellipse bounding the object. The purpose of this feature is to distinguish between objects with such symmetry (e.g. human, cars) to asymmetric ones (e.g. clutter, body organs). One way to capture this property is by calculating the histogram of I on both sides of the symmetric line and comparing between them using histogram intersection, Bhattacharyya distance or EMD. Equation 1shows the calculation of Luminance symmetry.

$$L_{sym} = \frac{1}{C}\frac{2}{w}\sqrt{\sum_{i=1}^{h}(\sum_{j=1}^{w/2} I(i,j).B(i,j) - \sum_{j=w/2+1}^{w} I(i,j).B(i,j))^2} \tag{1}$$

where h and w is the size of the bounding box, (i, j) of the object segmented from Background subtraction, C is the maximal luminance level possible, as illustrated in Fig. 2.B is the binary mask image, I is the original image in gray-scale.

3) Central Moments: Central moments – Early work by Hu [9] applied statistical moments to image analysis defining the Cartesian moments. Extending them to be invariant to translation and scale. We used the 7 scale-invariant Hu moments. on the I (The gray-scale version of the object's image).

4) ART moments: Angular radial transform (ART) is a moment-based image description method adopted in MPEG7 [10]as a region-based shape descriptor. It provides a compact and efficient way to express pixel distribution within a 2D object region; it can describe both connected and disconnected region shapes.We used here the ART with the standard configuration$nAngle = 12, nRadius = 6$ which gives in total $6 * 12 - 1$= 72 features.

5) Cumulants: To increase the inter-class variability. We masked the pixels of S0 using B0 to consider only foreground pixels and calculate three textural properties.

1) Mean value ($E\big[X\big]$) of the intensity , which is mostly high for bags (most bags have high contrast with the background) compared to clutters.
2) Standard deviation ($E\big[(X-\mu)^2\big]$) of the intensity histogram, which is mostly low for bags (their luminance is mostly homogenous) and high for clutters (luminance is often non-homogeneous).
3) Skewness ($\frac{E\big[(X-\mu)^3\big]}{E\big[(X-\mu)^2\big])^{3/2}}$) of the intensity histogram is mostly negative for objects and positive for remaining objects.

6) Horizontal and Vertical Projection: Horizontal (or vertical) projection is a histogram in which each bin, $HP_{i,+}$ (or$VP_{+,j}$) (Equation 2) corresponds to the sum of the pixels in row i (or column j). This feature capture histogram variation which discriminate between many object with low resolution property.

$$HP_{i,+} = \sum_{i} B^{'}(i,j), VP_{+,j} = \sum_{ji} B^{'}(i,j) \tag{2}$$

7) Morphological features: Finally we extract 4 more shape pieces of information . (1) Anthropometry (Ath) which is in the fixed relations between human body parts. (2) Compactness (Cmpct) which measures complexity of the shape, (3) Aspect ratio (AR) which is the ratio between the width and the height of the object, (4) Solidity (SD) which measures the portion of concave parts in the shapes. We used the following equation to evaluate these features.

$$Ath = \frac{H}{P}$$

$$Cmpct = \frac{Ar}{P^2}$$

$$AR = \frac{W}{H}$$

$$SD = \frac{Ar}{CHAr}$$

,where H , W are the width and the height the bounding box of the object, P is the perimeter of the object's contour. Ar if the contour area of the object, and $CHAr$ is the convex Hull area of the object.

C. Feature Configuration

In our study here after we extract the feature detailed in Section 4.2. Our system have 4 type of configuration for the features that is input to the learning phase.

1) Full length feature vectors (i.e. $X' = X$)
2) PCA features (i.e. $X' = PCA(X)$)
3) Selected features (i.e. $X' = SelectedFeatures(X)$)

The experimental results section more about the settings used to evaluate the proposed system.

D. Object classification

Finally, classification comes in. It is now the time to identify the class of the 2D image we have in hand. We have list of features X every row of this matrix represents feature vector for every object. The matrix X is of size $N \times M$, where N is the number of samples and M is the number of features. In addition to that we have ground truth labeling $Y_{N\times 1}$ acquired from VIRAT dataset, every values in this vector represents label for every object. So we use X and Y for learning classifier and then use that for classifying the objects. The result is the column vector $\hat{Y}_{N\times 1}$. We have used two different classification techniques *SVM* and *AdaBoost.*

For SVM, We used C-SVC version of SVM in our system[11]. The objective function of C-SVC is

$$\min_{\mathbf{w},\xi,b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \right\}$$

subject to

$$\begin{aligned} y_i(\mathbf{w}\cdot\mathbf{x_i} - b) &\geq 1 - \xi_i \\ \xi_i &\geq 0; \quad \forall i = 1,2,...,N \end{aligned}$$

We adopted the approach in [11]to perform multi-object classification.

For AdaBoost [7] and [8], this very well known machine learning technique is based on boosting the results of many weak classifiers to get a powerful one. In our case, we used AdaBoost of stump, by stump we mean the weakest classifier all over the world which is one level binary tree. There are different variants of AdaBoost such as Discrete AdaBoost, Real AdaBoost, LogitBoost, and Gentle AdaBoost [7]. On the other side, as any other classifier, AdaBoost has two versions, binary classifier version (*One-vs-All*), and multi-object classifier version. In this project, we use Real Binary AdaBoost of Stump.

We have trained binary classifier for every class $k \in 1,2,...,K$, and $k = 0$ for the objects that we could not be able to classify (*Others*). Then we test every classifier, and we get K classification results($\hat{Y}_k$), which is a $zero-one$ vector. The value $\hat{Y}_{kj} = 1$, if the object $j$is classified as class k, and zero other wise.

$$\begin{aligned} \hat{Y}_{kj} &= 1; \text{Classifier } k\text{recognizes Object } j \\ &= 0; \text{Other wise} \end{aligned}$$

Where is the class index, and $j = 1,2,...,N$ is the object index. For acquiring the best accuracy, we tried many combination for parameters like number of weak classifiers (W) and weight trim rate (ρ), as revealed in details in experiments and results section. Number of weak classifiers is the maximum number of weak classifiers could be used. To understand meaning of weight trim rateρ, we need to remember that AdaBoost is weighting the samples during training the weak classifiers, so that the miss-classified sample is assigned to higher weight than the correctly classified one. i.e, number of training samples with weight more. So ρ is percentage of samples that are used in training of the next stage's weak classifier, i.e, $(1-\rho)\%$ of the samples are assigned $zero-$weight when training the next stump.

Next step for AdaBoost is to merge those$K$ label vectors, easily we can create new vector $\hat{Y}$, where

$$\hat{Y}_j = k\hat{Y}_{kj}; \forall j = 1,2,...,N$$

We have conflict problem here when for certain j, $\hat{Y}_{kj} = 1$ for $k = u$ and for $k = v$. We resolve this conflict by comparing *confidence of classifier* $u\psi_u(j)$ with *confidence of classifier* $v\psi_v(j)$ and set $\hat{Y}_j$by the label of the higher confidence value, i.e, we can write the equation as

$$\hat{Y}_j = \arg\max_k \hat{Y}_{kj}\psi_k(j); \forall j = 1,2,...,N.$$

V. Experiments and results

This sections presents the experimental framework of our work and the accomplished experiments. To show contrast in our study, we performed the following object classification experiments. (1) Appearance based classification based on HOG features (2) PCA based classification on the 25000 Features. (3) Feature Selection Based Classification using SVM (4) Feature Selection Based Classification using AdaBoost. To do classification we have build up a tool to generate the features from the videos presented in section 5.1. Experiment1, 2, 3,4 are presented in Sections 5.2, 5.3, 5.4,5.5 respectively.

	Win Size	block size	cell size,strid	Bin Size
Human	64 x 128	16 × 16	8 × 8	9
Car	104 x 56	16 × 16	8 × 8	18
Vehicle	120 x 80	16 × 16	8 × 8	18
Bike	104 x 64	16 × 16	8 × 8	9
Object	64 x 64	16 × 16	8 × 8	18

Table I: HOG settings

A. *Feature Extraction Tool from Videos*

For training purposes, we extracted the features described in Section 3.2 from VIRAT dataset [6], [7]. We are choosing among six classes {Human, Car, Vehicle, Object, Bicycle and Others}. Vehicle class is any moving machine other than car such as van and truck; Object is anything man can carry like boxes and back bags; Others is any unrecognizable object by our system. We follow the following rules to extract the features and used the annotation to build up a balance and less noisy features used while learning. Hence, We defined the following constraints for the extraction of features . We used the same constraints in recognition mode except for the last one because the object class is unknown.

1) Detection Percentage ($Dp > 30\%$): Detection percentage of object i in the current frame is defined as the percentage of the contour area of object i (as an output of the background subtraction model) to the bounding box area of the detected object.

$$Dp_i = 100 * \frac{C_{area_i}}{BB_{area_i}}$$

2) Overlapping Percentage ($OP < 10\%$): Overlapping percentage of object i is defined as the percentage of sum of areas of intersection between the bounding box of object and all other object's bounding boxes to the bounding box area of object i.

$$OP = 100 * \frac{BB_{area_i}}{\sum_{i \neq j} \cap(BB_i, BB_j)_{area}}$$

3) Motion Constraints: The imposed constrained on motion is that we extract feature when distance that the object i moved is greater than a threshold th (th is 5 pixels in this experiment)

4) Object Instance Constraint: Per object Instance constraint, This constraint is followed while extracting the data for learning to ensure various object's features in training and testing. Extracted features for the given object instance is limited to at most 10 features if it's traveled distance is less than th.

B. *Appearance Features (HOG Experiment)*

In this experiment we use VIRAT [6] dataset annotations as input, and then we build up new dataset for each object following *CVPR2006-WorkShop* guidelines for multi-object classification. Table I shows HOG settings used for each object class.

Figure 7 shows samples of extracted dataset used for the evaluation based on the settings in Table .We used 5 fold cross validation C-SVM on 80%-20% Training-Test split and 71.4% of accuracy was recorded.

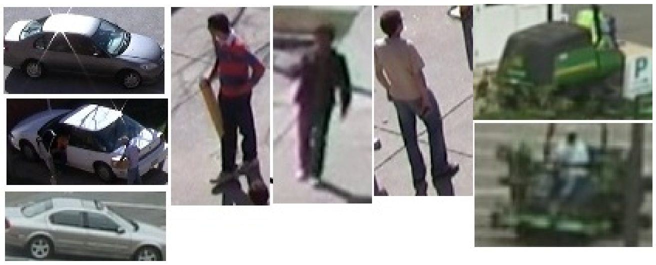


Figure 7: Extracted HOG patterns

C. *PCA Experiment*

The conclusion from the last experiment is that appearance based features (HOG in our case) performs bad in surveillance systems. The intuition behind that is low resolution of the detected object as apparent in Figure 7. In this experiment, we have computed first 30 eigen-vectors and projected the full length feature to evaluate the recognition performance on the new lower dimensional features.

$$y = A(x - \mu_x), A = [v_1, v_2, ..., v_k]^T,$$

where $v_1, v_2, ..., v_k$ are the first k eigen-vectors of $(X - \mu_x)(X - \mu_x)^T$. The dataset was split into two subsets (80% training and validation, 20% testing). Figure 1 shows 5-fold cross validation on C-SVM with different C-Values ranging from $v_1, v_2, ..., v_k$. From Figure 8, it's clear that the best C value is 1000, which is used to test on the remaining 20% of the data resulting in test accuracy of 89.9%. However, the overall accuracy seems satisfactory, there are two drawbacks of using PCA here. First full length feature vector (25000) has to be computed. Second we have to computer eigen-vectors for a big matrix 25000 × 25000. Both tasks are computationally intensive which violates the real time requirements of this system.

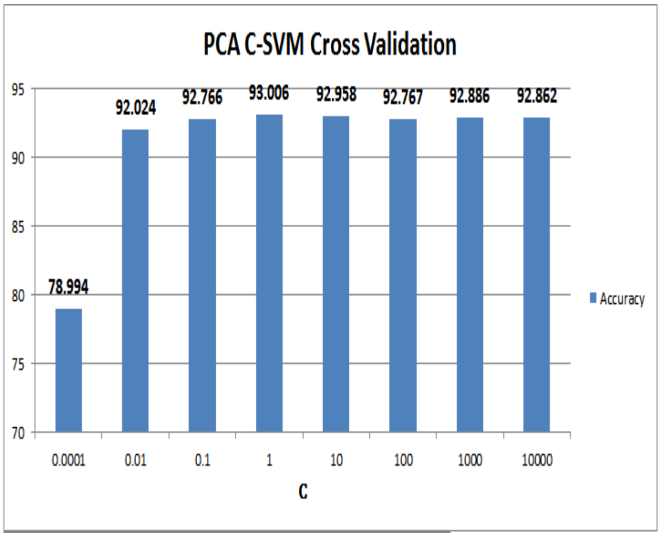
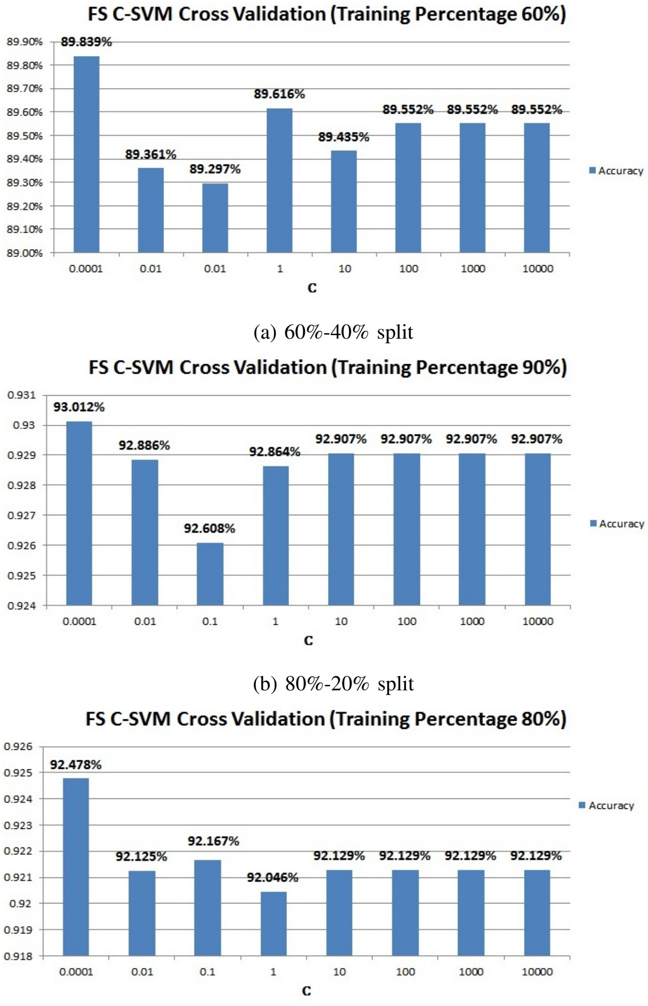


Figure 8: PCA Experiment on All Features (80% training percentage)

D. Selected Features SVM Experiment

The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. In have run an entropy based discretization approach on the features . An observation for the computed entropy is that Appearance features have little entropy and hence we performed the following experiment in which we have to compute only 142 features out of 25,000 features. Figure 9 shows the C-SVM performance on (90%-10% split) 5 fold cross validation. Best C value was100, which was used to evaluate the recognition accuracy on the test data. The recorded recognition rate was 99.3%. To show the effectiveness of our approach, we have computed the recognition accuracy of lower training percentage (80%-20% split, 60%-40% split). The recorded test accuracy are 96.5 and 87.5 on 80%-20% and 60%-40% splits, respectively (Table1II).



(a) 60%-40% split

(b) 80%-20% split

(c) 90%-10% split

Figure 9: Selected Feature SVM Validation Accuracy

Split	Test Accuracy
60%-40%	87.5
80%-20%	96.5
90%-10%	99.3

Table II: Selected Feature SVM Test Accuracy

E. Selected Features AdaBoost Experiment

As menstioned before, we have trained binary AdaBoost with stump for every class. Then using the confidence we merged those binary classification results to be mutli-object recognition. For every binary classifier k for recognizing objects in class k, we have learned the best parameters values.

We have train the classifiers using 5-fold cross validation for different values of weight trim rate $\rho \in 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 1.0$. Table III shows the training and testing accuracy results for every category for different values of ρ. The entire cells represent the average cross-validation error and last row lists the test accuracy.

ρ	Human (3942)	Car (4197)	Vehicle (286)	Objects (1188)	Bicycle (108)
0.10	0.594335	0.431907	0.029327	0.530660	0.010977
0.30	0.151745	0.277795	0.029327	0.530660	0.010977
0.50	0.151745	0.277795	0.029327	0.530660	0.603748
0.70	0.667836	0.690946	0.615973	0.877875	0.992978
0.90	0.993312	0.996228	1.000000	0.877875	1.000000
0.95	0.992281	0.994518	1.000000	0.877875	1.000000
1.00	0.991252	0.978054	1.000000	0.997941	0.998629
T.E	0.989977	0.995117	1.00000	0.997944	1.00000

Table III: AdaBoost: Cross Validation with ρ, and Test Error(T.E)

From Table III, we can notice many points. Choosing $\rho \in [0.90, 0.95]$ works good with most all classes. Recall, ρ is percentage of samples that are used in training the next stage's weak classifier. So this finding means that the algorithm tends to keep almost all sample for learning the upcoming weak classifier.

We have tried and we get many findings: first as we increase number of weak classifiers as we get better results, as in Table IV shows affect of changing weak classifiers count for the training and test accuracy for every category. But this increases the processing time. So by compromising these two constraints we find that choosing number of weak classifier close to dimensionality of the features vector space gives good enough results.

W	Human (3942)	Car (4197)	Vehicle (286)	Objects (1188)	Bicycle (108)
25	0.958678	0.952508	0.99606	0.997427	0.999656
100	0.984734	0.983714	1.00000	0.997941	1.00000
150	0.989881	0.990399	1.0000	0.997941	1.00000
200	0.993312	0.994518	1.0000	0.997941	1.00000
250	0.99571	0.997774	1.00000	0.997941	1.00000
300	0.998798	0.998973	1.00000	0.997941	1.00000
T.E	0.995888	0.998715	1.0000	0.997944	1.0000

Table IV: AdaBoost: Cross Validation with W, and Test Error(T.E)

From Table IV, we can see clearly that for the *balanced classes* (Human and Car), we need to high stumps count. Though, for unbalanced classes (Vehicle, Objects and Bicycle), small number of stumps is enough for getting the maximum accuracy. We mean by balanced here that for *One-vs-All* data, number of samples represent *One* and number of samples represent *All* are almost equal.

Final step, multi-label classification. After combining the results from all binary classifiers for producing the multi-label classifier, we get final accuracy=0.950782.

VI. Conclusion and future work

Due to the small size and low resolution of the frames, we have proved empirically that using visual features only like HOG features is not adequate for recognizing objects in videos captured from a surveillance system. We come up with a combination of several kind of features (luminance symmetry, central moments, ART moments, cumulants, horizontal/vertical projection and morphological features) in order to classify 2D image extracted from video. This combination of features is proved to be more efficient for object recognition. The classification techniques SVM and AdaBoost are doing good job for recognizing objects in surveillance system. Yet AdaBoost is doing slightly better than SVM. Also, we have proved empirically that using $C = 10^{-4}$ gives the best accuracy when using SVM. For AdaBoost, we have shown that using number of weak classifiers close to features vector size is the best ($W \in [100, 200]$) and using $\rho \in [0.90, 0.95]$ gives the best classification accuracy.

Due to the limited time frame, we still have many tasks waiting in the queue for future work. First, we want to activate the automation of extracting bounding boxes of the moving the objects using only background subtraction. Second, we are intended to use adequate feature selection technique for choosing the most efficient set of the extracted features, this helps in increasing the classification accuracy and decrease the processing time. Finally, we want to add activity recognition capability to our system. We find that the related literature are doing activity recognition base on extracting features for the trajectories of the moving object. We believe that our robust object recognition system can help in that.

References

[1] Dalal, Trigg, “Fast Human Detection by Boosting Histograms of Oriented Gradients”, CVPR, 2005

[2] A. Elgammal, D. Harwood, L. Davis, “Non-parametric Model for Background Subtraction”, 6th European Conference on Computer Vision, ECCV 2000.

[3] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, "Semantic Contours from Inverse Detectors" . ICCV 2011

[4] L. Chen, R. Feris, Y. Zhai, L. Brown, and A. Hampapur, “An Integrated System for Moving Object Classification in Surveillance Videos”, , Int. Conf. Adv. Video Signal-Based Surveillance, 2008.

[5] Raanan Yehezkel, Boaz Lachover,“Multiclass object classification for real time video surveillance systems”, Pattern Recognition Letters, 2011.

[6] Sangmin Oh et al, “A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video”, CVPR 2011.

[7] VIRAT Dataset http://www.viratdata.org/

[7] Friedman, J. H., Hastie, T. and Tibshirani, R. “Additive Logistic Regression: a Statistical View of Boosting”. Technical Report, Dept. of Statistics, Stanford University, 1998.

[8] Navneet Dalal , Bill Triggs. “Histograms of Oriented Gradients for Human Detection”. CVPR, 2005.

[9] Hu, M.K, “Visual pattern recognition by moment invariants”. IRE Trans. Inform. Theory IT-8, 179–187,1962.

[10] Manjunath, B., Salembier, P., Sikora, T., 2002. “Introduction to MPEG-7: Multimedia Content Description Interface”. Wiley & Sons.

[11] Fan, Rong-En and Lin, C. J, “A Study on Threshold Selection for Multi-label Classification”, National Taiwan University, 2007

[12] P. Felzenszwalb, R. Girshick, D. McAllester, “Cascade Object Detection with Deformable Part Models”, IEEE CVPR, 2010.

[8] Christopher M. Bishop, “Pattern Recognition And Machine Learning, Information Science and Statistics”, Springer, 2006

[14] S. L. Phung, A. Bouzerdoum, and D. Chai, "SKIN SEGMENTATION USING COLOR AND EDGE INFORMATION," in , Paris, 2003.

[15] F. Gasparin and R. Schettini, "Skin segmentation using multiple thresholding," in , San Jose, CA, USA, 2006

[16] T. Horprasert, D. Harwood, and L. S. Davis, "A Statistical Approach for Real-time Robust Background Subtraction and Shadow Detection," in , 1999.

[17] R. Tan, H. Huo, J. Qian, and T. Fang, "Traffic video segmentation using adaptive-K gaussian mixture model," in , 2006.

[18] S.-Y. Yang and C.-T. Hsu, "BACKGROUND MODELING FROM GMM LIKELIHOOD COMBINED WITH SPATIAL AND COLOR COHERENCY," in , 2006.

[19] Y. Zhang, S. J. Kiselewich, W. A. Bauson, and R. Hammoud, "Robust Moving Object Detection at Distance in the Visible Spectrum and Beyond Using A Moving Camera",CVPR , 2006.

[20] http://en.wikipedia.org/wiki/Connected_Component_Labeling

[21] Y. Dedeoğlu, “MOVING OBJECT DETECTION,TRACKING AND CLASSIFICATION FOR SMART VIDEO SURVEILLANCE (MSc Thesis)”. the Institute of Engineering and Science,bilkent university, 2004.

[22] M. B. Capellades, D. DeMenthon, and D. Doermann, "AN APPEARANCE BASED APPROACH FOR HUMAN AND OBJECT TRACKING," ICIP , 2003.

[23] J. H. Jianpeng Zhou, "Real Time Robust Human Detection and Tracking System," CVPR, 2005.

[24] V. Kellokumpu, M. Pietikäinen, and J. Heikkilä, "Human Activity Recognition Using Sequences of Postures," MVA , 2005.

[25] M. Zahedi, "Robust Appearance-based Sign Language Recognition," in , Aachen University, 2007.

[26] R. Akmeliawatil, M. P.-L. Ooi, and Y. Chow Ku, "Real-Time Malaysian Sign Language Translation using Colour Segmentation and Neural Network," in nstrumentation and Measurement Technology Conference , 2007.

[27] Raanan Yehezkel, Boaz Lachover, “Multiclass object classification for real time video surveillance systems”, Pattern Recognition Letters, 2011.

[28] Schmid C., Mohr R., Bauckhage C.: “Evaluation of interest point detectors.” Int. J. Comput. Vision 37, 2 (2000), 151–172. 15

[29] Harris C., Stephens M.: “A combined corner and edge detector”. In Proceedings of the 4th Alvey Vision Conference (1988), pp. 147–151. 15

[30] Matas J., Chum O., Urban M., Pajdla T.: “Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22”, 10, 761 – 767. British Machine Vision Computing 2002.

[31] Lowe D. G.: “Object recognition from local scale-invariant features”. In ICCV “, p. 1150. 15, 18, 41,1999

[32] Vogel J., Schiele B.: “A semantic typicality measure for natural scene categorization”. In DAGM-Symposium , pp. 195–203. 15,2004

[33] Li F.-F., Perona P.: “A bayesian hierarchical model for learning natural scene categories” In CVPR ’05: Proceedings of the 2005 IEEE Computer Society

[34] Vidal-Naquet M., Ullman S.: “Object recognition with informative features and linear classification”. In ICCV , pp. 281–288. 15,2003