ITS INSTITUTE
Intelligent Transportation Systems

Deployment of Practical Methods for Counting Bicycle and Pedestrian Use of a Transportation Facility

Final Report

Prepared by:

Guruprasad Somasundaram
Vassilios Morellas
Nikolaos Papanikolopoulos

Department of Computer Science and Engineering
University of Minnesota

CTS 12-01

CENTER FOR TRANSPORTATION STUDIES   UNIVERSITY OF MINNESOTA

Technical Report Documentation Page

1. Report No. CTS 12-01	2.	3. Recipients Accession No.

4. Title and Subtitle Deployment of Practical Methods for Counting Bicycle and Pedestrian Use of a Transportation Facility	5. Report Date January 2012 6.
7. Author(s) Guruprasad Somasundaram, Vassilios Morellas, Nikolaos Papanikolopoulos	8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Computer Science and Engineering University of Minnesota 200 Union Street SE Minneapolis, MN 55455	10. Project/Task/Work Unit No. CTS Project #2010080 11. Contract (C) or Grant (G) No.
12. Sponsoring Organization Name and Address Intelligent Transportation Systems Institute Center for Transportation Studies University of Minnesota 200 Transportation and Safety Building 511 Washington Avenue SE Minneapolis, MN 55455	13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code

15. Supplementary Notes
http://www.its.umn.edu/Publications/ResearchReports/

16. Abstract (Limit: 250 words)

The classification problem of distinguishing bicycles from pedestrians for traffic counting applications is the objective of this research project. The scenes that are typically involved are bicycle trails, bridges, and bicycle lanes. These locations have heavy traffic of mainly pedestrians and bicyclists. A vision-based system overcomes many of the shortcomings of existing technologies such as loop counters, buried pressure pads, infra-red counters, etc. These methods do not have distinctive profiles for bicycles and pedestrians. Also most of these technologies require expert installation and maintenance. Cameras are inexpensive and abundant and are relatively easy to use, but they tend to be useful as a counting system only when accompanied by powerful algorithms that analyze the images. We employ state-of-the-art algorithms for performing object classification to solve the problem of distinguishing bicyclists from pedestrians. We detail the challenges that are involved in this particular problem, and we propose solutions to address these challenges. We explore common approaches of global image analysis aided by motion information and compare the results with local image analysis in which we attempt to distinguish the individual parts of the composite object. We compare the classification accuracies of both approaches on real data and present detailed discussion on practical deployment factors.

17. Document Analysis/Descriptors Computer vision, Object classification, Image analysis, Bikeways, Cyclists, Pedestrians, Traffic counting	18. Availability Statement No restrictions. Document available from: National Technical Information Services, Alexandria, Virginia 22312

19. Security Class (this report) Unclassified	20. Security Class (this page) Unclassified	21. No. of Pages 35	22. Price

Deployment of Practical Methods for Counting Bicycle and Pedestrian Use of a Transportation Facility

Final Report

Prepared by:
Guruprasad Somasundaram
Vassilios Morellas
Nikolaos Papanikolopoulos

Department of Computer Science and Engineering
University of Minnesota

January 2012

Published by:

Intelligent Transportation Systems Institute
Center for Transportation Studies
200 Transportation and Safety Building
511 Washington Avenue SE
Minneapolis, Minnesota 55455

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. This report does not necessarily reflect the official views or policies of the University of Minnesota.

The authors, the University of Minnesota, and the U.S. Government do not endorse products or manufacturers. Any trade or manufacturers' names that may appear herein do so solely because they are considered essential to this report.

Acknowledgments

The authors wish to acknowledge those who made this research possible. The study was funded by the Intelligent Transportation Systems (ITS) Institute, a program of the University of Minnesota's Center for Transportation Studies (CTS). Financial support was provided by the United States Department of Transportation's Research and Innovative Technologies Administration (RITA).

Table of Contents

List of Figures

List of Tables

Executive Summary

The primary goal of this project is to develop a practical vision-based bicycle counting system that is capable of automatically processing video stream data of traffic scenes involving bicycle and pedestrian activity and estimating their traffic counts. Upon a successful deployment this can be potentially the cheapest and easiest method to do the counting. This is due to the fact that cameras are inexpensive and in many places we will be able to leverage the already existent security cameras, which can reduce incurred expenses further. While cutting costs is important, processing videos for getting bicycle counts is non-trivial and involves state-of-the-art machine learning and computer-vision algorithms. Based on practical experience from developing similar systems in the past for the Minnesota Department of Transportation (MnDOT), we have addressed some key issues from a scientific perspective to improve our algorithm software. Upon further testing, we evaluate the efficacy of our new algorithm and obtain improved performance. The new algorithm is more precise as it exploits the localized appearance characteristics of a bicyclist and a pedestrian. The particular challenge is that a bicyclist is a combination of a bicycle and a person. Also the algorithm is faster and can process more hours of video in a day, thereby saving time and hence costs. The new algorithm is not specific to bicyclists and pedestrians; it can also be extended to other vision-based recognition tasks such as vehicle class type (sedan, SUV, etc.) estimation, traffic estimation, etc. Albeit the parameters for the algorithm have been fine tuned for optimal performance with respect to bicyclist classification and counting.

Chapter 1 Introduction

The classification problem of distinguishing bicycles from pedestrians for traffic counting applications is the objective of this research project. The scenes that are typically involved are bicycle trails, bridges, and bicycle lanes. These locations have heavy traffic of mainly pedestrians and bicyclists. A vision-based system overcomes many of the shortcomings of existing technologies such as loop counters, buried pressure pads, infra-red counters, etc. These methods do not have distinctive profiles for bicycles and pedestrians. Also most of these technologies require expert installation and maintenance. Cameras are inexpensive and abundant and are relatively easier to use. Whereas cameras are cheap and easy to use, they tend to be useful as a counting system only with accompanying powerful algorithms to analyze the images.

We employ state-of-the-art algorithms for performing object classification to solve the problem of distinguishing bicyclists vs. pedestrians. The challenge in this problem is that the bicyclist is a composite combination of a bicycle and a person. We employ local image analysis in which we attempt to distinguish the individual parts of the composite object. We analyze the performance of the algorithm on real data and present detailed discussion on performance and accuracy.

Object classification has diverse applications. The techniques used to perform object classification are equally diverse and often depend on the target application for which the classification is performed. The accuracy of classification is thus subjective and depends on many parameters and computational resources. We will be considering the class of supervised learning methods in this paper and the method discussed here can be categorized under this domain. Usually learning methods have a training phase in which visual models of the different object classes are created. There are two key factors in training which are the object representation and the classifier and its training algorithm. Objects can be represented using several different features such as color, texture, shape, etc. This is one area of intensive research in computer vision. As part of this work we will examine the effect of learning the dictionaries using a few different feature representations. Classifiers are usually maximum-margin-based, like the support vector machine (SVM), which usually have high accuracies. SVMs are extremely popular not only in the computer vision domain but also in many other fields. Also the amount of labeled training samples available is important to the accuracy, which is a direct result from machine learning theory. For practical applications we would like to minimize the amount of training required while achieving a high degree of classification accuracy.

In recent times sparse signal models have become popular for image compression and reconstruction. Dictionaries representative of each object class can be learned. The underlying theory makes them suitable to incorporate discriminative components, wherein the learned dictionaries could be used to tell apart two different classes of signals. Mairal *et al.* [8] have effectively used this concept for texture classification with inspiring results. We employ similar methods to learn discriminative bicycle and pedestrian dictionaries and use them to classify bicyclists and pedestrians in traffic video images.

Chapter 2 Related Work

Choosing the right features for representation is often domain dependent and is very crucial for obtaining successful classification. Features can be classified broadly into interest point features, region features and shape features. Shape contexts [10] have been demonstrated to be a viable choice for matching and recognizing digits, letters, and 3D objects where shapes as well as pose of the objects are the discriminative factors. SIFT (scale invariant feature transform) [11] is one of the most effective interest point detectors, which uses scale space extrema as interest points and provides a localized high dimensional descriptor. While SIFT promises some great matching results, it lacks the simplicity of certain other features. It is computationally expensive and in low resolution images multiple scale spaces of gradients do not carry much information. A detailed report of interest point detectors which focus on scale and affine invariance of corner detectors is given in [12]. Some global image features are more suitable under certain circumstances such as human detection. PHOG (pyramidal histogram of oriented gradients) provides statistical information of edge directions which is understood to be a good indication of object and shape representation in images [13]. They also show good results for human detection. It is to be noted, that these features depend on the quality of the images and presence of sufficient gradient information. Techniques that work on standardized dataset images will not necessarily guarantee good results in practice, as we need to deal with problems like poor resolution, motion blur, occlusion, etc. The problems are escalated when the images are blob images obtained from tracking (See Figure 1).

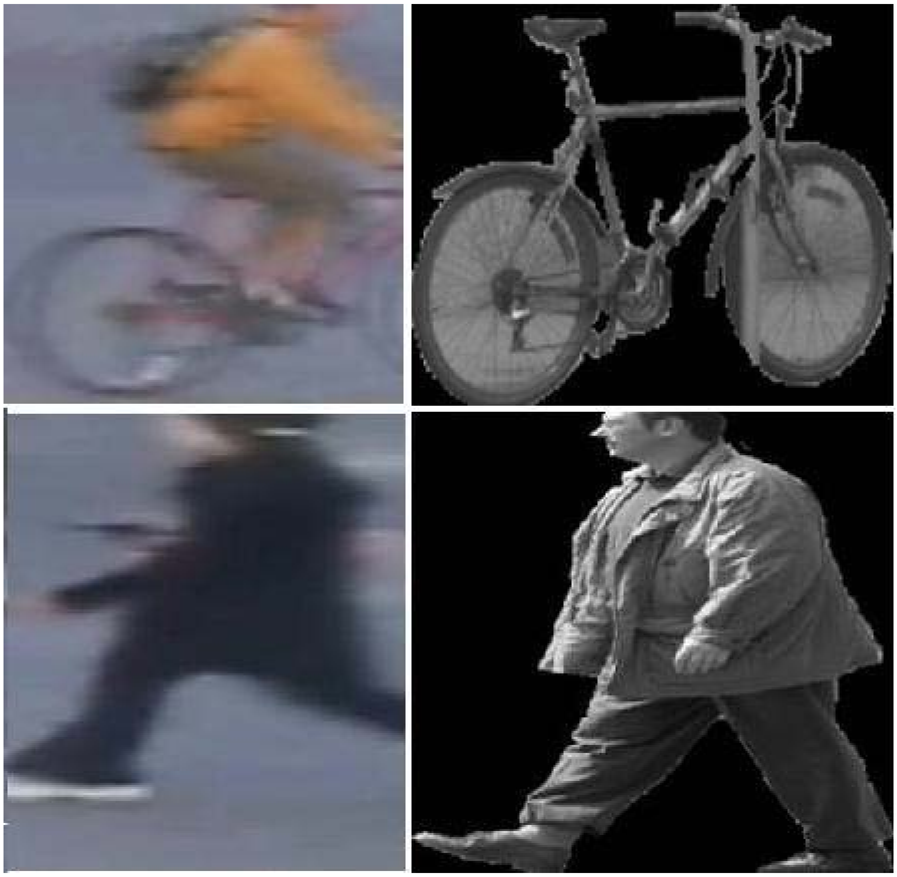


Figure 1 Comparison of bicycles and pedestrians between real blob images and images obtained from standardized datasets upon which prototypes are built.

Unsupervised learning methods for image classification are more popular in large scale applications like web search where human training is daunting except when there are tags available for images. Many of these methods are extensions of ideas from document classification based on LSI (latent semantic indexing).

SVM (support vector machine) based approaches have been prominent under supervised learning methods category [14]. Not limited to object classification, SVMs have been an attractive choice for many other classification and regression applications. SVMs have been used with many different object features with success and a detailed report is available in [3]. A similar approach which constructs different feature vocabulary sets as well as investigates the use of probabilistic latent semantic analysis, is presented in [15]. HOG features have been extended to construct part models which accommodate deformations in objects and hence improve accuracy [16]. Another approach to parts modeling is to learn the spatial contexts of the different parts by learning the weights of all possible configurations of parts across different object classes [17]. In papers [19], and [20], the use of LDA (Latent Dirichlet Allocation) for object and scene classification and annotation is discussed. Most of these methods use benchmark datasets such as the Caltech, Graz 02, etc.

Sparse signal models are an interesting area of research for a variety of different applications. They were applied for texture classification in images for the purpose of compression and reconstruction [8]. Sparse signal models are constructed using the K-SVD procedure and MOD (method of optimal directions) algorithms. We have derived the theory for our approach from the ideas presented by Mairal et al. [8] and Starck et al. [20] discuss the design of dictionaries for sparse representations of image content as texture and smooth regions, where they propose basis pursuit denoising with augmented dictionaries for separation of the two components in image content. We use a similar idea to recognize different image classes contained in composite images, using the image patches.

The associated literature presented here are important considerations while designing a classification method for the problem of bicyclists vs. pedestrians. Many researchers have approached the problem of object classification for many different object classes. The class of bicycles is typically considered in many standard datasets. However these approaches need extensions to work for the problem for the bicyclists vs. pedestrians.

Chapter 3 Theory and Implementation

We address the issue of classifying bicyclists vs. pedestrians as a problem of composite object class vs. simple object class problem. Here a composite object class is a combination of two or more simpler object classes. Hence a bicyclist which is an intricate combination of bicycle and person is rather complicated to build an appearance model for. Often motion information is difficult to obtain without calibration and we have to solely depend on appearance information to classify. In general we can observe the trend shown in Table 1.

Table 1 Dependency on appearance and motion characteristics as a function of the image quality.

Image Quality	Dependency on Motion characteristics	Dependency on Appearance characteristics
Low	High	Low
High	Low	High

In this project we developed a new localized appearance model which takes full advantage of the available information in the image. Hence better the quality of the image the better the method performs. This is a desirable result since high resolution cameras are cheap these days and that classification approaches are not bottlenecked by image quality any more. When we have high quality images, every localized area in an image is informative. Small regions in the image termed “patches” are used as the fundamental elements in creating the model for each object class. However the algorithms need to be efficient to handle high precision data, since often it results in a large amount of data. The method we present here handles high dimensional data efficiently and details of the complexity of the algorithm is presented in this chapter in the implementation section.

The underlying theory of this approach is based on the work by Mairal *et al.* [8]. (For the convenience of the reader we will stick to their notation.) Given a set of patches $\mathcal{X} = \{\boldsymbol{x}_l\}_{l=1}^{M}, \boldsymbol{x}_i \in R^n$, a reconstructive dictionary $\boldsymbol{D} \in \boldsymbol{R}^{n \times k}$ is learned adaptively from the data such that the respective decomposition α_l is sparse (*i.e.*, no more than L non-zero elements) by solving the optimization problem:

$$\min_{\alpha,\boldsymbol{D}} \sum_{l=1}^{M} \| \boldsymbol{x}_l - \boldsymbol{D}\alpha_l \|_2^2 \quad \text{s.t.} \quad \| \alpha_l \|_0 \leq L. \qquad (1)$$

The best possible reconstruction error for a given signal $\boldsymbol{x}$ and dictionary $\boldsymbol{D}$ is denoted as $\mathcal{R}^*(\boldsymbol{x},\boldsymbol{D})$, where

$$\mathcal{R}^*(\boldsymbol{x},\boldsymbol{D}) = \| \boldsymbol{x} - \boldsymbol{D}\alpha^*(\boldsymbol{x},\boldsymbol{D}) \|_2^2. \qquad (2)$$

Here $\alpha^*(\boldsymbol{x},\boldsymbol{D})$ is the optimal L-sparse decomposition for the pair$(\boldsymbol{x},\boldsymbol{D})$.

The K-SVD procedure was used to solve the dictionary learning problem in Eq. (1) in an iterative fashion efficiently with convergence guaranteed in few steps.

Suppose now that we have N different classes S_i of signals, $S_i = \{\boldsymbol{x}_l\}_{l \in S_i}$, and we would like to perform classification of these signals, with the help of dictionary learning. Mairal *et al.* [8] incorporate a discriminative component into the objective function in Eq. (1), and learn separate dictionaries, one per class, so that a signal belonging to one class is reconstructed poorly by a dictionary corresponding to another class. Thus the residual reconstruction error of a signal $\boldsymbol{x}$ by the dictionary belonging to class i, $\mathcal{R}^*(\boldsymbol{x},\boldsymbol{D}_i)$ is used as a discriminant for classification. The residual reconstruction error as a function of the sparsity level provides us a feature vector that can be used with any discriminative model. Here we use a simple fisher linear discriminant analysis as shown in Figure 4. The fisher linear discriminant analysis learns a set of weights for the reconstruction error as a feature. We employ this method for its simplicity and the relative ease with which it can be trained.

Training Requirements and Complexity Analysis

Typically dictionaries can be learned with very few iterations. The KSVD method is known to converge in <= 20 iterations. Convergence can be indicated by the relative change in the individual atoms of the dictionaries and can also be measured using the incoherence metric. Specifically the time taken for the KSVD algorithm is

T - K-SVD = R(2NL + K^2L + 7KL + K^3+ 4KN)+ 5NL^2

where R is the number of training signals, K is the number of iterations, N is the number of dimensions of the signal and L is the number of atoms in the dictionary. The dictionaries are however computed offline in batch mode. The classification step involves determining the reconstruction error which is carried out using the omp method. The omp method can be performed efficiently using the batch cholesky method. This method requires

T-OMP = 2KNL + 2K^2N + 2K(L + N) + K^3

per signal. In real time units, the tracking alone can be performed at 15 frames per second (0.06 seconds per image/ frame) and the classification can take up to 0.01 seconds per frame. We use efficient implementations of the OMP procedure by the SPAMS library [8]. The total is time is approximately 0.07 seconds per frame or 14.2 frames per second on an intel core i5 machine with 4GB of RAM running Windows 7.

Mixed Dictionaries to Detect Mixed Classes

The dictionaries are adapted in the above manner to discriminate between objects from N classes. However, when images from two different classes A and B co-occur in an image, for e.g., a person holding a bike, as shown in Figure 4, neither of the learned dictionaries D_A and D_B will be able to individually provide an efficient reconstruction. The primary goal of this experiment is to detect and classify this image as belonging to a class AB, without explicitly training for such a case. In this direction we construct a new dictionary D_{AB} by concatenating the dictionaries of the individual classes as: equation D $_{AB} = [\ D_A \mid D_B\ ]$. For our requirements class A is a bicycle and class B is a person, and class AB is a bicyclist. We extended this study to other mixed classes as shown in figure 4.

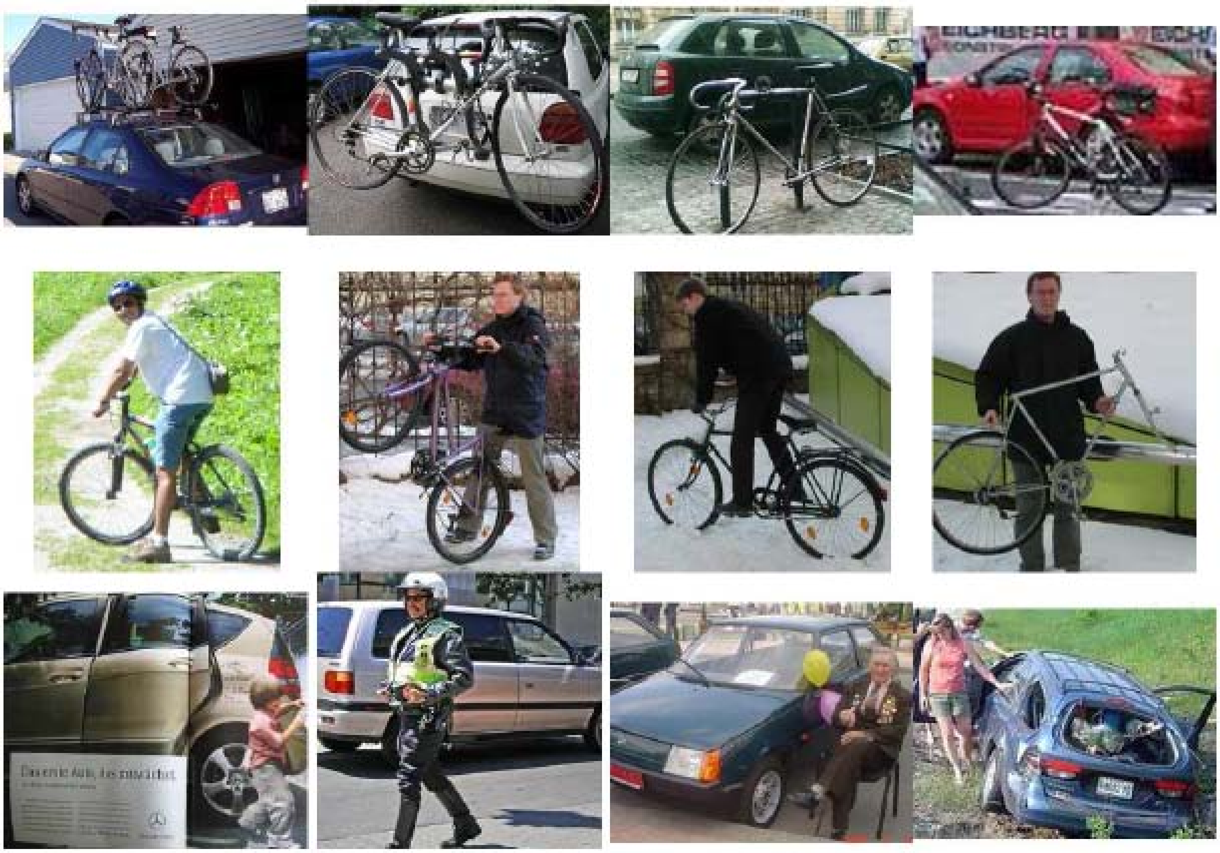


Figure 2 Example of mixed classes.

Let us denote by Ea, Eb and Eab respectively the normalized error curves generated by the dictionaries DA, DB and DAB. We compute the logarithm of the ratio of the differences of the curves Ea and Eb with respect to Eab. These differences are computed using either the l1 or l2 norm. If the image contains objects from two classes, as in Figure 3, then the curves Ea and Eb will both be roughly equally well separated from Eab, resulting in Ep ; p = 1; 2 being close to zero. However, if the image contains only one class, say A, then the Ea and Eab curves will be quite similar, whereas Eb will be very distinct from these two.

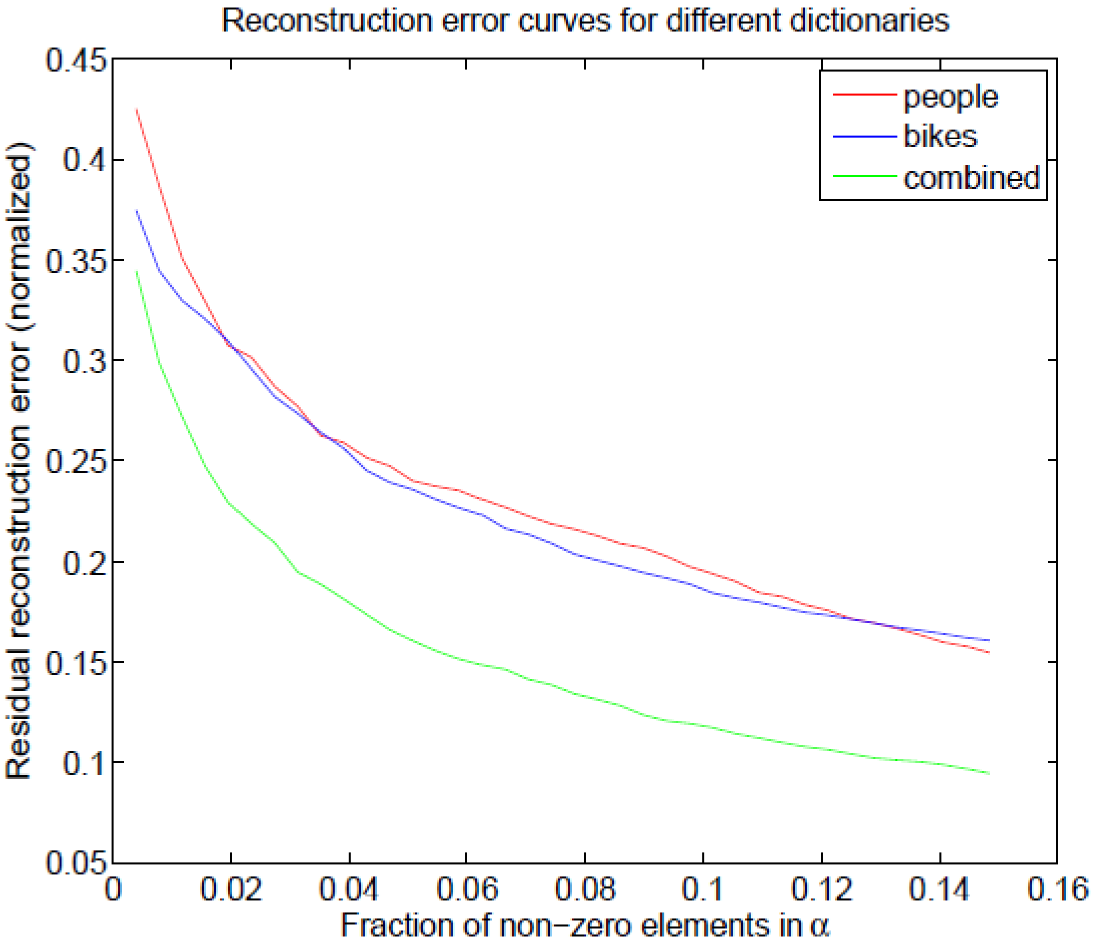


Figure 3 Reconstruction error curves for a bicyclist. We see that the reconstruction error is the lowest with the combined dictionary.

Implementation

An overview of the practical implementation of a classification and counting system is shown in Figure 2. The implementation was carried out in C++ using open source computer vision libraries such as OpenCV and VXL. This can be employed with live camera streams or with recorded videos. The first step of processing is background removal and separation of foreground objects. This is done using the mixture of Gaussians method of background modeling [23]. The separated blobs are associated across different frames using a bipartite graph based approach which is based on blob area overlap percentages. Although tracking blobs is not of primary interest to this problem, it provides useful information when combined with the calibration information. The scene was calibrated beforehand using the method described in [24].

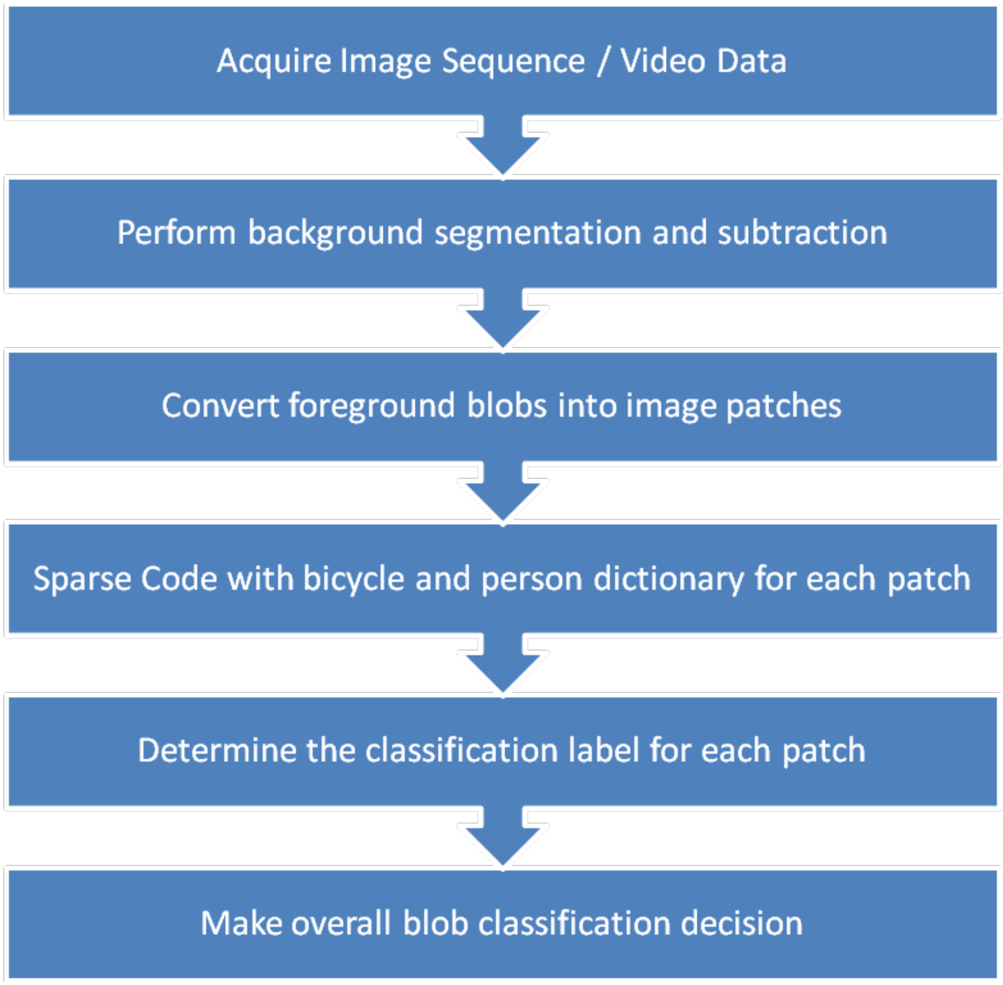

Acquire Image Sequence / Video Data
Perform background segmentation and subtraction
Convert foreground blobs into image patches
Sparse Code with bicycle and person dictionary for each patch
Determine the classification label for each patch
Make overall blob classification decision

Figure 4 A flow chart indicating the flow of operations in the software.

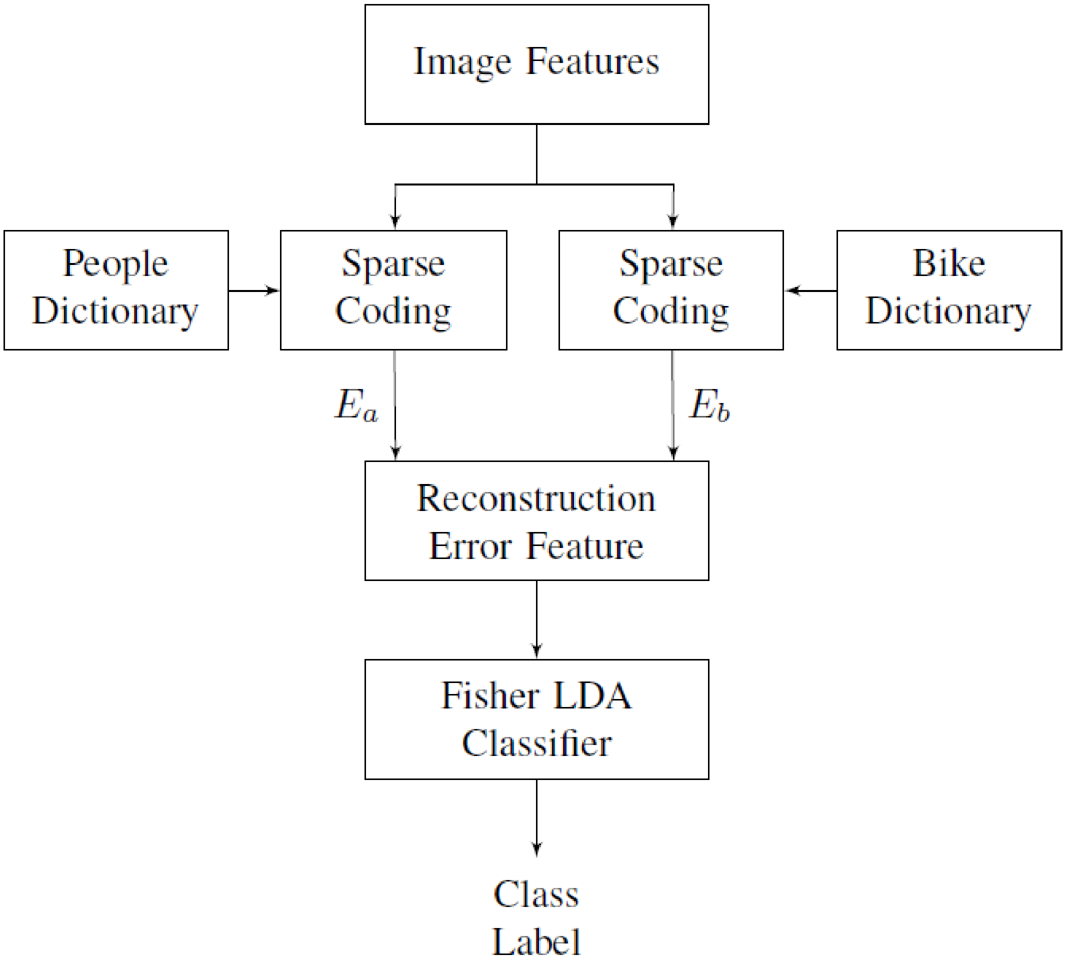


Figure 5 Classification of image patches using dictionaries.

Blob images tend to be small. For a typical video resolution of 320 X 240, blobs are approximately 25 X 45 pixels. However for some of our analysis we used a HD video camera thereby we could get higher resolution blobs to extract more meaningful patches for this analysis.

Figure 6 A Sony HD camera.

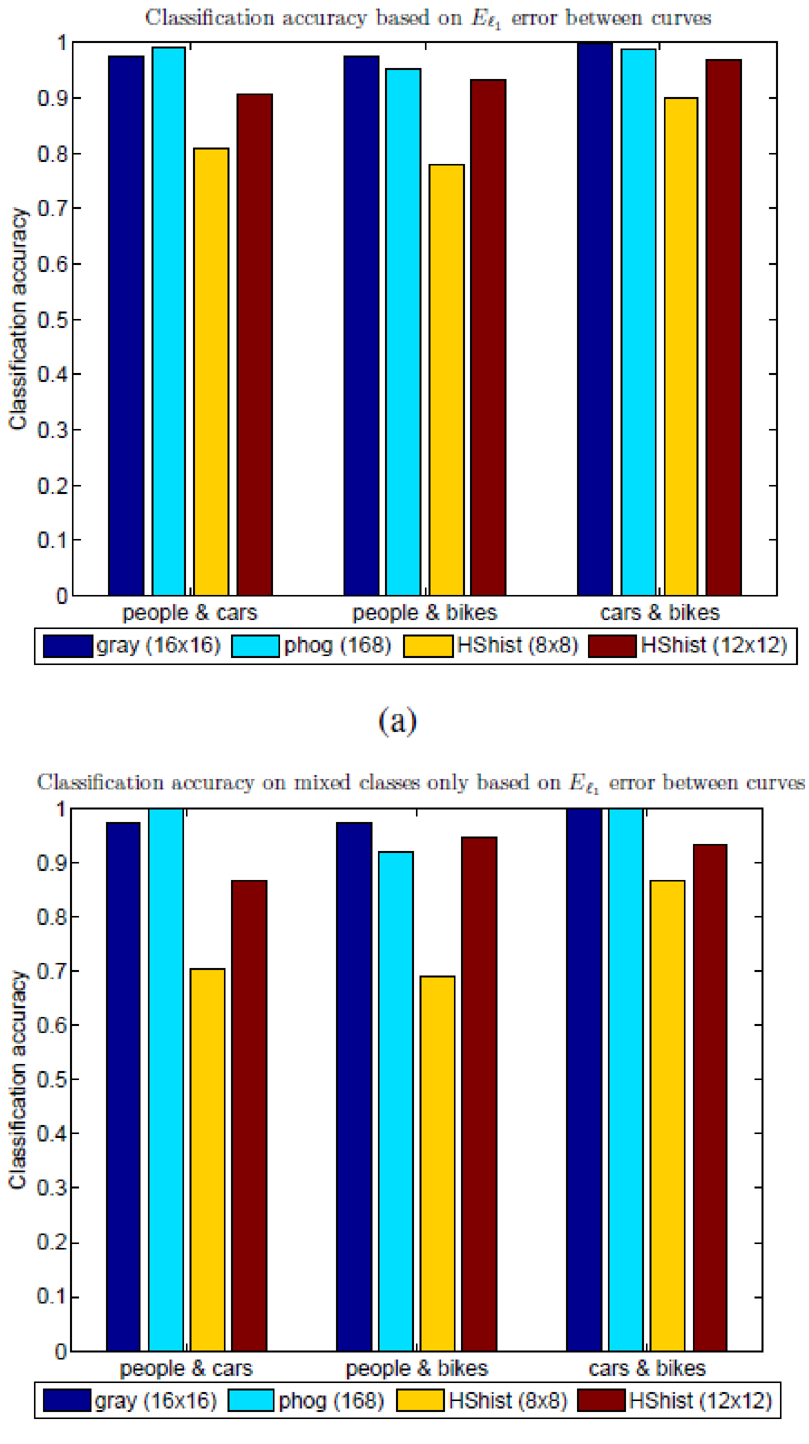
Another factor that influences the performance of the method is camera placement. Ideally, we would like to place the cameras in locations which can optimally observe the bicycle path with as much resolution as possible and minimizing occlusions at the same time. So an optimal choice would be directly perpendicular to the path of the pedestrians and cyclists. But this choice brings complete loss in counts due to occlusions and phenomena like parallel riding. But it was still considered the best overall single camera choice since it maximizes the resolution of the bicyclists or pedestrians for extracting appearance information given the physical constraints in placement.

We address the issue of self occlusion of the bicycle by the bicyclist using local image analysis based on dictionary learning. We have trained dictionaries to detect composite objects in thumbnail images and then discriminatively learned multiscale dictionaries to segment individual image patches.

Chapter 4 Experimental results on datasets and real data

The discriminative dictionary learning framework was used to learn dictionaries based on the people (p), cars (c) and bicycles (b) classes from the GRAZ02 dataset [21], [22]. Three different experiments were run based on pair wise combinations of these classes (N = 2). Test images consisting of mixed classes were obtained from the GRAZ01 dataset [27], [28], the INRIA Person dataset [29], and Flickr [30].

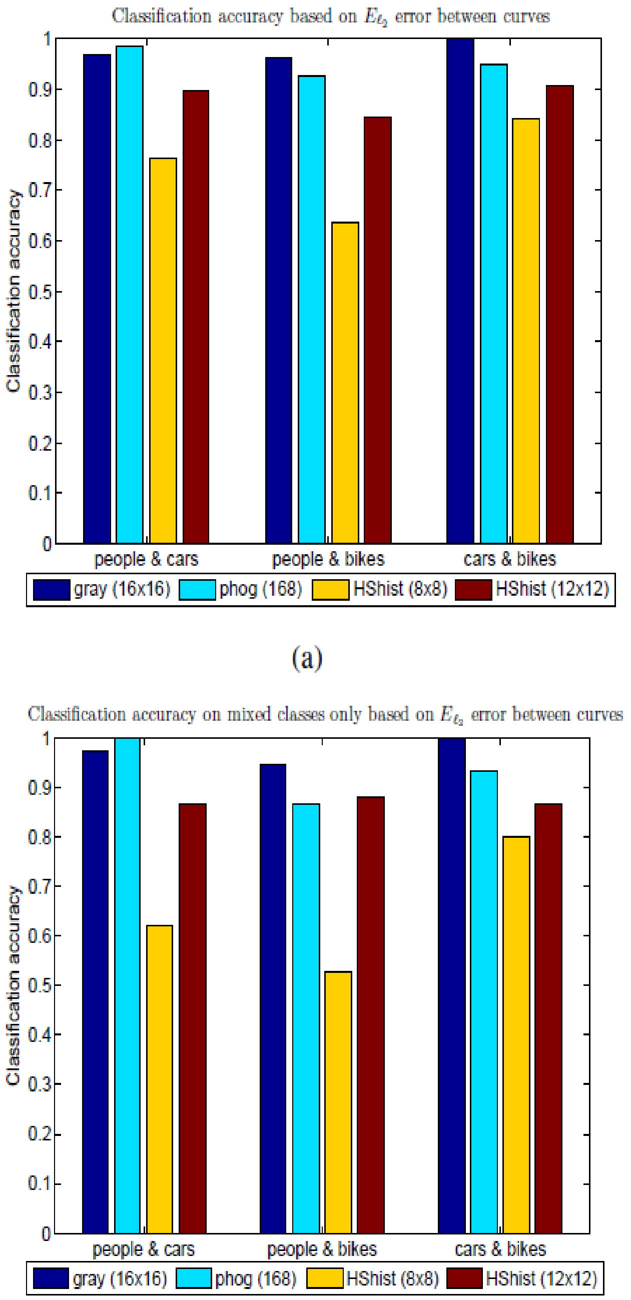
Two classes of images are selected for training from the GRAZ02 dataset, for e.g., people and cars. Discriminative dictionaries are learned based on the framework in [8], to yield Dp (for people) and Dc (for cars), and a combined dictionary Dpc is created by concatenating the individual dictionaries. For each image in the test set, the residual reconstruction error curves were plotted as explained in the previous section, and the quantities E1 and E2 are obtained. The images were classified using univariate Gaussian classifiers based on the E1 and E2 values and the accuracy was evaluated using leave-one-out cross-validation. Figures 7(a) and 7(b) show the classification accuracy based on the E1 difference between the individual curves and the combined curve. The results show that the dictionaries learned using gray scale and P-HOG features are better at recognizing mixed classes compared to the dictionaries learned from hue-saturation values. We believe that the hue and saturation values at such low scales (16 X 16) doesn't generate enough discriminative information. Traditionally the overall distribution of the hue and saturation values have been used for an entire image as a classification feature. Figures 8(a) and 8(b) show the corresponding results for the E2 difference. Using the E1 difference produced better results compared to the E2 difference. The corresponding confusion matrices for different class combinations using the gray scale intensity features are shown in Table 2.



(a)

(b)

Figure 7 Mixed class classification accuracies using different features and error norms. We notice greater than 90% accuracy with bicyclists vs pedestrians.

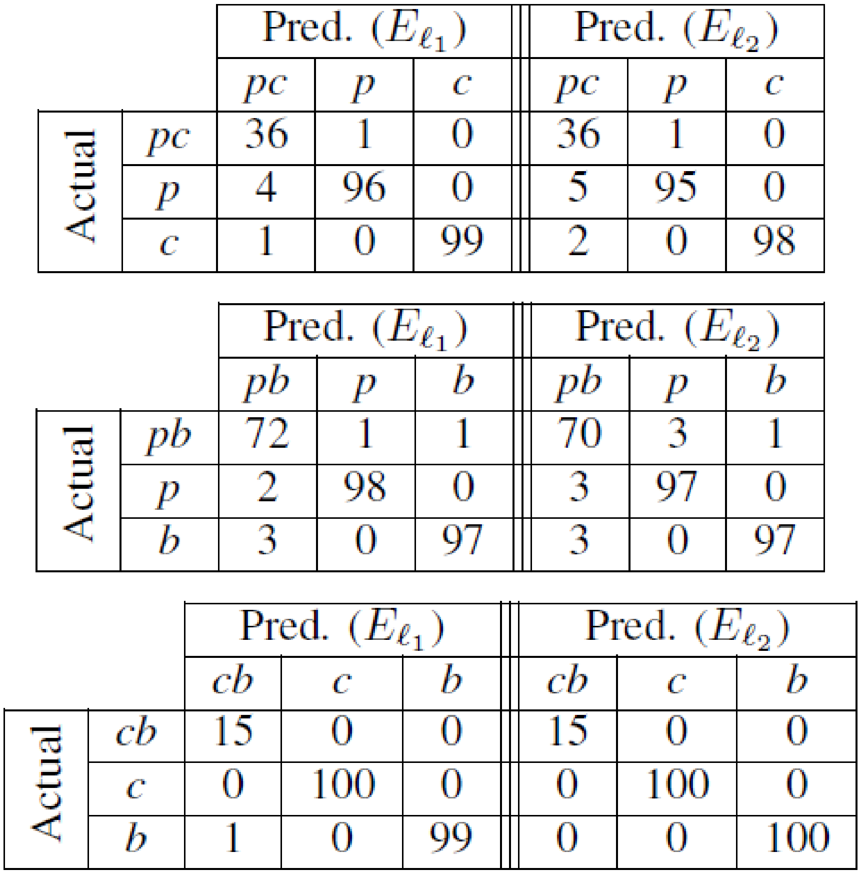


(a)

(b)

Figure 8 Mixed class classification accuracies using different features and error norms. We notice that l2 error is slightly inferior to l1 error based classification.

Table 2 Confusion matrices between mixed classes p: person c: car b: bicycle pc: person and car pb: bicyclist (person and bicycle) cb: car and bicycle.



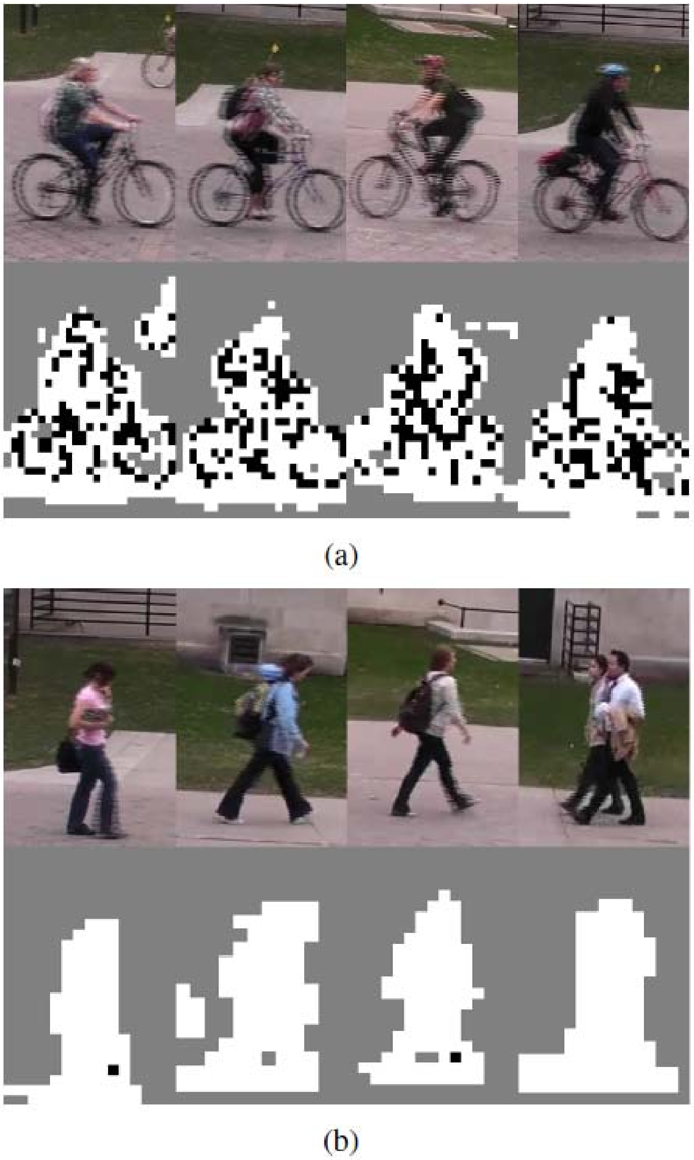
		Pred. (E_{ℓ_1})			Pred. (E_{ℓ_2})		
		pc	*p*	*c*	*pc*	*p*	*c*
Actual	*pc*	36	1	0	36	1	0
	p	4	96	0	5	95	0
	c	1	0	99	2	0	98

		Pred. (E_{ℓ_1})			Pred. (E_{ℓ_2})		
		pb	*p*	*b*	*pb*	*p*	*b*
Actual	*pb*	72	1	1	70	3	1
	p	2	98	0	3	97	0
	b	3	0	97	3	0	97

		Pred. (E_{ℓ_1})			Pred. (E_{ℓ_2})		
		cb	*c*	*b*	*cb*	*c*	*b*
Actual	*cb*	15	0	0	15	0	0
	c	0	100	0	0	100	0
	b	1	0	99	0	0	100

Results of Classification and Counting

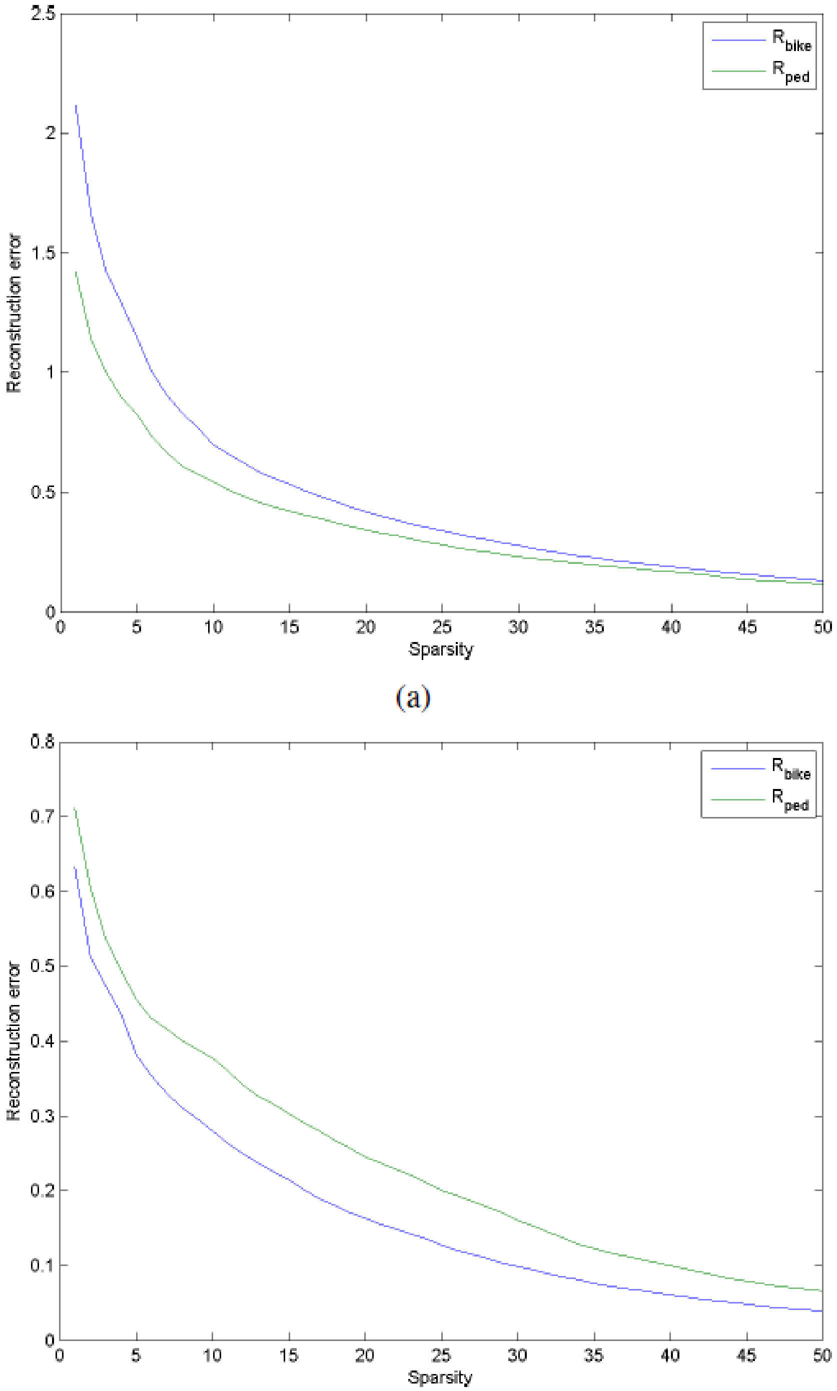
For this experiment, we acquired 2 hours of high quality video from 3 university walkway sites. We also used some high velocity bicycle traffic from the Gateway bicycle trail in uptown. The traffic analyzed is predominantly of bicyclists and pedestrians. The split in traffic was around 60% pedestrians and 40% bicyclists. The ground-truth for these blobs was obtained using manual labeling and then the classification of these blobs was done using the algorithm described in chapter 3. Two sample reconstruction error curves for patches originating from bicyclists and pedestrians are shown in Figure 10. Classifications label masks overlaid on sample blob images are shown right next to the true blob images in Figure 9. Note that the labels are for 16 X 16 patches. Here white corresponds to pedestrian patches, black to bicycle patches, and grey represents background, which was not considered for classification. The counting accuracy for bicyclists and pedestrians, as well as the overall accuracy, are shown in Table 3. From the overlaid class labels for each patch we observe that in some cases patches from the person riding the bicycle are classified as part of the bicycle. This factor has not biased the classification results. In fact we observe that we have very few pedestrian patches labeled as bicycle patches.



(a)

(b)

Figure 9 Sample blob images and corresponding classification overlays. Grey: background; white: person; black: bicycle.

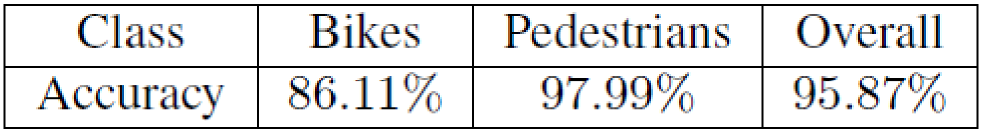


(a)

(b)

Figure 10 Reconstruction error curves with respect to bicycle and person dictionaries for a person patch (a) and bicycle patch (b).

Table 3 Overall classification accuracies



Class	Bikes	Pedestrians	Overall
Accuracy	86.11%	97.99%	95.87%

Chapter 5 Conclusions and Future Work

We discussed the problem of classifying bicyclists versus pedestrians. This problem is complicated since a bicyclist is an intricate combination of a bicycle and a person. The discriminating factor then is the presence of bicycle-like components, which are differentiated well by performing local image analysis using discriminative dictionaries. We also investigated the detection of mixed (composite) classes in traffic scenes and obtained significant success on different permutations of mixed classes without specifically learning models for mixed classes. Current work concerns extending this work to multiple classes of objects (3 or more). Work is in progress to integrate the crowd counting approach presented in [31] along with this appearance based method to improve counting in mixed groups.

References

[1] T. Lee and M. Lewicki, “Unsupervised image classification, segmentation, and enhancement using ICA mixture models,” *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 270-279, 2002.

[2] M. Omran, A. Salman, and A. Engelbrecht, “Dynamic clustering using particle swarm optimization with application in unsupervised image classification,” *5th World Enformatika Conference (ICCI)*, Prague, Czech Republic, pp. 199-204, 2005.

[3] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, New York, NY, June 2006.

[4] G. Somasundaram, V. Morellas, N. Papanikolopoulos, and L. Austin, “Counting pedestrians and bicycles in traffic scenes,” *Proceedings of the IEEE Intelligent Transportation Systems Conference*, St. Louis, MO, 2009.

[5] R. Sivalingam, G. Somasundaram, V. Morellas, N. Papanikolopoulos, O. Lotfallah, and Y. Park, “Dictionary learning based object detection and counting in traffic scenes,” *Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras*, Atlanta, GA, 2010.

[6] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing over complete dictionaries for sparse representation,” *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311-4322, 2006.

[7] K. Engan, S. Aase, and J. Husoy, “Frame based signal compression using method of optimal directions (MOD),” *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems*, vol. 4, pp. 1-4, July 1999.

[8] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, “Discriminative learned dictionaries for local image analysis,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, June 2008.

[9] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce, “Discriminative sparse image models for class-specific edge detection and image interpretation,” *Proceedings of the 10th European Conference on Computer Vision*, Berlin, Heidelberg, Germany, pp. 43-56, 2008.

[10] S. Belongie and J. Malik, “Matching with shape contexts,” *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 20-26, 2000.

[11] D. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the 7th IEEE International Conference on Computer Vision*, vol. 2, pp. 1150-1157, 1999.

[12] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63-86, 2004.

[13] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886-893, June 2005.

[14] C. Cortes and V. Vapnik, “Support vector networks,” *Machine Learning*, pp. 273-297, 1995.

[15] A. Bosch, A. Zisserman, and X. Munoz, “Scene classification using a hybrid generative/discriminative approach,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, 2008.

[16] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, June 2008.

[17] C. Desai, D. Ramanan, and C. Fowlkes, “Discriminative models for multiclass object layout,” *Proceedings of the International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2009.

[18] C. Wang, D. Blei, and L. Fei-Fei, “Simultaneous image classification and annotation,” *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.

[19] L.-J. Li, R. Socher, and L. Fei-Fei, “Towards total scene understanding: Classification, annotation and segmentation in an automatic framework,” *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR),* 2009.

[20] J.-L. Starck, M. Elad, and D. Donoho, “Image decomposition via the combination of sparse representations and a variational approach,” *IEEE Transactions on Image Processing*, vol. 14, pp. 1570-1582, 2005.

[21] M. Marszalek and C. Schmid, “Accurate object localization with shape masks,” *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, June 2007.

[22] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, “Generic object recognition with boosting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 416-431, 2006.

[23] C. Stauffer and W. Grimson, “Adaptive background mixture models for real-time tracking,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999.

[24] O. Masoud and N. Papanikolopoulos, “Using geometric primitives to calibrate traffic scenes,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, pp. 1878-1883, 2004.

[25] S. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397-3415, 1993.

[26] B. E. Trevor, T. Hastie, L. Johnstone, and R. Tibshirani, “Least angle regression,” *Annals of Statistics*, vol. 32, pp. 407-499, 2002.

[27] A. Opelt, M. Fussenegger, and P. Auer, “Generic object recognition with boosting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 416-431, 2006.

[28] M. Fussenegger, A. Opelt, A. Pinz, and P. Auer, “Object recognition using segmentation for feature detection,” *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 3, pp. 41-44, August 2004.

[29] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *Proceedings of the International Conference on Computer Vision & Pattern Recognition* (C. Schmid, S. Soatto, and C. Tomasi, eds.), vol. 2, pp. 6, June 2005.

[30] Flickr, 2010. http://www.flickr.com.

[31] D. Fehr, R. Sivalingam, V. Morellas, N. Papanikolopoulos, O. Lotfallah, and Y. Park, “Counting people in groups,” *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 152 -157, 2009.