Histogram of Oriented Gradients for Detection of Multiple Scene Properties

Maesen Churchill
Adela Fedor

I. Introduction

Content-based image retrieval is an interesting problem within the field of computer vision. There are many situations in which a user might want to query a large set of images to find only images with some set of properties. For example, in order to show photos from a recent vacation to a friend, one might want to be able to search through a large collection of personal photos to find scenes of people, on the beach.

It would be ideal if multiple properties of a scene could be detected from a single set of features. Ideally, a program would only have to make a one-time calculation of a feature vector upon image upload, and learn several properties of the image by applying this defined set of features to multiple classifiers. Furthermore, introducing a new category to this scheme could be done without recomputing all feature vectors for this set of images; one could simply impose a new, trained classifier to this set. Thus, it would be convenient if there were some canonical descriptor for different types of images.

For our project, we will test the feasibility of using Histogram of Oriented Gradient feature sets as such canonical descriptors. We will evaluate the performance of these descriptors at detecting multiple properties of images by attempting to classify images into one of four categories: scenes with humans, scenes with bikes, scenes with both humans and bikes, and scenes with neither humans nor bikes.

II. Previous Work

Histogram of Oriented Gradient (HOG) descriptors are proven to be effective at object detection, and in particular with human detection [1]. These features have been used to solve a variety of problems, including pedestrians recognition and tracking [7], hand gesture recognition for sign language translation and gesture-based technology interfaces [4], face recognition [3], and body part recognition for tracking [1]. HOG descriptors have also been used in tasks that are not human-centric, such as classification of vehicle orientation, a problem relevant for autonomous vehicles [5]. However, these classification tasks use a sliding window approach to obtain descriptions of areas localized to just the subject of detection. To detect multiple scene properties, descriptors would have to use the entire image as a detection window. Moreover, for each of these tasks, the parameters of the HOG features are tuned to optimize for the specific classification task. If using a single feature vector to learn several properties of an image is possible, we will have to use a fixed set of parameters.

III. HOG Descriptors for Image Classification

Part 1: HOG Descriptors for Human Detection

Summary

We first replicated Dalal and Trigg's method for extracting HOG features from images [2]. We used an SVM classifier to detect humans for an easy dataset.

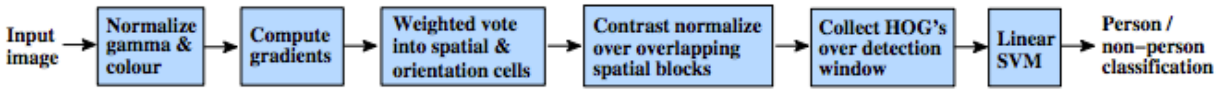
Input image
Normalize gamma & colour
Compute gradients
Weighted vote into spatial & orientation cells
Contrast normalize over overlapping spatial blocks
Collect HOG's over detection window
Linear SVM
Person / non−person classification

Figure 1: Pipeline for computation of HOG descriptors (Dalal and B. Triggs, 2005).

Dataset

Figure 2: Examples from the MIT pedestrian dataset.

We collected images from the MIT pedestrian database used by Dalal and Triggs for images containing humans. For negative images, we used the GRAZ01 portion of the INRIA Persons dataset. In total, we used 914 images containing humans, which we partitioned into 831 images for training and 93 images for testing. We also used 273 negative images, which we divided into 246 images for training and 27 images for testing. The images from the MIT database had dimensions 128 pixels by 64 pixels. The negative images had various dimensions, all around 200 pixels by 300 pixels. We cropped the negative images to match the size of the positives, mirroring Dalal and Triggs' strategy of creating a human detection window in the images.

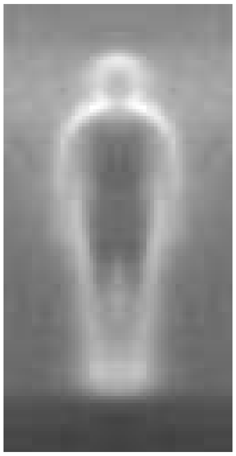

Figure 3: Average gradient computed by Dalal and Triggs [2].

Preprocessing and Gradient Computation

Dalal and Triggs computed the gradients for each color channel at each pixel. For the overall gradient, they used the color channel with the maximum gradient magnitude [2]. For simplicity, we make images gray scale and compute gradients using the library function provided by Matlab's Image Processing toolbox. Gradients are computed using Centered 1-D derivatives.

Weighted voting in cells
We partitioned each image into 6 by 6 pixel cells. For each cell, we computed a histogram of gradient orientation. We used a signed orientation (between 0 and 360 degrees) and had 9 spatial bins for each histogram. Votes were weighted by gradient magnitude.

Normalization over overlapping spatial blocks
In order to provide some invariance to illumination, histograms were concatenated together and then normalized across larger spatial blocks. We used blocks of 3 cells by 3 cells. Histograms were normalized according to L2 norm:

$$v \to v/\sqrt{\left|\left|v\right|\right|^2 + \epsilon^2}$$

For the overall descriptor, we concatenated the normalized histograms for dense, overlapping blocks across the entire image.

Support Vector Machine
For classification, we trained a soft (C = 0.01) SVM with a Gaussian kernel on our training examples.

Part 1: Results

We were able to classify our test set with 92% accuracy. Full results are shown below:

	Predicted Human	Predicted Non-human
Human	87	4
Non Human	5	22

Part 2: HOG Descriptors for detection of multiple image properties

Summary
We then assessed the accuracy of HOG descriptors at predicting: 1) whether the image has a human or not, and 2) whether the image has a bike or not, by applying two SVM classification schemes. Descriptors were computed across the entirety of images, not just a spatial window.

Dataset
We used a portion of the INRIA person dataset to create image sets for four object categories: those containing humans, those containing bikes, those containing humans and bikes, and those containing neither. In total, we had 323 images of humans (partitioned into 291 for training and 32 for testing), 373 images of bikes (partitioned into 336 for training and 37 for testing), 273 images of humans and bikes (partitioned into 246 for training and 27 for testing), and 210 images of neither (partitioned into 186 for training and 24 for testing).

Our dataset was diverse and challenging. It included images taken from a variety of viewpoints, with different illuminations and levels of occlusion. Though it is common practice to eliminate images that are particularly difficult [8], due to extreme occlusion, for example, we did not perform any data sanitation.

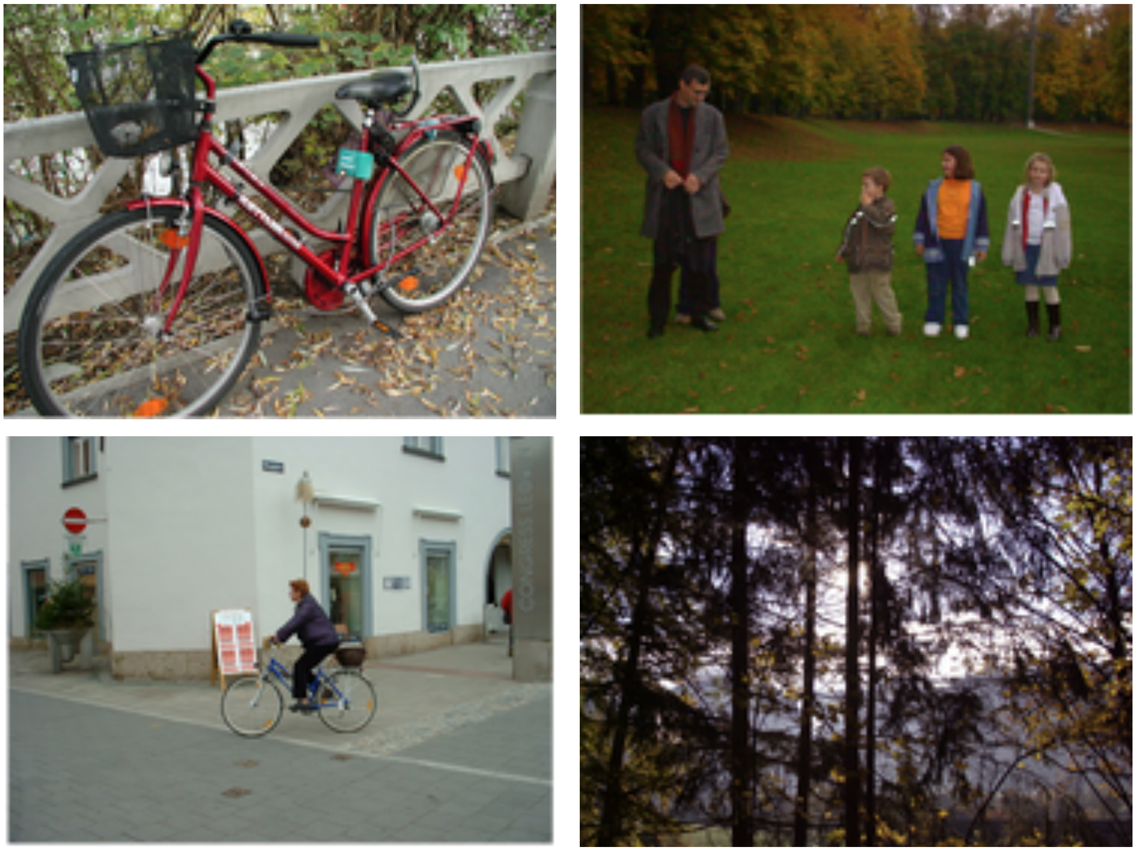

Figure 4: Samples from our dataset. Examples (from left to right) of bikes, humans, humans and bikes, and neither humans nor bikes.

Because our descriptors must serve as feature vectors for classifiers that predict the presence of multiple object categories, we did not clip our images or use a sliding detecting window at various scales. This marks a departure from Dalal and Triggs, who used a superset of our dataset, clipped to object-sized windows:

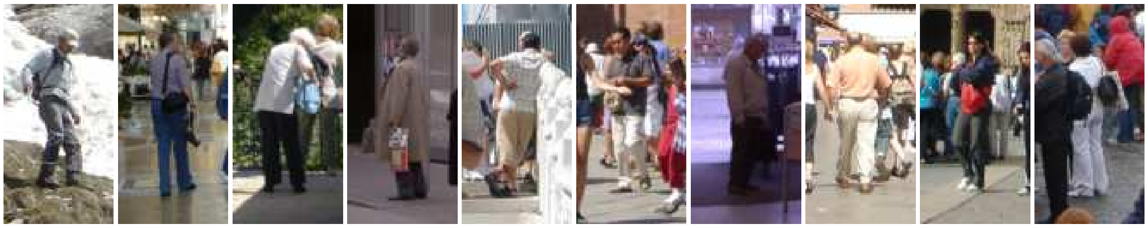

Figure 5: Images used by Dalal and Triggs, for comparison.

HOG Computation

For efficiency and accuracy, we used a Matlab library function for HOG computation during this part of our investigation. The HOG implementation that we used normalized non-overlapping blocks. To prevent over-fitting our classifiers, we used much larger blocks. We tried computing HOG descriptors for our 200 pixel by 300 pixel images with 3 and 6 blocks per image side.

Support Vector Machines

We used two soft (C=0.01) Support Vector Machines for classification (each with a Gaussian kernel). For the human classifier, images of humans made up the set of positive samples. For the bike classifier, images of bikes made up the set of positive samples. Thus, for both classifiers, images with bikes and humans are included as positive samples.

After classification by both SVMs, predictions for test images are assigned according to:

	Bike classifier predicted positive	Bike classifier predicted negative
Human classifier predicted positive	“Human and bike”	“Human”
Human classifier predicted negative	“Bike”	“Neither human nor bike”

IV. Results

Through this method, we achieved the highest success rates with 6 by 6 blocks. The success rate was 77% for human classification and 71% for bike classification. Overall, 58% of images were correctly classified into one of the four categories.

	Predicted: Human	Predicted: Bike	Predicted: Human & Bike	Predicted: Neither
Actual: Human	22	6	0	4
Actual: Bike	5	32	0	0
Actual: Human & Bike	10	5	12	0
Actual: Neither	8	15	1	0

These results demonstrate that our method does a fairly good job categorizing images containing either just humans or just bikes. Many of the incorrect classifications had obvious challenges that a HOG descriptor cannot account for, such as occlusion, scaling, illumination and orientation.

Scaling

Orientation

Illumination

Occlusion

However, our classification has very poor specificity. In particular, images with neither bikes nor humans were consistently labeled inaccurately.

A large set of humans make up the set of negative samples for the bike classifier. Likewise, bikes are abundant in the set of negative images for the human classifier. As a result, our model is likely over-fitting negative samples for bikes to be images of humans, and vice versa. To correct for this, we removed a large number of bike images from the negative samples of our human training set. We also removed a large number of humans from our bike training set. We then reran the classifier on descriptors with 6 blocks per side.

Using the new training sets, we achieved a success rate of about 76% for classifying humans correctly; however, the rate for classifying bikes correctly fell to about 58%.

	Predicted: Human	Predicted: Bike	Predicted: Human & Bike	Predicted: Neither
Actual: Human	3	1	16	12
Actual: Bike	8	5	2	22
Actual: Human & Bike	5	3	13	6
Actual: Neither	3	0	1	20

Clearly, specificity improves at the expense of sensitivity.

V. Discussion

Our initial classification scheme demonstrates that HOG features do have some descriptive power beyond the detection of a single object. Our dataset was very challenging, and the accuracy we attained was reasonable.

However, our investigation shows that classification schemes can easily become over-fit in a multiple object detection scenario. The performance of our method on the full training set shows that these descriptors would be well-suited to differentiate between two object categories, e.g. separate images of bikes from images of humans. However, the lack of specificity in this scheme makes it unsuitable for use in differentiating multiple separate categories.

Our second investigation with a more restricted training set gives a more realistic interpretation of the performance we could expect if we were to scale this method to several more object categories, since the set of negative samples for one category is not fit to match the description of the other category. Predictably, classification of bikes is less successful than classification of humans, since HOG descriptors are known to be best at classifying objects that take on a limited set of shapes when pictured from different angles.

Though the algorithm's performance was better than random guessing, the lack of a detection window diminished object detection success significantly. Furthermore, since HOG-based classification schemes depend on finding patterns in gradients, we would expect performance to worsen when more difficult object categories (i.e. object categories with more intra-class variability) were added. Furthermore, we would not expect good performance if HOG descriptors were used to classify images based on scene information (e.g. indoor/outdoor classification).

On this topic, future investigations could include multi-category classification using different types of descriptors, such as SIFT. One could also construct classification schemes for multiple object categories using Implicit Shape Models or ObjectBank heat maps. The performance of HOG descriptors could be used as a baseline for performance.

References

[1] Corvee, E., 2010. Body Parts Detection for People Tracking Using Trees of Histogram of Oriented Gradient Descriptors. In: IEEE pp. 469 – 475.

[2] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proc. CVPR 2005, vol. 1, pp. 886-893. <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1%467360>.

[3] Déniz, O., Bueno, G., Salido, J., De la Torre, F., 2011. Face recognition using Histograms of Oriented Gradients. In: Pattern Recognition Letters vol. 32 issue 12, http://www.sciencedirect.com/science/journal/01678655/32/12, pp. 1598–1603.

[4] Freeman, W. T., Roth, M., 1994. Orientation Histograms for Hand Gesture Recognition. In: IEEE.

[5] Huber, D., Morris, D.D., Hoffman, R., 2010. Visual classification of coarse vehicle orientation using Histogram of Oriented Gradients features. In: IEEE, pg. 921 – 928.

[6] Ludwig, O., 2011. HOG descriptor for Matlab. http:// http://www.mathworks.com/matlabcentral/fileexchange/28689-hog-descriptor-for-matlab/content/HOG.m. Universidade de Coimbra.

[7] Suard, F., Rakotomamonjy, A., Bensrhair, A., Broggi, A., 2006. Pedestrian Detection using Infrared images and Histograms of Oriented Gradients. In: IEEE, pp. 206 – 212.

[8] Szummer , M., R. W. Picard, 1997. Indoor-Outdoor Image Classification. In: IEEE, pp. 43-51.