3D Extended Histogram of Oriented Gradients (3DHOG) for Classification of Road Users in Urban Scenes

Norbert Buch
norbert.buch@kingston.ac.uk
James Orwell
j.orwell@kingston.ac.uk
Sergio A. Velastin
sergio.velastin@kingston.ac.uk

Digital Imaging Research Centre
Kingston University
Kingston upon Thames, UK

In recent years, there has been an increased scope for automatic analysis of urban traffic activity. Using general purpose surveillance cameras, the classification of vehicles is a demanding challenge (see Figure 1). In consultation with Transport for London, we use five generic categories to classify road users: Bus/Lorry; Van; Car/Taxi; Motorbike/Bicycle and Pedestrian.

Our contribution is three-fold. Firstly, 3D spatial models are introduced to define the location of interest points from which local features are extracted. The local features are constructed out of histograms of oriented gradients (HOG). The combination of 3D interest points and HOG is hence introduced as the novel 3DHOG feature. Performance is evaluated, comparing 3DHOG with FFT and histogram-based local features. The second contribution is a training and classification framework based on the 3DHOG feature which allows a variable number of interest points (previous approaches required a fixed number of interest points). Our third contribution is an extensive evaluation of the proposed method on real video benchmarking data (i-LIDS from UK Home Office) which is publicly available.

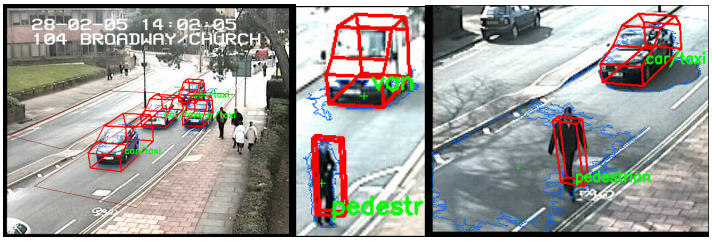

Figure 1 Example results from the i-LIDS data set with detected and classified road users. The blue outline is the initial foreground mask.

Classifying images or objects in images can be generally categorised either as top-down or bottom-up approach. Top down classification can be performed on motion silhouette measurement features [6], 3D models [7, 2], *etc.* in usual surveillance scenarios. In contrast, bottom up approaches are usually targeted at object categorisation and classification of still images with an extensive range of local features *e.g.* SIFT, SURF, GLOH, BFM, HOG [3]. The object recognition community moves towards surveillance applications *e.g.* [5]. Both approaches are combined by Dalal and Triggs [3], using local features with 2D fixed spatial constraints. This is used for pedestrian detection and for action recognition including a temporal extension in [4].

Our approach generalises the top down solution from HOG using a 2D search window [3] to 3D by 'wrapping' the camera image around the models. Using calibrated cameras the scale is determined directly, in contrast to the multiple scale search in [3]. By introducing a framework that deals with variable numbers of visible interest points, we can use a single model to detect objects from any angle. The algorithm uses texture to generate local features only and does not rely on potentially noisy motion information.

First we define the position of a set of interest points $\mathbf{P} = \{\mathbf{p}_j\}$ located on the faces of 3D models (Figure 2 left similar to [1]) of the objects to be classified. Then, for a candidate object (either during training or when classifying), we obtain image patches $\mathbf{I}_p$ for interest points $\mathbf{p}_k$ that are sufficiently visible. Finally, we calculate normalised feature vectors $\hat{\mathbf{f}}_k$ from those patches.

Three features computed from the patches $\mathbf{I}_p$ are compared: our novel 3DHOG, FFT and histogram. For 3DHOG, a Sobel kernel $[-1,0,1]$ is used to compute the gradient image from which gradients are calculated in the range $[0,2\pi]$. We use the visible part of 3D models to extract patches, which can be seen as '3D surface windows'.

Interest point appearances are modelled with single Gaussian distributions. A training set is used to estimate the mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ of every interest point $\mathbf{p}_j$. The Mahalanobis distance measure d_k is used to compare newly seen visible feature vectors $\hat{\mathbf{f}}_k$ with the model. After estimating the Gaussian models for

Figure 2 Extraction pipeline: One 3D model with interest points $\mathbf{P}$ followed by input image I, extracted image patches $\mathbf{I}_p$ and feature vectors $\hat{\mathbf{f}}_k$. The radius of cones indicates the weight q_j of point $\mathbf{p}_j$.

every interest point, the detection and localisation performance of every individual point can be improved by using a sigmoid function to calculate a match measure m_k

$$m_k = \frac{1}{1+e^{a(b-d_k)}}, \tag{1}$$

and by using weights q_k for the total match measure m of visible points

$$m = \sum_k m_k q_k \Big/ \sum_k q_k \,. \tag{2}$$

The classification framework uses background estimation with a Gaussian mixture model and shadow to generate a grid of 3D object hypotheses. The classifier sweeps through models and locations by scoring hypotheses based on equation (2) and finding the best matching model and position for objects in the scene.

Evaluation was performed on realistic (operational quality) videos from the i-LIDS data set licensed by the UK Home Office. All three algorithms are compared with state of the art classifiers. Out of the three features, the best performing algorithm is 3DHOG (Table 1) with a total recall of 81.1% at precision of 82% and classification accuracy of 92.1%. This compares well to recall of 88.2% at precision 89% for the motion silhouette baseline from [1] run on the same data set, but 3DHOG should be better dealing with noise and particularly occlusion in urban scenes.

ground truth / *detection*	bike	car/taxi	van	bus/lorry	FP
bike	1.00	.00	.00	.00	.44
car/taxi	.00	.83	.21	.03	.10
van	.00	.00	.67	.33	.08
bus/lorry	.00	.02	.02	.65	.00
FN	.00	.14	.10	.00	.00
count	27	361	48	40	
overlap	.70	.66	.73	.76	

ground truth / *detection*	bike	car/taxi	van	bus/lorry	FP
bike	.96	.03	.02	.00	.50
car/taxi	.00	.88	.12	.00	.05
van	.00	.02	.82	.00	.04
bus/lorry	.00	.01	.02	1.00	.03
FN	.04	.07	.02	.00	.00
count	28	361	57	29	
overlap	.73	.66	.70	.76	

Table 1 Left: Confusion matrix for 3DHOG detector and classifier. Right: Baseline algorithm (motion silhouette) from [1].

[1] N Buch, J Orwell, and S A Velastin. Detection and classification of vehicles for urban traffic scenes. In *VIE 2008*, pages 182–187.
[2] N Buch, J Orwell, and S A Velastin. Urban road user detection and classification using 3D wire frame models. *IET Compt. Vis. 2009*
[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In. *CVPR 2005.* pages 886–893, 2005.
[4] A Kläser, M Marszalek, and C Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008*, volume 2, pages 995 – 1004.
[5] B Leibe, K Schindler, N Cornelis and L Van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions on PAMI*, 30(10):1683–1698, 2008.
[6] B Morris, M Trivedi. Improved vehicle classification in long traffic video by cooperating tracker and classifier modules. In *AVSS 2006*.
[7] X Song and R. Nevatia. Detection and tracking of moving vehicles in crowded scenes. In *IEEE W. on WMVC '07.* pages 4–4, 2007.