Multidimensional Systems and Signal Processing (2019) 30:1697–1729
https://doi.org/10.1007/s11045-018-0621-1

A vehicle detection scheme based on two-dimensional HOG features in the DFT and DCT domains

Mohamed A. Naiel[1] · M. Omair Ahmad[1] · M. N. S. Swamy[1]

Received: 2 April 2018 / Revised: 19 October 2018 / Accepted: 21 October 2018 /
Published online: 9 November 2018

Abstract
Histogram of oriented gradients (HOG) are often used as features for object detection in images, since they are robust to changes in illumination and environmental conditions. However, these features are not invariant to changes in the resolution of input images. A 2D representation of these features, referred to as 2DHOG features, has been used since it preserves the relations among the neighboring pixels or cells. In this paper, a new vehicle detection scheme using transform-domain 2DHOG features is proposed. The method is based on extracting the 2DHOG features from the input image and applying to it 2D discrete Fourier or cosine transform. This is followed by a truncation process through which only the low frequency coefficients, referred to as the transform-domain 2DHOG (TD2DHOG) features, are retained. It is shown that the TD2DHOG features obtained from an image at the original resolution and a downsampled version from the same image are approximately the same within a multiplicative factor. This property is then utilized in our scheme for the detection of vehicles of various resolutions using a single classifier rather than multiple resolution-specific classifiers. Experimental results show that the use of the single classifier in the proposed detection scheme reduces drastically the training and storage cost over the use of a classifier pyramid, yet providing a detection accuracy similar to that obtained using TD2DHOG features with a classifier pyramid. Furthermore, the proposed method provides a detection accuracy that is similar or even better than that provided by the state-of-the-art techniques.

Keywords Vehicle detection · Transform-domain two-dimensional HOG · Classifier pyramid · Resolution-based feature approximation · Downsampling of transform-domain 2DHOG features

✉ M. Omair Ahmad
omair@ece.concordia.ca

Mohamed A. Naiel
m_naiel@ece.concordia.ca

M. N. S. Swamy
swamy@ece.concordia.ca

[1] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

1 Introduction

Detection of vehicles is used in many applications such as traffic surveillance, driver assistance systems, and autonomous vehicles. There has been a great deal of work carried out in this field during the past decade, and a survey of several techniques can be found in Buch et al. (2011), Dollár et al. (2012) and Sivaraman and Trivedi (2013a). For the purpose of object detection and recognition, there are several types of image features and representations, such as the histogram of oriented gradients (HOG) (Dalal and Triggs 2005), Haar-like features (Papageorgiou et al. 1998; Sivaraman and Trivedi 2013b), interest-points based features (Leibe et al. 2008), local binary patterns (Wang et al. 2009), and 3D voxel patterns (3DVP) (Xiang et al. 2015), that have been used. HOG features have been investigated widely and used in the state-of-the-art techniques for object detection and description (Dollár et al. 2012). Instead of the 1D vector representation of HOG (Dalal 2006; Dalal and Triggs 2005; Wu et al. 2014), several papers have adopted a 2D representation (Dollár et al. 2009; Felzenszwalb et al. 2010; Maji et al. 2008), since the latter preserves the relations among the neighboring pixels or cells. In order to distinguish the 2D representation from the 1D one, we will call it 2DHOG. Both the 1D and 2D representations of HOG capture the edge structure of the object and are robust against illumination changes and background clutters. However, neither of these representations is resolution invariant. Thus, detectors employing these representations require extracting HOG or 2DHOG at each scale from an image pyramid, thus requiring a costly multi-scale scanning in the testing mode (Dollár et al. 2009; Maji et al. 2008).

Recently, Dollár et al. (2014, 2010) proposed a feature approximation technique, where gradient histograms and color feature responses generated at one scale of an image pyramid can be used to approximate the feature responses at nearby scales. This method results in a speedup of extracting the features from the image pyramid over the methods of Dollár et al. (2009) and Maji et al. (2008), with only a small reduction in the detection accuracy. In this technique, the feature responses can be approximated with high accuracy within one octave of the scales of the image pyramid. Later, authors in Benenson et al. (2012) and Ohn-Bar and Trivedi (2015) enhanced the detection performance of Dollár et al. (2010) by constructing a classifier pyramid instead of an image pyramid. However, since the methods in Benenson et al. (2012) and Ohn-Bar and Trivedi (2015) are based on constructing a classifier pyramid with multiple classifiers trained at different sizes of the object (For example, in Benenson et al. 2012 the sizes of the object considered are 64 × 32, 128 × 64, 256 × 128, etc.), they require a high training and storage cost.

The part-based methods have received a great deal of attention from the research community, as these schemes can handle partial occlusion, and represent targets with several views (Felzenszwalb et al. 2010; Sivaraman and Trivedi 2013b; Takeuchi et al. 2010). For instance, (Felzenszwalb et al. 2010) have proposed a pictorial structure for HOG features, referred to as deformable part-based model (DPM). In this method, the locations of the parts are used as latent variables for a latent support vector machine (LSVM) classifier to find the optimal object position. Later, several other techniques have adopted DPM (Felzenszwalb et al. 2010) for vehicle detection (Li et al. 2014; Takeuchi et al. 2010; Wang et al. 2016). These methods provide high detection accuracy. However, these methods require convolutions of the features of a given level of the image pyramid with a number of part filters, which results in a high computational cost.

Some of the latest schemes in the area of object detection (Pepikj et al. 2013; Wang et al. 2015; Xiang et al. 2015) have attempted to solve the challenges of scale, aspect ratio or severe occlusion. For example, the method in Pepikj et al. (2013) has used a detection scheme based on the DPM detector (Felzenszwalb et al. 2010) and introduced a method for clustering the training data into a number of similar occlusion patterns. These patterns have been used with different occlusion strategies to train the LSVM classifier (Felzenszwalb et al. 2010). Later, Xiang et al. (2015) have combined 3DVP object representation, which encodes the appearance, 3D shape, view-point, the level of occlusion and truncation, with a boosting detector based on the detection scheme in Dollár et al. (2014) in order to learn from the occluded and non-occluded 3DVPs obtained from a training set. Recently, the method in Wang et al. (2015) has introduced region-based features with a coordinate normalization scheme, referred to as regionlet features, and a cascaded boosting classifier to tackle the challenges of detecting objects of different scales and aspect ratios. Even though these methods have been effective in tackling these challenges, they require high complexity either in the training mode, as in Wang et al. (2015) and Xiang et al. (2015), or in the testing mode, as in Pepikj et al. (2013).

The detection accuracy employing HOG or its variants in the spatial domain has started to saturate (Dollár et al. 2012). Recently, for the first time, the fast Fourier transform (FFT) has been used with 2DHOG in order to replace the costly convolution operation in the spatial domain by multiplication in the FFT domain (Dubout and Fleuret 2012). This scheme achieves a speedup over the spatial domain counterpart (Felzenszwalb et al. 2010). Later in Naiel et al. (2015), a method for approximating feature pyramids in the DFT domain instead of the spatial-domain has been introduced, resulting in a better feature approximation accuracy compared to the spatial-domain counterpart in Dollár et al. (2014). Despite the fact that both the methods in Dubout and Fleuret (2012) and Naiel et al. (2015) use a transform domain with 2DHOG, it is necessary to apply the corresponding inverse transform to classify the 2DHOG features in the spatial-domain. Thus, the methods in Dubout and Fleuret (2012) and Naiel et al. (2015) are based on training an object detector in the spatial-domain, which usually requires large storage and training cost. On the other hand, this paper develops a scheme that is able to classify the compressed and transformed 2DHOG features directly in the transform domain.

In this paper, we apply the 2D discrete Fourier transform (2DDFT) or the 2D discrete cosine transform (2DDCT) on block-partitioned 2DHOG, followed by a truncation process to retain only a fixed number of low frequency coefficients, which are referred to as TD2DHOG features. Further, using the 2DDFT downsampling theorem (Smith 2007) and considering the effect of image resampling on the 2DHOG features Dollár et al. (2014), it is shown that the TD2DHOG features obtained from an image at the original resolution and a downsampled version from the same image are approximately the same within a multiplicative factor, with a similar result holding true when 2DDCT is used. The use of TD2DHOG features simplifies the classifier training phase, since the classifier trained on high resolution vehicles can be used to detect the same or lower resolution vehicles in the test image, instead of training multiple classifiers, each being trained on vehicles with a specific resolution, as done in Benenson et al. (2012) and Naiel et al. (2014). Next, we employ the two-dimensional principal component analysis (2DPCA) (Yang et al. 2004) for feature extraction and dimensionality reduction. The design of the proposed scheme aims to solve the challenging problem of scale variation that is common in most vehicle detection datasets. Extensive experiments are conducted in order to evaluate the detection performance of the proposed technique and compare it with that of the state-of-the-art techniques.

The paper is organized as follows. In Sect. 2, we present a brief background about 2DHOG features, and the effect of image resampling on these features. In Sect. 3, we study the effect of downsampling a grayscale image on its DFT and DCT versions. In Sect. 4, a detailed description of extracting the TD2DHOG features is presented. Further, a model for the multiplicative factor that approximately relates the TD2DHOG features at two different resolutions of a given image is established. In Sect. 5, the model derived in Sect. 4 is used in proposing a scheme for vehicle detection of different resolutions using a single classifier rather than a classifier pyramid. In Sect. 6, we first validate experimentally the proposed model for the multiplicative factor in both the 2DDFT and 2DDCT domains. Then, the performance of the proposed vehicle detection scheme is studied by carrying out extensive experiments using a number of publicly available vehicle detection datasets and compared with that of the state-of-the-art techniques. Finally, Sect. 7 highlights the work of this paper.

2 Background

In this section, we present some background material required for the development of the proposed detection scheme using TD2DHOG features in subsequent sections.

2.1 Two-dimensional HOG features

2DHOG features are similar to the HOG features of Dalal and Triggs (2005), the difference being the way in which the features are represented, namely, in a 2D matrix format in the case of the former and a 1D vector format in the case of the latter. The 2DHOG features have been used in a number of papers (Dollár et al. 2009; Felzenszwalb et al. 2010; Maji et al. 2008).

Let us consider an image, I, of size $(M_1 \times M_2)$, and divide it into non-overlapping cells of size $(\eta_1 \times \eta_2)$ pixels. The 2DHOG features are computed from the input image as follows. First, we convolve the image I with the filter $L = [-1, 0, 1]$ and its transpose $L^\top$ to obtain the gradients $g_x(i, j)$ and $g_y(i, j)$, in the x and y directions, respectively, where i and j denote the pixel indices. Then, we compute the magnitude $\varGamma(i, j)$ and the orientation $\theta(i, j)$ of the gradient at (i, j) as

$$\begin{aligned} \varGamma(i, j) &= \sqrt{g_x(i, j)^2 + g_y(i, j)^2} \\ \theta(i, j) &= \arctan\big(g_y(i, j)/g_x(i, j)\big) \end{aligned} \tag{1}$$

Next, the orientation $\theta(i, j)$ at the (i, j)th pixel is quantized into β bins to obtain the corresponding quantized orientation $\hat{\theta}(i, j) \in \{\varOmega_l\}$, $\varOmega_l = (l-1)\dfrac{\pi}{\beta}$, $l = 1, 2, \ldots, \beta$. Then, the 2DHOG features for the lth layer, $h^l(\hat{i}, \hat{j})$, can be computed using the following equation

$$h^l(\hat{i}, \hat{j}) = \sum\nolimits_{i=(\hat{i}-1)\eta_1+1}^{\hat{i}\eta_1} \left(\sum\nolimits_{j=(\hat{j}-1)\eta_2+1}^{\hat{j}\eta_2} \varGamma(i, j)\delta_l(i, j) \right) \tag{2}$$

where

$$\delta_l(i, j) = \begin{cases} 1, & \text{if } \hat{\theta}(i, j) = \varOmega_l \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$\hat{i}$ and $\hat{j}$ being the cell indices, $1 \le \hat{i} \le \tilde{M}_1 = M_1/\eta_1$, $1 \le \hat{j} \le \tilde{M}_2 = M_2/\eta_2$, such that $\tilde{M}_1$ and $\tilde{M}_2$ are integers. Thus, the 2D representation for the HOG features results in β-layers, h^l $(l = 1, 2, \ldots, \beta)$, where the spatial relation between neighboring cells is maintained, and the size of each layer is $(\tilde{M}_1 \times \tilde{M}_2)$.

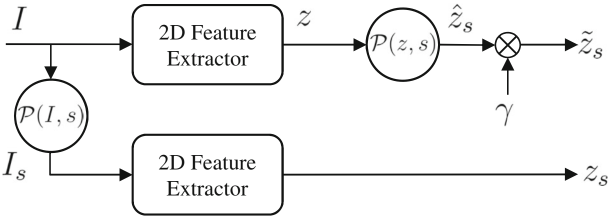

Fig. 1 Block diagram illustrating the approximate relationship between the resampled features of an image at a given resolution and the features extracted from a resampled version of the same image

2.2 Effect of image resampling on 2DHOG features

Statistics of resampled images in the spatial domain have been studied in Huang and Mumford (1999) and Ruderman (1994). Recently, the effect of image resampling on 2DHOG features in the spatial domain has been studied by Dollár et al. (2014, 2010). In this section, we give a brief description of the work in Dollár et al. (2014), which will be used later in studying the effect of image resampling on the features in the transform domain.

Let $I_s = \mathcal{P}(I, s)$ denote the input image I resampled by a factor s, where $s < 1$ represents downsampling, $s > 1$ represents upsampling, and $\mathcal{P}$ represents the resampling operator in the spatial domain. The exact channel features extracted from the image at the original resolution, and the same image at a different resolution can be represented by $z = \Lambda(I)$, and $z_s = \Lambda(I_s)$, respectively, where Λ denotes a 2D spatial-domain feature extractor. It has been shown in Dollár et al. (2014) that resampling the image I by a factor s, $I_s = \mathcal{P}(I, s)$, followed by computing the exact 2D channel features, $z_s = \Lambda(I_s)$, can be approximated by resampling the channel feature, z, followed by a multiplicative factor, γ, that is modeled by using the power law as

$$z_s = \Lambda(\mathcal{P}(I, s)) \approx \tilde{z}_s = \gamma \mathcal{P}(z, s) \tag{4}$$

where

$$\gamma = a_0 s^{-\lambda} \tag{5}$$

and a_0 and λ depend on the type of channel features, which could be gradient, color or 2DHOG, and are empirically determined. This relationship is illustrated by the block diagram of Fig. 1. The values of a_0 and λ are not necessarily the same for the case of upsampling and downsampling for the same type of channel features.

For object detection using a single detection window, one constructs an image pyramid encompassing different scales, and then extracts the features from every scale in the pyramid. The use of the approximation in (4) allows the features generated at one scale from the image pyramid to approximate the features at nearby scales, thus reducing the cost of feature computation.

3 Effect of downsampling a grayscale image on its transformed version

In this section, we study the effect of downsampling a grayscale image on its DFT and DCT versions, and these results are then used in Sect. 4 to investigate the effect of image downsampling on transform-domain 2DHOG features.

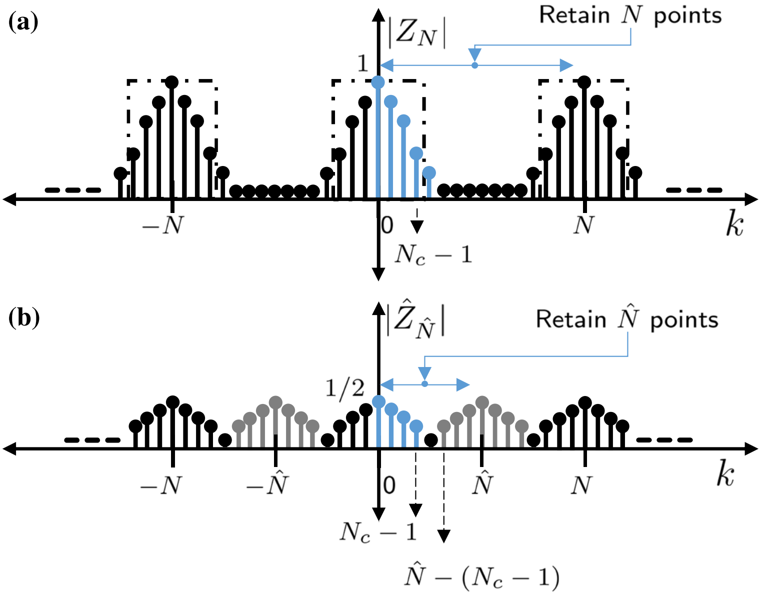

Fig. 2 **a** Magnitude of a signal in the DFT domain $Z_N[k]$, where a low pass filter with cutoff frequency N_c is used to bandlimit the signal. **b** Magnitude of the downsampled signal in the DFT domain $\hat{Z}_{\hat{N}}[k]$, where $N = 16$, $K = 2$, $\hat{N} = 8$, and $N_c = 4$ (Color figure online)

3.1 Effect on the DFT version

Let the N-point 1DDFT for the discrete time sequence, $z[n] \in \mathbb{R}$, be denoted as $Z_N[k]$, where $n = 0, 1, \ldots, N-1$, $k = 0, 1, \ldots, N-1$, N is an even integer multiple of K, and K being an integer. Let an ideal low pass filter of unity gain and a cutoff frequency $N_c \leq N/(2K)$ be used in order to bandlimit the signal. By downsampling z by K in the time domain, the downsampled signal $\hat{z}$ of length $\hat{N} = N/K$ is obtained. Then, the $\hat{N}$-point 1DDFT is employed on the downsampled signal, $\hat{z}$, in order to obtain the downsampled signal in the frequency domain, $\hat{Z}_{\hat{N}}$. Now, the relations between the original signal and its downsampled version in the time domain and that in the frequency domain are given by

$$\hat{z}[n] = z[Kn] \tag{6}$$

$$\hat{Z}_{\hat{N}}[k] = \frac{1}{K}\sum_{i=0}^{K-1} Z_N\left[k + i\hat{N}\right] \tag{7}$$

where $n = 0, 1, \ldots, \hat{N}-1$, and $k = 0, 1, \ldots, \hat{N}-1$. It is clear from (7) that the downsampled signal in the 1DDFT domain, $\hat{Z}_{\hat{N}}$, is represented by a sum of K shifted copies of the original signal in the 1DDFT domain, Z_N, scaled by the factor $1/K$ (Smith 2007). Figure 2 illustrates an example of this in the DFT domain, when $N = 16$, $\hat{N} = 8$, $K = 2$, and $N_c = 4$. Since the original signal is bandlimited, then for $k = 0, 1, \ldots, c_1 - 1$, $c_1 \leq N_c$, the contribution of the summation shown in (7) is only coming from the first copy of Z_N at $i = 0$, and so we have

$$Z_N[k] = K\hat{Z}_{\hat{N}}[k] \tag{8}$$

This result is supported by that presented in Bi and Mitra (2011). We now consider a 2D signal. Let $g \in \mathbb{R}^2$ represent a grayscale image in the spatial domain of size $(N_1 \times N_2)$, where N_1 and N_2 are even integer multiples of K_1 and K_2, respectively, K_1 and K_2 being integers. Assume that an ideal low pass filter of unity gain and cutoff frequencies $N_{c_1} \leq N_1/(2K_1)$ and $N_{c_2} \leq N_2/(2K_2)$ is used to bandlimit the original signal. Downsampling g by a factor K_1 in the y direction, and K_2 in the x direction results in $\hat{g}[n, m] = g[K_1 n, K_2 m]$ of size $(\hat{N}_1 \times \hat{N}_2)$, where n and m represent the spatial domain discrete sample indices, $0 \leq n \leq \hat{N}_1 - 1$, $0 \leq m \leq \hat{N}_2 - 1$, $\hat{N}_1 = N_1/K_1$ and $\hat{N}_2 = N_2/K_2$. We now take the 2DDFT of g and $\hat{g}$ to obtain G_{N_1,N_2} and $\hat{G}_{\hat{N}_1,\hat{N}_2}$ corresponding to the 2DDFT coefficients of the original image and that of its downsampled version, respectively. Similar to the case of 1DDFT, the relation between $G_{N_1,N_2}[u, v]$ and $\hat{G}_{\hat{N}_1,\hat{N}_2}[u, v]$ can be expressed as

$$\hat{G}_{\hat{N}_1,\hat{N}_2}[u, v] = \frac{1}{K_1 K_2} \sum_i \sum_j G_{N_1,N_2}[u + i\hat{N}_1, v + j\hat{N}_2] \tag{9}$$

where $u = 0, 1, \ldots, \hat{N}_1 - 1$, $v = 0, 1, \ldots, \hat{N}_2 - 1$, $i = 0, 1, \ldots, K_1 - 1$, and $j = 0, 1, \ldots, K_2 - 1$. It is seen from this equation that the downsampled image in the 2DDFT domain is represented by a sum of $K_1 \times K_2$ shifted copies of the original image in the 2DDFT domain and scaled by the factor $1/(K_1 K_2)$. Let c_1 and c_2 denote the maximum frequencies retained by the truncation operator. For $u = 0, 1, \ldots, c_1 - 1$, $v = 0, 1, \ldots, c_2 - 1$, $c_1 \leq N_{c_1}$, and $c_2 \leq N_{c_2}$ the contribution of the summation shown in (9) is from the copy corresponding to $i = j = 0$, and we can obtain the following relation

$$G_{N_1,N_2}[u, v] = K_1 K_2 \hat{G}_{\hat{N}_1,\hat{N}_2}[u, v] \tag{10}$$

From the above equation it is seen that the ratio between a grayscale image in the 2DDFT domain and that of its downsampled version is $K_1 K_2$.

3.2 Effect on the DCT version

In Ahmed et al. (1974) the N-point 1DDCT, X_N, for the discrete time sequence, $x \in \mathbb{R}$, is given by

$$X_N[k] = \hat{\Gamma}_N[k] \sum_{n=0}^{N-1} x[n] \cos \frac{\pi(2n+1)k}{2N} \tag{11}$$

where $\hat{\Gamma}_N[k] = \sqrt{1/N}$ for $k = 0$, and $\hat{\Gamma}_N[k] = \sqrt{2/N}$ for $0 < k \leq N - 1$. The N-point 1DDCT can be computed by 2N-point 1DDFT for a sequence, $y[n]$, as follows. First, let $x[n]$ be a bandlimited signal and $y[n]$ be defined as

$$y[n] = \begin{cases} x[n], & 0 \leq n \leq N - 1 \\ 0, & N \leq n \leq 2N - 1 \end{cases} \tag{12}$$

The 1DDFT is employed on y in order to obtain Y_{2N}. It has been shown in Ahmed et al. (1974) that the signal $X_N[k]$ in the 1DDCT domain is related to $Y_{2N}[k]$ by

$$X_N[k] = \hat{\Gamma}_N[k] \mathrm{Re}\left(Y_{2N}[k] e^{-j\frac{\pi k}{2N}} \right) \tag{13}$$

where $k = 0, 1, \ldots, N - 1$, and Re() is a function which returns the real part of an input complex number. Let an ideal low pass filter of gain unity and a cutoff frequency $N_c \leq N/K$ be used in order to bandlimit the signal Y_{2N}, where N is an even integer multiple of K,

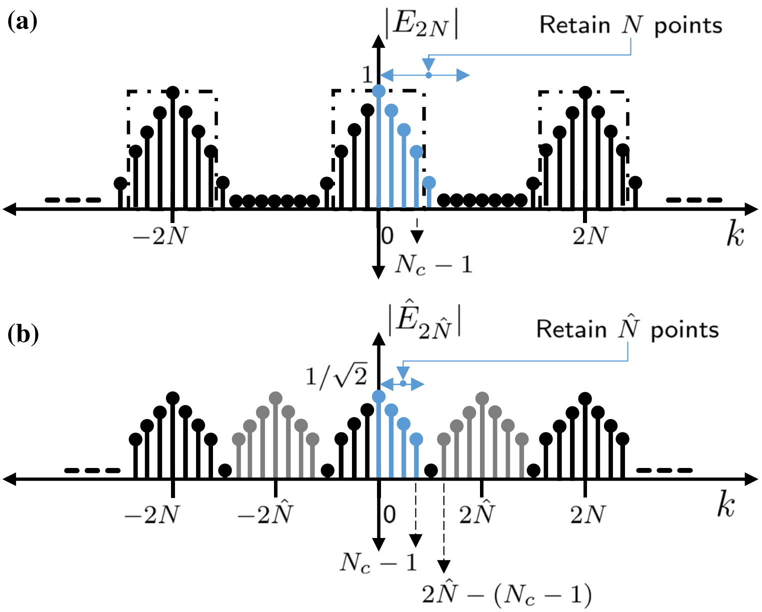

Fig. 3 **a** Magnitude of the signal $E_{2N}[k]$ defined as $E_{2N}[k]=Y_{2N}[k]e^{-j\frac{\pi k}{2N}}$, where a low pass filter with cutoff frequency N_c is used to bandlimit the signal. **b** Magnitude of the downsampled signal in the DFT domain $\hat{E}_{2\hat{N}}[k]$, where $N=8$, $K=2$, $\hat{N}=4$, and $N_c=4$ (Color figure online)

and K being an integer. Let $E_{2N}[k]$ be a 1D signal in the 1DDFT domain, and be defined as $E_{2N}[k]=Y_{2N}[k]e^{-j\frac{\pi k}{2N}}$. From the downsampling theorem given by (7), downsampling $E_{2N}[k]$ by a factor K in the 1DDFT domain is obtained as:

$$\hat{E}_{2\hat{N}}[k]=\frac{1}{K}\sum_{i=0}^{K-1}E_{2N}\left[k+i2\hat{N}\right] \tag{14}$$

where $\hat{E}_{2\hat{N}}$ is of length $2\hat{N}=2N/K$, and $k=0,1,\ldots,N-1$. Figure 3a and b illustrate an example for $E_{2N}[k]$ and $\hat{E}_{2\hat{N}}[k]$, respectively, where $N=8$, $K=2$, $\hat{N}=4$, and $N_c=4$. Now, the downsampled signal in the 1DDCT domain, $\hat{X}_{\hat{N}}$ of length $\hat{N}$, can be obtained as follows:

$$\hat{X}_{\hat{N}}[k]=\hat{\Gamma}_{\hat{N}}[k]\mathrm{Re}(\hat{E}_{2\hat{N}}[k]) \tag{15}$$

$$=\hat{\Gamma}_{\hat{N}}[k]\mathrm{Re}\left(\frac{1}{K}\sum_{i=0}^{K-1}Y_{2N}[k+i2\hat{N}]e^{-j\frac{\pi(k+i2\hat{N})}{2N}}\right) \tag{16}$$

Let c_1 denote the maximum frequency retained by the truncation operator. Since Y_{2N} is bandlimited to the maximum frequency $N_c\le N/K$, then for $k=0,1,\ldots,c_1-1$, where $c_1\le N_c$, the contribution of the summation shown in (16) is coming only from $i=0$ copy, and so we can simplify the above relation as

$$\hat{X}_{\hat{N}}[k] = \frac{1}{K}\hat{\Gamma}_{\hat{N}}[k]\mathrm{Re}\left(Y_{2N}[k]e^{-j\frac{\pi k}{2N}}\right) \tag{17}$$

$$= \frac{\hat{\Gamma}_{\hat{N}}[k]}{K\hat{\Gamma}_{N}[k]}\hat{\Gamma}_{N}[k]\mathrm{Re}\left(Y_{2N}[k]e^{-j\frac{\pi k}{2N}}\right) = \frac{\sqrt{1/\hat{N}}}{K\sqrt{1/N}}X_N[k] \tag{18}$$

$$= \frac{1}{\sqrt{K}}X_N[k] \tag{19}$$

Thus, the relation between a 1DDCT transformed signal and its downsampled version in the 1DDCT domain can be expressed as

$$X_N[k] = \sqrt{K}\hat{X}_{\hat{N}}[k] \tag{20}$$

where $0 \le k \le c_1 - 1$. Similar to the case of 1DDCT, the 2DDCT can be related to the 2DDFT. Let c_1 and c_2 denote the maximum frequencies retained by the truncation operator. Then, the relation between a grayscale image G_{N_1,N_2} in the 2DDCT domain and that of its downsampled version $\hat{G}_{\hat{N}_1,\hat{N}_2}$ can be represented as

$$G_{N_1,N_2}[u,v] = \sqrt{K_1K_2}\hat{G}_{\hat{N}_1,\hat{N}_2}[u,v] \tag{21}$$

where $\hat{N}_1 = N_1/K_1, \hat{N}_2 = N_2/K_2, u = 0, 1, \ldots, c_1 - 1, v = 0, 1, \ldots, c_2 - 1, c_1 \le N_1/K_1$, and $c_2 \le N_2/K_2$. In the appendix, the derivation of the above expression is provided.

4 Transform-domain 2DHOG features

In this section, we first define 2DHOG features in the transform domain. Then, utilizing the results derived in Sect. 3, we investigate the relationship between the transform-domain 2DHOG features obtained from an image of a given resolution and those obtained from a downsampled version of the same image.

4.1 Extraction of TD2DHOG features

Consider an input image I of size $(M_1 \times M_2)$. Let it be divided into non-overlapping cells of size $(\eta_1 \times \eta_2)$, where M_1 and M_2 are integer multiples of powers of 2, and η_1 and η_2 are integer powers of 2. Now, 2DHOG features are computed by following the steps explained in Sect. 2.1, resulting in β layers, where each layer corresponds to a certain quantized gradient orientation from 0° to 180°. The 2DHOG features of the l^{th} layer, denoted by h^l, is of size $(\tilde{M}_1 \times \tilde{M}_2)$, $\tilde{M}_1$ and $\tilde{M}_2$ being integer multiples of powers of 2. Each 2DHOG layer, h^l, is partitioned into a number of non-overlapping blocks, N_x and N_y in the x and y directions, respectively, where N_x and N_y are integers. Let $\boldsymbol{x}^l_{\imath\jmath}$, of size $(b \times b)$, represent the 2DHOG features of the $(\imath, \jmath)$th block of the lth layer, where $1 \le \imath \le N_y$, $1 \le \jmath \le N_x$, b being an integer power of 2. The block-partitioned 2DHOG features in the lth layer can be represented as

$$h^l = \begin{bmatrix} \boldsymbol{x}^l_{11} & \cdots & \boldsymbol{x}^l_{1N_x} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}^l_{N_y1} & \cdots & \boldsymbol{x}^l_{N_yN_x} \end{bmatrix} \tag{22}$$

This block partitioning is known to offer a robustness to partial occlusion (Wang et al. 2009; Wu et al. 2014). To illustrate let us consider an image of size 32 × 96, a cell size of 4 × 4, and

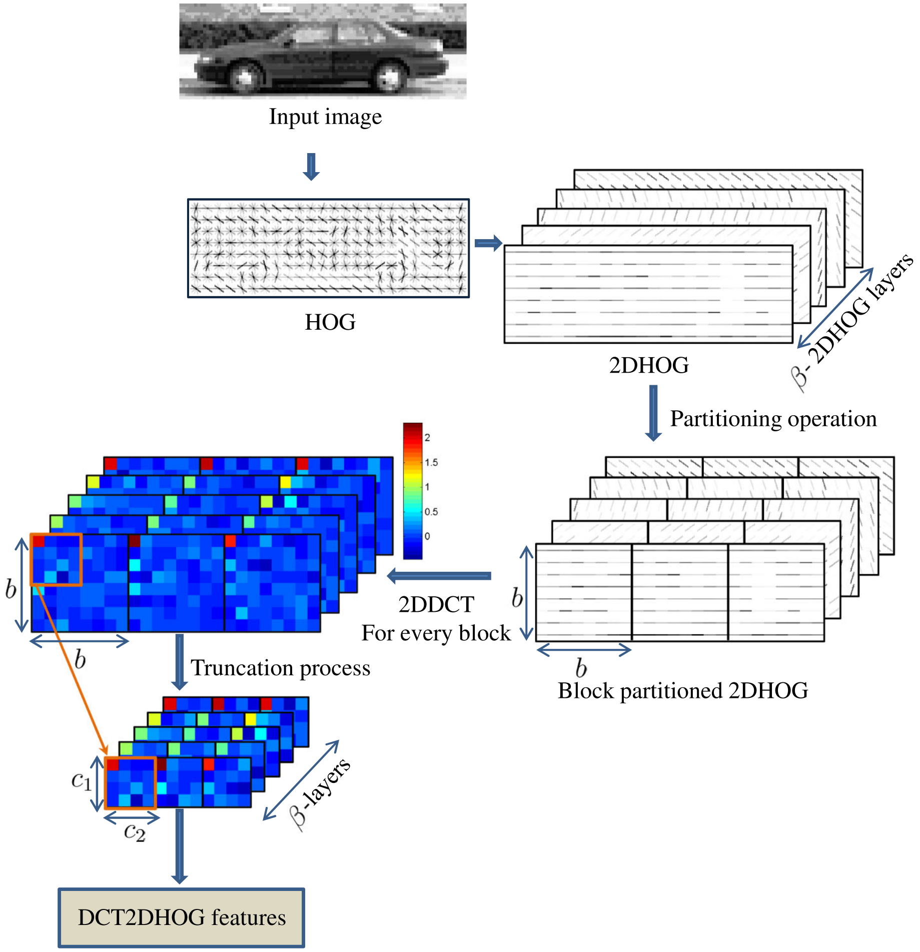

Fig. 4 Scheme for obtaining the DCT2DHOG features for an input car image of size 32 × 96 using $\beta = 5$, cell size 4 × 4, 2DDCT block size $b = 8$ and $c_1 = c_2 = 4$ (Color figure online)

$\beta = 5$. If $b = 8$, then $N_x = \tilde{M}_2/b = M_2/(\eta_2 b) = 3$, and $N_y = \tilde{M}_1/b = M_1/(\eta_1 b) = 1$. Hence, each of the five layers is partitioned into 3 blocks of size 8 × 8. However, if $b = 4$, then $N_x = 6$ and $N_y = 2$; that is, each of the layers is partitioned into 12 blocks of size 4 × 4.

Next, we apply the appropriate 2D transform, 2DDFT or 2DDCT, on each block resulting in 2DHOG of the corresponding block in the transform domain. Let $\mathbf{x}^l_{\iota\jmath} = T(\boldsymbol{x}^l_{\iota\jmath})$, where $T(.)$ represents the transform. The corresponding 2DHOG features in the transform domain can be represented as

$$H^l = \begin{bmatrix} \mathbf{x}^l_{11} & \dots & \mathbf{x}^l_{1N_x} \\ \vdots & \ddots & \vdots \\ \mathbf{x}^l_{N_y 1} & \dots & \mathbf{x}^l_{N_y N_x} \end{bmatrix} \tag{23}$$

Let $\boldsymbol{\phi}_{c_1 c_2}(.)$ denote the 2D truncation operator in the transform domain that truncates the coefficients corresponding to the frequencies greater than the frequencies c_1 and c_2. By

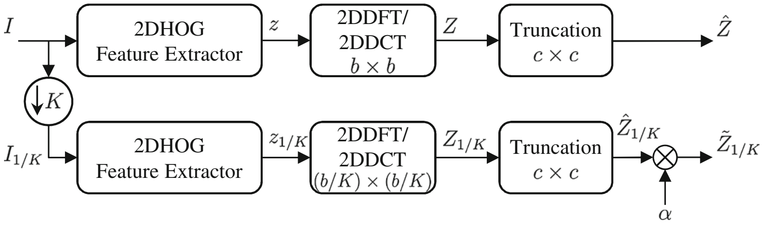

Fig. 5 Block diagram showing the effect of downsampling an input image by an integer factor K in both the x and y directions on the transform-domain 2DHOG features, where α is a multiplicative factor that allows the features extracted from the lower resolution image to approximate the features extracted from the image at the original resolution

applying $\phi_{c_1 c_2}(.)$ on each block, $\mathbf{x}^l_{\iota J}$, we can obtain the truncated features as $\hat{\mathbf{x}}^l_{\iota J} = \boldsymbol{\phi}_{c_1 c_2}(\mathbf{x}^l_{\iota J})$ of size $(c_1 \times c_2)$. Then, these features can be represented as

$$\hat{H}^l = \begin{bmatrix} \hat{\mathbf{x}}^l_{11} & \dots & \hat{\mathbf{x}}^l_{1N_x} \\ \vdots & \ddots & \vdots \\ \hat{\mathbf{x}}^l_{N_y 1} & \dots & \hat{\mathbf{x}}^l_{N_y N_x} \end{bmatrix} \tag{24}$$

where the size of $\hat{H}^l$ is $(\hat{M}_1 \times \hat{M}_2)$, $\hat{M}_1 = c_1 N_y$ and $\hat{M}_2 = c_2 N_x$. We call the above truncated transform-domain 2DHOG features given by $\hat{H}^l$ as TD2DHOG features. We refer to the TD2DHOG features as DFT2DHOG and DCT2DHOG features when the 2D transform used is 2DDFT and 2DDCT, respectively. The scheme for obtaining the DCT2DHOG features is illustrated in Fig. 4 for an image of size 32 × 96 with a cell size of 4 × 4, $\beta = 5$, and 2DDCT is employed with block size $b = 8$, and $c_1 = c_2 = 4$. It is noted that for this example the size of $\hat{H}^l$ is 4 × 12.

4.2 Effect of image downsampling on TD2DHOG features

In Sect. 3, we obtained the relation between the original image and its downsampled version when they are transformed by 2DDFT or 2DDCT. Now, in order to study the effect of image downsampling on the features in the transform domain, we use the block diagram shown in Fig. 5. For the original image I, a 2DHOG feature extraction operator $\Lambda(.)$ is employed to obtain $z = \Lambda(I)$. Then, we apply to z an appropriate 2D transform (2DDFT or 2DDCT), with a block size $b \times b$, followed by a truncation operation retaining the $c \times c$ low frequency coefficients for each block. The TD2DHOG features so obtained are denoted by $\hat{Z} = \hat{T}(z)$, where $\hat{T}$ represents the transform operation followed by the truncation operation. Let $I_{1/K}$ denote the image I downsampled by a factor K in both the x and y directions. Since $I_{1/K} = \mathcal{P}(I, 1/K)$, $\mathcal{P}$ representing the downsampling operator, the features extracted from the downsampled image are given by $z_{1/K} = \Lambda(\mathcal{P}(I, 1/K))$. We now obtain the features $\hat{Z}_{1/K} = \hat{T}_{1/K}(z_{1/K})$ in the transform domain, where the features $z_{1/K} = \Lambda(I_{1/K})$, and $\hat{T}_{1/K}$ represents the transform operation with a block size $(b/K) \times (b/K)$ followed by the truncation operation to retain the $(c \times c)$ low frequency coefficients.

The relationship between the transform coefficients of the features obtained from the image at the original resolution $\hat{Z}$ and that of its downsampled version $\hat{Z}_{1/K}$ can now be obtained as follows. Equations (4) and (5) are now used to approximate $z_{1/K}$ as

$$z_{1/K} \approx \mathcal{P}(z, 1/K)a'_0 K^{\lambda} \tag{25}$$

where a'_0 and λ are computed empirically for each type of channel features. Next, performing the transform operation $\hat{T}_{1/K}$ on both sides of (25), we obtain

$$\hat{T}_{1/K}(z_{1/K}) \approx \hat{T}_{1/K}(\mathcal{P}(z, 1/K))a'_0 K^{\lambda}$$

i.e.,

$$\hat{Z}_{1/K} \approx \hat{T}_{1/K}(\mathcal{P}(z, 1/K))a'_0 K^{\lambda} \tag{26}$$

Then, the ratio between the features in the transform domain obtained from the original image and its resampled version is

$$\frac{\hat{Z}}{\hat{Z}_{1/K}} \approx \frac{1}{a'_0 K^{\lambda}} \times \frac{\hat{T}(z)}{\hat{T}_{1/K}(\mathcal{P}(z, 1/K))} \tag{27}$$

where the first term, $1/(a'_0 K^{\lambda})$, represents the power law effect, while the second term, $\hat{T}(z)/\hat{T}_{1/K}(\mathcal{P}(z, 1/K))$, represents the transform domain resampling effect which is the ratio of the transform-domain coefficients of the channel feature, z, and that of its resampled version, $\mathcal{P}(z, 1/K)$.

Let $a_0 = 1/a'_0$ and assume the term $\hat{T}(z)/\hat{T}_{1/K}(\mathcal{P}(z, 1/K))$ can be represented by (10) and (21), in case of 2DDFT and 2DDCT, respectively. Then, the transform-domain coefficients of the original resolution, $\hat{Z}$, can be approximated by using the transform-domain coefficients at a lower resolution, $\hat{Z}_{1/K}$, as

$$\hat{Z} \approx \alpha(K)\hat{Z}_{1/K} \tag{28}$$

where

$$\alpha(K) = \begin{cases} a_0 K^{2-\lambda}, & \text{for 2DDFT} \\ a_0 K^{1-\lambda}, & \text{for 2DDCT} \end{cases} \tag{29}$$

In order to improve the approximation accuracy of expression in (28), we introduce an additive correction term a_1, such that α is of the form

$$\alpha(K) = \begin{cases} a_0 K^{2-\lambda} + a_1, & \text{for 2DDFT} \quad \text{(30a)} \\ a_0 K^{1-\lambda} + a_1, & \text{for 2DDCT} \quad \text{(30b)} \end{cases}$$

The constants a_0, a_1, and λ are computed empirically in the training mode for the 2DHOG channel. The usefulness of $\alpha(K)$ given by (30) lies in the fact that the features extracted from a lower resolution test image can be utilized to approximate the features of the test image extracted at a higher resolution by multiplying the former by $\alpha(K)$, which is a function of the downsampling factor, K, and the type of transform.

4.2.1 Estimation of a_0, a_1, and λ

Given a training set of N_t images, the parameters a_0, a_1, and λ for the 2DHOG channel can be estimated as follows. First, at each value of the downsampling factor, $K = 1, 2, 4, \ldots$, the multiplicative factor of the ith image sample, $\hat{\alpha}^i(K)$, is obtained as the factor that minimizes the mean square error (MSE) as

$$\min_{\hat{\alpha}^i(K)} \frac{1}{N_y N_x c^2 \beta} \sum_{l,j,k,u,v} (\hat{Z}^{i,j,k,l}[u, v] - \hat{\alpha}^i(K)\hat{Z}^{i,j,k,l}_{1/K}[u, v])^2 \tag{31}$$

where $i = 1, \ldots, N_t, 0 \leq u, v \leq c-1$, u and v are the frequency indices of the (j, k)th block, $1 \leq j \leq N_y$, $1 \leq k \leq N_x$, and $l = 1, 2, \ldots, \beta$. Then, the average value of the estimated multiplicative factor $\hat{\alpha}(K)$ is obtained as $\hat{\alpha}(K) = (1/N_t) \sum_{i=1}^{N_t} \hat{\alpha}^i(K)$. Finally, the values of the estimated multiplicative factor $\hat{\alpha}(K)$ are used to obtain the model parameters, a_0, a_1, and λ, of $\alpha(K)$ by using the least squares curve fitting. In Sect. 6.1, we compute empirically the values of a_0, a_1, and λ.

5 Scheme for vehicle detection

In this section, we propose a new vehicle detection scheme by using the results of the previous section concerning TD2DHOG features so as to employ a single classifier trained on vehicles of high resolution in order to detect vehicles of the same or lower resolution, instead of training multiple resolution-specific classifiers, as in Benenson et al. (2012) and Naiel et al. (2014). In order to detect vehicles of different resolutions in a given test image, an image pyramid of depth one octave is constructed, and TD2DHOG features are extracted at each scale from the image pyramid with blocks of different sizes. We now present our methods for training and testing of the proposed vehicle detection scheme.

5.1 Training mode

In order to take advantage of the fact that the transform-domain coefficients of the original resolution can be approximated by using the transform-domain coefficients at a lower resolution as given by (28), the training data is upsampled by a factor of R, R being an integer power of 2. Even though upsampling of the training data will cause an increase in the training cost, it has been observed from our experiments that training a classifier on TD2DHOG features obtained at a high resolution of images offers a detection accuracy higher than that achieved by the same classifier when trained on TD2DHOG features extracted from the same training set at a lower resolution. This is because of the fact that in the testing mode, going from a higher resolution to a lower resolution results in a smaller approximation error for TD2DHOG features than when going the other way around.

Figure 6a shows the training scheme for the proposed vehicle detector, where the training data is upsampled by a factor R in both the x and y directions. Let the set of the training data upsampled by R be denoted as $\mathcal{I}_R = \{I_{i,R}, i = 1, 2, \ldots, N_t\}$, where N_t denotes the number of training image samples. Then, the size of the ith training image sample is $(RM_1 \times RM_2)$. Assume the 2DHOG features of the lth layer, $h^l_{i,R}$, $(i = 1, 2, \ldots, N_t$ and $l = 1, 2, \ldots, \beta)$, are extracted by using the same cell size for all the resolutions $(\eta_1 \times \eta_2)$, then the size of the lth 2DHOG layer of the ith training image sample is $R\tilde{M}_1 \times R\tilde{M}_2$, i.e., increased by the same factor R. Similarly, the block size used to compute the corresponding TD2DHOG features is increased by the same factor R, i.e., $b_R = Rb_0$. We call b_0 as the *base block size*, which is defined as the block size at $R = 1$. Let $\hat{H}^l_{i,R}$, $i = 1, 2, \ldots, N_t$, denote the TD2DHOG features of the lth layer, where the size of $\hat{H}^l_{i,R}$ is $(\hat{M}_1 \times \hat{M}_2)$. It is important to note that, in the training phase we do not multiply TD2DHOG features by the multiplicative factor $\alpha(K)$, and we use the value of $\alpha(K)$ computed from (30) in the detection phase.

After the extraction of the TD2DHOG features, 2DPCA (Yang et al. 2004) is employed on each layer in order to maintain the relation between the neighboring blocks. Let the training data consist of N_{pos} and N_{neg} training image samples, corresponding to the posi-

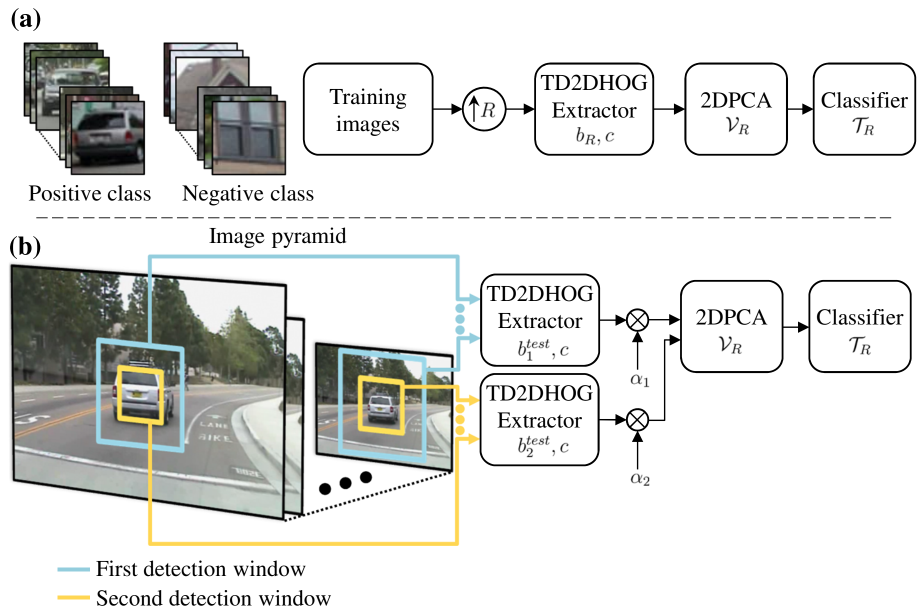

Fig. 6 **a** The scheme for training the proposed vehicle detector with training images of size 64 × 64, where R is the upsampling factor in both the x and y directions. **b** Proposed vehicle detection scheme for a sample test image, where the different colors in the image pyramid represent different scanning window sizes (here we have used only two window sizes, 128 × 128 and 64 × 64) (Color figure online)

tive and negative classes, respectively. The training data can be denoted as $\{(\hat{H}^l_{i,R}, y_i), i = 1, 2, \ldots, N_t\}, l = 1, 2, \ldots, \beta$, where $y_i \in \{+1, -1\}$ refers to the class label for the ith image sample. The covariance matrix, of size $(\hat{M}_2 \times \hat{M}_2)$, is first obtained for the TD2DHOG features of the lth layer as

$$Cov^l = \frac{1}{N_t}\sum_{i=1}^{N_t}(\hat{H}^l_{i,R} - \bar{H}^l_R)^\top(\hat{H}^l_{i,R} - \bar{H}^l_R) \tag{32}$$

where

$$\bar{H}^l_R = \frac{1}{N_t}\sum_{i=1}^{N_t}\hat{H}^l_{i,R} \tag{33}$$

Note that Cov^l is a nonnegative definite matrix. Next, we obtain the r_l eigenvectors of Cov^l that correspond to the r_l dominant eigenvalues. The number of eigenvectors, r_l, is chosen so that the sum of the magnitude of the retained eigenvalues represents at least 90% of the sum of the magnitude of all the eigenvalues. The eigenvectors are used to form the matrix V^l_R of size $(\hat{M}_2 \times r_l)$. Next, the TD2DHOG features of the lth layer of the ith training image sample are projected onto the constructed matrix V^l_R in order to obtain the matrix $Q^l_{i,R} = \hat{H}^l_{i,R}V^l_R$ of size $(\hat{M}_1 \times r_l)$, and $Q^l_{i,R}$ is vectorized[1] to obtain the corresponding feature vector $q^l_{i,R}$ of size $(1 \times \hat{M}_1 r_l)$. Then, for the ith training image sample, the feature

[1] The vectorization function is defined as Mat2Vec: $\mathbb{R}^{\mu\times\nu} \to \mathbb{R}^{\rho}$, where $\rho = \mu\nu$ is the dimension of the vector, and $(\mu \times \nu)$ is the order of the input matrix. The inverse of the vectorization function is defined as Vec2Mat: $\mathbb{R}^{\rho} \to \mathbb{R}^{\mu\times\nu}$.

vectors from different layers, $q^l_{i,R}$, are concatenated to obtain the feature vector, $f_{i,R}$, of size $(1 \times r)$, where $f_{i,R} = [q^1_{i,R}, \ldots, q^\beta_{i,R}]$ for $i = 1, 2, \ldots, N_t$.

Let the set of training features obtained after applying 2DPCA be denoted as $\mathcal{F}_R = \{f_{i,R}, i = 1, 2, \ldots, N_t\}$, and the set of the eigenvectors used to generate these features be denoted as $\mathcal{V}_R = \{V^l_R, l = 1, 2, \ldots, \beta\}$. Then, we train a classifier, $\mathcal{T}_R$, for the upsampling factor R by using the corresponding features $\mathcal{F}_R$. We use one of the two state-of-the-art classifiers: a support vector machine with fast histogram intersection kernel (FIKSVM) (Maji et al. 2008, 2013) or boosted decision tree classifier (BDTC) (Appel et al. 2013; Dollár 2016).

5.2 Testing mode

In the testing phase, we first obtain an image pyramid of depth of one octave from the given input test image. The test image at each scale of the image pyramid, is then scanned by using a number of detection windows of different sizes as $(\frac{RM_1}{K} \times \frac{RM_2}{K})$, where R is the upsampling factor at which the detector has been trained and $K = 1, 2, 4, \ldots,$ an integer power of 2. Figure 6b shows the proposed vehicle detection scheme when applied to a test image by assuming $R = 2$ and $K = 1$ and 2. Now for each detection window, we obtain the TD2DHOG features for different layers, $\{\hat{H}^l_{test}, l = 1, 2, \ldots, \beta\}$ by using a block size $b^{test} = \frac{b_R}{K}$; the size of each $\hat{H}^l_{test}$ is $(\hat{M}_1 \times \hat{M}_2)$. Then, the TD2DHOG features of each layer are multiplied by the multiplicative factor $\alpha(K)$ as

$$\tilde{H}^l_{test} = \alpha(K)\hat{H}^l_{test} \tag{34}$$

where $\tilde{H}^l_{test}$ is of size $(\hat{M}_1 \times \hat{M}_2)$, and $\alpha(K)$ is given by (30), which allows the TD2DHOG features obtained from a low resolution detection window to approximate the TD2DHOG features obtained at a higher resolution, indicating an approximate invariance of the TD2DHOG features within a multiplicative factor, when the image resolution is changed. Next, the TD2DHOG features of the lth layer, $\tilde{H}^l_{test}$, is projected onto the corresponding matrix V^l_R in order to obtain the matrix $Q^l_{test} = \tilde{H}^l_{test}V^l_R$ of size $(\hat{M}_1 \times r_l)$. Then, Q^l_{test} is vectorized to obtain the corresponding feature vector q^l_{test} of size $(1 \times \hat{M}_1 r_l)$. This is followed by concatenating the features, q^l_{test}, for different layers to obtain the feature vector, f_{test}, of size $(1 \times r)$, where $f_{test} = [q^1_{test}, \ldots, q^\beta_{test}]$.

Now, the trained classifier $\mathcal{T}_R$, namely, FIKSVM (Maji et al. 2008, 2013) or BDTC (Appel et al. 2013; Dollár 2016), is used to provide for each feature vector f_{test} a detection score corresponding to the input detection window. Finally, similar to Maji et al. (2008), a non-maximum suppression technique is used to combine several overlapped detections for the same object. This avoids detecting the same vehicle more than once, and allows detecting vehicles with different aspect ratios.

Figure 7a illustrates the scanning scheme for the proposed vehicle detector in the case of $R = 2$, and $K = 1$ and 2. Hence, in this example, the test image at each scale of the image pyramid is scanned by using two detection windows of sizes $(2M_1 \times 2M_2)$ and $(M_1 \times M_2)$. The proposed vehicle detector requires training a single classifier at the highest detection window size, namely, $(2M_1 \times 2M_2)$. The methods in Benenson et al. (2012) and Naiel et al. (2014) use a similar scanning strategy; however, they require constructing a classifier pyramid in order to classify detection windows of different sizes. It is to be noted that the scanning scheme used in several state-of-the-art object detectors (Dalal and Triggs 2005; Dollár et al. 2009; Maji et al. 2008) requires the extraction of features at each scale of an image pyramid of depth often more than one octave, even though the scheme employs one detection window

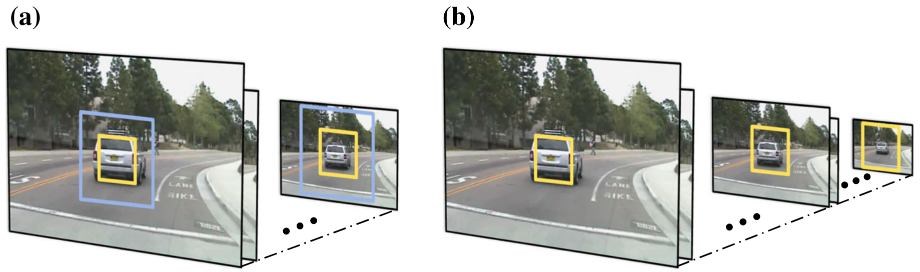

Fig. 7 **a** An illustration of the proposed scheme for scanning an image pyramid of depth one octave with two detection windows and a single classifier. **b** An illustration of the scheme for scanning an image pyramid of depth two octaves with one detection window and a single classifier (Color figure online)

and a single classifier. Figure 7b shows an example of this scanning scheme, when the image pyramid is of depth two octaves. The proposed vehicle detection scheme reduces the cost of training a classifier pyramid, as a single classifier trained on images of a given resolution can be used to detect vehicles of the same or lower resolutions. In addition, it reduces the storage requirements that are associated with training multiple resolution-specific classifiers.

6 Experimental results

We first carry out a number of experiments to validate, as mentioned in Sect. 4, the model for the multiplicative factor $\alpha(K)$ using the *UIUC car detection dataset* (Agarwal et al. 2004). Then, we study the performance of the proposed algorithm for vehicle detection in images using the *UIUC car detection dataset* (Agarwal et al. 2004), the *USC multi-view car detection dataset* (Kuo and Nevatia 2009), the *LISA 2010 dataset* (Sivaraman and Trivedi 2010) and the *HRI roadway dataset* (Gepperth et al. 2011). We also compare the performance of our algorithm with that of some of the existing methods.

The *UIUC car detection dataset* (Agarwal et al. 2004) consists of 1050 training images of size 40 × 100 divided into a set of 550 car images with side views, and a set of 500 other images, none of which is the image of a car with a side view. In order to facilitate the computation of the TD2DHOG features, the training images in this dataset are cropped by removing pixels from the first and last four rows and from the first and last two columns in order to reduce the size of each image from 40 × 100 to 32 × 96. The testing images in this dataset consist of 108 multi-scale images. The dataset consists of partially occluded cars, objects with low contrast, as well as highly textured background. Since the dataset includes a balanced number of positive and negative training images, the FIKSVM (Maji et al. 2013) is used as the baseline classifier for the proposed detector.

The *USC multi-view car detection dataset* (Kuo and Nevatia 2009) consists of cars with several views. The training data consists of 2462 positive training images of size 64 × 128, while the testing data consists of 196 images containing 410 cars of different sizes and views. In order to complete the training dataset, we collect 9512 negative training image samples from the *CBCL street scenes dataset* (Bileschi 2006). Since the USC dataset consists of cars with different views, BDTC (Appel et al. 2013; Dollár 2016) is chosen as the baseline classifier.

The *LISA 2010 dataset* (Sivaraman and Trivedi 2010) consists of test sequences of size 480 × 704 for rear view vehicles of different sizes, and this dataset has been captured under several illumination conditions. The first sequence (1600 frames) is taken on a high-density highway during a sunny day (H-dense), which includes vehicles in partial occlusions, heavy shadows, and some background structures are confused with the positive class, while the second (300 frames) on a medium-density highway on a sunny day (H-medium), where this sequence includes challenges similar to H-dense but at a lower density. The dataset does not include training data; therefore, we collect training images of size 64 × 64 from other datasets as follows: (1) 9013 images of vehicles in rear/front views from *KITTI dataset* (Geiger et al. 2012), and *USC multi-view car detection dataset* (Kuo and Nevatia 2009), and (2) 8415 negative image samples from *CBCL street scenes dataset* (Bileschi 2006). As in Sivaraman and Trivedi (2010), we collect a number of hard negative image samples from the test sequences (229 image samples from H-medium, and 806 image samples from H-dense). Due to the large number of training samples and the wide variation in the background structures, BDTC (Appel et al. 2013; Dollár 2016) is used as the baseline classifier on this dataset.

The *HRI roadway dataset* (Gepperth et al. 2011) consists of five test sequences of size 600 × 800 for vehicles on urban and highway areas. This dataset has been captured under several challenging weather and lighting conditions. Sequence I (908 frames) has been captured during a cloudy day, while Sequence II (917 frames) has been captured during a sunny day. Sequences III (611 frames), IV (411 frames) and V (830 frames) have been captured during a heavy rainy day, a dry midnight, and afternoon after a heavy snow, respectively. Since the HRI dataset does not have its own training set, in order to test the proposed scheme on a sequence of this dataset, the classifier in the proposed scheme is trained by employing the training set used in the case of *LISA 2010 dataset* along with the hard negative samples collected from the first 100 frames of this sequence of the HRI dataset.

6.1 Validation for the model of $\alpha(K)$

We now validate the model of $\alpha(K)$ given by (30) by making use of the block diagram of Fig. 5 and the scheme introduced in Sect. 4.2 for estimating the channel parameters a_0, a_1 and λ. For this purpose, we first consider the UIUC car detection dataset (Agarwal et al. 2004) and choose $N_t = 550$ car images. Since we do not have access to high resolution versions of these car images, they are upsampled by a factor $R = 8$. Now, we give the procedure to estimate the value of $\alpha(K)$ for the 2DHOG features in the 2DDFT domain. We first obtain the 2DHOG features of an upsampled image,[2] I_u, using the steps outlined in Sect. 2.1, assuming $\eta_1 = \eta_2 = 4$, and $\beta = 5, 7$ or 9. We then apply 2DDFT on block-partitioned 2DHOG features given by (22) for each of the layers, assuming the block size to be $b = Rb_0 = 8b_0$, $b_0 \in \{4, 8, 16\}$. This is followed by a truncation operation retaining the $(c \times c)$ low frequency coefficients, where $c = 4$, to obtain the 2DHOG features in the 2DDFT domain. Then, the whole operation is repeated after downsampling I_u by a factor K, $K = 1, 2, 4$, and 8, but with a block size of b/K. As explained in Sect. 4.2, the multiplicative factor of the ith image sample, $\hat{\alpha}^i(K)$, is obtained as the factor that minimizes the mean square error (MSE) given by (31). Then, the four values of the estimated multiplicative factor $\hat{\alpha}(K)$, $K = 1, 2, 4$, and 8, are used to obtain the model parameters, a_0, a_1, and λ, of $\alpha(K)$ by using the least squares curve fitting.[3] The above procedure is repeated to find the model parameters, a_0, a_1, and λ, of $\alpha(K)$ for the 2DHOG features in the 2DDCT domain.

[2] The toolbox (Dollár 2016) has been used to calculate the 2DHOG.

[3] The MATLAB function `lsqcurvefit` is used, http://www.mathworks.com/help/optim/ug/lsqcurvefit.html.

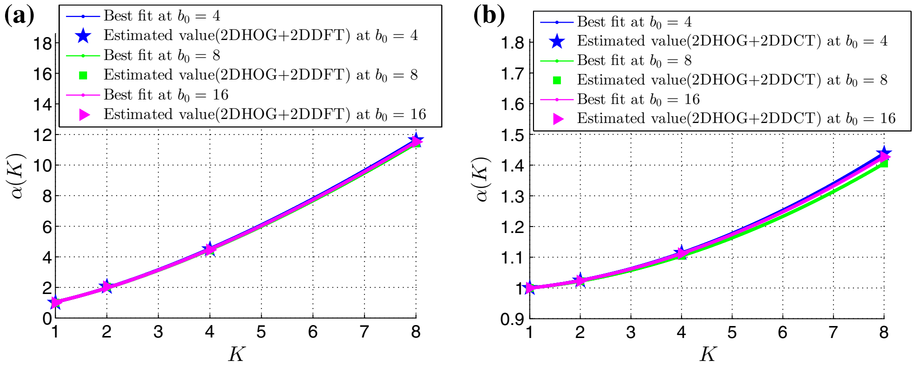

Fig. 8 The multiplicative factor $\alpha(K)$ for $K = 1, 2, 4, 8$, where (**a**) and (**b**) represent the case of the 2DHOG features in the 2DDFT and 2DDCT domains, respectively (Color figure online)

Table 1 summarizes the values of the parameters, a_0, a_1, and λ, for the above two cases for block size $b_0 = 4, 8, 16$ along with the corresponding mean square errors, when the number of layers, β, is 5, 7, or 9. It is seen from this table that irrespective of the transform used, the errors are insignificant. Figure 8 shows the plots of $\alpha(K)$ for the 2DHOG features for $\beta = 7$. It is seen from these plots that the proposed model is not sensitive to the block size b_0. It has been observed that $\alpha(K)$ is insensitive to b_0 for the other values of β also.

Similar studies have been conducted using $N_t = 1000$ positive training images from the USC multi-view car detection dataset, and $N_t = 1000$ positive training images, collected as mentioned earlier in this section, for the LISA 2010 dataset. It has been found that for both these datasets, $\alpha(K)$ is insensitive to b_0 irrespective of whether $\beta = 5, 7$ or 9.

It is to be noted that had we used the same model for $\alpha(K)$ as given by (30) also for the case of grayscale (GS) channel in the 2DDFT and 2DDCT domains and repeated the above procedures, we would obtain the values of a_0, a_1 and λ. These values for the 2DDFT and 2DDCT domains are also included in Table 1 using the UIUC car detection dataset. It is seen from this table that for the case of the grayscale channel, $\lambda \approx 0$, $a_0 \approx 1$ and $a_1 \approx 0$, and thus,

$$\alpha(K) \approx \begin{cases} K^2, & \text{for 2DDFT} \\ K, & \text{for 2DDCT} \end{cases} \tag{35}$$

Equation (35) has been found to be equally true in the case of the other two datasets, namely, the USC multi-view car detection dataset and the LISA 2010 dataset. It is seen that the two expressions on the right side of (35) are the same as that given by (10) and (21), respectively, when $K_1 = K_2 = K$. Thus, the proposed model for $\alpha(K)$ given by (30) for the TD2DHOG features is also valid for the grayscale images in the transform domain. These results show the versatility of the model for $\alpha(K)$ in representing channels other than the 2DHOG channel.

6.2 Vehicle detection using TD2DHOG features

In this section, we study the detection performance of the proposed scheme using the datasets mentioned earlier. Further, the detection performance of the proposed technique is compared with that of several state-of-the-art techniques. The 2DHOG is obtained assuming $\eta_1 = \eta_2 =$ 4 from which the TD2DHOG features are obtained. In case of using a single classifier, the

Table 1 The estimated channel parameters for grayscale image (GS) and 2DHOG features, where $b_0 = 4$, 8, or 16, and MSE refers to the mean square error of the curve fitting

		GS		2DHOG					
		2DDFT	2DDCT	2DDFT			2DDCT		
				$\beta = 5$	$\beta = 7$	$\beta = 9$	$\beta = 5$	$\beta = 7$	$\beta = 9$
$b_0 = 4$	λ	0.00635	− 0.00436	0.51538	0.53305	0.54992	− 0.79613	− 0.85311	− 0.87422
	a_0	1.00846	0.99210	0.51819	0.52523	0.53464	0.01179	0.00950	0.00834
	a_1	− 0.01189	0.00850	0.52819	0.52095	0.51050	0.98753	0.99027	0.99170
MSE		0.00001	0.00000	0.00251	0.00252	0.00243	0.00000	0.00000	0.00000
$b_0 = 8$	λ	0.00060	− 0.00085	0.51906	0.53351	0.54607	− 0.81072	− 0.87074	− 0.89513
	a_0	1.00055	0.99831	0.51119	0.51558	0.52167	0.01048	0.00846	0.00751
	a_1	− 0.00067	0.00183	0.53831	0.53411	0.52742	0.98901	0.99134	0.99245
MSE		0.00000	0.00000	0.00286	0.00291	0.00286	0.00000	0.00000	0.00000
$b_0 = 16$	λ	0.00036	0.00011	0.52324	0.53168	0.53758	− 0.79483	− 0.83676	− 0.85726
	a_0	1.00043	1.00014	0.51639	0.51853	0.52071	0.01107	0.00959	0.00883
	a_1	− 0.00057	− 0.00014	0.53153	0.52958	0.52731	0.98824	0.98991	0.99077
MSE		0.00000	0.00000	0.00269	0.00273	0.00273	0.00000	0.00000	0.00000

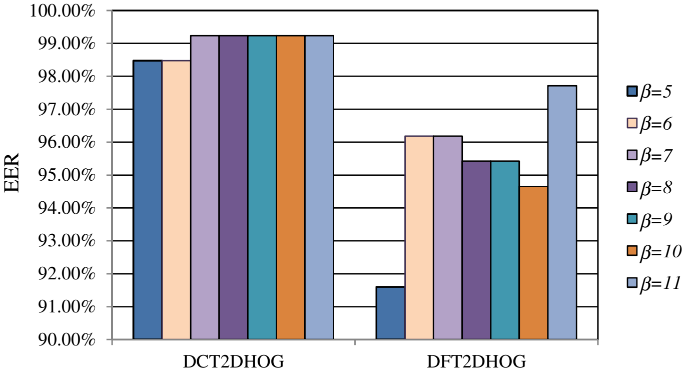

Fig. 9 Comparing the EER values of the DFT2DHOG-SC and DCT2DHOG-SC on UIUC dataset (Color figure online)

TD2DHOG features multiplied by the factor $\alpha(K)$ given by (30) are used, where the classifier is trained on TD2DHOG features obtained from training images upsampled by a factor R and used to classify images in the detection windows of the same or lower resolutions. We refer to this scheme using a single classifier (SC) as TD2DHOG-SC. Also, we consider the case of using multiple classifiers trained on TD2DHOG features at different values of R in order to classify images in the detection windows at the same resolution at which the classifier has been trained. We refer to this scheme using a classifier pyramid (CP) as TD2DHOG-CP. Unless specified otherwise, each octave of an image pyramid is considered to have 12 scales. Each scale is scanned by shifting the detection window(s) by $8R$ pixels in each of the x and y directions.

6.2.1 UIUC car detection dataset

On this dataset the equal error rate (EER) is used for evaluation, EER being the detection rate at the point of equal precision and recall; we use the methodology given in Agarwal et al. (2004) to calculate the precision and recall.

Choice of the transform: In this experiment, we evaluate the detection performance of the proposed TD2DHOG-SC by using 2DDFT or 2DDCT. The TD2DHOG features are obtained assuming $b^{train} = Rb_0$, $R = 2$, $b_0 = 4$, $c = 4$, $b^{test} = 4, 8$ and $\beta = 5, 6, \ldots, 11$. Figure 9 shows that DCT2DHOG-SC exhibits a better performance irrespective of β. Similar results have been obtained for other datasets, but are not included here in view of space constraints. In view of this, we will henceforth consider only DCT2DHOG features in all the experiments.

Choice of b_0, c, and β: We now study the performance of the proposed DCT2DHOG-SC for different values of b_0, c and β, in order to make an appropriate choice for these parameters. Figure 10 shows the EER values of the proposed DCT2DHOG-SC for $b_0 = 4, c = 2$ or 4; $b_0 = 8, c = 2, 4$ or 8 with $\beta = 5, 6, \ldots, 11$ and $b^{test} = b_0$ and $2b_0$. It is observed from this figure that the highest EER value is achieved at three different parameter settings: ($b_0 = 4, c = 4, \beta = 7$), ($b_0 = 4, c = 4, \beta = 9$), and ($b_0 = 8, c = 8, \beta = 7$). We choose the parameter setting $b_0 = 4, c = 4, \beta = 7$, since it retains the lowest number of eigenvectors compared to that of the other two parameter settings and thus it offers the lowest detection complexity. It has also been observed that in the case of DCT2DHOG-CP, the parameter setting $b_0 = 4, c = 4$ and $\beta = 7$ also provides the best EER value.

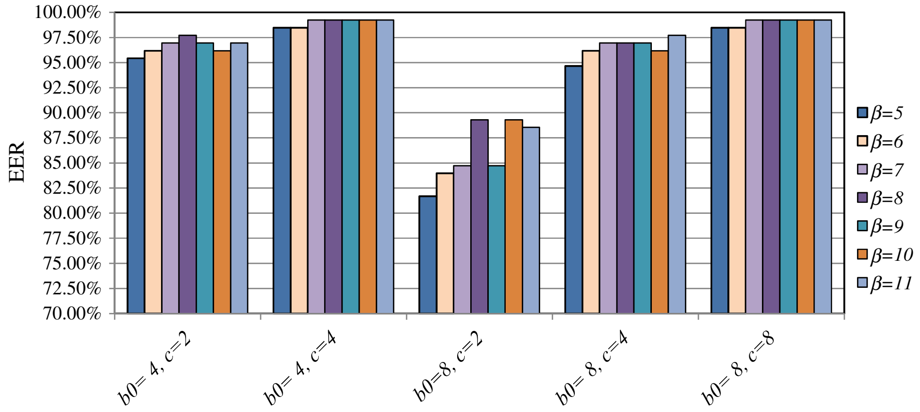

Fig. 10 EER value of the proposed scheme DCT2DHOG-SC at $c = 2, 4$ or 8 obtained on the UIUC dataset, where $\beta = 5, 6, \ldots,$ or 11, and the base block size $b_0 = 4$ or 8 (Color figure online)

Table 2 Equal Error Rate on UIUC car detection dataset

Method	EER
DCT2DHOG-SC($b_0 = 4, c = 4, \beta = 7$)	**99.28%**
DCT2DHOG-CP($b_0 = 4, c = 4, \beta = 7$)	**99.28%**
Lampert et al. (2008)	98.60%
Gall and Lempitsky (2009)	98.60%
Ohn-Bar and Trivedi (2015)*	98.56%
Wu et al. (2013)	97.80%
Dollár et al. (2014) (ACF-Exact)*	97.12%
Dollár et al. (2014) (ACF)*	95.68%
Maji et al. (2013)*	95.68%
Kuo and Nevatia (2009)	95.00%
Leibe et al. (2008)	95.00%
Mutch and Lowe (2008)	90.60%

*Denotes the results obtained by utilizing the code provided by the authors of the paper. The best and the second best results are shown in boldface and underscored, respectively

Performance evaluation: We first consider the case of the DCT2DHOG-SC scheme. In this case, the single classifier trained at $R = 2$ is used to classify the test images in detection windows with the same or lower resolutions (by making use of $\alpha(K)$, which is obtained using Table 1 and (30b)), where the test block sizes used are $b^{test} = 8$ and 4.

Now, we consider the case of DCT2DHOG-CP. In this case, we construct a classifier pyramid trained at $R = 1$ and 2. These two classifiers are used to classify the test images in detection windows of the corresponding two resolutions, where $b^{test} = 4$ and 8, respectively.

For each of the above cases, EER values are computed and are given in Table 2. The EER values corresponding to several state-of-the-art schemes, namely, the Gabor filter-based technique (Mutch and Lowe 2008), implicit shape model (Leibe et al. 2008), bag of words with spatial pyramid kernel (Lampert et al. 2008), discriminative parts with Hough forest (Gall and Lempitsky 2009), contour cue-based technique (Wu et al. 2013), HOG-based technique of Kuo and Nevatia (2009), aggregated channel feature (ACF) and ACF-Exact (Dollár et al.

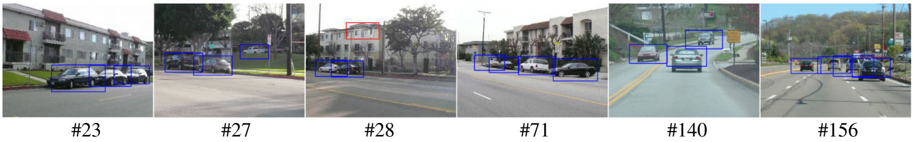

Fig. 11 Sample results for the proposed scheme when applied on USC multi-view car dataset, where colors represent: (blue) true positive, and (red) false positive (Color figure online)

2014), multi-resolution 2DHOG (Maji et al. 2013), and clustering appearance patterns based technique (Ohn-Bar and Trivedi 2015), are also included in Table 2. It is seen from this table that the performance of either of the two proposed schemes is better than that of the others in Dollár et al. (2014), Gall and Lempitsky (2009), Kuo and Nevatia (2009), Lampert et al. (2008), Leibe et al. (2008), Maji et al. (2013), Mutch and Lowe (2008), Ohn-Bar and Trivedi (2015) and Wu et al. (2013).

6.2.2 USC multi-view car detection dataset

For this dataset, as in Kuo and Nevatia (2009), the PASCAL visual object classes (VOC) criterion (Everingham et al. 2010, 2016) is used for the evaluation purpose with an overlap threshold of 0.5. To compare the performance of our method to that of some recent schemes, the average precision (AP) is used as an evaluation metric. In this dataset, the training images are upsampled by a factor of $R = 1$ and 2 in the case of using DCT2DHOG-CP, and by a factor of $R = 2$ in the case of using DCT2DHOG-SC. The performance of the proposed DCT2DHOG-SC scheme using this dataset is studied for $b_0 = 4$, $c = 4$; $b_0 = 8$, $c = 4$ or 8; and $\beta = 5, 7$ or 9, and $b^{test} = b_0$ and $2b_0$. It is observed that the highest AP value is achieved at two parameter settings, $b_0 = 8$, $c = 4$, $\beta = 9$ and $b_0 = 8$, $c = 8$, $\beta = 9$. We choose the parameter setting $b_0 = 8$, $c = 4$ and $\beta = 9$, since it retains a lower number of 2DDCT coefficients than that of the other parameter setting, and thus it provides a lower detection complexity. Therefore, this parameter setting is chosen for both the DCT2DHOG-SC and DCT2DHOG-CP schemes.

Figure 11 shows sample qualitative results for the proposed scheme on this dataset. It shows that the proposed scheme can detect cars in different views and resolutions. Table 3 shows that the performance of the proposed technique is better than that of the method in Ohn-Bar and Trivedi (2015) which is based on ACF and training multiple classifiers at different resolutions, the method in Kuo and Nevatia (2009) which uses HOG with Gentle AdaBoost, and that of the method in Wu and Nevatia (2007) which is based on using Edgelet feature with cluster boosted tree classifier, where the latter is evaluated using Kuo and Nevatia (2009). Further, the performance of the proposed method is slightly better than that of the implementations of the methods in Dollár et al. (2014), or that of the multi-resolution 2DHOG features presented in Maji et al. (2013) when used with BDTC. The proposed scheme achieves AP values of 90.44% in the case of DCT2DHOG-SC, and 89.92% in the case of DCT2DHOG-CP. Thus, DCT2DHOG with a single classifier exhibits a high detection performance, while requiring the training of only a single classifier, instead of multiple classifiers for each resolution.

6.2.3 LISA 2010

In this dataset, the same evaluation metrics presented in Sivaraman and Trivedi (2010) are used, namely, true positive rate (TPR) or recall, false detection rate (FDR) or (1-precision),

Table 3 Average Precision on USC Multi-view Car Dataset

Method	AP
DCT2DHOG-SC($b_0 = 8, c = 4, \beta = 9$)	**90.44%**
DCT2DHOG-CP($b_0 = 8, c = 4, \beta = 9$)	89.92%
Ohn-Bar and Trivedi (2015)*	86.71%
ACF-Exact (Dollár et al. 2014)*	89.31%
ACF (Dollár et al. 2014)*	89.64%
Multi-resolution 2DHOG (Maji et al. 2013)—BDTC*	89.38%
Kuo and Nevatia (2009)	85.61%
Wu and Nevatia (2007)	52.55%

∗Denotes the results obtained by utilizing the code provided by the authors of the paper. The best and the second best results are shown in boldface and underscored, respectively

average false positive per frame (AFP/F), average false positive per object (AFP/O), and average true positive per frame (ATP/F). These metrics are computed at the point of equal precision and recall. True positive detections are computed by using the PASCAL VOC criterion (Everingham et al. 2010, 2016) with an overlap threshold of 0.5.

On both the H-dense and H-medium sequences, the single classifier trained at $R = 2$ is used in the case of DCT2DHOG-SC and two classifiers trained at $R = 1$ and 2 are used in the case of DCT2DHOG-CP. As in our experiments on USC multi-view car detection dataset, the parameter setting chosen for both the DCT2DHOG-SC and DCT2DHOG-CP schemes on LISA 2010 dataset is $b_0 = 8$, $c = 4$, $\beta = 9$, and $b^{test} = 8$ and 16, since, these two datasets contain similar environmental conditions and the same type of classifier, namely, BDTC, is used in the detection process.

Table 4 gives the detection performance of the proposed method, from which it is clear that the performance of DCT2DHOG using a single classifier is almost as good as that of using classifier pyramid. Table 4 also lists the performance of some of the other methods, namely, the Haar-like features-based technique (Sivaraman and Trivedi 2010), ACF and ACF-Exact (Dollár et al. 2014), multi-resolution 2DHOG (Maji et al. 2013), and clustering appearance patterns based technique (Ohn-Bar and Trivedi 2015). From this table, it can be seen that the proposed scheme on H-medium sequence provides a performance better than that of the schemes of Dollár et al. (2014), Maji et al. (2013), Ohn-Bar and Trivedi (2015) and Sivaraman and Trivedi (2010), while for the H-dense sequence, our scheme provides 92.67% TPR at 6.03% FDR, which is better than that of the methods in Dollár et al. (2014), Maji et al. (2013) and Ohn-Bar and Trivedi (2015). The proposed method and the methods in Dollár et al. (2014), Maji et al. (2013) and Ohn-Bar and Trivedi (2015) are trained with hard negative samples collected from the *CBCL street scenes dataset* (Bileschi 2006), while the method in Sivaraman and Trivedi (2010) is trained on private data from sunny highway environment. The detection performance of the proposed scheme can be improved by using an online learning technique to incorporate the false positive samples in the learning process. Figure 12a shows sample qualitative results for the proposed scheme when applied on the H-dense sequence. As mentioned earlier, this sequence contains heavy shadows, vehicles in partial occlusions and some background structures are confused with the positive class. The proposed scheme can detect correctly 92.67% from the vehicles under these challenging conditions. Figure 12b shows the corresponding results for the H-medium sequence, which includes challenges similar to that of the H-dense sequence but at a lower density. It is

Table 4 The performance for the proposed scheme on LISA dataset

	Method	TPR	FDR	AFP/F	ATP/F	AFP/O
H-dense	DCT2DHOG-SC	92.67%	**6.03%**	**0.26**	4.06	**0.06**
	DCT2DHOG-CP	92.67%	**6.03%**	**0.26**	4.06	**0.06**
	Sivaraman and Trivedi (2010)	**93.50%**	7.10%	0.32	4.20	0.07
	Ohn-Bar and Trivedi (2015)*	87.54%	12.46%	0.61	**4.28**	0.12
	ACF-Exact (Dollár et al. 2014)*	87.43%	12.54%	0.55	3.83	0.13
	ACF (Dollár et al. 2014)*	86.75%	13.23%	0.58	3.80	0.13
	Multi-resolution 2DHOG (Maji et al. 2013)*	73.24%	26.76%	1.17	3.21	0.27
H-medium	DCT2DHOG-SC	98.11%	1.89%	0.06	2.94	**0.02**
	DCT2DHOG-CP	98.22%	**1.78%**	**0.05**	2.95	**0.02**
	Sivaraman and Trivedi (2010)	**98.80%**	10.30%	0.37	**3.18**	0.11
	Ohn-Bar and Trivedi (2015)*	96.11%	3.89%	0.12	2.88	0.04
	ACF-Exact (Dollár et al. 2014)*	93.11%	6.89%	0.21	2.79	0.07
	ACF (Dollár et al. 2014)*	94.33%	5.67%	0.17	2.83	0.06
	Multi-resolution 2DHOG (Maji et al. 2013)*	77.44%	19.70%	0.57	2.32	0.19

∗Denotes the results obtained by utilizing the code provided by the authors of the paper. The best and the second best results on each dataset are shown in boldface and underscored, respectively

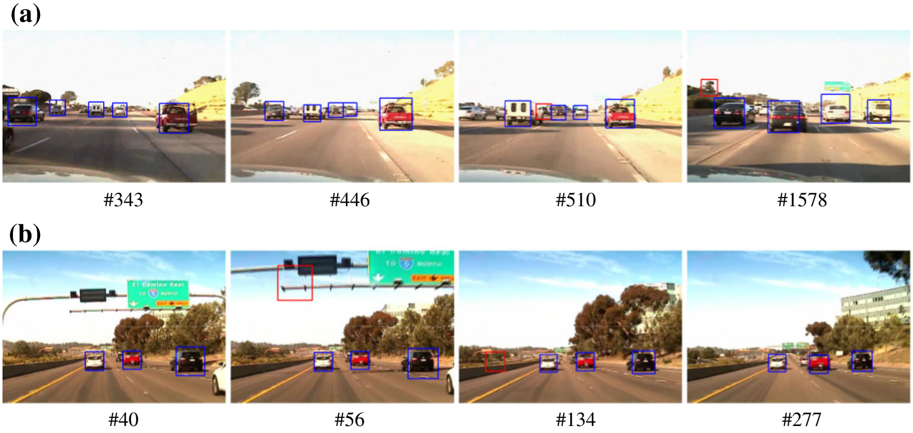

Fig. 12 Sample qualitative results for the proposed method on LISA 2010 dataset, such that (**a**) Highway-dense sequence, (**b**) Highway-medium or sunny sequence: (blue) true positive, and (red) false positive (Color figure online)

clear that the proposed technique can detect vehicles of various resolutions, under different illumination and background conditions.

6.2.4 HRI roadway dataset

For this dataset, the evaluation metrics presented in Sect. 6.2.3 are used. As in our experiments on the USC multi-view car detection dataset and LISA 2010 dataset, the single classifier trained at $R = 2$ is used in the case of DCT2DHOG-SC and two classifiers trained at $R = 1$ and 2 are used in the case of DCT2DHOG-CP for all the five test sequences of the HRI dataset.

Also, the same parameter setting is chosen for both the DCT2DHOG-SC and DCT2DHOG-CP schemes, namely, $b_0 = 8$; $c = 4$; $\beta = 9$, and $b_{test} = 8$ and 16. The choice of these parameters is made since these three datasets contain similar challenging conditions and the type of the classifier used is the same, namely, BDTC.

Table 5 shows the detection performance of DCT2DHOG-SC, DCT2DHOG-CP, and other state-of-the-art techniques, namely, the ACF and ACF-Exact (Dollár et al. 2014), multi-resolution 2DHOG (Maji et al. 2013), and clustering appearance patterns based method (Ohn-Bar and Trivedi 2015). From this table, it can be seen that for sequences I, II and IV either of the DCT2DHOG-SC and DCT2DHOG-CP schemes provides TPR values better than that in case of the schemes in Dollár et al. (2014), Maji et al. (2013) and Ohn-Bar and Trivedi (2015), whereas for the sequences III and V, the DCT2DHOG-SC scheme yields TPR values higher than that in case of DCT2DHOG-CP or when the schemes of Dollár et al. (2014) and (Maji et al. 2013) are used. In the work of Dollár et al. (2014) and Ohn-Bar and Trivedi (2015), the feature approximation is carried out in the spatial domain to handle the problem of variation in scale. However, in the proposed scheme, the problem of scale variation is addressed by carrying the feature approximation in the frequency domain rather than in the spatial domain.

6.2.5 Discussion

In this section, we present an evaluation of the proposed scheme in terms of the cost for the training and testing schemes. For a fair comparison, we use 2DPCA and FIKSVM or 2DPCA and BDTC as the main building blocks when 2DHOG or DCT2DHOG features are used. In the experiments that follow, the same values of η_1, η_2, b_0, c, and β that have been used to obtain the detection accuracy on the corresponding dataset are used. It should be noted that in practical situations, the choice of these parameters depends on the targeted vehicle view. In case the side view of the vehicles is of interest, the parameter settings recommended for obtaining DCT2DHOG features are $b_0 = 4$, $c = 4$, and $\beta = 7$ and FIKSVM can provide a fast and accurate classification scheme. In the case of detecting vehicles with different views, such as the situations that exist in urban and highway scenarios, the recommended parameter settings are $b_0 = 8$, $c = 4$, and $\beta = 9$ and BDTC is preferred, since it can be trained on a large number of samples and can capture large intra-class variations that exist within the positive class samples.

Training cost: In this experiment, we compare the training cost of the proposed DCT2DHOG against that of 2DHOG at six different resolutions. Table 6 lists the overall training time[4] of the proposed DCT2DHOG at six resolutions along with that of 2DHOG. It is seen from this table that the training time for the proposed scheme is less than that of 2DHOG by at least 49.79% when a classifier pyramid is used, and by at least 74.33% when a single classifier trained at $R = 2$ is employed. Table 7 gives the storage requirement of the proposed scheme and that of the 2DHOG-based scheme for classifiers trained at the six different resolutions considered. It is seen from this table that the storage requirement for the proposed scheme is lower than that of 2DHOG-based scheme in case of the UIUC dataset by 64.18% when the size of the detection window is 64 × 192, whereas both these schemes achieve the same storage for the cases of USC and LISA 2010 datasets. Note that the FIKSVM classifier is used for the UIUC dataset and BDTC is used for the USC and LISA 2010 datasets. It is observed from Tables 6 and 7, in order to detect vehicles of different resolutions, the proposed DCT2DHOG-SC

[4] Using modern computer of 2.9GHz CPU, and 8G RAM.

Table 5 The performance for the proposed scheme on HRI dataset

	Method	TPR	FDR	AFP/F	ATP/F	AFP/O
Sequence I	DCT2DHOG-SC	**78.13%**	**21.88%**	**0.16**	0.56	**0.22**
	DCT2DHOG-CP	**78.13%**	**21.88%**	**0.16**	0.56	**0.22**
	Ohn-Bar and Trivedi (2015)*	75.00%	25.00%	0.20	0.60	0.25
	ACF-Exact (Dollár et al. 2014)*	68.29%	31.71%	0.48	**1.04**	0.32
	ACF (Dollár et al. 2014)*	66.67%	33.33%	0.52	**1.04**	0.33
	Multi-resolution 2DHOG (Maji et al. 2013) - BDTC*	68.97%	31.03%	0.36	0.80	0.31
Sequence II	DCT2DHOG-SC	**67.86%**	**32.14%**	**0.20**	0.42	**0.32**
	DCT2DHOG-CP	**67.86%**	**32.14%**	**0.20**	0.42	**0.32**
	Ohn-Bar and Trivedi (2015)*	61.54%	38.46%	0.33	0.53	0.38
	ACF-Exact (Dollár et al. 2014)*	65.63%	34.38%	0.39	**0.75**	0.34
	ACF (Dollár et al. 2014)*	60.61%	39.39%	0.45	0.69	0.39
	Multi-resolution 2DHOG (Maji et al. 2013) - BDTC*	53.85%	46.15%	0.29	0.34	0.46
Sequence III	DCT2DHOG-SC	**72.73%**	27.27%	0.30	0.80	0.27
	DCT2DHOG-CP	66.67%	33.33%	0.37	0.73	0.33
	Ohn-Bar and Trivedi (2015)*	66.67%	33.33%	0.38	0.77	0.33
	ACF-Exact (Dollár et al. 2014)*	66.67%	20.00%	**0.29**	1.18	**0.17**
	ACF (Dollár et al. 2014)*	72.41%	**19.23%**	0.31	**1.31**	**0.17**
	Multi-resolution 2DHOG (Maji et al. 2013) - BDTC*	45.45%	54.55%	0.60	0.50	0.55
Sequence IV	DCT2DHOG-SC	73.33%	26.67%	0.20	0.55	0.27
	DCT2DHOG-CP	**80.00%**	**20.00%**	**0.15**	0.60	**0.20**
	Ohn-Bar and Trivedi (2015)*	65.63%	34.38%	0.55	**1.05**	0.34
	ACF-Exact (Dollár et al. 2014)*	63.16%	36.84%	0.50	0.86	0.37
	ACF (Dollár et al. 2014)*	63.16%	36.84%	0.50	0.86	0.37
	Multi-resolution 2DHOG (Maji et al. 2013) - BDTC*	73.33%	26.67%	0.20	0.55	0.27
Sequence V	DCT2DHOG-SC	66.67%	33.33%	0.22	0.44	0.33
	DCT2DHOG-CP	62.16%	37.84%	**0.02**	0.03	0.38
	Ohn-Bar and Trivedi (2015)*	**70.83%**	29.17%	0.26	0.63	0.29
	ACF-Exact (Dollár et al. 2014)*	64.00%	**23.81%**	0.36	**1.14**	0.20
	ACF (Dollár et al. 2014)*	61.54%	**23.81%**	0.33	1.07	**0.19**
	Multi-resolution 2DHOG (Maji et al. 2013) - BDTC*	51.85%	48.15%	0.32	0.34	0.48

∗Denotes the results obtained by utilizing the code provided by the authors of the paper. The best and the second best results on each dataset are shown in boldface and underscored, respectively

requires only a single classifier instead of multiple ones, resulting in a reduction in terms of the training cost by at least 44.63% and the storage requirement by at least 50.00% compared with that of DCT2DHOG-CP.

It is to be pointed out that the reduction in the training and storage costs is achieved by the proposed vehicle detector in comparison with that of the 2DHOG counterpart using a classifier pyramid with almost no loss in the detection accuracy.

Table 6 Feature extraction and classifier training times (in seconds) for the proposed DCT2DHOG method and for the 2DHOG method

Dataset		UIUC		USC		LISA 2010	
$M_1 \times M_2$		32 × 96	64 × 192	64 × 128	128 × 256	64 × 64	128 × 128
DCT2DHOG	FET	**8.00**	**9.72**	**245.87**	**283.91**	**85.03**	**107.22**
	CTT	**6.75**	**5.71**	**14.32**	**13.49**	**7.63**	**7.76**
	TT	**14.75**	**15.43**	**260.19**	**297.40**	**92.66**	**114.98**
2DHOG	FET	8.53	11.76	604.70	2133.36	291.74	806.56
	CTT	7.36	32.46	54.14	170.76	52.01	141.11
	TT	15.89	44.22	658.84	2304.11	343.74	947.67
Reduction in TT (CP)		49.79%		81.18%		83.92%	
Reduction in TT (SC)		74.33%		89.96%		91.10%	

Boldface fonts denote the best results

FET time in seconds for feature extraction, *CTT* time in seconds for training a classifier, *TT* average training time in seconds, Reduction in TT (CP) and (SC) refer to the amount of reduction in TT of DCT2DHOG-CP method over 2DHOG method, and DCT2DHOG-SC method over 2DHOG method, respectively

Table 7 Storage requirements (in MByte) for the proposed DCT2DHOG method and for the 2DHOG methods

Dataset	UIUC		USC		LISA 2010	
$M_1 \times M_2$	32 × 96	64 × 192	64 × 128	128 × 256	64 × 64	128 × 128
DCT2DHOG	1.51	2.16	0.21	0.21	0.21	0.21
2DHOG	1.51	6.03	0.21	0.21	0.21	0.21
Reduction in storage (CP)	51.33%		0.00%		0.00%	
Reduction in storage (SC)	71.35%		50.00%		50.00%	

Reduction in storage (CP) and (SC) refer to the amount of reduction in storage of DCT2DHOG-CP method over 2DHOG method, and DCT2DHOG-SC method over 2DHOG method, respectively

Table 8 Average feature extraction and detection time in seconds for Methods A, B and C applied to three datasets

Dataset		UIUC	USC	LISA 2010
Range of vehicle size		32 × 96 to 128 × 384	64 × 128 to 256 × 512	64 × 64 to 256 × 256
Number of detection windows per frame		1398	1141	1365
Method A	FET	**0.061**	**0.077**	**0.059**
	DT	**0.143**	**0.212**	**0.218**
Method B	FET	0.064	0.112	0.122
	DT	0.174	0.397	0.475
Method C	FET	0.073	0.130	0.137
	DT	0.301	0.376	0.375
Min. reduction in FET		4.69%	31.25%	51.64%
Min. reduction in DT		17.82%	43.62%	41.87%

Boldface fonts denote the best results

FET feature extraction time in second, *DT* detection time in second, Min. reduction in FET and DT refer to the minimum amount of reduction in FET and DT of Method A over those of Methods B and C

Detection time: Table 8 gives a comparison of the feature extraction time as well as the detection time (in seconds) of the proposed transform-domain based detector (Method A) with that of the spatial-domain counterparts (Methods B and C) on the three vehicle detection datasets, UIUC (Agarwal et al. 2004), USC (Kuo and Nevatia 2009) and LISA 2010 (Sivaraman and Trivedi 2010). We use test images of size 480 × 640. We assume that each octave of an image pyramid consists of 8 scales, and that each scale is scanned by shifting the detection window(s) by 16 pixels in each of the x and y directions. This generates 1398, 1141 and 1365 detection windows per frame for UIUC, USC and LISA 2010 datasets, respectively.

Method A in Table 8 corresponds to the proposed method, where the DCT2DHOG-2DPCA features are used to train a single classifier at $R = 2$. Further, two detection windows of different sizes are used to scan an image pyramid of depth one octave and the same classifier is used to classify DCT2DHOG-2DPCA features obtained from images within these detection windows after incorporating the multiplicative factor $\alpha(K)$ given by (30b).

Method B corresponds to the traditional method that uses a single classifier trained on features obtained in the spatial domain, namely, 2DHOG-2DPCA features, at $R = 1$. Further, it uses a single detection window to scan an image pyramid of depth two octaves. Then, the 2DHOG-2DPCA features obtained from an image within a detection window are classified by the trained classifier.

Method C corresponds to a spatial domain method which uses 2DHOG-2DPCA features to train two classifiers at $R = 1$, and 2. Further, two detection windows of different sizes are used to scan an image pyramid of depth one octave. Then, the two classifiers trained at $R = 1$ and 2 are used to classify images within the detection windows of the same resolution at which the classifier is trained.

For the UIUC dataset, the first detection window is of size 32 × 96 and the second one of size 64 × 192. For this dataset, the range of vehicle size that can be detected by using the method A, B or C is 32 × 96 to 128 × 384. For USC and LISA 2010 datasets the corresponding window sizes are 64 × 128 and 128 × 256, and 64 × 64 and 128 × 128, respectively.

It is seen from Table 8 that the proposed transform-based method provides a minimum of 4.69% reduction in the feature extraction time and a minimum of 17.82% reduction in the detection time over that of the two spatial-domain methods B and C for the UIUC dataset and very much higher reductions for the other two datasets.

Finally, it is worth mentioning that the classification time of the proposed method represents on average about 65% of the total detection time. Thus, further gains in the detection speed could be achieved by reducing the classification time.

7 Conclusion

In this paper, we have introduced transform domain features of two-dimensional histogram of oriented gradients of images, referred to as TD2DHOG features. Then, we have studied the effect of image downsampling on the TD2DHOG features. It has been shown that the TD2DHOG features obtained from a high resolution image can be approximated by using the TD2DHOG features obtained from the image at a lower resolution by multiplying the latter by a factor that depends on the downsampling factor. A model for this multiplicative factor has been proposed and validated experimentally in the case of 2DDFT and 2DDCT domains. Next, a novel vehicle detection scheme using these TD2DHOG features has been proposed. It has been shown that the use of TD2DHOG features reduce the cost of training a classifier pyramid, since a single classifier can be used to detect vehicles of the same or lower resolution at which the classifier has been trained, instead of training multiple resolution-specific classifiers.

Experimental results have shown that when the proposed TD2DHOG features are used with the multiplying factor and a single classifier for vehicle detection, it provides a detection accuracy similar to that obtained using these features with a classifier pyramid; however, the use of a single classifier has a significant advantage over the use of a classifier pyramid in that the former results in substantial savings in training and storage costs. In addition, the proposed method provides a detection accuracy that is similar or even better than that provided by the state-of-the-art techniques.

Acknowledgements M. A. Naiel would like to acknowledge the support from Concordia University to conduct this research. This work is supported by research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Regroupement Stratégeique en Microsystémes du Québec (ReSMiQ) awarded to M. O. Ahmad and M. N. S. Swamy.

Appendix: Derivation of Equation (21)

The 2DDCT for a grayscale image in the spatial domain, $x \in \mathbb{R}^2$, is given by

$$X_{N,M}[u,v] = \hat{\Gamma}_N[u]\hat{\Gamma}_M[v] \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} x[n,m] \cos\left(\frac{\pi(2n+1)u}{2N}\right) \times \cos\left(\frac{\pi(2m+1)v}{2M}\right) \tag{1}$$

where $0 \le u \le N-1$, $0 \le v \le M-1$, $\hat{\Gamma}_N[k] = \sqrt{1/N}$ for $k=0$ and $\hat{\Gamma}_N[k] = \sqrt{2/N}$ for $0 < k \le N-1$. Let N and M be even multiples of K_1, and K_2, respectively, where K_1 and K_2 are the downsampling factors in the y and the x directions, respectively. Let $x[n,m]$ be a bandlimited sequence, and the sequence $y \in \mathbb{R}^2$, of size $(2N \times 2M)$, be defined as

$$y[n,m] = \begin{cases} x[n,m], & 0 \le n \le N-1, 0 \le m \le M-1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The N × M-point 2DDCT can be computed by 2N × 2M-point 2DDFT for a sequence, $y[n,m]$, as follows. First, the 2DDFT is employed on $y[n,m]$ in order to obtain $Y_{2N,2M}$. Similar to the 1DDCT case, the relation between the signal in the 2DDCT domain $X_{N,M}[u,v]$, and $Y_{2N,2M}[u,v]$ can be expressed as

$$X_{N,M}[u,v] = \hat{\Gamma}_N[u]\hat{\Gamma}_M[v]\mathrm{Re}\left(Y_{2N,2M}[u,v]e^{-j\left(\frac{\pi u}{2N}+\frac{\pi v}{2M}\right)}\right) \tag{3}$$

where $0 \le u \le N-1$, $0 \le v \le M-1$. Let c_1, c_2 denote the maximum frequencies retained by the truncation operator, where $c_1 < \hat{N}$, $c_2 < \hat{M}$, $\hat{N} = N/K_1$, and $\hat{M} = M/K_2$. Assume $Y_{2N,2M}$ is bandlimited to the maximum frequencies $(\hat{N}, \hat{M})$. Then, the downsampled signal in the 2DDCT domain, $\hat{X}_{\hat{N},\hat{M}}$, can be obtained as

$$\hat{X}_{\hat{N},\hat{M}}[u,v] = \frac{1}{K_1K_2}\hat{\Gamma}_{\hat{N}}[u]\hat{\Gamma}_{\hat{M}}[v]\mathrm{Re}\left(Y_{2N,2M}[u,v]e^{-j\left(\frac{\pi u}{2N}+\frac{\pi v}{2M}\right)}\right) \tag{4}$$

$$= \frac{\hat{\Gamma}_{\hat{N}}[u]\hat{\Gamma}_{\hat{M}}[v]}{K_1K_2\hat{\Gamma}_N[u]\hat{\Gamma}_M[v]}\hat{\Gamma}_N[u]\hat{\Gamma}_M[v]\mathrm{Re}(Y_{2N,2M}[u,v] \times e^{-j\left(\frac{\pi u}{2N}+\frac{\pi v}{2M}\right)}) \tag{5}$$

$$= \frac{\sqrt{1/\hat{N}}\sqrt{1/\hat{M}}}{K_1 K_2 \sqrt{1/N}\sqrt{1/M}} X_{N,M}[u, v] \tag{6}$$

$$= \frac{1}{\sqrt{K_1 K_2}} X_{N,M}[u, v] \tag{7}$$

where $0 \leq u \leq c_1 - 1$ and $0 \leq v \leq c_2 - 1$. Thus, the relation between the 2DDCT coefficients of the original image and that of the downsampled version is given by

$$X_{N,M}[u, v] = \sqrt{K_1 K_2} \hat{X}_{\hat{N},\hat{M}}[u, v] \tag{8}$$

References

Agarwal, S., Awan, A., & Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 26*(11), 1475–1490. http://cogcomp.org/page/resource_view/13/. Accessed November 1, 2018.

Ahmed, N., Natarajan, T., & Rao, K. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, *C–23*(1), 90–93.

Appel, R., Fuchs, T., Dollár, P., & Perona, P. (2013). Quickly boosting decision trees—Pruning underachieving features early. In *Proceedings of the international conference on machine learning (ICML)* (pp. 594–602).

Benenson, R., Mathias, M., Timofte, R., & Gool, L. V. (2012). Pedestrian detection at 100 frames per second. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2903–2910).

Bi, G., & Mitra, S. (2011). Sampling rate conversion in the frequency domain [dsp tips and tricks]. *IEEE Signal Processing Magazine*, *28*(3), 140–144.

Bileschi, S. (2006). *StreetScenes: Towards scene understanding in still images*. Ph.D. thesis, Massachusetts Institute of Technology. CBCL dataset link: http://cbcl.mit.edu/software-datasets/streetscenes. Accessed November 1, 2018.

Buch, N., Velastin, S. A., & Orwell, J. (2011). A review of computer vision techniques for the analysis of urban traffic. *IEEE Transactions on Intelligent Transportation Systems (ITS)*, *12*(3), 920–939.

Dalal, N. (2006). *Finding people in images and videos*. Ph.D. thesis, Institut National Polytechnique de Grenoble.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)* (vol. 1, pp. 886–893).

Dollár, P. (2016). Piotr's Image and Video Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html. Retrieved, November 1, 2018.

Dollár, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *36*(8), 1532–1545.

Dollár, P., Belongie, S., & Perona, P. (2010). The fastest pedestrian detector in the west. In *Proceedings of the British machine vision conference (BMVC)* (pp. 68.1–68.11).

Dollár, P., Tu, Z., Perona, P., & Belongie, S. (2009). Integral channel features. In *Proceedings of the British machine vision conference (BMVC)* (pp. 91.1–91.11).

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *34*(4), 743–761.

Dubout, C., & Fleuret, F. (2012). Exact acceleration of linear object detectors. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 301–311).

Everingham, M., Gool, L. V., Williams, C. K., Winn, J., & Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., & Zisserman, A. (2016). The Pascal visual object classes challenge 2007 (VOC2007) results. http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html. Retrieved, November 1, 2018.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *32*(9), 1627–1645.

Gall, J., & Lempitsky, V. (2009). Class-specific hough forests for object detection. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1022–1029).

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings conference on computer vision and pattern recognition (CVPR)* (pp. 3354–3361).

Gepperth, A., Rebhan, S., Hasler, S., & Fritsch, J. (2011). Biased competition in visual processing hierarchies: A learning approach using multiple cues. *Cognitive Computation*, *3*(1), 146–166.

Huang, J., & Mumford, D. (1999). Statistics of natural images and models. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)* (vol. 1, pp. 541–547).

Kuo, C. H., & Nevatia, R. (2009). Robust multi-view car detection using unsupervised sub-categorization. In *Proceedings of the IEEE workshop on applications of computer vision (WACV)* (pp. 1–8).

Lampert, C. H., Blaschko, M., & Hofmann, T. (2008). Beyond sliding windows: Object localization by efficient subwindow search. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).

Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, *77*(1), 259–289.

Li, B., Wu, T., & Zhu, S. C. (2014). Integrating context and occlusion for car detection by hierarchical and-or model. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 652–667).

Maji, S., Berg, A. C., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).

Maji, S., Berg, A. C., & Malik, J. (2013). Efficient classification for additive kernel SVMs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *35*(1), 66–77.

Mutch, J., & Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, *80*(1), 45–57.

Naiel, M. A., Ahmad, M. O., & Swamy, M. (2015). Vehicle detection using approximation of feature pyramids in the DFT domain. In *Proceedings of the international conference on image analysis and recognition. (ICIAR)* (pp. 429–436). Springer.

Naiel, M. A., Ahmad, M. O., & Swamy, M. N. S. (2014). Vehicle detection using TD2DHOG features. In *Proceedings of New circuits and systems conference (NewCAS)* (pp. 389–392).

Ohn-Bar, E., & Trivedi, M. M. (2015). Learning to detect vehicles by clustering appearance patterns. *IEEE Transactions on Intelligent Transportation Systems (ITS)*, *16*(5), 2511–2521.

Papageorgiou, C. P., Oren, M., & Poggio, T. (1998). A general framework for object detection. In *Proceedings of the sixth IEEE international conference on computer vision (ICCV)* (pp. 555–562).

Pepikj, B., Stark, M., Gehler, P., & Schiele, B. (2013). Occlusion patterns for object class detection. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3286–3293).

Ruderman, D. L. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, *5*, 517–548.

Sivaraman, S., & Trivedi, M. (2010). A general active-learning framework for on-road vehicle recognition and tracking. *The IEEE intelligent transportation systems (ITS)*, *11*(2), 267–276.

Sivaraman, S., & Trivedi, M. (2013a). Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *The IEEE intelligent transportation systems (ITS)*, *14*(4), 1773–1795.

Sivaraman, S., & Trivedi, M. (2013b). Vehicle detection by independent parts for urban driver assistance. *The IEEE intelligent transportation systems (ITS)*, *14*(4), 1597–1608.

Smith, J. O. (2007). *Mathematics of the discrete Fourier transform (DFT)* (2nd ed.). W3K Publishing.

Takeuchi, A., Mita, S., & McAllester, D. (2010). On-road vehicle tracking using deformable object model and particle filter with integrated likelihoods. In *Proceedings of the IEEE intelligent vehicles symposium (IV)* (pp. 1014–1021).

Wang, C., Fang, Y., Zhao, H., Guo, C., Mita, S., & Zha, H. (2016). Probabilistic inference for occluded and multiview on-road vehicle detection. *IEEE Transactions on Intelligent Transportation Systems (ITS)* *17*(1).

Wang, X., Han, T. X., & Yan, S. (2009). An HOG-LBP human detector with partial occlusion handling. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 32–39).

Wang, X., Yang, M., Zhu, S., & Lin, Y. (2015). Regionlets for generic object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *37*(10), 2071–2084.

Wu, B., & Nevatia, R. (2007). Cluster boosted tree classifier for multi-view, multi-pose object detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 1–8).

Wu, B. F., Kao, C. C., Jen, C. L., Li, Y. F., Chen, Y. H., & Juang, J. H. (2014). A relative-discriminative-histogram-of-oriented-gradients-based particle filter approach to vehicle occlusion handling and tracking. *IEEE Transactions on Industrial Electronics*, *61*, 4228–4237.

Wu, J., Liu, N., Geyer, C., & Rehg, J. (2013). C^4: A real-time object detection framework. *IEEE Transactions on Image Processing*, *22*(10), 4096–4107.

Xiang, Y., Choi, W., Lin, Y., & Savarese, S. (2015). Data-driven 3D voxel patterns for object category recognition. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1903–1911).

Yang, J., Zhang, D., Frangi, A. F., & Yang, J. Y. (2004). Two dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *26*(1), 131–137.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Mohamed A. Naiel received the B.Sc. degree in electrical engineering from Tanta University, Tanta, Egypt in June 2006, the M.Sc. degree in communication and information technology from Nile University, Giza, Egypt in June 2010, and the Ph.D. degree in electrical and computer engineering from Concordia University, Montreal, QC, Canada in May 2017. From 2011 to 2017, he was a research assistant with the center for signal processing and communications, Concordia University. He is currently a Postdoctoral Fellow at the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include image and video processing, computer vision, human action recognition, object detection and recognition, and multi-object tracking.

M. Omair Ahmad received the B.Eng. degree in electrical engineering from Sir George Williams University, Montreal, QC, Canada, and the Ph.D. degree in electrical engineering from Concordia University. From 1978 to 1979, he was a Faculty Member with New York University College, Buffalo, NY, USA. In 1979, he joined the faculty of Concordia University as an Assistant Professor of computer science. He joined subsequently the Department of Electrical and Computer Engineering, Concordia University, where he was the Chair with the Department from 2002 to 2005, and is currently a Professor. He is also the Concordia University Research Chair (Tier I) in multimedia signal processing. He was a Founding Researcher of Micronet, as a Canadian Network of Centers of Excellence, from 1990 to 2004. He has authored in the area of signal processing and holds four patents. His current research interests include the areas of image and speech processing, biomedical signal processing, watermarking, biometrics, video signal processing and object detection and tracking, deep learning techniques in signal processing, and fast signal transforms and algorithms. In 1988, he was a member of the Admission and Advancement Committee of the IEEE. He was a recipient of numerous honors and awards, including the Wighton Fellowship from the Sandford Fleming Foundation, an inductee to Provosts Circle of Distinction for Career Achievements, Guest Professor of Southeast University, Nainjing, China, and the Award of Excellence in Doctoral Supervision from the Faculty of Engineering and Computer Science, Concordia University. He was the Local Arrangements Chairman of the 1984 IEEE International Symposium on Circuits and Systems. He has served as the Program Co-Chair for the 1995 IEEE International Conference on Neural Networks and Signal Processing, the 2003 IEEE International Conference on Neural Networks and Signal Processing, and the 2004 IEEE International Midwest Symposium on Circuits and Systems. He was a General Co-Chair of the 2008 IEEE International Conference on Neural Networks and Signal Processing. He is the Chair of the Montreal Chapter IEEE Circuits and Systems Society. He was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART I: FUNDAMENTAL THEORY AND APPLICATIONS from 1999 to 2001. Dr. Ahmad is a Fellow of the Institute of Electrical and Electronics Engineers.

M. N. S. Swamy received the B.Sc. degree (Hons.) in mathematics from the University of Mysore, Mysore, India, in 1954, the Diploma degree in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1957, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Saskatchewan, Saskatoon, SK, Canada, in 1960 and 1963, respectively. He was conferred the title of Honorary Professor by the National Chiao Tong University, Hsinchu, Taiwan, in 2009. He is currently a Research Professor with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada, where he served as the founding Chair of the Department of Electrical Engineering from 1970 to 1977, and the Dean of engineering and computer science from 1977 to 1993. During that time, he developed the faculty into a research-oriented one from what was primarily an undergraduate faculty. Since 2001, he has been the Concordia Chair (Tier I) in signal processing. He has also taught with the Department of Electrical Engineering, Technical University of Nova Scotia, Halifax, NS, Canada, the University of Calgary, Calgary, AB, Canada, and the Department of Mathematics, University of Saskatchewan. He has published in the areas of number theory, circuits, systems, and signal processing, and holds five patents. He has co-authored nine books and five book chapters. He was a Founding Member of Micronet, Ottawa, Canada, as a Canadian Network of Centers of Excellence from 1990 to 2004, and also its Coordinator of Concordia University. He is a fellow of the Institute of Electrical and Electronics Engineers, the Institute of Electrical Engineers, U.K, the Engineering Institute of Canada, the Institution of Engineers, India, and the Institution of Electronic and Telecommunication Engineers, India. He was inducted in 2009 to the Provosts Circle of Distinction for career achievements. He was a recipient of many IEEE-CAS Society awards, including the Education Award in 2000, the Golden Jubilee Medal in 2000, and the 1986 Guillemin-Cauer Best Paper Award. He served as a Program Chair for the 1973 IEEE Circuits and Systems (CAS) Symposium, a General Chair of the 1984 IEEE CAS Symposium, a Vice Chair of the 1999 IEEE CAS Symposium, and a member of the Board of Governors of the CAS Society. He served as the Editor- in-Chief for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I from 1999 to 2001, and an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1985 to 1987. He has served the IEEE in various capacities such as the President Elect in 2003, President in 2004, Past-President in 2005, and Vice President (publications) from 2001 to 2002, Vice President in 1976. He has been the Editor-in-Chief of the journal Circuits, Systems and Signal Processing (CSSP) since 1999. Recently, CSSP has instituted a Best Paper Award in his name.