Comparative Analysis for Detecting Objects Under Cast Shadows in Video Images

Jorge Scandaliaris, Michael Villamizar and Alberto Sanfeliu
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
{*jscandal,mvillami,sanfeliu*}*@iri.upc.edu*

Abstract

Cast shadows add additional difficulties on detecting objects because they locally modify image intensity and color. Shadows may appear or disappear in an image when the object, the camera, or both are free to move through a scene. This work evaluates the performance of an object detection method based on boosted HOG paired with three different image representations in outdoor video sequences. We follow and extend on the taxonomy from van de Sande [7] with considerations on the constraints assumed by each descriptor on the spatial variation of the illumination. We show that the intrinsic image representation consistently gives the best results. This proves the usefulness of this representation for object detection in varying illumination conditions, and supports the idea that in practice local assumptions in the descriptors can be violated.

1. Introduction and related work

Object detection is still a hard problem that have raised much interest in the research community. Techniques based on Histograms of Oriented Gradients (HOGs) have received a lot of attention since its introduction by Dalal [2]. These image descriptors are translation, rotation, and scale invariant. They are also partially robust against certain types of illumination changes thanks to the normalizations involved in their construction.

The majority of descriptors used today are intensity based, although recently color descriptors have been proposed to increase illumination invariance and discriminant power. Burghouts [1] compare the discriminative power and invariance of local color and gray-value descriptors and evaluate their invariance in the presence of shadows and highlights. They show that C-SIFT, a shadow invariant descriptor plugged into the SIFT descriptor, outperforms other methods that combine color and SIFT. Van de Sande [7] also addressed the issue of evaluating a large number of color descriptors based on histograms, color moments and moment invariants, and color SIFT. They studied the invariant properties of the descriptors with respect to photometric transformations analytically and experimentally. They also experimentally assessed the distinctiveness of the color descriptors using two benchmarks from the image and video domain and concluded that invariance to light intensity and light color changes affect object recognition, and that the descriptors with the best overall performers were C-SIFT, rgSIFT, OpponentSIFT and RGB-SIFT.

In this work we use HOGs paired with several different image representations for object detection, and evaluate their relative performance in outdoor video sequences. We share some ground with [1, 7] in the use of color-based invariant image representations to cope with illumination changes, and because the HOG descriptor is similar to the SIFT descriptor. Moreover, we have included specifically the RGB-based HOG descriptor to be able to establish some, at least qualitative, comparisons and extend some of their conclusions. We focus, however, in a more specific problem, as our aim is to be able to perform robust object detection from images acquired from a mobile platform in urban outdoor settings. These images typically show a high degree of variability in the illumination conditions, e.g. the sun position might vary from being behind the camera to being at front of it, presence of self and cast shadows, over and under exposure during transitions from dark to bright areas and vice versa, among others. These conditions were the motivation to explore image representations with better invariance properties.

Our results show that the intrinsic image representation proposed by Finlayson [3] consistently gives

This work has been partially funded by the Spanish Ministry of Science and Innovation under projects UbROB DPI2007-61452, and MIPRCV Consolider Ingenio 2010 CSD2007-00018. The first author is funded by the Technical University of Catalonia (UPC).

Table 1. Invariance of descriptors against illumination changes. '+' denotes sensitivity and '-' invariance. Letters indicate the spatial region assumed constant for the invariance to hold: 'p', pixel; and 'd', region used in the descriptor calculation.

	Intensity change	Intensity shift	Intensity change + intensity shift	Color change	Color change + color shift
G-HOG	-d	-d	-d	+	+
RGB-HOG	-d	-d	-d	-d	-d
II-HOG	-p	-d	-d	-p	+

the best performance when tested on images from sequences acquired in an outdoor environment at different times of the day. This added invariance, however, comes at the price of relying on some camera properties. The implications of this dependence, however, are reduced by the existence of a method that estimates the required parameters directly from images [3].

2. Image representations and descriptors invariance to illumination changes

We use three image representations: intensity or gray value, RGB, and the intrinsic image representation proposed by Finlayson [3]. From each of these image representations we compute an HOG descriptor, which we will refer to by G-HOG, RGB-HOG, II-HOG, respectively.

Next, we analyze the image representations and descriptors following the taxonomy introduced by van de Sande [7] with some additional considerations regarding the constraints imposed by each descriptor on the spatial variation of the illumination. In their analysis, they implicitly assume that the illumination is spatially constant, at least within the image region used in the calculation of the descriptor. Our experience tell us that this is not always true, and thus we include this factor into our analysis. Some of the invariant properties of the descriptors evaluated arise from the image representation they are based on, while others are due to the way the descriptor is constructed. Although it might not seem evident at first, this has some important consequences. The invariance properties derived by van de Sande assume the diagonal-offset model proposed by Finlayson [4] and Lambertian reflectance.

G-HOG descriptor. According to [7], HOG descriptors in general are invariant to light intensity shifts due to use of the gradient. They are also invariant to light intensity changes, and to light intensity changes plus light intensity shifts, due to normalization. These properties hold true as long as the particular photometric changes do not occur within the descriptor region. The fact that these descriptors are *local* in relation to object or image size does not mean that there can not be illumination changes within the descriptor region.

RGB-HOG descriptor. The RGB-HOG descriptor gains invariance to light color change and to light color change plus light color shifts because three independent HOGs, one for each channel, are computed independently including normalization, and stacked together. Again, these invariant properties assume that there are no illumination changes within the descriptor region.

Intrinsic image representation. The image representation proposed by Finlayson [3] is derived from a transformation of the RGB color space formed by dividing each band by the geometric mean, $\sqrt[3]{R \times G \times B}$, and then calculating their logarithm:

$$\rho_k = \log(\frac{R_k}{(\prod_{i=1}^{3} R_i)^{1/3}}),\ k = 1, 2, 3 \qquad (1)$$

All 3-vector $\underline{\rho}$ lie on a plane orthogonal to $\underline{u} = 1/\sqrt{3}(1, 1, 1)$. The redundant dimension is removed by transforming 3-vectors $\underline{\rho}$ into a coordinate system *in* the plane using a 2×3 matrix $\boldsymbol{U}$ (see [3] for details)

$$\underline{\chi} \equiv \boldsymbol{U}\underline{\rho},\ \underline{\chi}\ is\ 2 \times 1. \qquad (2)$$

It can be shown [3] that under the assumption of Planckian illumination, narrow band camera sensitivities and Lambertian surfaces $\underline{\chi}$ has the form

$$\underline{\chi} = \underline{s} + \frac{1}{T}\underline{e}, \qquad (3)$$

where $\underline{s}$ depends on surface and the camera, $\underline{e}$ is independent of surface, but which again depends on the camera, and T is the illuminant color temperature. As a consequence, changes in T result in shifts in the same direction for all surfaces. An invariant to illumination color changes can be obtained by projecting $\underline{\chi}$ into the direction $\underline{e}^{\perp}$ orthogonal to $\underline{e}$, obtaining a single scalar

$$I = \exp(\chi_1 \cos\theta + \chi_2 \sin\theta) \qquad (4)$$

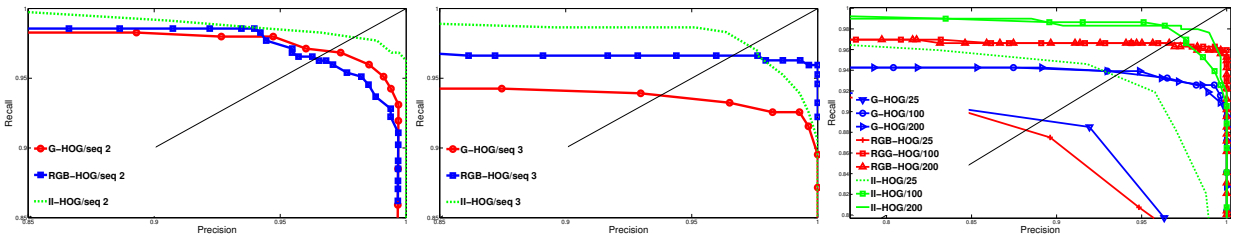

Figure 1. Detection performances. Left: Sequence two. Middle: Sequence three. Right: Number of features.

where the exponentiation removes the effect of the logarithm. This image representation is invariant to all photometric quantities at the pixel level, with the exception to light intensity and color shifts. The II-HOG descriptor gains invariance against light intensity shifts thanks to the gradient in the HOG, but it is not invariant to light color changes plus light color shifts. The differential characteristic of the II-HOG with respect to the other descriptors analyzed is its invariance against illumination intensity *and* illumination color changes at a pixel level. Table 1 summarizes the invariance properties of the descriptors just discussed.

3. Computation of HOG-based detector

The computation of the object detector is based on a boosting algorithm in order to obtain an efficient and robust detector. The goal is to construct a strong classifier, H, by the selection and combination of weak classifiers, h, where each one relies on one HOG-based feature evaluated at coordinates (u, v). Then, the target object is described by a set of local features (local HOGs) evaluated in defined locations which have been obtained via the boosting mechanism. In this work we use the Real Adaboost version because it deals with confidence-rated weak classifiers instead of boolean ones [6]. This is an useful aspect when dealing with our features characterized by histograms of oriented gradients. The boosted classifier is then defined by the combination of T weak classifiers,

$$H(x) = \sum_{t=1}^{T} h_t(x) > \beta, \qquad (5)$$

being x a test image sample and β the detector threshold. Weak classifiers map the sample space X to real-valued space $\Re^n$ whose dimension n is determined by the HOG feature dimension. For comparison purposes, our local HOGs consist of 4x4 spatial cells and 8 gradient orientation bins, yielding a 128-dimensional vector ($n = 128$) similar to SIFT descriptor [5]. Additionally, each cell is formed by 4x4 pixel-gradients.

4. Experiments

Experimental evaluation of the three HOG-based detectors was carried out over three sequences of images acquired from a mobile platform in an outdoor setting at different times of the day. The sequences consist of one person walking in an urban setting exposed to cast shadows and abrupt illumination changes. In all of them the person closes a loop loosely following a path around some raised garden beds. There are pose, scale and illumination changes of the person in front of the camera. In Figure 2 we can see some image examples. In all experiments the first image sequence is used for training the detectors while sequences two and three are used for testing.

For evaluation, test images are labeled with bounding boxes, indicating the location of the person. These bounding boxes, B_g, represent the ground truth. The quality of the results is measured by the overlap ratio of detections, also defined by bounding boxes, B_d, and the ground truth. If this ratio exceeds 50 percent, the detection is considered as a true positive, otherwise, it is considered as a false positive. The overlap ratio is computed as $\frac{|B_g \cap B_d|}{|B_g \cup B_d|} > 0.5$. Finally, the performance of the detector is measured by using a Recall-Precision curve that is computed by true and false positive rates evaluated for various detector thresholds, β.

Evaluation of sensitivity to number of features. In this experiment the detector performance when the number of features selected by the boosting algorithm varied was evaluated in sequence three. We considered 25, 100 and 200 features. Results for each of the descriptors (G-HOG, RGB-HOG and II-HOG) are shown in Figure 1. They show that increasing the number of features the detection performance increases for all approaches. Furthermore, the detector based on II-HOG

Figure 2. Detection results. Cyan rectangles are correct detections and red ones are their ground truth.

outperforms the other ones at the same number of features and the difference in performance increases as the number of selected features decreases. For instance, the detector using an II-HOG representation with 100 features achieves better detection results than the other methods using 200 features. Detectors based on G-HOG and RGB-HOG are more sensitive to the reduction of the number of selected features, requiring more features to achieve a comparable performance to the detector based on II-HOG. One possible explanation to this behavior is that the detectors based on G-HOG and RGB-HOG compensate for the illumination variations by relying on exhaustive training and more features for building the boosted classifier, while the detector based on II-HOG benefits from the better invariant properties of the underlying image representation. This also suggests that the assumption of the illumination being spatially constant in the descriptor region is violated in practice, which then gives relevance to the II-HOG's invariant properties at the pixel. The detector based on II-HOG is computationally more efficient because it requires less object features to achieve good detection rates.

Evaluation under image conditions changes. The HOG-based detectors are also tested with the aim of measuring their performance under different illumination conditions and the presence of cast shadows. To this end, the detectors were evaluated over sequences two and three which present unknown image conditions given that the sequences were acquired with a couple of hours difference between each other. Figure 1 shows detection performances for all the approaches. Results show that II-HOG is consistently better than the other approaches in both sequences, achieving an ERR (Equal Error Rate) of 97.9% and 97.3% for sequences two and three, respectively. G-HOG and RGB-HOG achieve 96.9% and 96.4% for sequence two and 93.7% and 96.6% for sequence three, respectively. This experiment has been carried out using a boosted classifier with 200 HOG-based features. Figure 2 shows some detection results for the approach based on II-HOG, where the extreme illumination conditions present are evidenced.

5. Conclusions

We have evaluated the detection performance of HOG descriptors based on three different image representations under abrupt illumination changes. The descriptor based on the intrinsic image representation consistently outperformed the other descriptors. The RGB-HOG and the G-HOG improve their detection rates at the expense of requiring a larger number of features to achieve comparable performance to the II-HOG descriptor. This supports two conclusions: first, that the intrinsic image representation proves to be a useful image representation for object detection when the illumination conditions vary considerably; and second, that in practice the illumination invariance assumption of local descriptors can be violated.

References

[1] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local colour invariants. *CVIU*, 2009.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005.

[3] G. D. Finlayson, M. S. Drew, and C. Lu. Intrinsic images by entropy minimization. *ECCV*, 2004.

[4] G. D. Finlayson, S. D. Hordley, and R. Xu. Convex programming colour constancy with a diagonal-offset model. *ICIP*, 2005.

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[6] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 1999.

[7] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 2010.