

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. In this work we use an object detection method based on boosted HOG paired with three different image representations, and we evaluate their relative performance. 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 performance when tested on images from sequences acquired in an outdoor environment at different times of the day. This proves the usefulness of this representation for object detection in varying illumination conditions, and supports the idea that local assumptions in the descriptors can, in practice, be violated.

1. Introduction and related work

Object detection is still a hard problem that have raised much interest in the research community. This is reflected by the large amount of literature on this topic. Techniques based on Histograms of Oriented Gradients (HOGs) have received a lot of attention since its introduction by Dalal *et al.* [2]. The key idea behind HOGs is that local object shape and appearance can be captured by local histograms of image gradient orientations, calculated over a group of pixels referred to as a cell. The combined histogram entries of several cells form the HOG descriptor, which is usually normalized to gain partial invariance to illumination changes. The original work by Dalal *et al.* used a dense and overlapping tiling of cells within the detection window with local contrast normalization for pedestrian detection in static images. In a similar way, SIFT features [5] compute fixed HOG descriptors in a grid of 4x4 cells and 8 gradient orientations around interest points. These image descriptors are translation, rotation, and scale invariant. They are



Figure 1. Image sequence examples.

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 *et al.* [1] compare the discriminative power and invariance of local color descriptors to Gray-value descriptors and evaluate the invariance of local color descriptors in the presence of shadows and highlights. They concluded that the shadow invariant descriptor, referred to as C-colour, performs the best, and show that when plugged into the SIFT descriptor, and then termed C-SIFT, it outperforms other methods that combine color and SIFT. Van de Sande *et al.* [7] also addressed the issue of evaluating a large number of color descriptors based on histograms, color moments and moment invariants, and color SIFT. Among the color SIFT descriptors evaluated, they included C-SIFT. They studied the descriptors analytically using a taxonomy based on the invariant properties with respect to photometric transformations and verified these results on a dataset with known illumination conditions. They also experimentally assessed the distinctiveness of the color descriptors using two benchmarks from the

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	-

image and video domain. The conclusion was 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. RGB-SIFT, in particular, is invariant to all illumination changes considered in their work.

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 shares many similarities with 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 have being the motivation to explore image representations with better invariance properties.

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

posed by Finlayson *et al.* [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. Then, we present and analyze each image representation making explicit the assumptions they rely upon, together with their invariance properties.

We follow the taxonomy introduced by van de Sande *et al.* [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 region of pixels where the descriptor is being calculated. 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 *et al.* assume the diagonal-offset model proposed by Finlayson *et al.* [4] and Lambertian reflectance.

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. This is important for our analysis. 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 pixel computed 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, the invariance implies no illumination changes within the descriptor region.

Intrinsic image representation. The image representation proposed by Finlayson [3] is derived from a

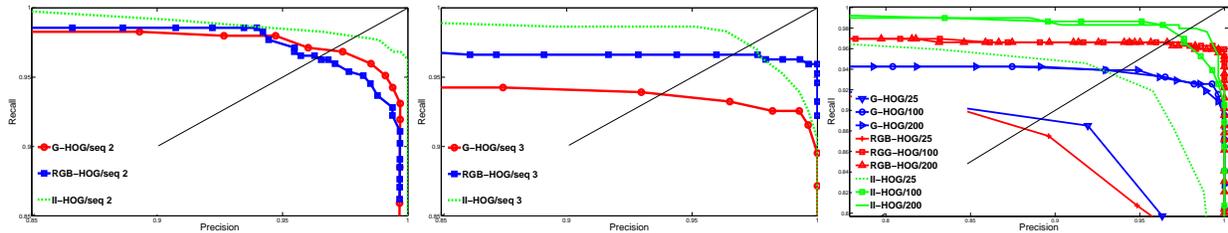


Figure 2. Detection performances. Left: Sequence two. Middle: Sequence three. Right: Number of features.

transformation of the RGB color space formed by dividing each band by the geometric mean, $\sqrt[3]{R \times G \times B}$, and the calculating logarithm

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

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 rotating 3-vectors $\underline{\rho}$ into a coordinate system *in* the plane using a 2×3 matrix \underline{U} (see [3] for details)

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

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

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

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

$$I' = \chi_1 \cos \theta + \chi_2 \sin \theta. \quad (4)$$

To remove the effect of the logarithm, the last step in the derivation of the intrinsic image is to exponentiate

$$I = \exp(I'). \quad (5)$$

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 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,

$$H(x) = \sum_{t=1}^T h_t(x) > \beta, \quad (6)$$

being x a test image sample and β the detector threshold. Weak classifiers map the sample space X to real-valued space \mathbb{R}^n whose dimension n is determined by the HOG feature dimension. For comparison purposes, our local HOGs consist of 4×4 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 4×4 pixel-gradients.

4. Experiments

Experimental evaluation of the three image representations in addition to a HOG-based detector was carried out over three sequences of video images acquired from a mobile platform in outdoor settings and taken 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 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 1 we can see some image examples. In all experiments the first image sequence is used for training the HOG-based detector while sequences two and three are used for testing.

For evaluation, test images are labeled with bounding boxes, indicating location of the walking person within images. These bounding boxes (B_g) represent the ground truth and the quality of results are measured by their overlap ratio with detections defined by bounding boxes (B_d) as well. If this ratio exceeds 50 percent,



Figure 3. Detection results. Cyan rectangles are correct detections and red ones are their ground truth.

a true positive detection is considered, otherwise, such detection is considered as a false positive. This 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. The detector performance has been evaluated in the sequence three in terms of the number of HOG-based features selected by the boosting algorithm. For this experiment we have considered 25, 100 and 200 features. In Figure ?? are shown the detection performances for each one of the image representations (G-HOG, RGB-HOG and II-HOG). They show that increasing the number of features the detection performance increases for all approaches. Furthermore, the II-HOG-based detector outperforms the other ones at the same number of features. This detector achieves remarkable rates, while the other methods yield lower detections rates. For instance, the detector using an II-HOG representation with 100 features achieves better detection results than the other methods using 200 features. It proves that 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 II-HOG. One possible explanation to this is that G-HOG and RGB-HOG are compensating the illumination variations by using an exhaustive training and more number of HOG-based features for building the boosted classifier.

By other hand, because the II-based detector requires less object features to achieve a good detection rate, this approach is more efficient because it reduces the computational burden caused by the evaluation of object features in the recognition stage.

Evaluation under image conditions changes. The HOG-based detector is also tested with the aim of measuring its performance under different illumination conditions and cast shadows. To this end, the detector is evaluated over sequences two and three which present unknown image conditions, given that the sequences were acquired with a couple of hours of difference between each other. Figure 2 shows detection performances for each one of the proposed approaches. Re-

sults 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. In Figure 3 are shown some detection results for the II-HOG based approach that show that it is able to deal with extreme illumination conditions with remarkable detection rates.

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 the illumination invariance assumption of local descriptors can in practice 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.