2009 IEEE International Conference on Robotics and Automation
Kobe International Conference Center
Kobe, Japan, May 12-17, 2009

Combining Color-Based Invariant Gradient Detector with HoG Descriptors for Robust Image Detection in Scenes under Cast Shadows

Michael Villamizar[1,2], Jorge Scandaliaris[1], Alberto Sanfeliu[1,2] and Juan Andrade-Cetto[1]
[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC
[2]Department of Automatic Control, UPC

***Abstract*— In this work we present a robust detection method in outdoor scenes under cast shadows using color based invariant gradients in combination with HoG local features. The method achieves good detection rates in urban scene classification and person detection outperforming traditional methods based on intensity gradient detectors which are sensible to illumination variations but not to cast shadows. The method uses color based invariant gradients that emphasize material changes and extract relevant and invariant features for detection while neglecting shadow contours. This method allows to train and detect objects and scenes independently of scene illumination, cast and self shadows. Moreover, it allows to do training in one shot, that is, when the robot visits the scene for the first time.**

I. Introduction

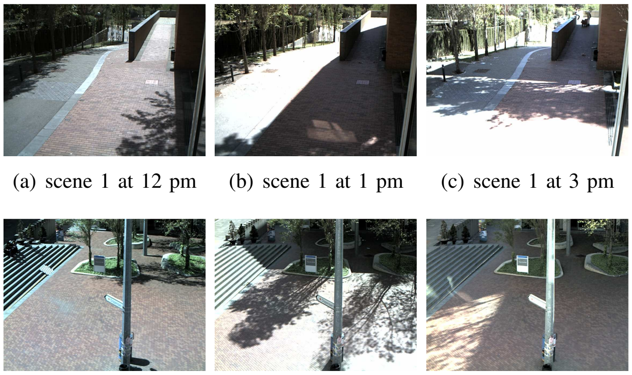
This work proposes a robust detection method for robotics applications, such as people and object detection or scene classification under cast shadows. In outdoor vision tasks, illumination conditions constraint the detector performance due to varying features produced by cast shadows. This method is focused to object detection in urban settings, within the European Project URUS [1], where varying shadows are present and illumination conditions are extreme (Figure 1). In these environments, detection becomes a challenging task for robots and network robot systems.

Recently, several techniques based on Histograms of Oriented Gradients [2]–[6] have been developed, which are robust and reliable for representing image local features. The key point in using HoG descriptors is to capture or encode feature appearance layout where each histogram cell contains an oriented gradient distribution for pixels within this cell. Although, successful detection results with HoG based detectors are sensible to cast shadows because they depend on intensity gradients. In cast shadows scenes, a lot of false gradients are present making difficult to train good reliable descriptors and perturbing object gradients in the detection phase, see Figure 2. Dalal and Triggs [3] proposed to use HoG descriptors for pedestrian detection in static images and videos. They use an overlapping local contrast normalization in order to improve detection performance giving certain invariance to illumination and shadows. In Bosch *et al.* [2] pyramidal Histograms of Oriented Gradients are used for object categorization. These pyramidal descriptors encode features and their spatial layout at several resolution levels, allowing robustness to small feature shifts. Finer histogram levels are weighted more than coarser ones, since finer resolutions have more detailed feature shape description. This idea is inspired by image pyramid representation of scenes [5]. The descriptor matches measure the appearance and spatial correspondences of features, i.e. oriented gradients. Pyramidal descriptors are computed on Regions of Interest (ROIs) in order to suppress background clutter and occlusions. This spatial pyramidal representation is an extension to the Dalal and Triggs method where Histograms of Oriented Gradients are restricted to finer resolutions. In the same way, SIFT features [6] compute fixed HoG descriptors in a grid of 4×4 cells and 8 gradient orientations around key points. In Scandaliaris *et al.* [7], we proposed to use color based invariant contours and compared them against simple intensity contours in object detection domain under shadows. The proposed method outperformed classical methods and showed that invariant contours can be extracted. The method is based on a physical model of the image formation process and strives to remove the effects of shadows, producing a contour image invariant to shadows. Instead of calculating the gradient modulus from the color images, we detect contours that correspond to material changes using a modification to the approach proposed by Gevers *et al.* [8] based on a combination of photometric invariant contours and automatic local noise-adaptive thresholding. Using these invariant contours we compute and select discriminative Haar-like features to build a simple but fast object detector. However, a large number of contour Haar features had to be computed because of the limited discriminative power of these features.

In this paper, we use pyramidal Histograms of Oriented Gradients in order to have more discriminative and robust descriptors with which to describe objects or scenes, and to face up the drawback of cast shadows using color based invariant gradients (referred to as invariant gradients from now on) [7] that reduce illumination and shadow effects and improve detection results in robotics applications. The descriptors extracted during the learning phase (see section IV) encode

This research was partially supported by the Spanish Ministry of Innovation and Science under projects Consolider Ingenio 2010 CSD2007-00018, and projects DPI 2007-614452 and DPI 2008-06022; by the URUS project IST-045062 of the European Union; by the Generalitat of Catalonia's Department of Innovation, University and Industry and the European Social Fund to JS; and by the Technical University of Catalonia (UPC) to MV.

978-1-4244-2789-5/09/$25.00 ©2009 IEEE



(a) scene 1 at 12 pm (b) scene 1 at 1 pm (c) scene 1 at 3 pm

(d) scene 5 at 12 pm (e) scene 5 at 4 pm (f) scene 5 at 6 pm

Fig. 1. Barcelona Robot Lab. Change of illumination conditions at different instances of daylight

relevant and invariant features useful for detection. As cast shadows are reduced, this method uses few sample images to train the detector. Therefore, it is possible to train the detector using just one illumination condition, being able to detect object or scenes with different illumination conditions. Then, the process of learning (Boosting of HoG features) is limited to the first time the robots visit the area (this is true when the illumination conditions do not change dramatically). To validate our method, we have carried out experiments in scene classification and person detection under cast shadows and compare them against using intensity gradient based HoG descriptors (see section V). Our results outperform intensity gradient based HoG descriptors and achieve good detection rates.

II. Color based invariant gradient detector

In this work we assume the dichromatic reflection model [9] for the physical interaction between the light incident over the scene, the surfaces of the scene and the camera. Moreover, we assume that the illumination source is white or spectrally smooth and the interface reflectance is neutral. Under these assumptions the reflection model is given by:

$$V_k = G_b(\vec{n},\vec{s})E\int_\lambda B(\lambda)F_k(\lambda)\,d\lambda + G_i(\vec{n},\vec{s},\vec{v})ESF \qquad (1)$$

where V_k is the kth sensor response, G_b and G_i are geometric terms denoting the geometric dependencies of the body and surface reflection component. That is, surface normal, $\vec{n}$, illumination direction, $\vec{s}$, and viewing direction, $\vec{v}$. $B(\lambda)$ is the surface albedo, E denotes the illumination source, and S denotes the Fresnel reflectance, both assumed independent of λ. $F_k(\lambda)$ denotes the kth sensor spectral sensitivity and $F = \int_\lambda F_k(\lambda)\,d\lambda$.

A. Color Models

The invariant gradient detector that we propose is based on the work of Gevers [8] and the modifications proposed in [7]. This detector uses three color models that have different and complementary properties regarding their response: *RGB*, *c1c2c3* and *o1o2*. In the *RGB* color model $\{R,G,B\}$ values correspond directly with V_k in (1). The *c1c2c3* color model is defined by

TABLE I

Color model sensitivity to parameters of the image formation process. + denotes sensitivity and - invariance of the color model to a particular parameter.

	shadow	geometry	material	highlights
RGB	+	+	+	+
c1c2c3	-	-	+	+
o1o2	+	+	+	-

$$c1(R,G,B) = \arctan(R/\max(G,B)) \qquad (2)$$

$$c2(R,G,B) = \arctan(G/\max(R,B)) \qquad (3)$$

$$c3(R,G,B) = \arctan(B/\max(R,G)) \qquad (4)$$

and the *o1o2* color model is defined by

$$o1(R,G,B) = (R-G)/2 \qquad (5)$$

$$o2(R,G,B) = (R+G)/4 - B/2 \qquad (6)$$

It follows from (1) that the *RGB* color model is sensitive to all parameters of the dichromatic reflection model, the *c1c2c3* color model depends only on the sensor spectral sensitivities and the surface albedo or material for dull objects, being independent of shadows and geometry (E and G_b in the model) and the *o1o2* color model is invariant to highlights for shiny objects under the same assumptions. *o1o2* is still dependent on geometry (G_b). These results can be seen in Table I.

B. Color Invariant Gradient

The invariant gradient is computed by calculating the x and y derivatives for each channel of the three aforementioned color models using Gaussian derivatives and the gradient magnitude for each color model is computed using the Euclidean metric over the various channel derivatives:

$$\nabla C = \sqrt{\sum_{i=1}^{N}\left[\left(\frac{\partial c_i}{\partial x}\right)^2 + \left(\frac{\partial c_i}{\partial y}\right)^2\right]} \qquad (7)$$

with C representing each color model, N being their dimensionality, and c_i the particular color channels.

However, the presence of noise in the images can lead to maxima in the gradient modulus that are not related to the parameters of the image formation process. One way to eliminate these maxima is propagating the uncertainties associated to the color models as well as the different gradient moduli. In order to do this, we calculate the gradient magnitude of the *RGB*, *c1c2c3* and *o1o2* color models, and then, we propagate (see [7] for details) the *RGB* uncertainties, assumed to be known, through the various color models up to the gradient magnitudes, using the uncertainties associated with the gradient magnitude of each color space, σ_{∇_C}. Finally we define the gradient product

$$M = \nabla RGB \cdot \nabla c1c2c3 \cdot \nabla o1o2 \qquad (8)$$

M will have a maximum value when the gradient moduli of all color models are simultaneously maximum, and will have low values when the gradient modulus of any of the color models is low. By looking at Table I it is evident that the response of M emphasizes material changes in the image, in contrast to those in shadows, geometry and highlights.

Then, the uncertainty in the function M is also computed to yield

$$\sigma_M \leq \sum_j \frac{\partial M}{\partial(\nabla C_j)} \sigma_{\nabla C_j} \tag{9}$$

where the summation is over the three color models, RGB, $c1c2c3$ and $o1o2$, and with $\sigma_{\nabla C}$, dropping the j subindex, being calculated for each color model using

$$\sigma_{\nabla C} \leq \frac{\sum_i \left[\left|\frac{\partial c_i}{\partial x}\right| \cdot \sigma_{\frac{\partial c_i}{\partial x}} + \left|\frac{\partial c_i}{\partial y}\right| \cdot \sigma_{\frac{\partial c_i}{\partial y}}\right]}{\sqrt{\sum_i \left[\left(\frac{\partial c_i}{\partial x}\right)^2 + \left(\frac{\partial c_i}{\partial y}\right)^2\right]}} \tag{10}$$

with C representing each color model and c_i the particular color channels. The uncertainties $\sigma_{\frac{\partial c_i}{\partial x}}$ and $\sigma_{\frac{\partial c_i}{\partial y}}$ are computed by approximating the derivatives by filtering with a mask, Gaussian in this case.

Finally, we obtain a local noise-adaptive threshold for removing noisy measurements from M.

$$M' = \begin{cases} M & M > 3\sigma_M \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The final result emphasizes the contribution of material changes and at the same time reduces that of shadow-geometry and highlights on the input images.

III. Pyramidal HoG descriptor

Methods based on Histograms of Oriented Gradients (HoG) have shown successful results in object detection and classification [2], [4], [5], [10]. Although they have been tested in outdoor scenes for object detection, such as cars, these methods have not been tested under extreme illumination conditions, such as varying shadows. We address detection under these conditions and compare the results by using the invariant gradient for detecting the main features of the scene object contours that do not belong to a cast shadow. In this work, the pyramidal HoG [2] is used as local descriptor using its spatial histogram resolution pyramid, see Figure 3. With this representation, features (oriented gradients) can be matched at several spatial grid resolutions, which improves the detection of features, unlike other methods which have a fixed spatial histogram resolution [3], [6]. This implementation allows a certain invariance to image transformations, i.e. feature shifts. The pyramidal HoG descriptor is similar to the well known SIFT descriptor [6], however the last one has fixed spatial grid resolution (4×4) with descriptors located around key points (blobs). We opt for localizing local descriptors in multiple scales and locations.

We use the pyramid match kernel similarity measure between two HoG descriptors H_x, H_y [11]. This measure is defined as a weighted sum of feature matches that occur in each resolution level. Feature matching in each level is carried out using histogram intersection [12] and its level weight assigned according to histogram resolution. Matches in coarser levels have lower weight than finer ones. This technique is robust to clutter since additional features do not affect the pyramid matching. Additionally, pyramid matching computation is linear in the number of local features [11]. The matching can be expressed as :

$$k^L(H_x,H_y) = \frac{1}{2^L} I^0(H_x,H_y) + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l(H_x,H_y) \tag{12}$$

where $I^l(H_x,H_y) = \sum_{i=1}^{D} min(H_x^l(i), H_y^l(i))$ is the intersection measure between descriptor histograms H_x and H_y, of dimension D, at level l.

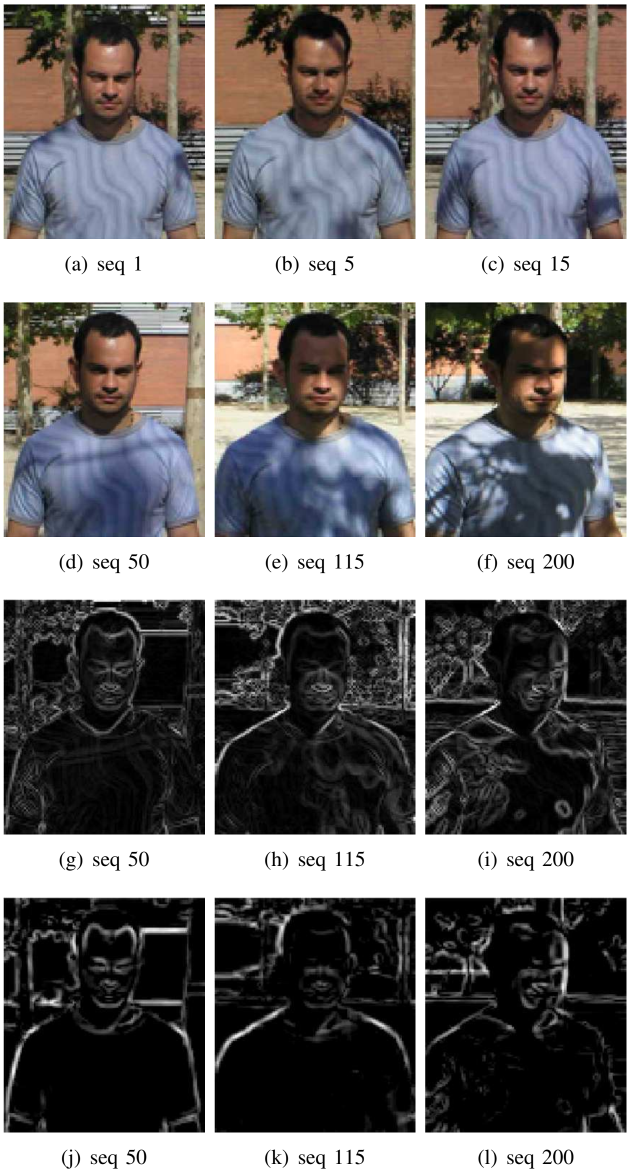
IV. Implementation Details

In order to compute and select local descriptors a boosting algorithm is used [13]. This algorithm selects the most discriminative HoG descriptors to build a robust classifier by means of a weighted linear combination of them. At each iteration a weak classifier is selected which better classifies the training images from positive and negative samples. Each weak classifier is defined by one pyramidal HoG descriptor and its location, scale and threshold. The threshold is calculated automatically inside the boosting algorithm as the threshold where the classification rate is maximum. In our experiments 280 weak classifiers are boosted to form a strong classifier.

The minimum and maximum descriptor scales are 16×16 pixels and 0.6 of training image size, respectively. The training image size depends on the target to detect and its aspect ratio. In our case, for person detection, 120×100 images are used and for scene classification 180×240 images are used. Our experiments show that scene classification can be efficiently done in low resolution images.

V. Experiments

This section describes experiments carried out to show the proposed detector performance under cast shadows and diverse illumination conditions, and to compare these results with the same method but using intensity gradient descriptors. False gradients resulting from cast shadows affect the detector performance and make difficult to train the detector. In the experiments, it is possible to observe how the proposed method extracts and selects more robust features to shadows than using traditional intensity gradient methods. These intensity gradient based descriptors can only detect objects or scenes under similar illumination conditions as the training images. The experiments are performed on scene classification and person detection that are typical applications in robotic vision systems and are part of the URUS project [1].



(a) seq 1 (b) seq 5 (c) seq 15

(d) seq 50 (e) seq 115 (f) seq 200

(g) seq 50 (h) seq 115 (i) seq 200

(j) seq 50 (k) seq 115 (l) seq 200

Fig. 2. Person image sequence. a-c) training samples d-f) test samples g-i) intensity contours j-l) shadow invariant contours

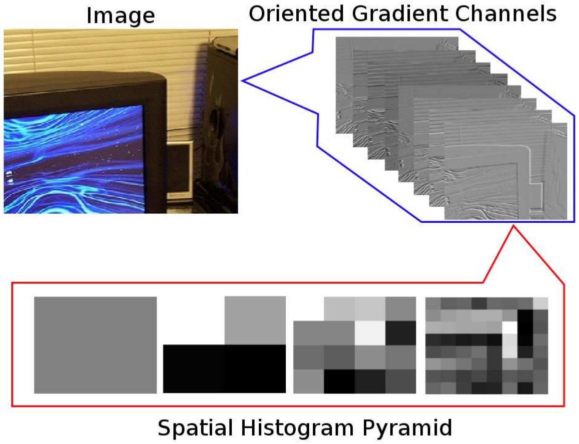
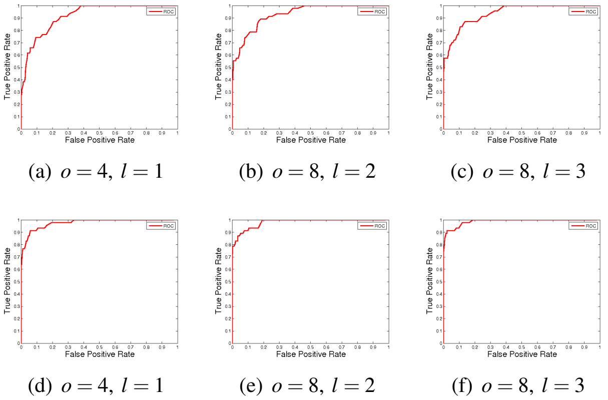


Fig. 3. Pyramidal HoG descriptor



(a) $o = 4$, $l = 1$ (b) $o = 8$, $l = 2$ (c) $o = 8$, $l = 3$

(d) $o = 4$, $l = 1$ (e) $o = 8$, $l = 2$ (f) $o = 8$, $l = 3$

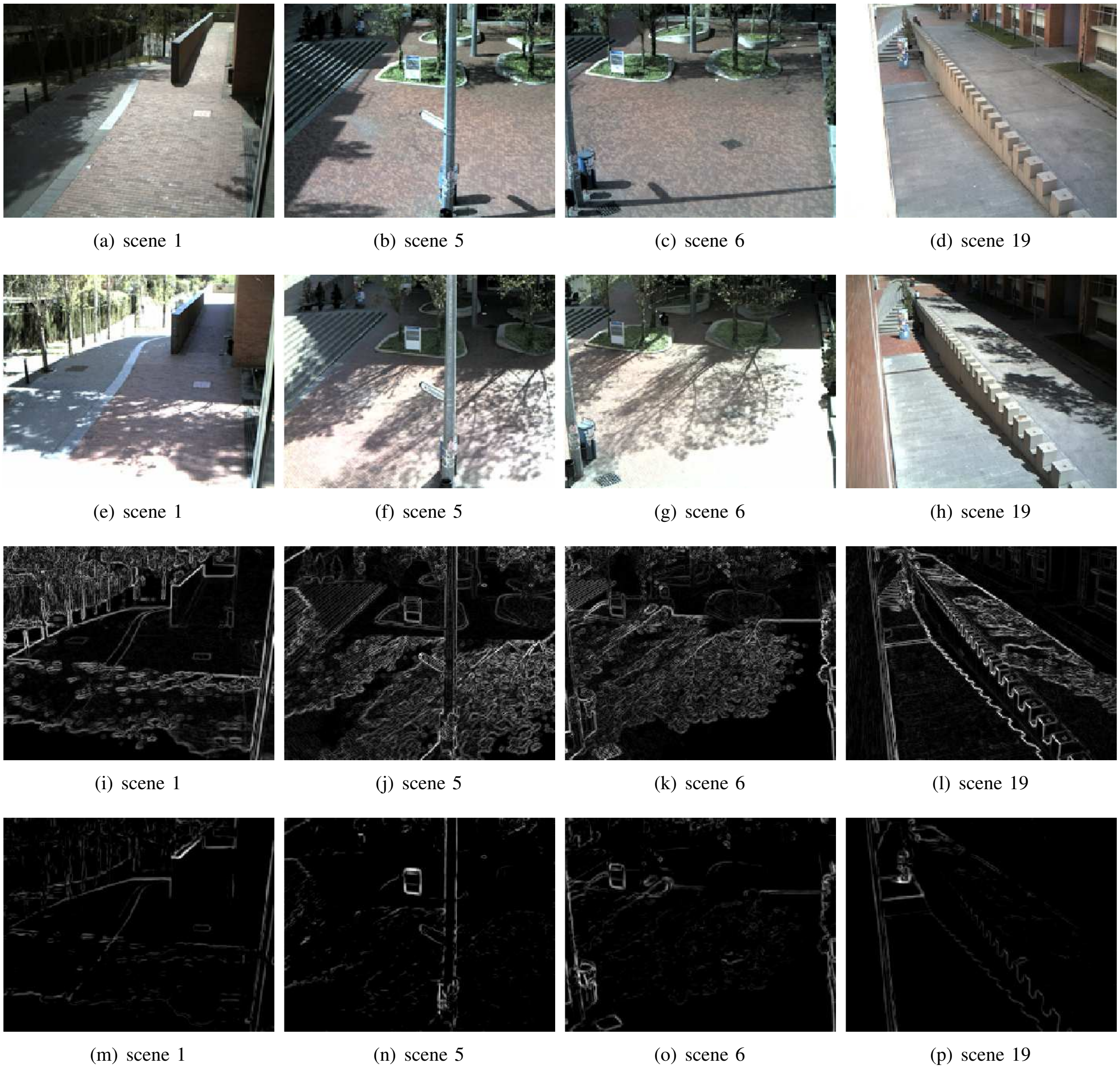
Fig. 4. Person detection ROC curves. a, b, c) Intensity gradient method. d, e, f) The proposed invariant gradients method. o: number of gradient orientations; l: number of levels in the pyramid.

A. *Person detection*

In this experiment, we have a sequence of images of a person that moves through a scene with shadows caused by trees and buildings. The images show the upper body part of the person. We have used the first 20 images for training and the rest of the images, about 100, for tests. These images have cast shadows (see Figure 2 d-f), the upper body presents some shifts and also some 3D body rotations. The experiments were carried out as follows: the same sequence of images has been filtered by an intensity gradient and a invariant gradient (Figure 2). Then a HoG descriptor has been used in both sequences of filtered images, and the measure distance described in section III and boosting classifier have been used for detection on the scenes. The results are shown in the detection curves (ROC) shown in Figure 4. The axes represent true positive ratio and false positive ratio; both ratios are in the interval [0,1]. As we can see, our proposed invariant gradient outperforms the traditional gradient filter method. Moreover, the maximum detection rate of the HoG descriptor is obtained when we use 8 gradient orientations and 3 levels in the pyramid representation (16×16 cell grid).

B. *Scene Classification*

The same method has also been used for urban scenes. In this case the images were taken by four cameras of the Barcelona Robot Lab (experimental site for urban robots at UPC, Barcelona). Figure 5 shows some examples of these images. The training of each detector was done with the images captured from one of the four scenes in a short time interval in the morning. In this case, we used about 10 images for each camera taken within an interval of 5 minutes. The aim is to classify the images from a negative image dataset and test images from the four scenes in order to measure scene detector performance and discrimination among scenes. 500 negative samples were extracted from images with high contrast. Around 100 test images per scene were selected from the scene camera sequence to test the detectors. The results are shown in Figures 6 and 7. These curves were obtained testing each learned scene



(a) scene 1 (b) scene 5 (c) scene 6 (d) scene 19

(e) scene 1 (f) scene 5 (g) scene 6 (h) scene 19

(i) scene 1 (j) scene 5 (k) scene 6 (l) scene 19

(m) scene 1 (n) scene 5 (o) scene 6 (p) scene 19

Fig. 5. Barcelona Robot Lab scenes. a-d) training images e-h) test images i-l) intensity contours m-p) shadow invariant contours

detector (rows) with negative images and test images from the four scenes (columns). What we expect is that curves in the diagonal must have maximum detection with minimum false positive rate. We can see that this is true for Figure 7 for where the invariant gradient detector was used. This fact shows that learned scene detectors respond better with images of their own class. Therefore, it is possible to perform robust and reliable classification in images with varying cast shadows, even having similiar scenes patterns. As this method is based on local descriptors, it can withstand mild occlusions. In cases with a moderate amount of unknowns elements (bikes, people, etc.), some descriptors would fail but other carry on with the supporting decision, thanks to the boosted classifier (section IV).

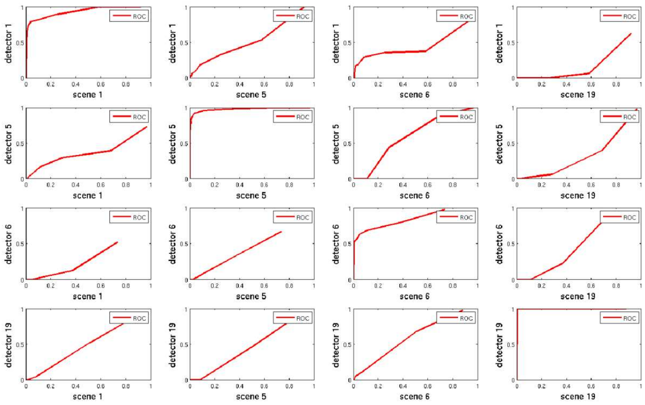


Fig. 6. Scene detection ROC curves (intensity gradient detector)

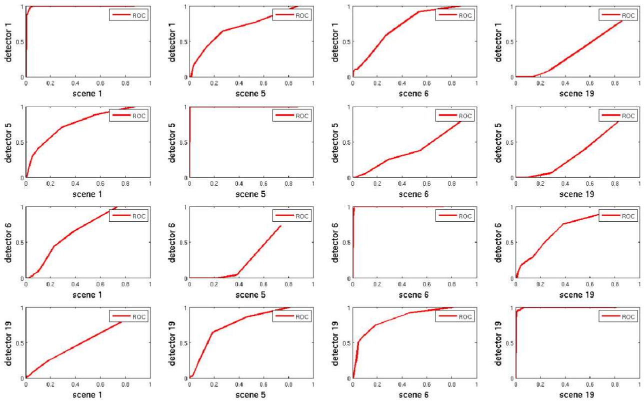


Fig. 7. Scene detection ROC curves (invariant gradient detector)

VI. Conclusions

In this paper, we have shown that detection performance in outdoor scenes under cast shadows improves when combining invariant gradients with pyramidal HoG descriptors. The method has been tested in person detection and scene classification achieving high detection rates and outperforming the pyramidal HoG descriptors based on intensity gradients. The descriptors based on the invariant gradients are more robust to shadows and changes in illumination conditions, and thus the proposed method allows for training with a small number of sample images taken at any time of the day.

References

[1] A. Sanfeliu and J. Andrade-Cetto, “Ubiquitous networking robotics in urban settings,” in *Proc. IEEE/RSJ IROS Workshop Network Robot Syst.*, Beijing, Oct. 2006, pp. 14–18.

[2] A. Bosch, A. Zisserman, and X. Munoz, “Image classification using ROIs and multiple kernel learning,” *Int. J. Comput. Vision*, 2008, submitted.

[3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, vol. 1, San Diego, Jun. 2005, pp. 886–893.

[4] I. Laptev, “Improvements of object detection using boosted histograms,” in *Proc. British Machine Vision Conf.*, vol. 3, Oxford, Sep. 2005, pp. 949–958.

[5] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, vol. 2, New York, Jun. 2006, pp. 2169–2178.

[6] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, pp. 91–110, 2004.

[7] J. Scandaliaris, M. Villamizar, J. Andrade-Cetto, and A. Sanfeliu, “Robust color contour object detection invariant to shadows,” in *Progress in Pattern Recognition, Image Analysis and Applications*, ser. Lect. Notes Comput. Sci., vol. 4756. Viña del Mar: Springer-Verlag, Nov. 2007.

[8] T. Gevers and H. Stokman, “Classifying color edges in video into shadow-geometry, highlight, or material transitions,” *IEEE Trans. Multimedia*, vol. 5, no. 2, pp. 237–243, Jun. 2003.

[9] S. A. Shafer, “Using color to separate reflection components,” *Color Res. Appl.*, vol. 10, no. 4, pp. 210–218, 1985.

[10] O. Chum and A. Zisserman, “An exemplar model for learning object classes,” in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, Minneapolis, Jun. 2007, pp. 1–8.

[11] K. Grauman and T. Darrell, “The pyramid match kernel: discriminative classification with sets of image features,” in *Proc. IEEE Int. Conf. Comput. Vision*, vol. 2, Beijing, Oct. 2005, pp. 1458–1465.

[12] M. Swain and D. Ballard, “Color indexing,” *Int. J. Comput. Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[13] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, vol. 1, Hawaii, Jun. 2001, pp. 511–518.