Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2011, Article ID 684819, 16 pages
doi:10.1155/2011/684819

Research Article

Contextual Information and Covariance Descriptors for People Surveillance: An Application for Safety of Construction Workers

Giovanni Gualdi,[1] Andrea Prati,[2] and Rita Cucchiara[1]

[1] *DII, University of Modena and Reggio Emilia, 41122 Modena, Italy*
[2] *DISMI, University of Modena and Reggio Emilia, 42122 Reggio Emilia, Italy*

Correspondence should be addressed to Andrea Prati, andrea.prati@unimore.it

Received 30 April 2010; Revised 7 October 2010; Accepted 10 December 2010

Academic Editor: Luigi Di Stefano

In computer science, contextual information can be used both to reduce computations and to increase accuracy. This paper discusses how it can be exploited for people surveillance in very cluttered environments in terms of perspective (i.e., weak scene calibration) and appearance of the objects of interest (i.e., relevance feedback on the training of a classifier). These techniques are applied to a pedestrian detector that uses a LogitBoost classifier, appropriately modified to work with covariance descriptors which lie on Riemannian manifolds. On each detected pedestrian, a similar classifier is employed to obtain a precise localization of the head. Two novelties on the algorithms are proposed in this case: polar image transformations to better exploit the circular feature of the head appearance and multispectral image derivatives that catch not only luminance but also chrominance variations. The complete approach has been tested on the surveillance of a construction site to detect workers that do not wear the hard hat: in such scenarios, the complexity and dynamics are very high, making pedestrian detection a real challenge.

1. Introduction

The research in computer vision and pattern recognition is often challenged by two conflicting goals, that is, strong requirements on accuracy of results, that should ideally reproduce or even improve the outcome of the human vision system, and tight constraints in the response time. For this reason, it is always helpful to exploit contextual information provided as prior or additional knowledge (learned either before or at run time). The use of context, indeed, has the twofold advantage to save computational time (by reducing the hypotheses' search space) and to increase the accuracy (by removing potential sources of errors, such as distractors). For this reason, the exploitation of contextual information in computer vision is an emerging field [1] and has been proposed in several scopes. Indeed, as the human visual perception is correlated to its context belief or knowledge (e.g., a moving object in a soccer field is unlikely to be associated to a bike), similarly computer vision models gain from considering contextual information.

With these premises, this paper discusses how to include and model the contextual information in a generic framework for people surveillance, applied in large and complex open areas, like construction working sites. These areas are typically very cluttered, with several people and machineries moving all around (see Figure 1). Thus, motion-based segmentation and tracking are seriously challenged and do not guarantee a sufficient degree of reliability. To make things even worse, the construction working sites are continuously evolving, and the lack of fixed reference points makes it very difficult to exploit precise geometric calibration and models, that would help in scene understanding.

Indeed, for many surveillance purposes just object detection is needed and tracking is not necessary; this paper aims at showing that in such challenging conditions it is possible to obtain a reliable and general-purpose people surveillance through *appearance-based object detection and classification* and *context exploitation*.

On one side, the appearance is a feature that can be used regardless of the state of motion that lies behind the generating object, and this condition is very useful in

(a) (b)

Figure 1: Examples from a construction working site.

challenging contexts such as construction sites. In particular, we focus our attention on the detection of pedestrians (i.e., standing people) for two reasons: pedestrian detection is actually a very active research topic [2–6] and people are the main objects of interest in surveillance and security. Nevertheless, since the pedestrian is an extremely articulated object equipped with a significant number of degrees of freedom, it is clear that what is proposed for pedestrians can be easily extended to other classes of objects.

On the other side, the contextual information, that is autonomously collected by the system through a learning stage, demonstrates to provide successful results. More specifically, we propose to exploit contextual information in two manners: (i) a *relevance feedback* (RF) strategy which enriches the pedestrian detection phase by replacing the final stages of a cascade of classifiers (that have been trained for generic pedestrian detection), with new stages trained on positive and negative samples that are (semi) automatically extracted from a specific context only; (ii) a *weak scene* (auto) *calibration* which roughly estimates the scene perspective in order to discard out-of-scale detections.

The general structure of our framework, depicted in Figure 2, is divided in two parts, namely (A) learning and (B) exploiting the context. Step (A) makes use of video data and of general-purpose models to extract new and refined models that better fit the domain-specific data (Figure 3). In our proposal, the general purpose models are embodied by a set of detectors (pedestrian detector, head detector, etc.) trained on generic and context-free (unbiased) training datasets. The context models learned during Step (A) are stored and then used during the Step (B), where video data *coming from the same domain* is processed using both general-purpose and context-dependent models to produce video analysis for surveillance purposes (Figure 4).

For the sake of validating the proposed approach, without limiting its scope, we specifically designed and deployed a test system to support worker's safety in construction sites, detecting the presence of workers that do not wear the hard hat; we propose here a pedestrian detector for cluttered environments based on the pedestrian classifier with covariance descriptors, initially proposed by Tuzel et al. [6] and described in Section 3.1; on each detected pedestrian, a head localization step is performed to obtain the accurate head position of the targets; even if this module exploits a classifier similar to the one used for pedestrian detection, we propose an innovative approach, that is, the use of polar image transformations and multispectral image derivatives to better exploit the head appearance. On the located heads, a final hard-hat detection algorithm based on color features is performed (see Section 3.2).

The context learning step (see Section 4 and Figure 3) is made of two components: relevance feedback for enriching the training set (Section 4.1) and weak scene calibration (Section 4.2). The domain-specific video surveillance (see Section 5 and Figure 4) exploits the learned context to effectively and efficiently produce an appearance-based people detection, devoted at supporting workers safety in construction sites.

Summarizing, the main novelties of the paper are as follows:

(1) the use of contextual information to improve the classification accuracy of general-purpose boosted classifiers (trained on context-free samples), with context-dependent training data that are autonomously extracted;

(2) the use of contextual information to infer the perspective of the scene, in order to increase speed and accuracy of the detection process;

(3) the use of polar transformations and multispectral derivatives to improve the classification performance of the covariance-descriptor LogitBoost for head detection.

The first and second tasks aim at proving that it is not necessary to design specific detectors for specific applications, but good detection results can be also achieved with general-purpose classifiers with the inclusion of specific contextual information in the process. The third task extends

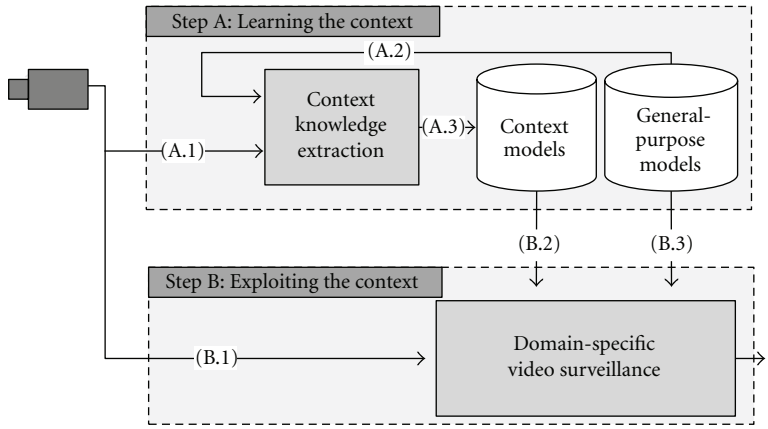


Figure 2: Scheme proposed to exploit context information.

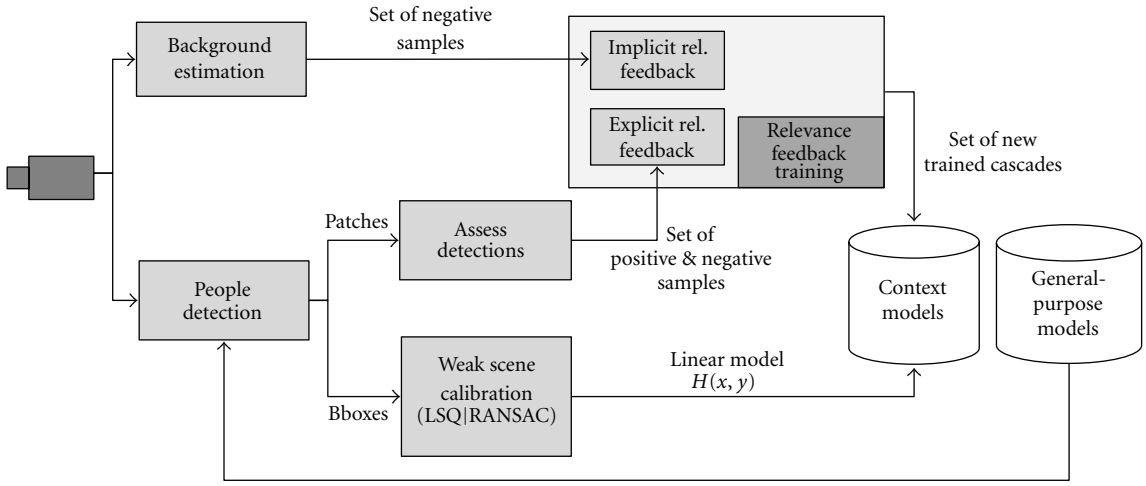


Figure 3: The scheme of the context learning step through domain-specific data. The meaning and use of function $H(x, y)$ are provided in Section 4.2.

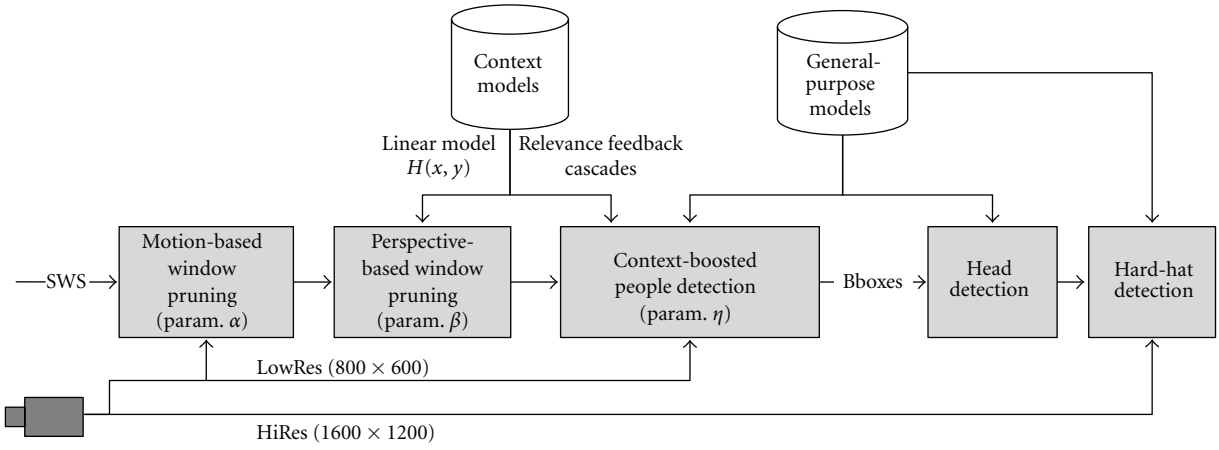


Figure 4: The scheme of the domain-specific video surveillance. The meaning and use of parameters α, β, and η are provided in Section 5.1. The use of multiresolution video data is described in Section 3.2.

the use of covariance descriptors for circular objects, such as heads and hard hats.

2. Related Works

Two classes of approaches have been followed in the literature for people detection [2]. The first one makes use of a model of the human body by looking for body parts in the image and then imposing certain geometrical constraints on them [3]. Their relevant limitation is that they require a sufficiently high image resolution for detecting body parts, and this is not appropriate in contexts like open areas overlooked by long-view cameras. The second class of proposals is based on applying a full-body human detector for all possible subwindows in a given image [4–6]. Then, a dense feature representation can be used, as in [4] where a linear SVM classifier is applied to both densely sampled histograms of oriented gradients (HOGs) and histograms of differential optical flow features inside the detection window. As founding block of our proposal, we adopt the pedestrian classifier based on covariance descriptors proposed by Tuzel et al. [6], for three main reasons: first, it is demonstrated to perform better in per-window classification with respect to the popular HoG SVMs [7]; second, it is based on a rejection cascade of boosting classifier; this architecture benefits of the property that a very reduced portion of the rejection cascade is used when classifying those patches whose appearance strongly differs from the trained model (reducing therefore the computational load); third, the covariance descriptor is very flexible and can be modeled according to the different application contexts; for example, we will exploit two different configurations of the covariance matrix depending on the object to classify (pedestrians or heads).

The context has been used before in computer vision and object recognition [8], especially in unfavorable circumstances, where viewing quality is poor (due to blurring, noise, occlusions, or distractors) and to model several contextual relationships: between-objects relationships, especially in object segmentation [9, 10], objects and surroundings, exploiting perspective and 3D [11, 12], and objects and scene [13], using the statistics of low-level features. The statistical relationship between people and objects in home environments has been exploited in [14] for object recognition, with Markov Logic Networks for incorporating user activities (such as sitting on a chair or watching the TV) as context information. Vice versa, Moore et al. [15] use existing objects in the scene to recognize activities, exploiting Bayesian networks to model their relationship. A more modern way of exploiting context for action recognition is in [16]. Co-occurrence graphs, modeling relations between contextual cues (spoken words or pauses), and visual head gestures are used in [17] for selecting relevant contextual features and inferring the visual features that are more likely in multiparty interactions. Context has also been used for incremental learning, in order to boost performance of object classification, detection, and recognition trained on generic or poor training data [18, 19]. Regarding this latter context exploitation, many of the works propose an "online update" of the trained classifiers through the confidence measure (or margin or rank) of the classifier in order to update the training set, as in [18] (in other words, they use the classifier itself as a mean for self-updating). The underlying assumption is that high (low) confidence detection can be considered surely true positives (negatives); in fact, this consideration is correct from a statistical point of view, but in practice it can be very risky since "strong misclassification" (i.e., true negative with very high confidence or true positives with very low confidences) is definitely not rare event in real-world scenarios, and these occurrences would inject misclassified video data in the updated training data, compromising the classifier result that might drift toward a wrongful classification. In this paper, we propose to update/retrain an object classifier through "domain-specific" visual data that are orthogonal to the features used by the classifiers or by means of manually supervised video data.

Eventually, the head localization could exploit multiview face detection techniques [20–22]; however, the limitation of all these detectors is that a portion of the face must be visible; this part can be even very limited but not null, and the limit is often defined as the profile face. This case clearly does not apply to our working conditions, where the person could even face the opposite direction of the camera or could be frontal but wear a scarf or a protective face gear. With these premises, head localization can be performed with circle detection due to its circular shape. Locating circles in images has been deeply explored in the literature, for both robotic or industrial applications [23] and object classification, for example, traffic sign recognition [24] or 3D object reconstruction [25]. All the proposed methods present two main shortcomings for our purposes: first, they rely on fitting the pixel values or edge points with a certain parametric function. This can be difficult to generalize and is heavily affected by the unfavorable correlation between strong false positives and weak true positives (weak signal problem), that is, a typical limit of parametric approaches such as Hough transforms. For this reason, [24] proposes to measure a curve's distinctiveness through a one-parameter family of curves; in this way, the Hough transform becomes one dimensional and much more accurate since the feature accounts for both the current hypothesis and the curves in the hypothesis's immediate vicinity. The second shortcoming regards the computational complexity. Most of these methods are highly time consuming, tackling the problem as an optimized search in highly dimensional spaces [25]. In [23], the problem is formulated as a maximum likelihood estimator, and the method is proved to be fast and accurate also in the case of occlusions, but it relies on the good extraction of the points describing the curves, being therefore sensitive to noise. A more generic exploitation of the defining shape of objects is proposed in [26]; however, even this method relies on discriminative gradients, and in case they are degraded by visual clutter, low video resolution, and compression artifacts, the results can be quite unreliable. Most of the other head detectors approaches that do not rely on shapes, exploit color features (typically of hair and

skin [27–29]). Differently from all these works, that exploit gradients or color in an exclusive manner, the approach we propose uses appearance-based features, specifically colors and gradients, in a unified manner, through a covariance matrix representation. This approach will be demonstrated to make circle (head) detection suitable also in case where the objects to classify are not easily modeled by parametric curves or precise edges cannot be extracted due to the complexity of the scenes.

3. Covariance Descriptors for Object Classification

In this section, we briefly introduce the use of covariance descriptors for pedestrian classification (Section 3.1), and we propose a new related descriptor suitable for circular objects, such as heads and hard hats (Section 3.2).

3.1. Pedestrian Classification with Covariance Descriptors. Without digging in details, the classifier proposed in [6] is based on a rejection cascade of LogitBoost classifiers (the strong classifiers), each composed of a sequence of logistic regressors (the weak classifiers). The original LogitBoost classifier [30] is modified to account for the fact that covariance matrices do not lie on Euclidean space but on Riemannian manifolds. Each logistic regressor is trained to best separate the covariance descriptors that are computed on *randomly sampled subwindows* of the positive (pedestrian) or negative (nonpedestrian) training images.

Given an input image I and the following 8-dimensional set F of features (defined over each pixel of I):

$$F = \left[x, y, |I_x|, \left|I_y\right|, \sqrt{I_x^2 + I_y^2}, |I_{xx}|, \left|I_{yy}\right|, \arctan \frac{\left|I_y\right|}{|I_x|} \right]^T, \tag{1}$$

where x and y are the pixel coordinates, I_x, I_y and I_{xx}, I_{yy} are, respectively, the first, and the second-order derivatives of the image, it is then possible to compute the covariance matrix of the set of features F for any axis-oriented rectangular patch of I. Regardless of the specific composition of F, this matrix is referred as *covariance descriptor*; it is proved to be a very informative descriptor for several computer vision tasks and, moreover, extremely well suited for pedestrian classification. Furthermore, since the subwindows associated with the logistic regressors are rectangular and axis oriented, the covariance descriptor can be efficiently computed using integral images [31].

The main issue related with covariance matrix approaches is that they lie on a Riemannian manifold, and in order to apply any traditional classifier in a successful manner, it is necessary to map the manifold over an Euclidean space. A detection procedure over a single patch involves the mapping of several covariance matrices (approx. 350) onto the Euclidean space via the inverse of the exponential map [6]

$$\log_\mu(Y) = \mu^{1/2} \log\left(\mu^{-1/2} Y \mu^{1/2}\right) \mu^{1/2}. \tag{2}$$

This manifold-specific operator maps a covariance matrix from the Riemannian manifold to the Euclidean space of symmetric matrices, defined as the space tangent to the Riemannian manifold in μ, that is, the weighted mean of the covariance matrix of the positive training samples. Each matrix logarithm operation (2) requires at least one SVD of an 8×8 matrix, and such operation is known to be computationally demanding.

The original algorithm in [6] proposes to train a cascade of LogitBoost classifier, exploiting logistic regressors as weak classifiers. They also suggest to use the well-known INRIA pedestrian database [7] for training data.

A typical cascade of this classifier is made of 25 stages: each stage is designed to reject approximately 35% of negative samples coming from the preceding stages, and therefore the accumulated rejection ratio over the negative samples at the 25th stage is approximately ($1–0.65^{25}$). We implemented, trained, and tested this precise configuration, that we name as the *general-purpose pedestrian classifier* for the rest of the paper.

Given the binary classifier, the pedestrian detection is performed through a sliding window paradigm, that is based on the idea of passing to the classifier all possible subwindows of an image; since all classifiers in general tend to trigger multiple detections over a single positive object, the detection step is followed by a typical mean-shift-based nonmaximal suppression [32].

This approach is very suitable for pedestrian classification in general-purpose images, but its performances degrade in complex scenes, and thus we enhanced it as it will be fully described in Sections 4 and 5.

3.2. Head and Hard-Hat Detectors. A similar descriptor can be exploited for specific body parts: for instance, detecting the head would help in understanding the presence or absence of the hard hat on the workers, and this is the topic of the present section, where we propose a specific version of the covariance-descriptor classifier for circular objects.

A sliding-window head detection is applied to the upper part of the detected body with a twofold purpose: first, it validates the correctness of the pedestrian detection, that is, the head detector must return exactly one detection otherwise the window is rejected as a false positive produced by the pedestrian detector; second, it locates with precision the position and the scale of the person's head. Since the visual features qualifying head shapes (from any viewing directions) are not strongly discriminative with respect to generic circular shapes, the performance of the classifier is boosted exploiting images with a resolution (indicated as HiRes in Figure 4) that is at least doubled with respect to the one used for the pedestrian detection. This allows the classifier to catch those features that would be lost at lower resolutions. To this aim, the system employs 2 MP (two mega-pixels) cameras and each frame is grabbed at full resolution (typically 1600×1200). The first part of the processing aimed to pedestrian detection is performed on a downsampled version of the frame. Depending on the depth of the viewed scene, the downsampling factor typically

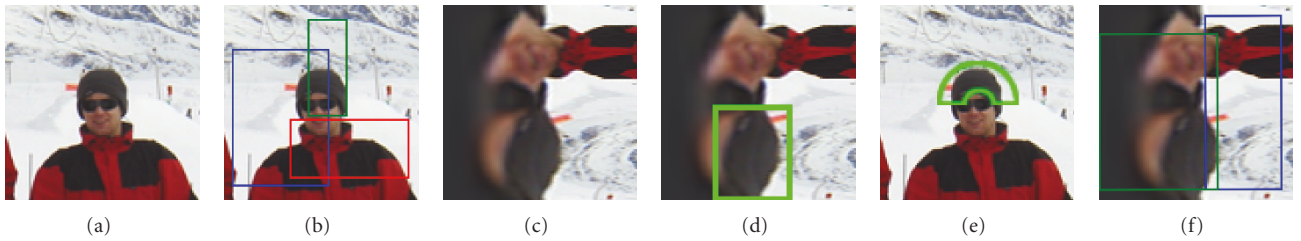


Figure 5: (a) an image I used for head classification; (b) examples of rectangular patches used by weak classifiers according to the original proposal [6]; (c) polar transformation I_p of image I with respect to the center of image (a); (d) rectangular axis-oriented patch on polar image; (e) transformation of the image (d) and its patch onto the original image; (f) examples of rectangular patches used by weak classifiers learned over polar images.

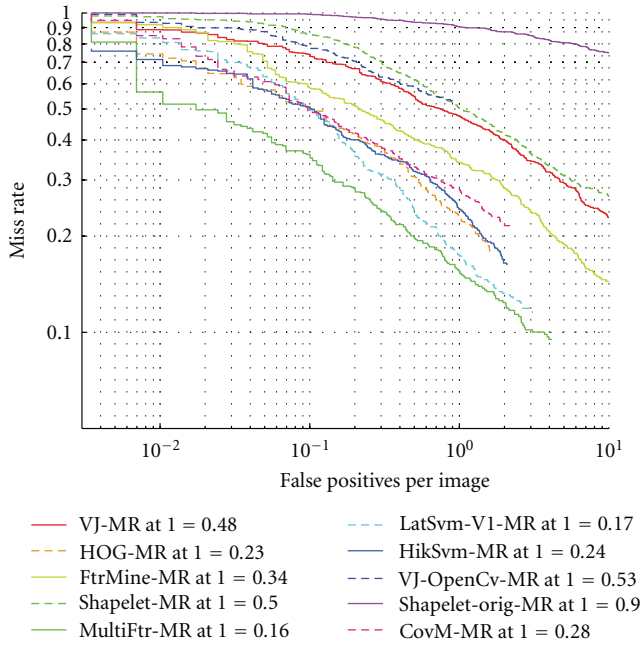


Figure 6: Results of pedestrian detectors on the INRIA pedestrian dataset. MR at 1 means the MR measured at FPPI = 1. The method we employ in our system is marked as CovM. The plots are automatically generated by the tool described in [33]. Please refer to it for other methods.

ranges from 2 to 4. Then, the full resolution is exploited for head and hard-hat detection. This use of multiple resolution images resembles the electronic zooming of EPTZ cameras (Electronic PTZ, [34]), that exploit mega-pixel sensors to reproduce a virtual behavior of pan, tilt, and zoom even if the camera and its focal length are static. In our case, the fundamental advantage of electronic zooming with respect to traditional optical zooming is that the wide (zoom out) and the telescopic (zoom in) images refer exactly to the same time instant; therefore, the spatial detection of the head can be reliably correlated to the pedestrian detection results; conversely, traditional zooming would introduce a time gap between the grabbing of the wide and of the telescopic images (time gap due to the mechanical movement of the optics), and this would introduce uncertainty on any spatial correlation between head and pedestrian detection.

When the covariance descriptor classifier is applied to objects with nonrectangular shape (e.g., holes, heads, wheels, etc.), the performances in terms of classification accuracy degrade due to the inclusion of nondiscriminative pixels within the rectangular patches used by the classifier (see Figures 5(a) and 5(b)).

Aiming to classify circular features, the use of patches with generic circular shapes would catch variations more accurately than just using axis-oriented rectangular shapes. Indeed, using circles or annulus would exclude from the covariance matrix computation all the pixels that do not strictly belong to the circular shape to recognize (see Figures 5(a) and 5(e)). Even if this technique would yield more accurate classification results, the use of nonrectangular or nonaxis-oriented patches would hinder the use of integral images, that are strongly exploited by the classifier for fast covariance matrix computations [31]. This limitation can be solved by the use of polar images; given $I(x, y)$ (i.e., the input image to classify, Figure 5(a)), its polar transformation $I_p(\rho, \vartheta)$ (Figure 5(c)) is computed defining a reference point $C = (x_C, y_C)$ (i.e., the center of the transformation on I), and ρ, ϑ (resp., modulus and angle) as

$$\rho = \sqrt{(x - x_C)^2 + (y - y_C)^2},$$
$$\vartheta = \arctan\left(\frac{y - y_C}{x - x_C}\right). \tag{3}$$

Through this change of variables, the original image I is warped onto I_p, using linear or bicubic blending and outliers filling.

Indeed, given an image and its polar transformation with respect to the image center, any slice of annulus on the original image (centered in the image center) can be represented as an axis-oriented rectangular patch in the polar transformation. Therefore, the polar transformation creates a bridge between the circular patches (useful for classification purposes) and the rectangular patches (needed by the intrinsic classifier architecture); given an image to classify, as first step the polar image transformation is computed and then the weak classifiers are applied on it; specifically, each of them operates on a rectangular patch over the polar image, that represents a slice of annulus over the original

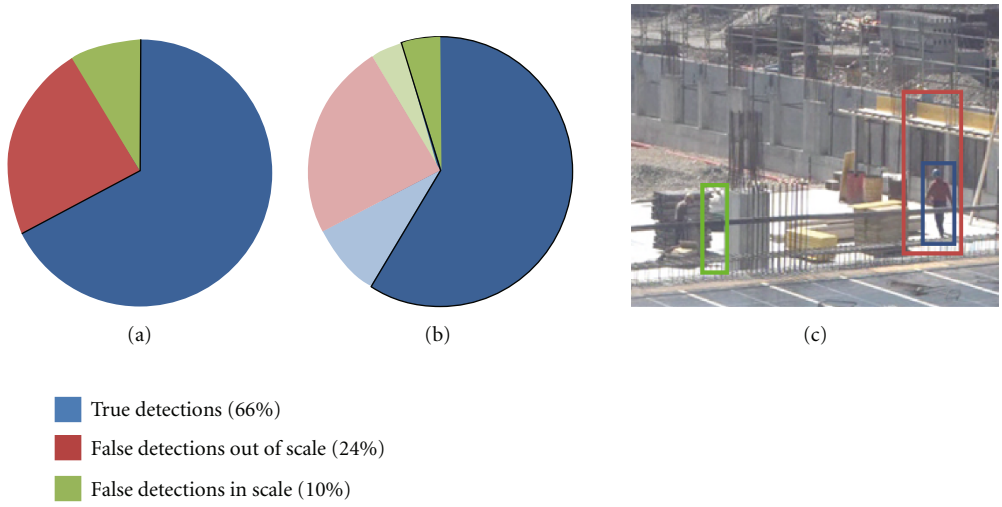


Figure 7: Weak scene calibration through LSQ and RANSAC. (a) Distribution of detections on the training video, (b) consensus set, and (c) visual example of the three types of detections.

image (see Figures 5(d) and 5(e)); this procedure generates a classifier, that we name "polar classifier" (in juxtaposition with the traditional "Euclidean classifier"), more suited to circular shape classification.

A second improvement can be obtained by making use of color information. In fact, in appearance-based object classification, it is common to avoid the use of chrominance since in most cases color does not convey any discriminative information (e.g., in the classification of pedestrians, vehicles, textures, etc.). Instead, since chrominance can be successfully used to compute more accurate edge derivatives with respect to luminance data only [35], we claim that the chrominance can be exploited for image derivative computation in the covariance descriptor, especially for the head classifiers, where the hard-hat positive patches are qualified by strong chrominance. In order to compute covariance descriptors sensitive to luminance and chrominance, we exploit multidimensional gradient methods and define these directional derivatives for the RGB color space

$$I_x^{\mathrm{RGB}} = \sqrt{\left|\frac{\partial R}{\partial x}\right|^2 + \left|\frac{\partial G}{\partial x}\right|^2 + \left|\frac{\partial B}{\partial x}\right|^2},$$
$$I_{xx}^{\mathrm{RGB}} = \sqrt{\left|\frac{\partial^2 R}{\partial x^2}\right|^2 + \left|\frac{\partial^2 G}{\partial x^2}\right|^2 + \left|\frac{\partial^2 B}{\partial x^2}\right|^2}, \tag{4}$$

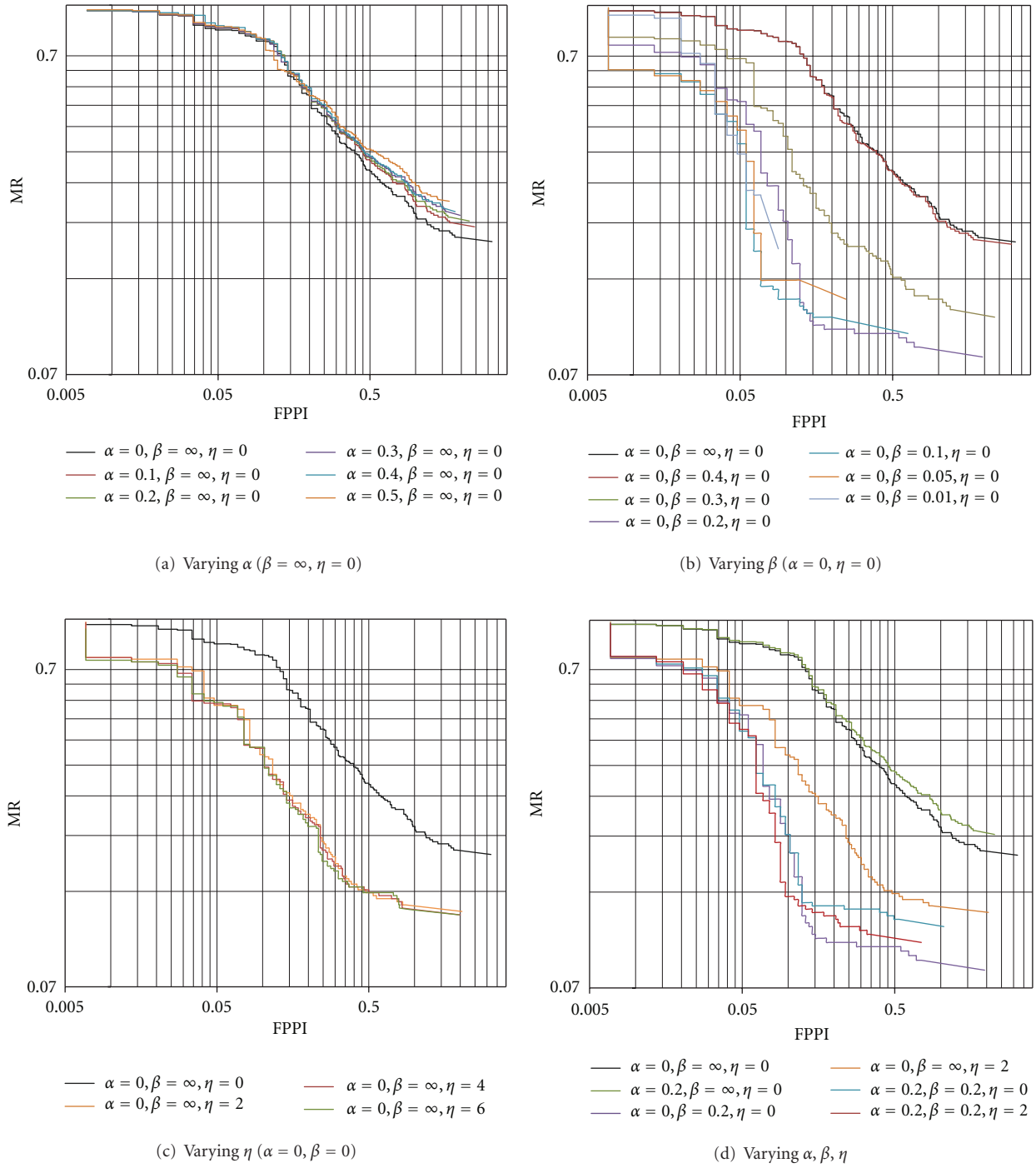
and similarly for I_y^{RGB}, I_{yy}^{RGB} and for Lab color space I_x^{Lab}, I_{xx}^{Lab}, I_y^{Lab}, I_{yy}^{Lab}; at this point, it is straightforward to extend (1) to RGB and Lab color spaces

$$F^{\mathrm{RGB}} = \left[x, y, \left|I_x^{\mathrm{RGB}}\right|, \left|I_y^{\mathrm{RGB}}\right|, \sqrt{\left(I_x^{\mathrm{RGB}}\right)^2 + \left(I_y^{\mathrm{RGB}}\right)^2}, \left|I_{xx}^{\mathrm{RGB}}\right|, \left|I_{yy}^{\mathrm{RGB}}\right|, \arctan\frac{\left|I_y^{\mathrm{RGB}}\right|}{\left|I_x^{\mathrm{RGB}}\right|}\right],$$
$$F^{\mathrm{Lab}} = \left[x, y, \left|I_x^{\mathrm{Lab}}\right|, \left|I_y^{\mathrm{Lab}}\right|, \sqrt{\left(I_x^{\mathrm{Lab}}\right)^2 + \left(I_y^{\mathrm{Lab}}\right)^2}, \left|I_{xx}^{\mathrm{Lab}}\right|, \left|I_{yy}^{\mathrm{Lab}}\right|, \arctan\frac{\left|I_y^{\mathrm{Lab}}\right|}{\left|I_x^{\mathrm{Lab}}\right|}\right]. \tag{5}$$

After having located the position of the head, the classification of hard hats versus heads (see Figure 4) is performed using a minimum distance classifier trained on the average Lab color computed through a Gaussian kernel centered in the centered-upper part of the head; this simple approach yields accurate results, as demonstrated in Section 6.

4. Learn the Context

4.1. Relevance Feedback for Additional Training. Recent pedestrian classifiers in general have reached remarkable performances, since miss rates (MRs) of approximately 5% are obtained yielding approximately 1 false positive every 100 K tested windows (10^{-5} False Positives Per Window (FPPW), [4, 6]); however, when these techniques are applied to exhaustive, multiscale people search through sliding window approach, the performances quickly drop; indeed, even tolerating 1 false positive per image (FPPI) on average, it is very challenging to obtain MR lower than 20% [33]; this drop of performance is due to the exhaustive search over the image, that generates a *large quantity of false positives* caused by video clutter and distractors that were not present in the negative training set of the classifier. Therefore, our proposal is to limit this rate of false positives, through the exploitation of context visual data. In particular, we design a procedure to enrich the general-purpose training dataset, that has been formerly used to train the pedestrain classifier, with context-dependent additional data; this *biased* training set is used



(a) Varying α ($\beta = \infty$, $\eta = 0$)

(b) Varying β ($\alpha = 0$, $\eta = 0$)

(c) Varying η ($\alpha = 0$, $\beta = 0$)

(d) Varying α, β, η

Figure 8: DET curves at different α, β, η.

to produce a new classifier that is more robust, within the specific context, to clutter and distractors.

As depicted in Figure 3, the relevance feedback training is fed with two different contributions; the first, called *implicit RF,* is totally autonomous, a background estimator provides a set of background images that do not contain moving objects by definition (specifically people); this video data is then suitable to enrich the negative training set. The second, called *explicit RF*, requires a user assessment; after having run the general-purpose pedestrian classifier on a video sequence, the assessor is requested to separate true from false positives, which are, respectively, used to enrich the positive and the negative training sets.

Since the training phase of the whole pedestrian classifier can be very time (and memory) consuming, it is advisable to use a pedestrian classifier based on a rejection cascade; indeed, its multistage architecture allows to limit the retraining step to the latest stages of the classifier; this choice

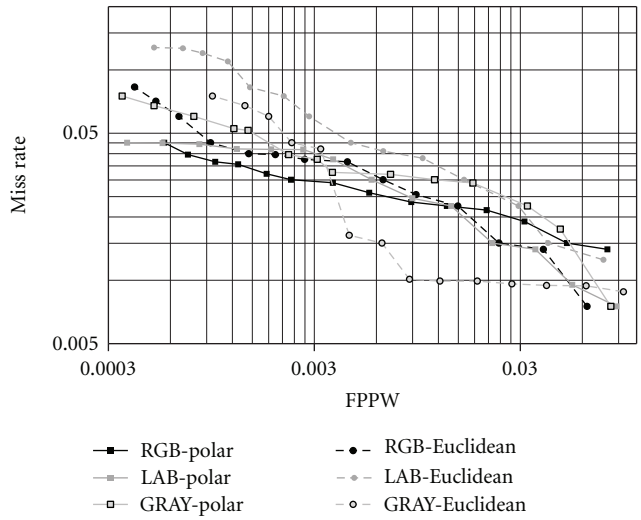


Figure 9: DET curves on the Head Image Dataset. Each marker represents the performance up to a cascade level. The markers at the bottom-right corner represent the results using up to 5th cascade; adding more cascades, the markers move toward the upper-left corner, up to the 18th cascade. The more cascades are introduced, the lower the FPPW, the higher the miss rate.

Table 1: Values of α, β, and percentage of windows rejected; the variation on η is not considered here since it does not affect the window pruning.

	Base	Varying α					
α	0	0.1	0.2	0.3	0.4	0.5	
β	∞	∞					
%	0%	78%	83%	87%	90%	92%	
	Base	Varying β					
α	0	0					
β	∞	0.4	0.3	0.2	0.1	0.05	0.01
%	0%	45%	58%	72%	86%	93%	99%
	Base	Optimal					
α	0	0.2					
β	∞	0.2					
%	0%	92%					

decreases the time required for the context-dependent retraining, that is, just a small fraction of the time required to retrain the whole classifier. As an example, in the proposed test case (Section 6), the retraining involves from 2 to 6 stages only, requiring, respectively, 7% and 19% of the training time that would be necessary for the whole classifier.

Retraining only the latest stages of the cascades makes it obviously impossible to raise the performance on the false negatives (because of the nature of rejection cascades, what was wrongly rejected by the first stages cannot be recovered then), conversely it is still possible to act in a strong manner on the reduction of the false positives, that is exactly the problem to tackle.

4.2. Weak Scene Calibration. Let us name the set of all possible windows as *Sliding Windows Set* or *SWS*; this set spans over the whole space of window states (typically position and scale), and its cardinality depends on the size of the image, on the range of scales to check, and on the stride of scattering of the windows. Since we make no assumption on the observed scene, the size of humans is totally unknown and the range of the searched scales is fairly wide (typically tens of scale steps), regarding the strides, to obtain a successful detection process, the SWS must be rich enough so that at least one window targets each pedestrian in the image and this depends on the region of attraction of the classifier (typical stride for position is 4 to 8 pixels, for scale is 1.05 to 1.2). The cardinality of SWS is therefore very high (approximately 50 K/100 K windows for each frame) and since each window is passed through a classification procedure, maintaining real-time processing becomes a critical issue. Thus, the SWS should be pruned *before* the classification step, by means of context information; we propose here to exploit the perspective of the observed scene.

Hoiem et al. [11] define a statistical framework to automatically retrieve the scene perspective in order to focus the detection tasks at the right scales. Borrowing the geometric consideration in that paper, we assume the following hypotheses:

(1) all the people move on the same ground plane;

(2) people are in standing position, and all the observed people are assumed to have consistent physical height;

(3) camera tilt is small to moderate, and camera roll is zero or image is rectified;

(4) camera intrinsic parameters are typical of rectilinear cameras (zero skew, unit aspect ratio, and typical focal length).

Hypothesis (2) comes with the definition of pedestrians, and focusing our attention to adult people detection, we can assume without loss of generality that the difference on people height is negligible; hypothesis (3) is satisfied because in our context the cameras are installed with very low tilt in order to observe wide areas and given an initial system configuration. By employing cameras with fixed focal length and by compensating the other camera parameters with an intrinsic calibration, hypothesis (4) is satisfied too.

Finally, in case hypothesis (1) is satisfied, it is correct to approximate the height (in pixels) of the human silhouette with a linear function H in the image coordinates (x, y), that represent the point of contact of the person with the ground plane: defining h_w as the height of the person in world coordinates, h_c as the height of the camera from the ground plane, f as the focal length, θ as the tilt angle of the camera with respect to the ground plane, y_t and y_b as the y image coordinates of the top and bottom of the pedestrian,

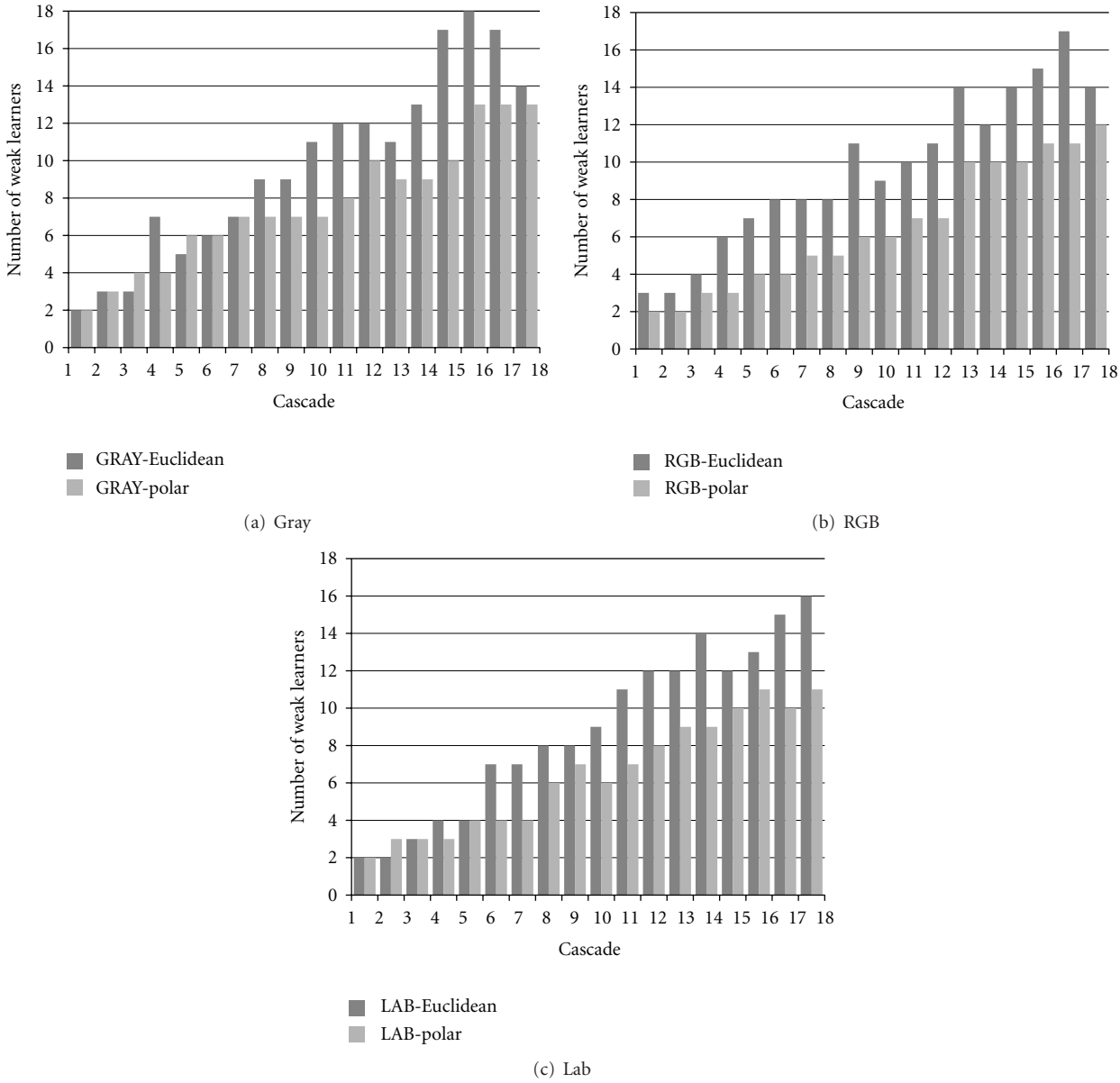


Figure 10: Number of weak classifiers per cascade for the 6 classifiers: (a) gray (Euclidean/polar), (b) RGB (Euclidean/polar), and (c) Lab (Euclidean/polar).

and y_c and y_0 as the y image coordinates, respectively, of the optical center and of the vanishing point, we can write [11]

$$h_w = \frac{\left(\dfrac{fh_c(f\sin\theta - (y_c - y_t)\cos\theta)}{f\sin\theta - (y_c - y_b)\cos\theta} - fh_c\right)}{(y_c - y_t)\sin\theta + f\cos\theta}. \quad (6)$$

Given hypothesis (3), we can introduce the following approximations: $\cos\theta \approx 1$, $\sin\theta \approx \theta$, $\theta \approx (y_c - y_0)/f$, and $(y_c - y_0)(y_c - y_t)/f^2 \approx 0$, and (6) can be simplified as follows:

$$h_w = h_c \frac{y_t - y_b}{y_0 - y_b}. \quad (7)$$

Refer to [11] for the details on the errors introduced by these approximations. Given hypothesis (2), h_w is a constant; therefore through (7), the height of a person in pixel $(y_t - y_b)$ is linearly correlated with y_b, that is, the contact of the person with the ground plane. To tolerate camera roll or a slanting ground plane, it is enough to introduce in the formula also the x coordinates.

By estimating the parameters of the $H(x, y)$ function, we can prune the SWS by discarding all the windows whose height significantly differs from the estimated function.

In case the hypothesis (1) is violated (e.g., construction workers on scaffoldings move on multiple parallel planes), it is still possible to perform perspective pruning by partitioning the image in areas and accepting the rougher assumption that the height (in pixels) of the people inside each area is almost constant.

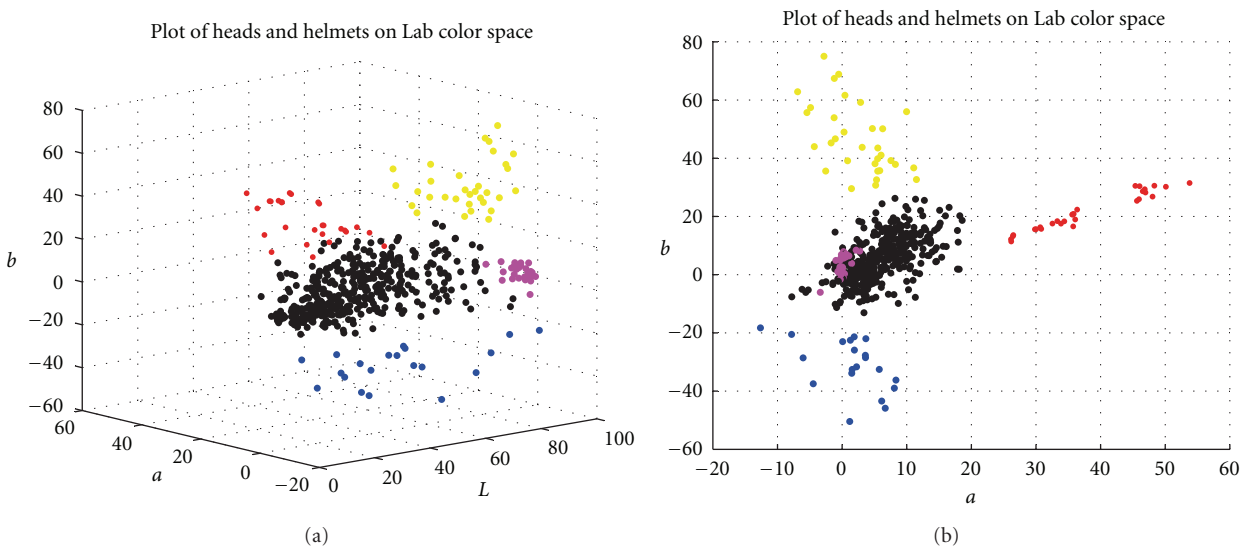


Figure 11: Scattering of the average Lab color of the 527 patches of heads and hard hats. (a) 3d view; (b) 2d view of ab; black dots: bare heads or heads with headgears; blue, red, yellow dots: heads with helmet of corresponding color; magenta: white helmets.

Figure 12: Example snapshots. (a, b) Pedestrian detection; (c–f) head detection; blue box: head position blindly estimated from the pedestrian detection; red box: head position obtained with the head detector; final detections of (g, h) bare heads and of hard hats (i, j).

Table 2: Aggregated results of the complete system.

(a)

	System config ($\alpha\|\beta\|\eta$)	Speedup		Ground Truth Data		Pedestrian detection performance		
		Euclid. head D.	Polar head D.	Ped with hard-hat	Ped w/out hard-hat	Detected	Missed	False Pos.
1	0.0|∞|0	1.00	1.04	5199 on 16646 (31.2%)	11447 on 16646 (68.8%)	3732 (22.4%)	12914 (77.6%)	612
2	0.2|∞|0	3.59	4.11			3536 (21.2%)	13110 (78.8%)	
3	0.0|0.2|0	2.67	2.95			13553 (81.4%)	3093 (18.6%)	
4	0.0|∞|2	0.96	0.99			10460 (62.8%)	6186 (37.2%)	
5	0.2|0.2|0	5.00	6.05			13620 (81.8%)	3026 (18.2%)	
6	0.2|0.2|2	4.89	5.90			14388 (86.4%)	2258 (13.6%)	

(b)

	System config ($\alpha\|\beta\|\eta$)	Head detector				Hard-hat detector			
		Pedestrian validator		Head detection performance		Ped. with hard-hat		Ped. w/out hard-hat	
		Euclid. head D.	Polar head D.	Euclid. head D.	Polar head D.	Euclid. head D.	Polar head D.	Euclid. head D.	Polar head D.
1	0.0|∞|0	318 on 612 (52.0%)	499 on 612 (81.5%)	3451 (20.7%)	3564 (21.4%)	2291 (20.0%)	2365 (20.7%)	1045 (20.1%)	1080 (20.8%)
2	0.2|∞|0			3270 (19.6%)	3376 (20.3%)	2170 (19.0%)	2241 (19.6%)	991 (19.1%)	1023 (19.7%)
3	0.0|0.2|0			12534 (75.3%)	12942 (77.7%)	8319 (72.7%)	8590 (75.0%)	3797 (73.0%)	3920 (75.4%)
4	0.0|∞|2			9674 (58.1%)	9988 (60.0%)	6421 (56.1%)	6630 (57.9%)	2931 (56.4%)	3026 (58.2%)
5	0.2|0.2|0			12596 (75.7%)	13005 (78.1%)	8360 (73.0%)	8632 (75.4%)	3816 (73.4%)	3940 (75.8%)
6	0.2|0.2|2			13305 (79.9%)	13738 (82.5%)	8831 (77.1%)	9119 (79.7%)	4031 (77.5%)	4161 (80.0%)

The novelty we introduce in this passage is related to the cue we propose to exploit in order to learn the perspective; differently from [11], that recovers the perspective using probabilistic estimates of 3D geometry, both in terms of surfaces and world coordinates, we propose to obtain the automatic weak scene geometry calibration (i.e., $H(x, y)$) exploiting the responses of the pedestrian classifier only (the procedure potentially works with any classifier).

During the context learning phase (see Figure 3, *weak scene calibration*), the people detector is run over a video that must contain, among other objects, also some people; all the bounding boxes detected as positives are passed to an LSQ (Least SQuare) estimator that, through RANSAC, discards the outliers (due to out-of-scale false detections) and retains a consensus set made of the windows which contribute to the correct estimation of H. Detailed results are provided in Section 6. In case the estimated model is not sustained by a wide enough consensus set or the error of the inliers relative to the model is large, we deduce that the hypothesis (1) is violated and the system configures itself to detect pedestrians on multiple parallel planes, as aforementioned.

5. Exploit the Context

5.1. Domain-Specific Video Surveillance. The domain-specific video surveillance layer performs people detection exploiting both general-purpose and context-dependent models (see Figure 4). The first block, named *motion-based window pruning*, reduces the cardinality of the SWS, focusing the people detection on the regions where motion has been detected at present or in the recent past. To this aim, we first extract the instant Motion Detection (MD_t); in case the camera is fixed, it is enough to employ a frame differencing approach, but a more sophisticated approach based on background suppression is used [36]. If a PTZ camera with patrolling motion is employed, the frame differencing is preceded by a motion compensation step, that is based on a projective transformation whose parameters are obtained from the frame-to-frame matching of visual features (see [37]). Then, to account for the accumulation of motion in time (and, thus, considering also regions where the motion was present in the recent past), we exploit the *Motion History Image* (MHI_t) introduced in [38]

$$\mathrm{MHI}_t(i,j) = \begin{cases} \tau & \text{if } \mathrm{MD}_t(i,j) = 1, \\ \max\left(0, \mathrm{MHI}_{t-1}(i,j) - 1\right) & \text{otherwise,} \end{cases} \tag{8}$$

where the parameter τ represents the duration period over which the motion is integrated.

For each frame, the SWS is pruned of all the windows with motion ratio lower than a threshold α. Motion ratio is computed as the count of nonzero MHI pixels inside the window divided by the window area. This provides a good tradeoff between searching all over the image versus limiting the search to current moving regions only. Even if the motion information is not extremely accurate and generates an MHI that is redundant (typical in outdoor scenarios with moving cameras), the recall of the system is not affected since the appearance-based pedestrian detector does not depend on the motion segmentation.

A further pruning is performed exploiting the perspective model of pedestrian height $H(x, y)$ (Section 4.2). Since this model contains several approximations (i.e., height of people approximated to a constant value, geometric assumptions on the camera viewing direction, errors due to automatic estimation, etc.), the perspective pruning is controlled as follows: (x, y) is the estimated feet position of the potential pedestrian contained in a window to be classified; if the gap between the height was estimated with the perspective model $H(x, y)$ and the window height is beyond threshold β, the window is pruned. To obtain a normalized measure, the gap is divided by $H(x, y)$.

The windows which survive motion and perspective pruning are passed to the pedestrian classifier described in Section 3; in our domain-specific classifier, we train η additional stages with the context-dependent training data as described in Section 4.1; the first 25 stages, that belong to the general-purpose classifier, yield approximately a rejection ratio of $(1–0.65^{25})$ on generic negatives; the last retrained η stages generate a further rejection ratio $(1–0.65^{\eta})$, that is specialized in rejecting context-specific clutter and distractors. The threshold η can be chosen according to the classification complexity of the visual context and to the time that is available for retraining the additional cascades.

Summarizing, we employ three parameters that are tuned depending on the degree of trust that is granted to the observed context. SWS pruning is regulated through α and β: the first exploits the motion of the objects, the second uses the estimated perspective. The classifier is biased towards a view-dependent pedestrian detection through η. The effect of these parameters on the system performance is thoroughly analyzed in the following section.

6. Experimental Results

We tested the described approach over videos recorded in a construction working site of approximately 25000 m^2, over a time span of 3 months; the scenario changed from an open field with some machineries to a roughly completed building. The videos were grabbed at 3 fps, 1600×1200, for a total of 34 minutes and 6120 frames of test-set video with annotated ground truth of pedestrian bounding boxes (available at request). The pedestrian classifier is trained on the INRIA pedestrian dataset [4], while for the head classifier, we generated a *Head Image Dataset* made of 1162 positives and 2438 negatives for training, and 266 positives and 906 negatives for testing (the authors are going to make the dataset publicly available). The positive set is made of patches of fixed size (96×96) containing heads with and without headgears, at any viewing direction and placed in the patch center. The classifier used to separate bare heads from heads with hard hats is trained using the Lab color space on 527 patches (399 heads and 128 hard hats).

The accuracy of pedestrian detection is measured as Miss Rate (MR) versus False Positives Per Images (FPPIs), while the head classifier is measured on windows basis, comparing MR versus False Positives Per Windows (FPPW); the latter measurement is preferred for measuring classifier performance, while the former is used for assessing object detection on images or video; in both cases, we plot the performances using Detection Error Trade-Off (or DET) curves [39], showing the trend of the MR (i.e., the reciprocal of the detection rate) versus a false positive rate, varying a system parameter (typically detection confidence or number of stages employed in a classifier).

The matching of the bounding box found by the detector (BB_{dt}) with the bounding box in the ground truth (BB_{gt}) as defined in the PASCAL object detection challenge [40] which states that the ratio between the area of overlap of BB_{dt} with BB_{gt} and the area of merge of the two BBs must be greater than 50%; multiple detections of the same ground-truthed person, as well as a single-detection matching multiple ground-truthed people, are affecting the performance in terms of MR and FPPI.

The overall performance of our implementation of Tuzel's pedestrian detector [6] is shown in Figure 6, that highlights how a few, more recent pedestrian detectors perform better than the one employed in our system. These results do not affect the claims of our paper, since these are not related to the specific performance of the employed pedestrian classifier; however, they demonstrate that the experimental results on pedestrian detection (that are following in this section) are close to optimal and can be slightly improved employing more modern classifiers.

Regarding the weak scene calibration, Figure 7(a) shows the distribution of the pedestrian detection results with respect to perspective considerations. Indeed, the positive detections are made of true and false positives; while the former are in scale by definition, the latter can be in scale or out of scale (Figures 7(a) and 7(c)); all these positives are passed to the LSQ and RANSAC; as shown in Figure 7(b), the extracted consensus set excludes all the out-of-scale false positives, proving the effectiveness of the learned model.

The effect of motion and perspective pruning on the accuracy of pedestrian detection is evaluated in Figures 8(a) and 8(b), where the DET curve is plotted at different values of α and β, that are used to tune the degree of window-pruning exploiting, respectively, motion and weak calibration; with $\alpha = 0$ and $\beta = \infty$, the pruning is completely inhibited (i.e., traditional sliding window spanning over the whole space of windows states), and it is gradually enabled increasing α and decreasing β. The increase of α slightly affects the detection accuracy (Figure 8(a)), conversely the decrease of β significantly improves it (Figure 8(b)), since out-of-scale windows are rejected. However, β should be tuned in order to be tolerant with respect to the several approximations introduced by the weak calibration. Indeed, there is a critical boundary for β (between 0.05 and 0.1); moving below that value, the accuracy degrades (the perspective pruning becomes too strict). Table 1 shows the percentage of pruned windows with respect to the complete SWS (i.e., $\alpha = 0$ and $\beta = \infty$). As expected, the higher the α and the lower the β, the stronger is the window pruning and, therefore, the reduction of computational load.

The performance of the additional cascades trained with relevance feedback approach is evaluated in Figure 8(c). The higher is the parameter η, the more additional cascades are

trained, the longer training time is required, and the higher is the gain in accuracy; however, the significant improvement is from $\eta = 0$ to $\eta = 2$; adding more cascades does not significantly modify the accuracy. In these tests, we used a very limited additional training set (only 1 background image coming from the implicit RF and 100 patches coming from the explicit RF; both are extracted from a validation set), but even such limited additional training data generate a significant gain. The validation set is a video sequence recorded with the same camera in the same viewing position but at different time of the test-set video.

An optimal trade-off between improvement of accuracy and reduction of computational load is obtained with $\alpha = 0.2$, $\beta = 0.2$, and $\eta = 2$, (see Figure 8(d)); taking as reference FPPI = 0.1, this setup processes on average 13433 windows per frame and generates an MR = 0.14, outperforming the traditional pedestrian detection that, without exploiting, any contextual information ($\alpha = 0$, $\beta = \infty$, $\eta = 0$) processes 168387 windows per frame (12.5 times higher) and generates an MR = 0.78 (3.7 times higher).

Regarding the classification of circular features by means of polar transformation and multispectral derivatives, Figure 9 plots the results of the classifiers applied over the Head Images Dataset; to perform head detection in our final application, we trained a rejection cascade made of 18 LogitBoost classifiers. We verified that adding more stages just saturated classification performance. At the first stages of the rejection cascades (right side of the DET graph), the Euclidean configuration provides better results. However, the working point of such classifiers is to be sought at the very left hand side, where the most of the stages are employed and the lowest false positives rates are obtained. Regardless of the chosen image derivative, the last cascades of the polar classifiers always yield better results with respect to the Euclidean classifiers. Moreover, the use of polar transformation generates lighter classifiers that will benefit the detection process with a lower computational load (on average, over the three color spaces, polar classifiers use 23% less weak classifiers; see Figure 10). The use of color brings further increase in performance; overall best performances at the lowest false positives rates are obtained with the polar classifier using Lab image derivatives, that has a *miss rate* (MR) of 4.5% (w.r.t. 7.5% of the Euclidean classifier over gray values), a *False Positives Per Window* (FPPW) of 0.037% (w.r.t. 0.135%), and 33.5% less weak classifiers.

Finally, the two classes, bare heads (or headgears), and hard hats are clearly separated in the Lab color space (see Figure 11), and a simple minimum distance classifier obtains satisfying performances, since both precision and recall are above 90% (MR = 10%). Most of the errors are generated by misclassifications of white hard hats and of white-haired persons. Indeed, removing the white hard-hat class, the classifier reaches precision and recall of approximately 97% (MR = 3%) using only chrominance information (L can be discarded).

The aggregated results of the whole system, as depicted in Figure 4, are summarized in Tables 2(a) and 2(b). The six rows represent the configurations of the 6 parameters α, β, and η as proposed in Figure 8(d); the first row shows the system with no use of motion, perspective, or relevance feedback. Rows 2, 3, and 4, respectively, test the impact of motion, perspective, and relevance feedback independently among each others. Row 5 tests motion and perspective together and eventually row 6, that yields the best results both as speedup and accuracy and puts them all together. For the sake of evaluation, we configured the pedestrian detector to work at FPPI = 0.1, therefore producing the same number of false positives regardless of the configuration of α, β, and η.

Both Euclidean and polar head detectors have been configured to exploit the whole cascade made of 18 stages. For Euclidean configuration, we employed the derivatives on luminance only (as in (1)), resembling the traditional proposal of Tuzel's classifier. For the polar configuration, we employed the derivatives on Lab color spaces (as in (5)). The head detector is exploited twice: first, as pedestrian validator to reduce the number of pedestrian false positives (as described in Section 3.2), then to localize the precise head position; percentage numbers in the columns referring to the head detector refer to the number of detected head with respect to the total number of heads (i.e., pedestrians) in the dataset. Eventually, regarding the hard-hat detector, we employed here the version without the white hard-hat class. The percentage numbers refer to the number of detected hard hats (bare heads) with respect to the total number of persons with (without) hard hat. Figure 12 shows examples of the correct outcome of the complete system. The whole system, configured with α, β, and η as in Row 6 of Tables 2(a) and 2(b) is able to process in real time a 1600×1200 video stream at approximately 1 fps on a single core of a modern desktop PC.

7. Conclusions

The paper introduces a framework that exploits context visual information to enhance object classifiers trained on generic and unbiased datasets; specifically, the proposal is to infer scene perspective through the response of a generic object (e.g., pedestrian) detector and to refine the generic classifier through an additional training step based on a context-dependent dataset. On top of these two techniques, the system also exploits motion, to further speed up the detection process, and multispectral derivatives, to increase the accuracy of the covariance-descriptor classifier.

After having detected pedestrians, a head detector is employed to obtain the precise head position. The head appearance is dominated by a circular shape and the paper proposes the use of polar image transformation to better exploit this feature during classification. Furthermore, the use of multispectral image derivatives provides better classification results with respect to luminance derivatives. The experimental results are evaluated in the scenario of construction working sites, where a prototype to support worker's safety in construction sites has been deployed; in particular, the system detects workers that do not wear the compulsory hard hat.

Acknowledgments

This work is currently under development and improvement within the Project THIS (no. JLS/2009/CIPS/AG/C1-028), with the support of the Prevention, Preparedness and Consequence Management of Terrorism and other Security-related Risks Programme European Commission—Directorate-General Justice, Freedom and Security. This Project is also partially funded by Regione Emilia-Romagna under the PRRIITT funding scheme and in collaboration with the company Bridge.129 SpA.

References

[1] H. Aghajan, R. Braspenning, Y. Ivanov et al., "Use of context in vision processing: an introduction to the ucvp 2009 workshop," in *Proceedings of the Workshop on Use of Context in Vision Processing (UCVP '09)*, pp. 1–3, ACM, New York, NY, USA, 2009.

[2] D. M. Gavrila, "The visual analysis of human movement: a survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[3] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[4] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proceedings of the 9th European Conference on Computer Vision (ECCV '06)*, vol. 3952 of *Lecture Notes in Computer Science*, pp. 428–441, 2006.

[5] J. Tao and J.-M. Odobez, "Fast human detection from videos using covariance features," in *Proceedings of the ECCV Visual Surveillance Workshop (ECCV-VS '08)*, 2008.

[6] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 886–893, June 2005.

[8] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.

[9] A. Gupta and L. S. Davis, "Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifiers," in *Proceedings of the 10th European Conference on Computer Vision (ECCV '08)*, vol. 5302 of *Lecture Notes in Computer Science*, pp. 16–29, 2008.

[10] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proseedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, October 2007.

[11] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 3–15, 2008.

[12] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool, "Dynamic 3D scene analysis from a moving vehicle," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07)*, June 2007.

[13] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[14] C. Wu and H. Aghajan, "Using context with statistical relational models: object recognition from observing user activity in home environment," in *Proceedings of the Workshop on Use of Context in Vision Processing (UCVP '09)*, pp. 1–6, 2009.

[15] D. J. Moore, I. A. Essa, and M. H. Hayes, "Exploiting human actions and object context for recognition tasks," in *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV '99)*, pp. 80–86, September 1999.

[16] A. Gupta and L. S. Davis, "Objects in action: an approach for combining action understanding and object perception," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[17] L. P. Morency, "Co-occurrence graphs: contextual representation for head gesture recognition during multi-party interactions," in *Proceedings of the Workshop on Use of Context in Vision Processing (UCVP '09)*, ACM, November 2009.

[18] A. Kembhavi, B. Siddiquie, K. Cornelis, R. Mieziank, S. McCloskey, and L. Davis, "Scene it or not? incremental multiple kernellearning for object detection," in *Proceedings of the International Conference on Computer Vision*, 2009.

[19] L. J. Li, G. Wang, and F. F. Li, "Optimol: automatic online picture collection via incremental model learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, 2007.

[20] M. Viola, M. J. Jones, and P. Viola, "Fast multi-view face detection," in *Proceedings of the Computer Vision and Pattern Recognition*, 2003.

[21] S. Li, L. Zhu, Z. Zhang et al., "Statistical learning of multi-view face detection," in *Proceedings of the 7th European Conference on Computer Vision (ECCV '02)*, 2002.

[22] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 446–453, October 2005.

[23] I. Frosio and N. A. Borghese, "Real-time accurate circle fitting with occlusions," *Pattern Recognition*, vol. 41, no. 3, pp. 1041–1055, 2008.

[24] Y. C. Cheng, "The distinctiveness of a curve in a parameterized neighborhood: extraction and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1215–1222, 2006.

[25] K. Kanatani and N. Ohta, "Automatic detection of circular objects by ellipse growing," *International Journal of Image and Graphics*, vol. 36, 2001.

[26] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[27] Z. Zhang, H. Gunes, and M. Piccardi, "Head detection for video surveillance based on categorical hair and skin colour models," in *Proceedings of the 16th IEEE International Conference on Image Processing (ICIP '09)*, pp. 1137–1140, November 2009.

[28] M. Zhao, D. hua Sun, and W. mei Fan, "Hair-color model and adaptive contour templates based head detection," in *Proceedings of the 8th World Congress on Intelligent Control and Automation (WCICA '10)*, pp. 6104–6108, July 2010.

[29] J. Garcia, N. da Vitoria Lobo, M. Shah, and J. Feinstein, "Automatic detection of heads in colored images," in *Proceedings of the 2nd Canadian Conference on Computer and Robot Vision*, pp. 276–281, May 2005.

[30] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[31] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: a fast descriptor for detection and classification," in *Proceedings of the 9th European Conference on Computer Vision*, pp. 589–600, 2006.

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

[33] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: a benchmark," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 304–311, June 2009.

[34] F. Bashir and F. Porikli, "Collaborative tracking of objects in eptz cameras," in *Visual Communications and Image Processing*, vol. 6508 of *Proceedings of SPIE*, 2007.

[35] M. A. Ruzon and C. Tomasi, "Color edge detection with the compass operator," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99)*, vol. 2, pp. 160–166, June 1999.

[36] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337–1342, 2003.

[37] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[38] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[39] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV '99)*, pp. 1895–1896, 1997.

[40] J. Ponce, T. Berg, D. Everingham et al., *Dataset Issues in Object Recognition*, Springer, 2006.