Perspective and Appearance Context for People Surveillance in Open Areas

Giovanni Gualdi
D.I.I. Univ. of Modena
and Reggio Emilia, Italy
giovanni.gualdi@unimore.it

Andrea Prati
D.I.S.M.I. Univ. of Modena
and Reggio Emilia, Italy
andrea.prati@unimore.it

Rita Cucchiara
D.I.I. Univ. of Modena
and Reggio Emilia, Italy
rita.cucchiara@unimore.it

Abstract

Contextual information can be used both to reduce computations and to increase accuracy and this paper presents how it can be exploited for people surveillance 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 exploits covariance descriptors through a LogitBoost classifier on Riemannian manifolds. The approach has been tested on a construction working site where complexity and dynamics are very high, making human detection a real challenge. The experimental results demonstrate the improvements achieved by the proposed approach.

1. Introduction

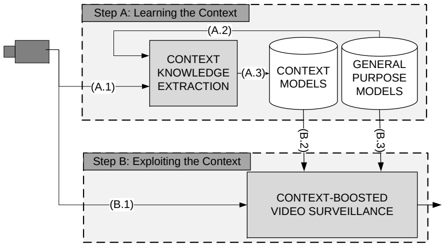
The research in computer vision and pattern recognition must face very difficult problems where reproducing the accuracy of the human vision system is challenging and the accuracy requirement is often coupled with tight time constraints. For this reason, it is always helpful to exploit contextual information 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).

In the field of computer vision, the context has been explicitly used for 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. For instance, the awareness of the scene can be exploited for assessing the likelihood of a certain event or of the presence of a given object, or for completely discarding a set of hypotheses, reducing search time and computational burden (e.g., if we know that we are observing a kitchen, the event of "running" or the presence of a car are improbable).

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. 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.

The paper aims at showing that in such challenging conditions it is possible to avoid ad-hoc solutions: the adoption of state-of-the-art machine-learning approaches, enhanced with contextual information are proved to provide successful results. More specifically, we exploit two contextual types of information: (i) a *relevance feedback* (RF) strategy which enriches the human detection phase by replacing the final stages of a cascade of classifiers (that have been trained for generic human 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 is depicted in Fig. 1: it is divided in two parts, namely (A) learning and (B) exploiting the context. (A) makes use of video data and of general-purpose models to extract new and refined models that better fit the observed visual context (Fig. 2). These new models are stored and then used during the (B) step, where video data *coming from the same context* is processed using both general-purpose and context-dependent models to produce video analysis for surveillance purposes



Step A: Learning the Context
(A.2)
CONTEXT KNOWLEDGE EXTRACTION
(A.3)
CONTEXT MODELS
GENERAL PURPOSE MODELS
(A.1)
(B.2)
(B.3)
Step B: Exploiting the Context
(B.1)
CONTEXT-BOOSTED VIDEO SURVEILLANCE

Figure 1. Scheme proposed to exploit context information.

978-1-4244-7028-0/10/$26.00 ©2010 IEEE

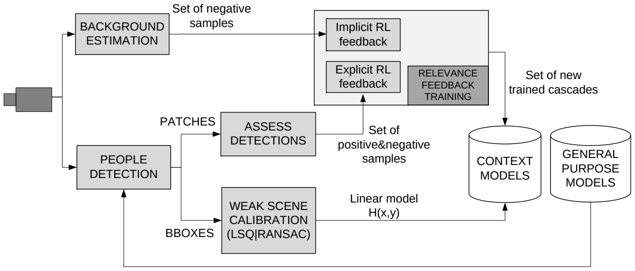


Figure 2. The scheme of the context learning step.

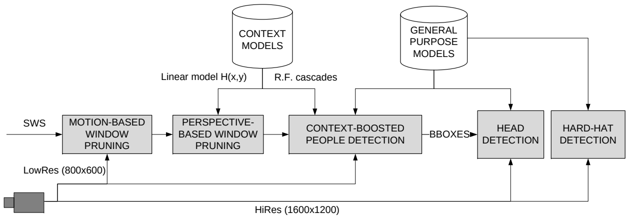


Figure 3. The scheme of the context-boosted video surveillance.

(Fig. 3). Without limiting the scope of the proposal, for the sake of validating the proposed approach, we specifically designed and deployed a test system for appearance-based people detection to support worker's safety in construction sites, detecting the presence of workers without hard-hat.

The context learning step (see Sec. 3 and Fig. 2), is made of two components: relevance feedback for enriching the training set (Sec. 3.1) and weak scene calibration (Sec. 3.2). The context-boosted video surveillance (see Sec. 4 and Fig. 3), exploits the learned context to effectively and efficiently produce an appearance-based people detection (Sec. 4.1) followed by head and hard-hat detectors (Sec. 4.2).

2. Related Works

Two classes of approaches have been followed in the literature for people detection [4]. 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]. One relevant limitation of these approaches 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 sub windows in a given image [2, 11, 13]. Then, a dense feature representation can be used, as in [2] where a linear SVM classifier is applied to both densely sampled histograms of oriented gradients (HOG) and histograms of differential optical flow features inside the detection window. The method by Tuzel *et al.*[13], based on covariance descriptors, has proved to yield superior performance with respect to HOG.

The context has been used before in computer vision, mainly for improving object recognition, especially in unfavorable circumstances where viewing quality is poor (due to blurring, noise, occlusions or distractors). Torralba in [12] proposed a simple yet effective framework for modeling the relationship between context and object properties 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. Markov Logic Networks are used to incorporate user activities (such as sitting on a chair or watching the TV) as context information. Vice versa, Moore *et al.*[7] use existing objects in the scene to recognize activities, exploiting Bayesian networks to model their relationship. Co-occurrence graphs, modeling relations between contextual cues (spoken words or pauses) and visual head gestures, are used in [8] for selecting relevant contextual features and inferring the visual features that are more likely in multi-party interactions.

3. Learn the Context

3.1. Relevance Feedback for Additional Training

People classifiers have reached remarkable performances, since miss rates (MR) of approximately 5% are obtained yielding approximately 1 false positive every 100K tested windows (10^{-5} False Positives Per Window - FPPW, [2, 13]); however, when these techniques are applied to exhaustive, multi-scale 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%; this drop of performance is due to the exhaustive search over the image, that highlights a *large quantity of false positives* generated by video clutter and distractors that was 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 the specific, 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 to produce a new classifier that is more robust, within the specific context, to clutter and distractors.

As depicted in Fig. 2, 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 an 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 (typically a cascade of boosting classifiers as in [13]): its multi-stage architecture allows to limit the re-training step to the latest stages of the classifier; this choice decreases the time required for the context-dependent re-training of approximately 80%. Re-training 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.

3.2. Weak Scene Calibration

We propose here to exploit the responses of the pedestrian classifier applied to the context in order to obtain an automatic weak scene geometry calibration.

Let's 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 (30 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 (e.g. 50K windows on VGA resolution for each frame) and it becomes critical to maintain real-time processing given the fact that each window is passed through a classification procedure.

Hoiem *et al.*[5] propose a statistical framework to automatically retrieve the scene perspective in order to focus the detection tasks at the right scales. With a similar approach, we make the following hypotheses: (1) all the people move on the same ground plane; (2) people are in standing position; (3) camera tilt is small to moderate; (4) camera roll is zero or image is rectified; (5) camera intrinsic parameters are typical of rectilinear cameras (zero skew, unit aspect ratio, typical focal length); (6) all the observed people are assumed to have consistent physical height.

Hypothesis (2) comes with the definition of pedestrians; hypothesis (3) is satisfied because in our context the cameras are installed with very low tilt in order to observe wide views. Hypothesis (4) is fulfilled through initial system configuration. By employing cameras with fixed focal length and by compensating the other camera parameters with an intrinsic calibration hypothesis (5) is satisfied too. By focusing our attention to adult people detection we can assume without loss of generality that the difference on people height is negligible (hypothesis (6)).

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 (see also by eq. 7 in [5]). By estimating the parameters of this 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 (for instance 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.

Differently from [5], that recovers the perspective using a probabilistic framework, we use a LSQ (Least SQuare) estimator. During the context learning phase, 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 the LSQ 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 parameter estimation. Detailed results are provided in Section 5.

4. Exploit the Context

4.1. Context-Boosted Video Surveillance

The context-boosted video surveillance layer performs people detection exploiting both general purpose and context-dependent models (see Fig. 3). The first block, named "motion-based window pruning", reduce 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. 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 [6]). 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 [1].

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 non-zero MHI pixels inside the window divided by the window area. This provides a good trade-off between searching all over the image and limiting the search to current moving regions only. Even if the motion information is not extremely accurate (typical in outdoor scenarios with moving cameras), the system recall 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 $H(x, y)$ (Sec. 3.2). 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: be (x, y) the estimated feet position of the potential pedestrian contained in a window to be classified; if the gap between the height 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. As mentioned in Sec. 1, a classifier based on an ensemble of cascades should be preferred, and for this reason we employ the pedestrian classifier [13], that is based on a rejection cascade of Logit-Boost classifiers (the strong classifiers), each composed of a sequence of logistic regressors (the weak classifiers). Given an input image I and the following 8-D set F of features (defined over each pixel of I):

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

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 rectangular patch of I.

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 i^{th} stage is approximately $(1 - 0.65^i)$. In our context-boosted classifier, we train η additional stages with the context-dependent training data as described in Sec. 3.1; the first 25 standard stages yield approximately a rejection ratio of $(1 - 0.65^{25})$ on generic negatives; the last re-trained 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 re-training the cascades.

4.2. Head and Hard-Hat Detectors

Once a pedestrian has been successfully detected, in order to evaluate the presence or absence of the hard-hat on the workers, we employ a covariance-descriptor classifier trained on a dataset of heads observed from any viewing direction. A sliding window search is applied to the upper part of the detected body and it achieves a two-fold purpose: first, it validates the correctness of the patch, i.e. a patch is a valid pedestrian iff the head detector returns one detection; 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 w.r.t. to generic circular shapes, the performance of the classifier are boosted exploiting images with a resolution (indicated as HiRes in Fig. 3) that is at least doubled w.r.t. the one used for the pedestrian detection. This allows the classifier to catch those features that would be lost at lower resolutions.

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 color can be successfully used to compute more accurate edges w.r.t. luminance images [10], we claim that the use of chrominance for image derivative computation can improve the classification results also, 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 multi-dimensional gradient methods and define these directional derivatives for the RGB color space:

$$\begin{aligned} I_x^{RGB} &= \sqrt{\left|\frac{\partial R}{\partial x}\right|^2 + \left|\frac{\partial G}{\partial x}\right|^2 + \left|\frac{\partial B}{\partial x}\right|^2}; \\ I_{xx}^{RGB} &= \sqrt{\left|\frac{\partial^2 R}{\partial x^2}\right|^2 + \left|\frac{\partial^2 G}{\partial x^2}\right|^2 + \left|\frac{\partial^2 B}{\partial x^2}\right|^2}; \end{aligned} \tag{2}$$

and similarly for I_y^{RGB}, I_{yy}^{RGB} and for Lab color space; it is then straightforward to extend equation 1 to RGB and Lab color spaces, obtaining F^{RGB} and F^{Lab}.

The classification of hard-hats vs heads 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.

5. Experimental Results

We tested the described approach over videos recorded in a construction working site of approximately $25000m^2$, 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 3fps, 1600x1200, for a total of 34 minutes of ground-truthed video (available at request). The pedestrian classifier is trained on the INRIA pedestrian dataset [2], 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 (96x96) containing heads with and without head-covers, at any viewing direction and placed in the patch center. The classifier used to separate bare heads from heads

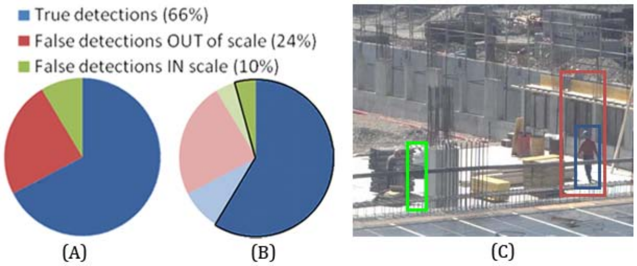


Figure 4. Weak scene calibration through LSQ and RANSAC. (A) distribution of detections on the training video, (B) consensus set, (C) visual example of the three types of detections.

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 MR vs FPPI, while the head classifier is measured on windows basis, comparing MR vs FPPW; the latter measurement is preferred for measuring classifier performance, while the former is used for assessing object detection on images or video: in this case we measure 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 [9] 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.

Regarding the weak scene calibration, Fig. 4 shows the results obtained on the videos used for learning the weak scene calibration. The positive detections are made of true and false positives; the latter can be in-scale or out-of-scale (Fig. 4A): all these positives are passed to the block of LSQ and RANSAC: the extracted consensus set (Fig. 4B) 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 Fig. 5a and Fig. 5b, where MR vs FPPI is plotted at different values of α and β. These two parameters are used to tune the degree of window-pruning exploiting motion and weak-calibration. The increase of α slightly degrades performances (Fig. 5a), while the decrease of β significantly improves accuracy (Fig. 5b), since out-of-scale windows are rejected. However, β should be tuned in order to be tolerant w.r.t. 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). Tab. 1 shows the percentage of pruned windows w.r.t. the complete SWS (i.e. $\alpha = 0$, $\beta = \infty$). 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 Fig. 5c. The higher is the parameter η, the more additional cascades are trained, the longer training time is required, 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), but even such limited additional training data generate a significant gain.

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 Fig. 5d): taking as reference $FPPI = 0.1$, this set-up processes on average 13433 windows per frame and generates a $MR = 0.14$, 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 a $MR = 0.78$ (3.7 times higher).

	Base	**Varying α**				
α	0	0.1	0.2	0.3	0.4	0.5
β	∞	∞				
%	0%	78%	83%	87%	90%	92%

	Optimal	**Varying β**					
α	0.2	0					
β	0.2	0.4	0.3	0.2	0.1	0.05	0.01
%	92%	45%	58%	72%	86%	93%	99%

Table 1. Values of α, β, and percentage of windows rejected.

Regarding the performance of the head classifiers on luminance and multi-spectral derivatives (recall Eq. 1 and 2), see Fig. 6. While in the earlier stages the luminance derivative perform better, the RGB color space ends up being definitively more effective in reducing the false positives rate: using 18 cascades, at similar MR, the RGB derivative yields $FPPW = 3.8 \cdot 10^{-4}$, while the luminance derivatives $FPPW = 1.1 \cdot 10^{-3}$ (2.9 times higher).

Finally, the two classes, bare heads (or headgears) and

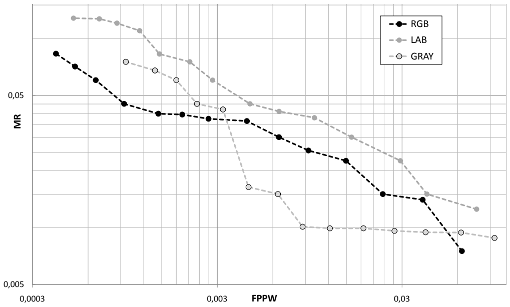
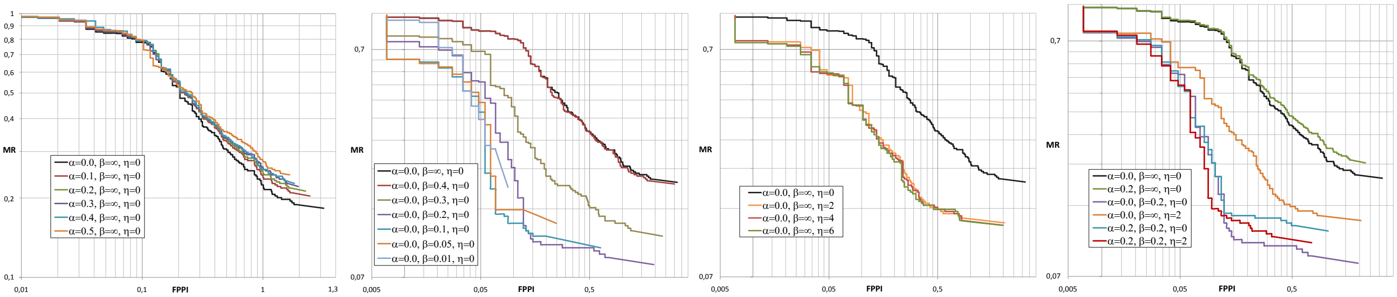


Figure 6. MR vs FPPW on the Head Image Dataset. Each marker represents the performance up to a cascade level, starting from the 5^{th} cascade on the bottom-right corner up to the 18^{th}.



(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 5. Miss Rate vs False Positives Per Image at different α, β, η.

hard-hats, are clearly separated in the Lab color space, and a simple minimum distance classifier obtains satisfying performances, since both precision and recall are above 90%. The most of the errors are generated by misclassifications of white hard-hats and of white-haired persons. Indeed, removing the white hard-hat patches, the classifier reaches precision and recall of approximately 97% using only chrominance information (L can be discarded).

Fig. 7 shows examples of the correct outcome of the complete system for hard-hat detector in construction working sites. The whole system is able to process in real-time a 1600x1200 video stream at approximately 1fps.

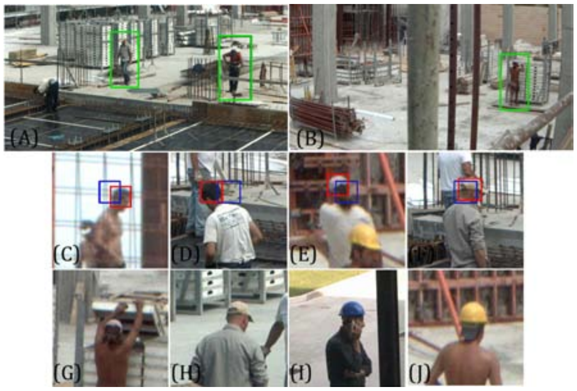


Figure 7. Example snapshots. (a,b) pedestrian detection; (c-f) head detection: blue box=head position blindly estimated from the pedestrian detection; red box=head position obtained with the head detector; final detections of (g,h) bare heads, and of hard-hats(i,j).

6. 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 bases on a context-dependent dataset. On top of these two techniques, the system also exploits motion, to further speed up the detection process, and multi-spectral derivatives, to increase the accuracy of the covariance-descriptor classifier. 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.

References

[1] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *T-PAMI*, 23(3), Mar. 2001. 3

[2] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *ECCV'06*. 2, 4

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int'l Journal of Computer Vision*, 61(1):55–79, 2005. 2

[4] D. M. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73:82–98, Jan 1999. 2

[5] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *Int'l Journal of Computer Vision*, 80, 2008. 3

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 3

[7] D. Moore, I. Essa, and M. Hayes. Exploiting human actions and object context for recognition tasks. *ICCV*, 1999. 2

[8] L.-P. Morency. Co-occurrence graphs: contextual representation for head gesture recognition during multi-party interactions. In *UCVP '09*. ACM, 2009. 2

[9] J. Ponce, T. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, C. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman. *Dataset issues in object recognition*, page 2948. Springer, 2006. 5

[10] M. Ruzon and C. Tomasi. Color edge detection with the compass operator. In *CVPR*, volume 2, 1999. 4

[11] J. Tao and J.-M. Odobez. Fast human detection from videos using covariance features. In *Workshop on Visual Surveillance (VS) at ECCV 2008*, 2008. 2

[12] A. B. Torralba. Contextual priming for object detection. *Int'l Journal of Computer Vision*, 53(2):169–191, 2003. 2

[13] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *T-PAMI*, 30(10):1713–1727, Oct. 2008. 2, 3, 4

[14] C. Wu and H. Aghajan. Using context with statistical relational models: object recognition from observing user activity in home environment. In *UCVP '09*, pages 1–6. 2