Covariance Descriptors on Moving Regions for Human Detection in Very Complex Outdoor Scenes

Giovanni Gualdi
D.I.I - Univ. of
Modena and Reggio Emilia, Italy

Andrea Prati
D.I.S.M.I - Univ. of
Modena and Reggio Emilia, Italy

Rita Cucchiara
D.I.I - Univ. of
Modena and Reggio Emilia, Italy

Abstract—**The detection of humans in very complex scenes can be very challenging, due to the performance degradation of classical motion detection and tracking approaches. An alternative approach is the detection of human-like patterns over the whole image. The present paper follows this line by extending Tuzel *et al.*'s technique [1] based on covariance descriptors and LogitBoost algorithm applied over Riemannian manifolds. Our proposal represents a significant extension of it by: (a) exploiting motion information to focus the attention over areas where motion is present or was present in the recent past; (b) enriching the human classifier by additional, dedicated cascades trained on positive and negative samples taken from the specific scene; (c) using a rough estimation of the scene perspective, to reduce false detections and improve system performance. This approach is suitable in multi-camera scenarios, since the monolithic block for human-detection remains the same for the whole system, whereas the parameter tuning and set-up of the three proposed extensions (the only camera-dependent parts of the system), are automatically computed for each camera. 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.**

I. Introduction

Speaking of modern automatic video surveillance systems, vendors wish to be provided with flexible and plug-and-play systems (working properly in every scenario and requiring a minimal setup of parameters), which should be able to adapt to predictable and unpredictable situations. Unfortunately, despite the tremendous advances that automatic video surveillance algorithms achieved in the last years, the state of the art is still far from that point. Specifically, there is an everyday effort in customizing the algorithms to the specific domain, by finding the best trade-off between flexibility (to avoid solutions which work on a very specific situation only) and accuracy.

For these reasons, special-purpose solutions tailored to the setup and the application are often adopted. This paper deals with one of these challenging cases, by addressing the use of video surveillance technology in a very complex outdoor scene like a construction working site. This environment is typically very cluttered, with several people and machineries moving all around. Thus, motion-based segmentation and tracking are very problematic and cannot guarantee a sufficient degree of reliability in every situation. To make things even worse, the construction working sites are continuously evolving and the lack of fixed reference points makes it very difficult (if not unfeasible) to learn and exploit geometric calibration and models, that would help in scene understanding.

Finally, these large outdoor scenes imply the use of multiple distributed cameras/sensors to cover the whole area: in such scenarios it is not desirable to adopt solutions that require specific training and tuning on each different view offered by the many cameras. Therefore effective solutions in such large multi-camera scenarios should be as most view-independent as possible, or, in other words, the view-dependent parts should be demanded to algorithms that are able to automatically tune themselves for proper computations.

Fig. 1 shows two examples that clearly depict the complexity of the visual environment in all the aforementioned aspects.

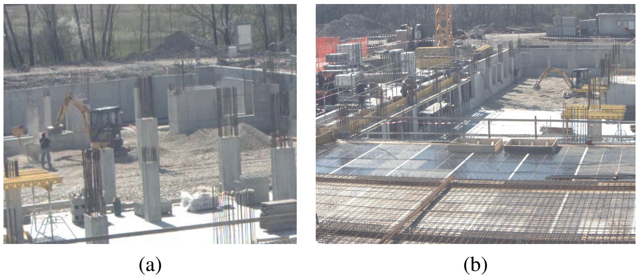

(a) (b)

Fig. 1. Examples from a construction working site

Given the current regulations in many countries in the field of security of working places, one interesting application could be to identify the workers which are not wearing the protective helmet in construction sites. This task can be achieved by detecting all the people in the scene and then extracting those wearing the helmet. Given the limits of motion-based detection and tracking solutions, this paper proposes, as a first step, to detect people in the scene with an approach based on the *covariance descriptors* proposed by Tuzel *et al.*in [1], where segmentation and tracking of people are not required. This approach exploits the covariance matrix computed over simple visual features, in order to learn the class of humans by using the LogitBoost classifier [2], trained on publicly available datasets of humans.

Although the classifier proposed by Tuzel provides remarkable human detection performance over generic contexts, it is likely to be distracted by scenario-specific visual clutter (such as the pillars in Fig. 1a) and to miss people occluded

978-1-4244-4620-9/09/$25.00 ©2009 IEEE

by scenario-specific objects (such as the scaffoldings in Fig. 1b): therefore we propose three additional steps to take into account the complexity of our scenes.

We exploit motion information as focus of attention for human detection; differently from other approaches we avoid to use motion-based people segmentation since it is not possible to rely on clear segmentation in such cluttered scenarios. For this same reason, we purposely avoid Yao and Odobez approach [3], that describes how to include motion information as a feature in the covariance descriptors for a human classifier: this approach implies a robust detection of the motion. Moreover, motion observation can be view-dependent, making the training set less expandable to different scenes and view points. Our proposal is to drive the search for humans on the areas where motion is present or was present in the recent past. This provides a good trade-off between searching all over the image (that has the drawback of heavy computational load) and limiting the search to current moving regions only (that hinders to detect still people in the scene).

As an additional contribution, we also use a rough estimation of the scene perspective to reduce false detections. More specifically, since the scene geometry of the observed scene is rather dynamic, we developed an automatic calibration procedure which estimates (through RANSAC) the reasonable height of a standing person in function of his position in the scene. This estimate is used to discard detections which are significantly different from it. The use of RANSAC allows the system to use an unsupervised approach for the perspective learning.

Finally, in order to account for the specific situation observed by a given camera, we also enrich the human detection phase by replacing the final stages of the cascaded LogitBoost, trained for generic human detection, with dedicated cascades trained on positive and negative samples automatically (or semi automatically) taken from that specific view only.

II. Related Works

As mentioned above, detecting humans in complex images is a challenging problem. The approaches proposed in the literature can be summarized in two main classes [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 [5], [6], [7], [8], [9]. 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 construction working sites overlooked by long view cameras.

The second class of proposals, often called holistic approaches [4], [10], is based on applying a full-body human detector for all possible sub windows in a given image [1], [3], [10], [11], [12], [13], [14], [15]. In [11], a human detector based on geometrical pixel-value structures of human appearances is proposed. Unfortunately, the color-based detection method is usually not robust enough for human detection. Then, a dense feature representation can be used, such as that proposed in [12] where a SVM classifier is learned using Haar wavelets as descriptors. Multiple classifiers are trained to detect human parts and the responses in the detection window are compared to give the final classification. Similarly, but applied to videos instead of still image, the work in [13] proposes an efficient detector applicable to videos using a cascade of Adaboost classifiers relying also on Haar wavelet descriptors but extracted from spatio-temporal video differences. By using Adaboost (or other boosting techniques), the most discriminative features can be retained.

Another example of dense representation used for human detection is reported in [14], where a linear SVM classifier is applied to both densely sampled histograms of orientation gradient (HOG) and histograms of differential optical flow features inside the detection window. The method by Tuzel *et al.*[1] is based on a cascade of LogitBoost classifiers that uses covariance features as human descriptors. More precisely, sub windows of the detection windows are represented by the covariance matrix of image features, such as spatial location, 1st and 2nd derivatives (magnitude, orientation). The LogitBoost classifier is modified by mapping the covariance matrix features (lying in a Riemannian manifold) in a vector space. This approach has proved to yield superior performance with respect to other proposals. Yao and Odobez [3] extended this approach to speed up the computation and take into account the temporal information. First, they proposed to build weak classifiers based on subsets of the complete image feature space. This corresponds to explore the covariance between features in small groups rather than altogether for each weak classifier. In addition, the means of the features inside the sub window are used to allow faster rejection. Finally, they proposed to include background suppression results (i.e. motion information) as features in the classifiers.

Other approaches tackle the problem mainly from the performance viewpoint, i.e. proposing fast solutions to human detection. For instance, a near-real-time detection performances were achieved by training a cascade model using HOG features [15]. A recent paper [10] proposed to employ both intensity-based (those used in [13]) and gradient-based (called *Edge Orientation Histogram* - EOG) features and Real Adaboost algorithm to select critical features. Instead of using the standard boosted cascade, a novel cascaded structure is proposed in which both stage-wise classification and interstage cross-reference information are used.

Finally, the paper in [16] presents a very nice statistical framework to model the relationship between objects and scene geometry, by modeling the interdependences between objects, surface orientations, and camera viewpoints. This framework effectively evaluate the correct perspective of objects in the scene, in a way similar to our proposal.

III. System Description

We are developing a complete system for distributed monitoring of construction working sites that will make use of different sensors, including RFIDs, to verify the presence of unauthorized personnel in the site. In this scenario, RFID sensors will provide a delocalized identification of people in a certain area, while cameras will provide their detection and localization, but not their identification. A statistical framework will fuse these data to infer the presence (and position) of unauthorized people. Moreover, the system needs to detect people which are not wearing the protective helmet. For both these functionalities a first, essential step is to detect humans in a complex outdoor scene.

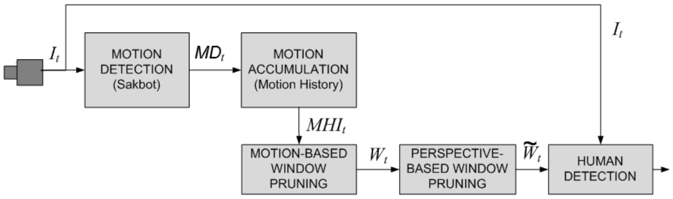

I_t
MOTION DETECTION (Sakbot)
MD_t
MOTION ACCUMULATION (Motion History)
I_t
MHI_t
MOTION-BASED WINDOW PRUNING
W_t
PERSPECTIVE-BASED WINDOW PRUNING
$\widetilde{W}_t$
HUMAN DETECTION

Fig. 2. Scheme of the proposed approach

The overall scheme of our approach for human detection is sketched in Fig. 2. The basic idea is to merge results achieved by a standard motion detection algorithm with the human detection achieved through a sliding-window approach. The former approach has the advantage to focus the attention on moving regions, avoiding to be distracted by areas which are for sure not humans (such as buildings or the sky). However, humans can also be still and motion is not always a reliable cue in complex scenes. Therefore, we propose a hybrid approach which aims at exploiting the advantages of both.

Instantaneous motion is computed on segmented moving objects detected using background suppression. The *Sakbot* (Statistical And Knowledge-Based Object Tracker) approach proposed in [17] is exploited to build a background model and to segment only real moving objects, distinguished from other artifacts. Obviously, a correct segmentation of moving regions is a crucial task. The instantaneous motion at time t $MD_t\left(i,j\right)$ (MD as motion detection) is computed by thresholding the difference with the background model. 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) introduced by Bobick and Davis in [18], defined as

$$MHI_t\left(i,j\right)=\begin{cases}\tau & \text{if } MD_t\left(i,j\right)=1\\ \max\left(0,MHI_{t-1}\left(i,j\right)-1\right) & \text{otherwise}\end{cases} \quad (1)$$

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

The full sliding-window set (SWS) over a frame is pruned at first exploiting MHI_t, i.e. retaining a number windows is proportional to the amount of motion present in the area (W_t in fig. 2), then using perspective constraints, i.e. removing all the windows whose size is not compliant with the perspective model ($\widetilde{W}_t$ in fig. 2). After these procedures, the human detection algorithm is issued over the survived windows only.

The next section will describe how to prune the window set by exploiting both motion (Section IV-A) and perspective (Section IV-B) constraints, while Section V will describe the human detection algorithm. Section VI presents the experimental results.

IV. Window Selection

As described in the previous section, our approach is based on a sliding-window scheme, that basically scans all possible windows, searching for the best position of humans in terms of similarity. In fact, if it were possible to obtain reliable object tracking in the scene, it would be possible to compute very focused human detection over the bounding box of the tracked objects. Unfortunately, it is not possible to rely on object tracking in such complex scenarios, therefore we are obliged to adopt a brute force approach like the sliding window over the whole image.

Since the size of humans in the scene is unknown, the sliding-window scheme scales the size of the windows to a certain degree. Differently from Tuzel's approach [1], that performs sliding window at 5 different scales only (increasing the size of 20% from step to step), denoting the assumption of an a-priori knowledge about the human size inside the image, we have to search at very different scales, making no assumption on the observed scene. This choice is unavoidable due to the large variability of the depth of view in our typical scenario. Consequently, we employ 21 steps, with an increase of 10% in size from step to step. At any given scale, we choose a window from the preceding one with a displacement in pixels that is not static (typical values are from 2 to 4 pixels), but is function of the size of the window, in order to obtain smaller (bigger) displacements for small (big) windows.

Such loose conditions over the sliding window generation have the drawback to generate a very redundant window set. Considering that a detection procedure must be issued on each window, it is desirable to reduce the window set as much as possible, in order to limit the computational burden. Therefore we employ some pruning techniques that have also the advantage of increasing the accuracy of the human detector by reducing the number of false positives per image.

A. Motion-based pruning

By exploiting the motion detection MD_t provided by Sakbot, a straightforward approach for pruning useless windows would be to remove all the windows that contain no (or low degree of) motion. This sharp window rejection has the drawback to prune windows in areas where the motion is fragmented (because of background camouflage) or where there is absence of motion just in the very last frame. Employing a more conservative approach based on the motion history image MHI_t solves the problem; however, if the τ (see eq. 1) is set from moderate to high values (i.e. 5-30 seconds), and if there are objects that move all around the image frame, the reduction of the number of windows is very limited. As a solution to this, it is reasonable to relate the number of retained windows in an area with the amount of motion history present in that area, by increasing the amount of discarded windows together with the "age" of the motion recorded within the window (i.e. windows containing "older" motion will be discarded more easily than windows containing "fresher" motion).

The precise procedure is presented in Algorithm 1. At first, the motion ratio MR_t^i for the window w^i at frame t is computed as the ratio of pixels with motion ($MHI_t(p) > 0$) within the window, with respect to its area. If the motion ratio is too low, the window is discarded; otherwise there is a further verification, that depends on the maximum (i.e. most recent) motion found inside the window: a random pruning technique is performed depending on such value. As Tuzel *et al.*clearly described in [1], the LogitBoost detector based on covariance features is quite insensitive to small translations and scales, generating several true detections around a true human silhouette. Therefore, pruning a variable number of windows around a human silhouette will not affect the overall detection performance, since from the many encompassing windows, some of them will very likely survive from the pruning and trigger a true-detection over the human detector.

Algorithm 1 Motion-based window pruning

```
Require: Sliding-window set $SWS$, Frame $I_t$, Motion History Image $MHI_t^*$
  for all window $w^i \in SWS$ do
    $MR_t^i = \frac{\#\left\{pixel\ p | p \in w^i \wedge MHI_t(p) > 0\right\}}{\#\{pixel\ q | q \in w^i\}}$
    if $MR_t^i < 0.5$ then
      prune $w^i$
    else
      $v = \max_{p \in w^i} \ MHI_t\left(p\right)$
      Sample $u$ from uniform distribution $U(u|0, \tau)$
      if $u > v$ then
        prune $w^i$
      else
        $w^i \rightarrow W_t$
      end if
    end if
  end for
```

* = We assume that $MHI_t\left(p\right) \in [0, ..., \tau]$: 0 means no motion in the history of the pixel, τ means motion in the last observed frame

B. Perspective-based pruning

As mentioned above, we avoided to define an expected size of the human silhouette (in standing position) inside the observed scene; to this aim we generated a sliding-window set with a very wide range over the scales. Hoiem *et al.*[16] propose a statistical framework to automatically retrieve the scene perspective in order to focus the detection tasks at the right scales. Following 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 (3) is satisfied because in our context the cameras are installed with very low tilt in order to observe wide views. Moreover, 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, assuming for the sake of simplicity that hypothesis (1) is a-priori satisfied, it is correct to approximate the height (in pixels) of the human silhouette with a linear function in the image coordinates (x, y) of the point of contact with the ground plane (confirmed also by eq. 7 in [16]). By estimating the parameters of this function, we can further prune the sliding-window set W_t by discarding all the windows whose height strongly differs from the estimated function and obtain the set $\widetilde{W_t}$ (see Fig. 2).

Differently from [16], that recovers the perspective using a probabilistic framework, we use a LSQ (Least SQuare) estimator with outliers rejection based on RANSAC. During the training phase, the motion-based pruning (see Section IV-A) and the human detector (described in the next Section V) are run over a video that must contain, among other objects, also some people: all the windows that the human detector classify as positives are be passed to a LSQ estimator and to the RANSAC iterative method, that is capable to discard the possible outliers (due to out of scale false detections) and retain only the windows which contribute to the correct parameter estimation. Detailed results are given in Section VI.

In the case the hypothesis (1) is violated in the observed context (for instance workers on scaffoldings, see Fig. 1b), it is still possible to perform some perspective pruning by partitioning the image in areas and accept the rougher assumption that the height (in pixels) of the people inside each area is almost constant.

V. Human Detection over Riemannian Manifolds

As mentioned in Section II, the detection over Riemannian manifolds using covariance descriptors provides remarkable results [1]. Given an input image I, this work proposes to extract at each pixel location (x, y) the following 8-dimensional set F of features:

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

where I_x,I_y and I_{xx},I_{yy} are respectively the first and the second-order derivatives of the image.

Given a rectangular window, we can compute the covariance matrix of the features F inside the window, that was demonstrated to be a very informative descriptor which encodes information about the variance of the features, their correlations with each other and the spatial layout. Furthermore, it can be efficiently computed using integral images [19]. The authors propose to learn a cascade of rejecting LogitBoost classifiers [2], where the weak learners in each cascade are trained to classify the covariance features computed on randomly sampled portion inside the window. However, since covariance matrices do not lie in a vector space but in the Riemannian manifold of symmetric positive definite matrices, Tuzel *et al.*proposed a few modifications to the original LogitBoost algorithm to specifically account for the Riemannian geometry.

A. Additional Learning with Relevance Feedback

We tested the original 30-cascades algorithm proposed by [1] using linear functions as weak learners and trained over the INRIA pedestrian database [14] ("INRIA-based detector" from now on). Confirming the robustness of the approach, the detection rates on construction working sites are good, especially on positive detections; in fact the appearance of human silhouettes in this specific context does not differ much from the one coming from a more generic scenario: at the low resolution used to observe the humans in our working system (just a few pixels in width and height), the only noticeable difference is the presence of a protective helmet: being the features mainly based on image derivatives (neither color, nor luminance), the visual appearance of a helmet does not differ from human hair or just a cap. For this reason, the miss rate reported in [1] is approximately confirmed in our context also.

On the opposite, the performance on negative samples is seriously challenged: as a matter of facts, the extremely cluttered scenarios of construction sites produce a rate of false positives that is higher than the one produced on the INRIA test bed (the False Positives Per Window - FPPW [1] - are increased by a factor of 6 approximately). Motion- and perspective-based pruning reduce but do not nullify this performance degradation: we additionally tackle this problem training a few extra cascades in a relevance feedback (RF from now on) fashion [20], that will replace a portion of the INRIA-based detector; it is important to notice here that we do not aim to replace the whole training of the INRIA-based detector, since it provides by itself remarkable performance even in our challenging scenario, but we adapt and improve a few stages of the cascaded detector with additional scenario- and view-dependent video data.

As described by [1], the performance of the 30-cascades of the INRIA-based detector is asymptotic (see Fig. 3), resulting in limited rejecting abilities of the last cascades; therefore it is reasonable to perform a re-training of the last cascades only, using additional RF training data, that is generated using a twofold procedure:

1) *implicit (assessment free) RF feeding*: the pool of negative examples is enriched with background images from our construction working site that are automatically extracted using Sakbot. In order to be sure of the absence of humans inside the scene, we accept background images only when no motion is detected during a time gap of 10 minutes. In our test bed of videos from construction working sites, we verified that it is feasible to extract several person-free images on daily basis;
2) *explicit (with assessment) RF feeding*: from the pool of detections obtained with the INRIA-based detector (run on a few videos from a specific view and using motion history and perspective pruning), an user provides an assessment, extracting some true positives and false positives, that will respectively enrich the positive and negative training data.

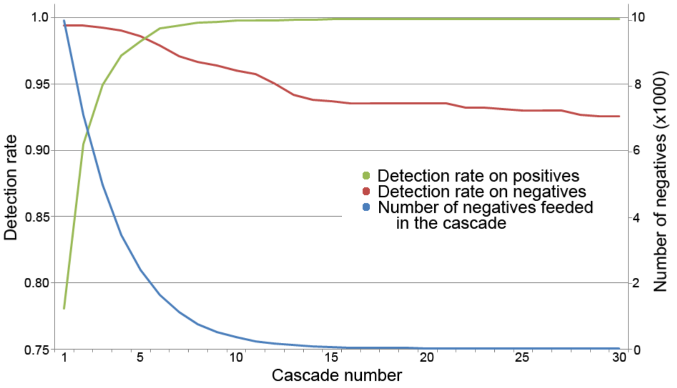


Fig. 3. Asymptotic behavior of the last cascades of [1] classifier, learned on INRIA pedestrian dataset.

Differently from the implicit feeding, that produces only generic negative training data, it is very important to note that the explicit feeding enriches also the positive training data: the additional negative samples extracted here have a strongly-informative content, being all of them (false) positive detections according to the INRIA-based detector. Therefore, any additional cascade learned using these specific data as negative training data will generate a hard split inside the space of positive detections modeled so far. By reinforcing the positive training data with samples (snapshots of workers) taken from the same context of the false positives, the cascade learner can better model the logistic function of the LogitBoost.

These re-trained cascades using RF training data are placed at the final stages of the rejection cascade process of the INRIA-based detector; therefore it is not possible, at this point, to raise the performance over the true-detections (what was wrongly rejected by the first cascades cannot be recovered then, because of the nature of rejection cascades), but conversely it is still possible to act in a strong manner on the reduction of the false-detections, that is exactly our goal.

The learning time of a complete 30-cascades LogitBoost detector on Riemannian manifolds is very long (in the order of days of computation); since the cardinality of the miss-classified negative samples decreases exponentially at each cascade (see Fig. 3), the longest time is spent on learning the first cascades. For this same reason, the re-training of the last stages of the LogitBoost classifier using RF training data is computed in a time (in the order of a couple hours of computation), that is sufficiently short to consider it as a symptomatic operation, in the meaning that it could be re-computed whenever the viewed scene is subject to remarkable changes (common situation in construction working sites) and the performance of the system begins to degrade. In the case it is not possible to provide explicit assessment, additional learning on implicit data only can be performed, making the system totally autonomous and suitable for multi-camera systems.

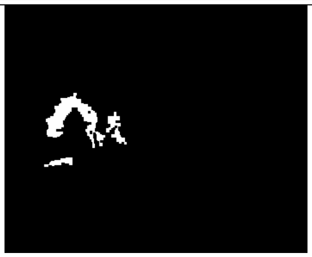
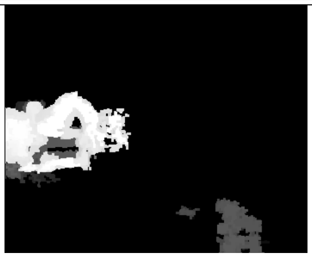
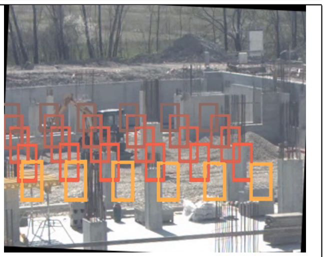
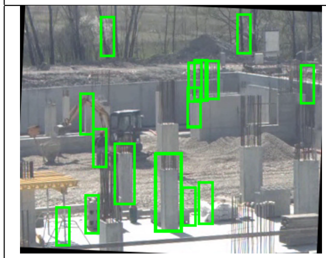
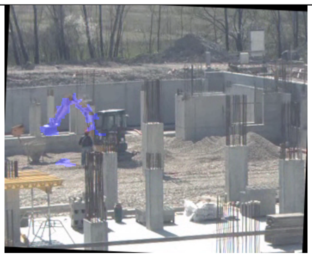
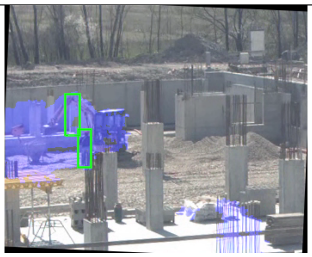
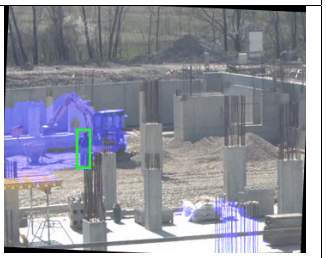
Input image I_t	**Motion Detection** MD_t	**Motion history** **image** MHI_t	**Motion history** MHI_t **+ Perspective**
			
			
283798	425	35012	3127

TABLE I

QUALITATIVE EVALUATION OF HUMAN DETECTION, USING MOTION-BASED AND PERSPECTIVE-BASED PRUNING. TOP ROW, FROM LEFT TO RIGHT: ORIGINAL SNAPSHOT; INSTANT MOTION DETECTION; MOTION HISTORY; PERSPECTIVE DEPICTION. MIDDLE ROW: HUMAN DETECTION RESULTS, MAKING USE OF THE CORRESPONDING MASK; FROM LEFT TO RIGHT: USE OF NO MASK, OF INSTANT MOTION, OF MOTION HISTORY; OF MOTION HISTORY AND PERSPECTIVE. BOTTOM ROW: NUMBER OF WINDOWS THAT SURVIVED THE PRUNING AND WERE FED TO THE HUMAN CLASSIFIER (REPRESENTING COMPUTATIONAL LOAD FOR CLASSIFICATION).

VI. EXPERIMENTAL RESULTS

We tested the described approach over three videos at 352x288 resolution recorded at a construction working site, in the following different scenarios:

- *video 1, 1740 frames*: objects moving over a single plane; approximately a similar number of humans and other kind of objects (cranes, bulldozers,...) in the scene;
- *video 2, 2760 frames*: objects moving over a single plane; prevalence of other kind of objects (cranes, bulldozers,...) w.r.t. humans;
- *video 3, 1740 frames*: objects moving over several planes (scaffolding scene); prevalence of humans w.r.t. other objects. Serious presence of occlusions.

In table I we present a qualitative summarization of the effect of motion-based and perspective-based pruning, where it is possible to appreciate how the two pruning techniques improve both performance (the number of windows to analyze per frame gets strongly reduced) and precision of the detection (since pruning basically reduces the number of false positives coming out from the human detector). As explained in section IV-A, the use of pruning based on the instant motion detection might negatively affects the recall, and the example here gives a clear depiction of the problem: the man close to the bulldozer is still now (MD_t is zero over him), but moved in a recent past (MHI_t is non zero over him); this is reflected in the corresponding detections. The number on analyzed windows presented in the third row of table I refers to the mean computed over several frames and the order of magnitude is confirmed all along the three videos: pruning based on MHI_t reduces sliding windows by approximately a factor 10, pruning on perspective by a further factor 10.

Regarding the perspective-based pruning, table II reports the accuracy of the RANSAC-based method during the perspective-learning phase (described in section IV-B). The first row reports the number of windows that successfully passed the human classifier and that were used to feed the perspective learner; please notice that some of the false positive windows were showing a very similar scale to what a human would have shown in that same position. The consensus set provided by the RANSAC, that is then used by the LSQ to estimate the model parameters, excludes all the false detections that are out of scale; this fact, together with the standard deviation of the error (low for in-scale detections, much higher for out-of-scale detections) gives proof that a linear model for perspective estimation is correct and it can be successfully learned by means of the classification output only, avoiding any camera calibration.

	False positives		**True positives**
	Out of scale	**In scale**	**In scale**
# of windows	110	40	310
Consensus set	0	20	270
Std. error deviation	21.57	8.7	1.5

TABLE II

RESULTS OF LSQ AND RANSAC FOR PERSPECTIVE LEARNING OVER A 1000 FRAMES LONG VIDEO. THE LAST ROW REFERS TO THE STANDARD DEVIATION OF THE DIFFERENCES BETWEEN THE ESTIMATED AND THE ACTUAL HEIGHTS OF THE DETECTION WINDOWS.

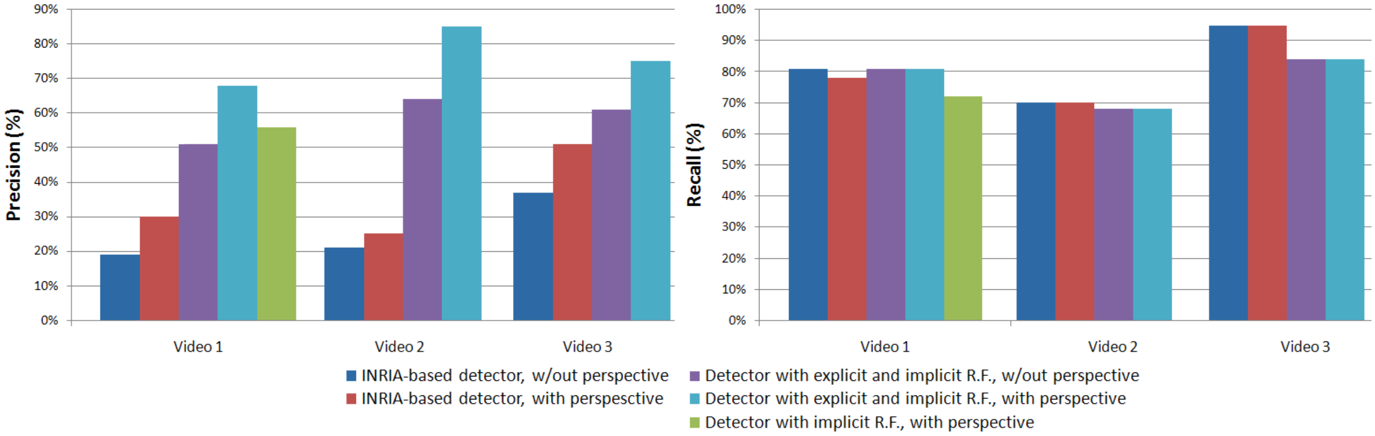


Fig. 4. Precision and recall at object level.

The results of the measurements of human detection performance on the three videos are shown in Fig. 4, that depicts the *precision* and the *recall* of the system with several setups: these two metrics were calculated measuring true/false positives and false negatives of the human detection task at object level. The use of perspective-based pruning reduces the number of false detections, therefore improving the precision. Please consider that in video 3 the humans are moving on scaffoldings, therefore the hypothesis (1) of section IV-B is not satisfied and we model the perspective not as a function of the image coordinates but as an automatically estimated *constant*: regardless of the roughness of the approach, the precision improves significantly (+14%). The perspective pruning can rarely cause the removal of true-detection windows also: it happens when the window to be detected is beyond the border of the range admitted by the perspective and this explains the slight decrease of recall (video 1, first two bins of the recall histogram of Fig. 4).

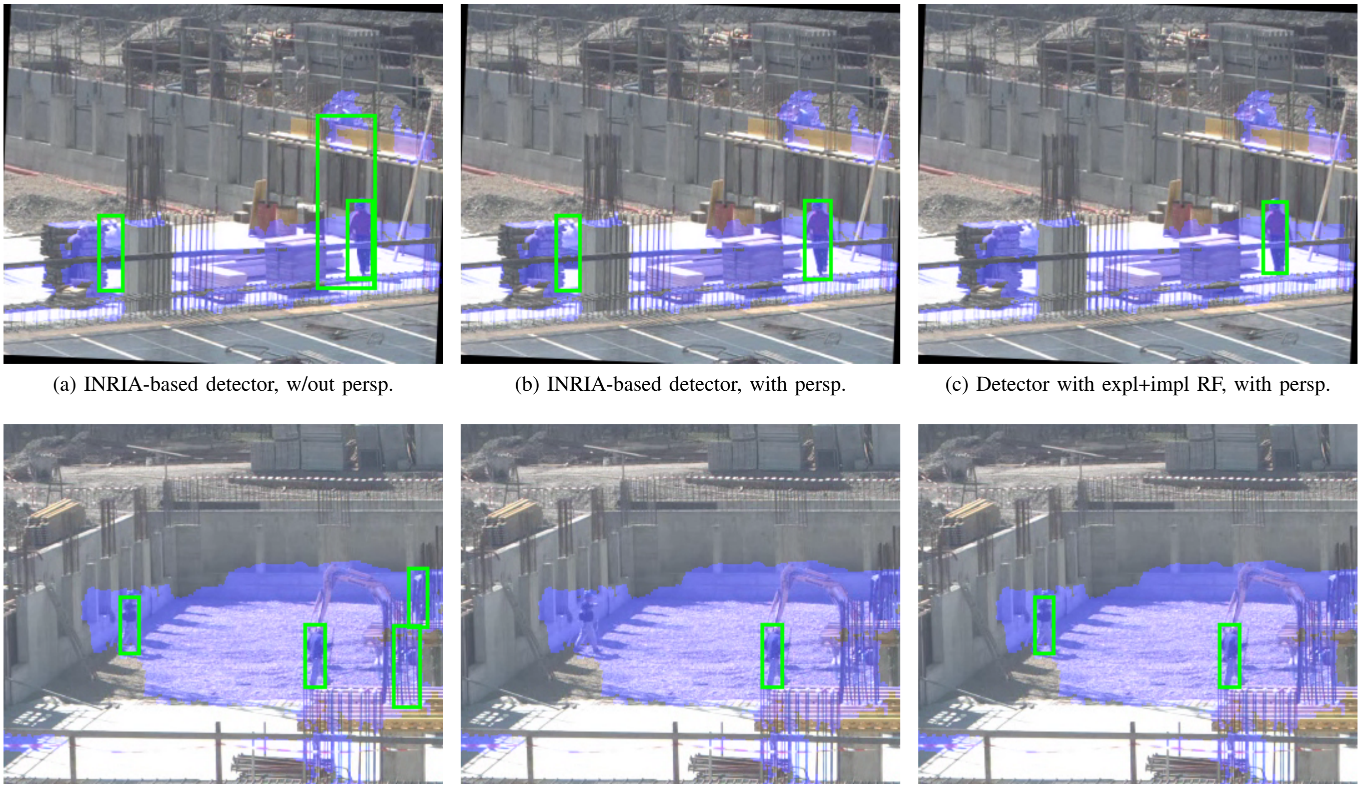
The additional learning with RF gives a strong improvement over precision: when using both explicit and implicit R.F. data, the detection gains approximately +35% in video 1, +60% in video 2, +23% in video 3 w.r.t. the INRIA-based detector. Notice that the stronger is the presence of moving (or recently moved) objects different from humans in the scene (from highest to lowest: video 2,1,3), the higher is the gain in precision thanks to additional RF learning: this can be explained because, as we mentioned in section V-A, the RF cascades are placed at the final stages of a rejection cascade process and they will improve the performance over (the removal of) false-positives and not over (the inclusion) of false-negatives. For this same reason, our approach does not affect the recall, that remains basically the same of the INRIA-based detector. Only in video 3 there is a slight decrease of recall, due to the fact that the implicit feeding trained the additional cascades in a very strong way over the scaffolding, and in such video the humans are often integrated into or partially hidden by the scaffoldings.

Regarding the implicit learning performance, the additional learning based on implicit feeding only, increases the precision, even if with a lower degree w.r.t. the additional learning with both explicit and implicit feeding (experiments were performed on video 1 only). On the opposite there is a slight negative effect on recall: as explained in section V-A, the additionally learned cascades cannot rely on positive samples taken from the construction working site in order to best shape the positive-detection space.

The system used for our test replaced the last 6 cascades of the INRIA-based classifier with 6 cascades learned with RF training data. The implicit feeding was composed of 6 negative background images extracted from videos recorded at the same camera position used in videos 1, 2 and 3; the explicit feeding was composed of 500 true-positive detections and 1000 false-positive detections extracted again from videos recorded in the same day. Figure 5 provides a visual outcome of the detections over video 3 and 1. In sub-figures 5-a,b,c it is depicted the advantage of perspective-pruning first, and of false-positive removal thanks to the RF cascades then. Subfigures 5-d,e,f show the behavior of the RF cascades with respect to the INRIA-based detector: the implicit RF cascades (fig. 5-e) remove the false positives together with a true-positive; the explicit and implicit RF cascades instead (fig. 5-f) are able to remove false positives without affecting the true positives.

VII. Conclusions and Acknowledgments

The present paper extends human detection based on covariance descriptors to very complex outdoor scenes, especially focusing on construction working sites. The proposed approach makes use of the human detector proposed by [1], trained with a generic pedestrian database, enriched with techniques to improve the performance and the accuracy of the detection. At first, the sliding window set, generated under very loose requirements on the expected scale of the humans, is pruned making use of motion and perspective constraints. Then, the human detector, based on a cascade of LogitBoost classifiers, is enriched in a relevance-feedback fashion in order to increase the accuracy of the system on false detections. Experimental results prove the validity of the proposed approaches. The parameters of the perspective pruning and the relevance-feedback cascades can be learned in a totally autonomous



(a) INRIA-based detector, w/out persp. (b) INRIA-based detector, with persp. (c) Detector with expl+impl RF, with persp.

(d) INRIA-based detector, with persp. (e) Detector with impl only RF, with persp. (f) Detector with expl+impl RF, with persp.

Fig. 5. Example of human detection in video 3 (a,b,c) and in video 1 (d,e,f). The blue shadow highlights the pixels of the image where $MHI_t > 0$.

way, making the system very attractive for those massive multi-camera systems where it is not reasonable to require human intervention for the parameters' tuning for each field of view. The authors are specially thankful to Oncel Tuzel and to Jean-Marc Odobez for their prompt help and availability. This research activity was partially funded by Bride129 S.p.A.

References

[1] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Trans. on PAMI*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[2] 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.

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

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

[5] S. Ioffe and D. A. Forsyth, "Probabilistic methods for finding people," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 45–68, 2001.

[6] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *ECCV04*, 2004, pp. Vol I: 69–82.

[7] 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.

[8] G. Gualdi, A. Albarelli, A. Prati, A. Torsello, M. Pelillo, and R. Cucchiara, "Using dominant sets for object tracking with freely moving camera," in *Proceedings of 8th Int'l Workshop on Visual Surveillance*, Marseille, France, Oct. 2008.

[9] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Trans. on PAMI*, vol. 29, no. 1, pp. 65–81, 2007.

[10] Y.-T. Chen and C.-S. Chen, "Fast human detection using a novel boosted cascading structure with meta stages," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1452–1464, 2008.

[11] A. Utsumi and N. Tetsutani, "Human detection using geometrical pixel value structures," in *FGR*, 2002, pp. 39–44.

[12] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, 2001.

[13] P. A. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[14] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV (2)*, 2006, pp. 428–441.

[15] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *CVPR (2)*, 2006, pp. 1491–1498.

[16] 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.

[17] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts and shadows in video streams," *IEEE Trans. on PAMI*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[18] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on PAMI*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[19] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *In Proc. 9th European Conf. on Computer Vision*, 2006, pp. 589–600.

[20] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," *Journal of the American Society for Information Sctence*, vol. 41, no. 4, p. 288297, 1990.