PRINCIPLED DETECTION-BY-CLASSIFICATION FROM MULTIPLE VIEWS

Jérôme Berclaz*, François Fleuret*† and Pascal Fua*
*Computer Vision Laboratory, EPFL, Lausanne, Switzerland
†IDIAP Research Institute, Martigny, Switzerland

Keywords: People detection, Classification, Bayesian framework.

Abstract: Machine-learning based classification techniques have been shown to be effective at detecting objects in complex scenes. However, the final results are often obtained from the alarms produced by the classifiers through a post-processing which typically relies on *ad hoc* heuristics. Spatially close alarms are assumed to be triggered by the same target and grouped together.
Here we replace those heuristics by a principled Bayesian approach, which uses knowledge about both the classifier response model and the scene geometry to combine multiple classification answers. We demonstrate its effectiveness for multi-view pedestrian detection.
We estimate the marginal probabilities of presence of people at any location in a scene, given the responses of classifiers evaluated in each view. Our approach naturally takes into account both the occlusions and the very low metric accuracy of the classifiers due to their invariance to translation and scale. Results show our method produces one order of magnitude fewer false positives than a method that is representative of typical state-of-the-art approaches. Moreover, the framework we propose is generic and could be applied to any detection-by-classification task.

1 INTRODUCTION

Detection in images is often treated as a repeated classification problem. Given a two-class classifier which predicts "target present" or "target not present" from an input signal and a candidate *pose* (such as location or scale), detection is achieved by applying it for any possible pose and collecting the ones associated to positive responses. Such schemes often yield multiple responses for every single true positive and therefore require post-processing to refine the outcome.

This step is usually *ad hoc* and involves grouping and averaging similar poses corresponding to positive classifications. Such a procedure is standard for detecting faces (Viola and Jones, 2001; Fleuret and Geman, 2002), cars (Zhao and Nevatia, 2001) and pedestrians (Viola et al., 2003; Leibe et al., 2005). Some people tracking approaches also introduce temporal consistency to combine the classifier responses in a stochastic manner (Okuma et al., 2004).

In this paper, we propose a statistically consistent Bayesian approach for processing answers from repeated classification algorithms. As opposed to simple grouping-and-averaging or non-maximum suppression schemes that are usually applied for this step, our method takes into account knowledge about both the classifier response model and the scene geometry, which yields a more accurate detection with less false positives.

We demonstrate our approach on the problem of multi-people detection using several widely spaced cameras, as illustrated by Fig. 1. In this application, a classifier is repeatedly applied to every possible 3D pose in different camera views, which results in one map of classifier answers per camera view. The several maps of classifier answers are then post-processed and combined by our algorithm to yield the final detection.

At the heart of our approach is a sophisticated application of Bayes' law. Using a model of the responses of a classifier given the true occupancy, we infer a posterior probability on the occupancy given the classifier responses. We will show that this lets us combine the multiple and noisy classifier responses in separate camera views and infer accurate world coordinates for our detections.

Our main contribution is thus a principled approach for processing detection-by-classification results and generating a final accurate detection out of it. When applied to the problem of multi-people detection using several cameras, our approach produces one order of magnitude fewer false positives than a

Berclaz J., Fleuret F. and Fua P. (2008).
PRINCIPLED DETECTION-BY-CLASSIFICATION FROM MULTIPLE VIEWS.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 375-382
DOI: 10.5220/0001081003750382

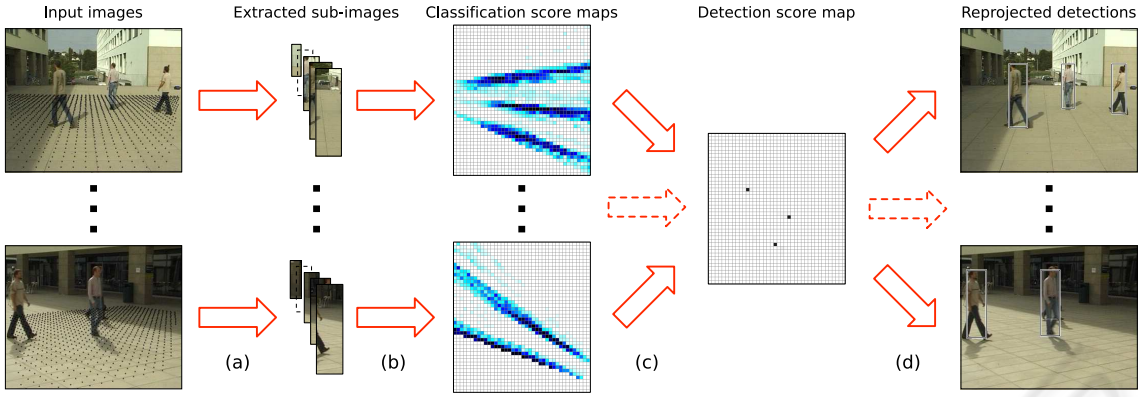


Figure 1: Overview of the detection process. Video sequences are acquired by widely separated and calibrated cameras. The ground plane of the tracked area is discretized into a finite number of locations, depicted by the black dots in the leftmost column. (a) We first extract from each image the rectangular sub-images that correspond to the average height of a person at each of these locations. (b) We apply a classifier trained to recognize pedestrians to each sub-image to estimate probabilities of occupancy in the ground plane from each view *independently*. (c) We use the algorithm that is at the core of this paper to combine the individual classification score maps into a single detection score map. (d) We reproject into the original images a person-sized rectangle located at local maxima of the probability estimate.

baseline method, that is representative of what is typically done by state-of-the-art methods. Moreover, the framework we propose is generic and could be used with any detection-by-classification application, whether single or multi view, for which a model of the classifier response is available.

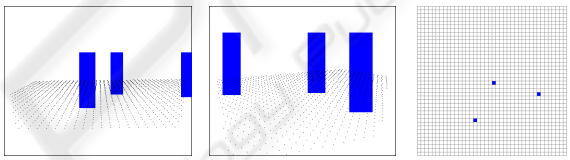


Figure 2: Correspondence between camera views (left and center pictures) and top view (right picture) is made through rectangles computed with ground plane homographies. We call $I_c(i)$ the rectangle on camera view c that has the average shape and position of a pedestrian standing at location i of the ground plane.

2 RELATED WORK

We address a problem usually solved by simple *ad hoc* solutions. Therefore, even though our framework for processing detection-by-classification results is generic, we compare it here to pedestrian detection algorithms, which is the application we chose to demonstrate our method in this paper. Some of the multi-view pedestrian detection works we reference below are close in spirit to our framework.

Until recently, most approaches to locating people in video relied on recursive frame-to-frame pose estimation. While effective in some cases, these techniques usually require manual initialization and reinitialization if the tracking fails. As a result, there is now increasing interest for techniques that can detect people in individual frames.

A popular approach (Viola et al., 2003; Okuma et al., 2004; Dalal and Triggs, 2005) is to use classification-based techniques to decide whether or not image windows depict a person. Such global approaches tend to be very occlusion sensitive and bag-of-features approaches have proved more effective at detecting pedestrians monocularly in crowded scenes (Leibe et al., 2005).

However, with the exceptions of (Khan and Shah, 2006; Mittal and Davis, 2003), we are not aware of many attempts at combining the output of detectors across views to overcome the problems created by occlusions in a principled way. In (Khan and Shah, 2006), the algorithm classifies individual pixels as background or part of a moving object and combines these results across views by assuming independence given the presence of a pedestrian at a certain ground location. Hence, this scheme does not use a generic pedestrian detector based on a high-level model of silhouettes and textures. Neither does it explicitly model the fact that a detection in one view is influenced by the presence of distant pedestrians creating occlusions, which, as we will see, can trigger many false alarms. By contrast, the M_2Tracker (Mittal and Davis, 2003) explicitly models the relation between mutliple pedestrians and the image at the pixel level, thus naturally taking occlusions into account. However, this approach relies on temporal consistency, and since it is based on a tight integration between the handling of occlusions and a color-based appearance model, it can not be generalized to use a generic pedestrian vs. background classifier.

Table 1: Notation.

C	number of cameras.
G	number of locations in the ground plane ($\simeq 1000$).
X_k	boolean random variable standing for the occupancy of location k on the ground plane.
$\mathbf{I}_c$	input image from camera c.
$I_c(i)$	rectangular human size sub-window cropped from camera view c at ground location i.
$\delta_c(i,j)$	horizontal distance between the centers of $I_c(i)$ and $I_c(j)$ on camera view c.
$n_c(i)$	neighborhood of i on camera c, $\{j \neq i, I_c(j) \cap I_c(i) \neq \emptyset\}$.
$T_c(i)$	sum of the responses of the binary decision trees at ground location i in camera view c, thus an integer value in $\{0,...,N_T\}$ where N_T is the number of decision trees.
$\mathbf{T}$	vector of all the $T_c(i)$.
Q	the product law with the same marginals as the real posterior distribution $P(\cdot\|\mathbf{T})$. $Q(\mathbf{X}) = \prod_{i=1}^{G} Q(X_i)$.
E_Q	expectation under $\mathbf{X} \sim Q$. $E_Q(x) = \int xQ(x)dx$.
q_k	the marginal probability of Q, i.e. $Q(X_k = 1)$.
$\|.\|$	area of a sub-image.

In contrast to the approaches described above, our method relies on classifiers applied on separate views independently. We explicitly integrate occlusion effects between alarms and quantitative knowledge about the classifier insensitivity to pose change into a sound Bayesian framework to combine the multiple classifier answers and yield the final detection.

3 ALGORITHM

We start by giving an overview of our algorithm, before going into more details in the following subsections. We use notations summarized in Table 1.

In our setup, an area of interest is filmed by C widely separated and calibrated cameras. We discretize the ground plane into a regular grid of G locations separated by 25cm (Elfes, 1989), and compute homographies that relate the ground plane to its projections in the camera views. This way, we can determine, for every camera view c and every location i, the sub-image $I_c(i)$, which roughly corresponds to the average size of a person that would be standing at location i of the ground plane, as shown on Fig. 2. Our algorithm involves two main steps:

1. For each camera c and ground plane location i, the algorithm extracts sub-image $I_c(i)$. Classifiers based on decision trees are then applied to every sub-image $I_c(i)$, as shown on Fig. 3. These classifiers have been trained at recognizing pedestrians, and their answer on sub-image $I_c(i)$ can be interpreted as a rough probability of occupancy of ground plane location i, given the sub-image. This first step thus produces as many *classification score maps* (see third column of Fig. 1) as there are cameras and is described in §3.1.
2. The several classification score maps, generated during step 1, are now combined into a final probability of occupancy map (called hereafter *detection score map*), such as the one of the fourth column of Fig. 1. This represents an estimate of $P(X_i = 1 | \mathbf{I}_1, \ldots, \mathbf{I}_C)$, the true marginal of the probabilities of presence at every location, given the full signal.

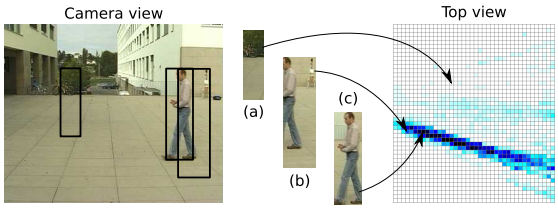


Figure 3: Generation of the *classification score maps*. Images (a), (b) and (c) show sub-windows extracted from the camera view at 3 random locations of the ground plane. Classifiers are applied to sub-images $I_c(i)$ corresponding to every ground plane location i. Images depicting background (a) produce a low classification score for the corresponding location. Images showing badly centered pedestrian (b) produce a slightly higher score and images featuring a well centered pedestrian (c) receive high score.

We compare two approaches for the second step. Section §3.2 describes the one, which is representative of what is usually done by state-of-the-art methods. We refer to it as the *baseline* because it combines the individual classification score maps without taking into account the interactions between presence of pedestrian due to occlusion. By constrast, the second approach takes into account potential occlusions and knowledge about the classifier behavior and yields a substantial increase in performance. It is at the core of our contribution and is discussed in §3.3.

3.1 Classification Score Maps

We introduce the classifier we use for single-view pedestrian detection and to compute our classification score maps.

3.1.1 Classifier as a Pedestian Detector

During a learning step, we create a set of decision trees dedicated to the classification of rectangular images into two classes: “person” or “background”. The binary decision trees we use as classifiers are based on thresholded Haar wavelets operating on grayscale images (Viola and Jones, 2001). They are trained using a few thousands of images of different sizes, each of which represents either a pedestrian correctly cen-

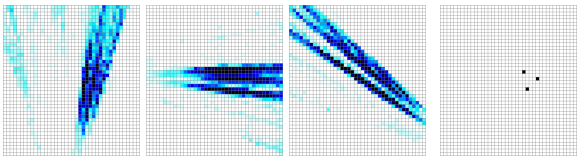


Figure 4: The 3 images on the left show the *classification score maps* of a scene viewed under three different angles. The right image represents the corresponding ground truth.

tered in the rectangular frame, or *background*, which could be anything else.

More specifically, for every tree, several hundreds of features of different scales, orientations and aspect ratios are generated randomly and applied to our training set. The one that best separates the two populations according to Shanon's entropy is kept as the root node and the training set is split and then dropped into two similarly-constructed sub-nodes (Breiman et al., 1984). This process is repeated until either the *person* and *background* sets are completely separated or it reaches the tree maximum depth $d = 5$. Our classifier consists of a forest (Breiman, 1996) of $N_T = 21$ decision trees built in this manner.

3.1.2 Computing Classification Score Maps

The algorithm iterates through every camera and ground location, extracts a sub-image corresponding to the rectangular shape of human size, and takes its score to be the number of trees classifying the sub-image as "person" (Fig. 3).

If we see the individual tree responses as many i.i.d. samples of the response of an ideal classifier, the classification score in location i is an estimate of the probability for such a classifier to respond that i is actually occupied given the subimage at that location. Hence, it is a good indicator of the actual occupancy.

This produces, for each camera, a map such as the ones depicted by the third column of Fig. 1 or by the three left pictures in Fig. 4, which assigns a voting score to every ground location. As shown on those figures, detected pedestrians appear as "cone shapes" in the axis of the camera, on the classification score maps. This is due to the high tolerance in scale and limited tolerance in translation of the classifiers, and hinders precise people location. Hence the need of an extra step, which combines classification score maps from different camera views into one accurate detection score map. Sections §3.2 and §3.3 present two possible methods for this operation.

3.2 Baseline Approach

The *baseline approach* consists of multiplying the responses of the trees from different viewpoints. This is essentially what the product rule used in (Khan and Shah, 2006) does. It is more sophisticated than a crude clustering and averaging in separated views, since it assumes the conditional independence between the different views, given the true occupancy. Recall that $T_c(i)$ is an integer standing for the sum of the trees' answers at location i on camera view c, and $\mathbf{T}$ is the vector of all $T_c(i)$. Formally, we have

$$P(X_i = \alpha \,|\, \mathbf{T}) = P(X_i = \alpha \,|\, T_1(i), \ldots, T_C(i)) \quad (1)$$

$$= \frac{P(X_i = \alpha)}{P(T_1(i), \ldots, T_C(i))} P(T_1(i), \ldots, T_C(i) \,|\, X_i = \alpha) \quad (2)$$

$$= \frac{P(X_i = \alpha)}{P(T_1(i), \ldots, T_C(i))} \prod_c P(T_c(i) \,|\, X_i = \alpha). \quad (3)$$

Equality (1) is true under the assumption that only the responses of the trees at location i bring information about the occupancy at that location, equality (2) is directly Bayes' law, and equality (3) is true under the assumption that given the occupancy of location i, the tree's responses at that location from different camera views are independent.

We then model the probability of the trees' response at a certain point given that it is occupied ($\alpha = 1$) by a density proportional to the number of trees responding at that point, and the probability of response when the location is empty ($\alpha = 0$) by a constant response. This leads to a final rule that multiplies the responses of the trees from the different viewpoints to estimate a score increasing with the probability of occupancy at that point.

3.3 Principled Approach

The baseline method of the previous section assumes that, given the true occupancy at a certain location, the responses of the trees at that point for different viewpoints are independent from each other, and are not influenced by occupancy at other locations. As shown in Section §4, it usually triggers many false alarms. By contrast, our principled approach relies on an assumption of conditional independence of the tree responses at any location i, given the occupancy of the full grid $(X_1, \ldots, X_G)$, and not anymore X_i alone. Such an assumption is far more realistic, and leads to an algorithm which takes into account the long-range influence of both the occlusions between pedestrians and the presence of an individual on the classification score maps, due to the invariance of the classifiers.

3.3.1 Conditional Marginals

We want to compute numerically, at every location i of the ground plane, $P(X_i \,|\, \mathbf{T})$ the conditional mar-

ginal probability of presence given the response of the classifiers at all locations. We will show that computing this quantity requires $P(\mathbf{T}|\mathbf{X})$, the tree response model given the ground occupancy. It is learnt by applying the classifier on sequences for which we have a ground truth, and is described in §3.3.2. As explained below, there is no possible analytical way to obtain $P(X_i|\mathbf{T})$ given our underlying assumptions, hence the need to evaluate it numerically through an iterative process. At each new iteration, the marginal probabilities of presence $P(X_i|\mathbf{T})$ for all ground locations i are reevaluated using their previous estimate, until convergence.

Let $\mathbf{X}_{j\neq i}$ denote the vector $(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_G)$, Q the product law with the same marginals as the posterior $\forall i,\ Q(X_i=1)=P(X_i=1|\mathbf{T})$ and E_Q the expectation under $\mathbf{X}\sim Q$, as summarized in Table 1. To obtain a tractable form for $q_i^\alpha = P(X_i=\alpha|\mathbf{T})$, we first marginalize $\mathbf{X}_{j\neq i}$

$$\begin{aligned} q_i^\alpha &= \sum_{\mathbf{X}_{j\neq i}} P(X_i=\alpha|\mathbf{T},\mathbf{X}_{j\neq i})P(\mathbf{X}_{j\neq i}|\mathbf{T}) \\ &= E[P(X_i=\alpha|\mathbf{T},\mathbf{X}_{j\neq i})|\mathbf{T}], \end{aligned} \quad (4)$$

where $\mathbf{T}$ is equal to the observed trees' answers and the only random quantity in the expectation is $\mathbf{X}$. We then apply Bayes' law to make the model of the trees' answers given the true occupancy state appear

$$q_i^\alpha = E\left[\frac{P(\mathbf{T}|X_i=\alpha,\mathbf{X}_{j\neq i})P(X_i=\alpha,\mathbf{X}_{j\neq i})}{P(\mathbf{X}_{j\neq i}|\mathbf{T})P(\mathbf{T})}|\mathbf{T}\right]. \quad (5)$$

However, there is no analytical expression for (5), and we thus have to estimate the expectation numerically by sampling the $\mathbf{X}_{j\neq i}$ and averaging the corresponding probability. To this end, we substitute the expectation under the true posterior law by a re-weighted expectation under a product law Q with the conditional marginals as marginal

$$\begin{aligned} & q_i^\alpha \\ &= E_Q\left[\frac{P(\mathbf{T}|X_i=\alpha,\mathbf{X}_{j\neq i})P(X_i=\alpha,\mathbf{X}_{j\neq i})}{P(\mathbf{X}_{j\neq i}|\mathbf{T})P(\mathbf{T})}\frac{P(\mathbf{X}_{j\neq i}|\mathbf{T})}{Q(\mathbf{X}_{j\neq i})}\right] \\ &= E_Q\left[\frac{P(\mathbf{T}|X_i=\alpha,\mathbf{X}_{j\neq i})}{P(\mathbf{T})}\frac{P(X_i=\alpha,\mathbf{X}_{j\neq i})}{Q(\mathbf{X}_{j\neq i})}\right]. \end{aligned} \quad (6)$$

Such a formulation ensures that, when we estimate the expectation numerically, the sampling of $\mathbf{X}_{j\neq i}$ will accumulate on the occupancy configurations consistent with the tree responses, thus leading to a far better estimate of the averaging with a reasonable number of samples. Finally we simplify the expression by assuming that the prior distribution is a product law (i.e. $P(\mathbf{X})=\prod_{i=1}^G P(X_i)$)

$$q_i^\alpha = \frac{P(X_i=\alpha)}{P(\mathbf{T})}E_Q\left[P(\mathbf{T}|X_i=\alpha,\mathbf{X}_{j\neq i})\prod_{j\neq i}\frac{P(X_j)}{Q(X_j)}\right]. \quad (7)$$

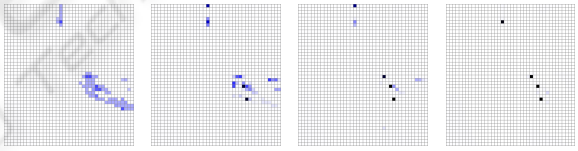
We end up with an expression of each marginal as a function of the other marginals, thus a large system of equations to solve.

This result is intuitive: the conditional marginal probability of presence at location i given the trees' answers can be computed by fixing X_i, sampling all the other X_j according to the current estimate of Q, and averaging the corresponding probability that the trees respond what they actually respond. The more the value associated to X_i makes the actual tree responses likely, the highest its probability.

We get rid of the unknown $P(\mathbf{T})$ quantity by computing

$$P(X_i=1|\mathbf{T}) = \frac{P(\mathbf{T})P(X_i=1|\mathbf{T})}{P(\mathbf{T})P(X_i=0|\mathbf{T})+P(\mathbf{T})P(X_i=1|\mathbf{T})}$$

In the end, we obtain a large number of equations relating the $P(X_i=1|\mathbf{T})$. We can iterate these equations to estimate the conditional marginals. After initialization of all q_is to a prior value, each of these equations can be evaluated numerically by sampling according to a product law Q with the current estimates as marginals. Experimental results show that with such a choice, since the sampling accumulates on the configurations consistent with the observations, a few tens of iterations are sufficient to provide good numerical precision. Fig. 5 shows four iterations of the detection score map convergence process.



iteration #2 iteration #5 iteration #8 iteration #10

Figure 5: Example of convergence of a detection score map during the iterative estimation.

3.3.2 Tree Response Model

At the core of Equation (7) above is $P(\mathbf{T}|\mathbf{X})$, the responses of the trees given the true occupancy state, where $\mathbf{X}=(X_1,\ldots,X_G)$. It must account for effects such as occlusion and classifier invariance. Assuming that the trees' responses are independent given the true state, we write

$$P(\mathbf{T}|\mathbf{X}) = \prod_{c,i} P(T_c(i)|\mathbf{X}). \quad (8)$$

As shown in Fig. 6, the trees' response at position i can only be influenced by ground location j, whose corresponding sub-image $I_c(j)$ intersects the $I_c(i)$. We call such locations the *neighborhood* $n_c(i)$ of i on camera view c. Thus, Equation (8) becomes

$$P(\mathbf{T}|\mathbf{X}) = \prod_{c,i} P(T_c(i)|X_i,\mathbf{X}_{n_c(i)}), \quad (9)$$

where we simply ignore positions outside $n_c(i)$. The classifier response at location i can thus be interpreted as a function of the presence of individuals in the neighborhood of i, as opposed to the whole scene.

In the rest of the section, we show how to express (9) numerically in some simple particular cases, and we then extend it to the general case, thus deriving a model for the classifier response.

Empty Neighborhood. If the neighborhood of i is empty (Fig. 8, (a) and (b)), the trees' answer in i depends only on the occupancy of i. Precisely $\forall \alpha \in \{0,1\}$:

$$P(T_c(i) = t \,|\, X_i = \alpha, \forall j \in n_c(i), X_j = 0) = \mu_\alpha(t). \quad (10)$$

The functionals μ_0 and μ_1 are modeled as histograms estimated on training samples, and shown on Fig. 7.a.

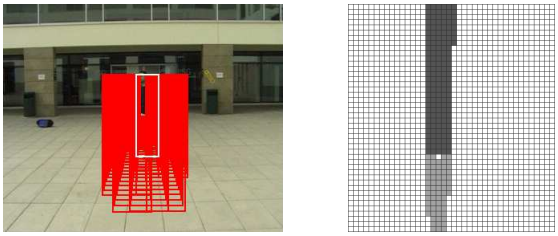


Figure 6: Left image shows the neighborhood $n_c(i)$ in camera view and right image shows it in top view.

One Individual in the Neighborhood. We now consider the case where only one person is present in the neighborhood of i, at location j. If location i is empty, sub-image $I_c(i)$ will contain some body parts of the person present at location j, in addition to background. This influences the classifier answer in i, in a way that depends on the "distance" between $I_c(i)$ and $I_c(j)$ in the image.

To characterize this pseudo-distance between sub-images, we define functions $\alpha(i,j) = \sqrt{\|I_c(i)\|/\|I_c(j)\|}$ and $\beta(i,j) = \delta_c(i,j)/\sqrt{\|I_c(i)\|}$, where $\alpha(i,j)$ quantifies the size ratio between $I_c(i)$ and $I_c(j)$, and $\beta(i,j)$ their misalignment. $\delta_c(i,j)$ is described in Table 1.

With this, we obtain the tree response model $\mu'_0(t,\alpha(i,j),\beta(i,j))$, which is computed as histograms from the training samples. It is plotted on Fig. 7 (c).

We finally model the case where location i is occupied, with one person present in its neighborhood at location j. For this purpose, we have to distinguish positions from the neighborhood located "behind" i – that is, further from the camera than i – and those located closer to it. We denote the former set by $n_c^-(i)$ and the latter by $n_c^+(i)$ and illustrate them geometrically in Fig. 8.

When i is occupied, positions from $n_c^-(i)$ do not influence the classifier answer on $I_c(i)$, but positions from $n_c^+(i)$ do. As for the previous case, we define a pseudo-distance function $\gamma(i,j) = \|I_c(i) \cap I_c(j)\|/\|I_c(i)\| \cdot (1 - \|I_c(i) \cap I_c(j)\|/\|I_c(j)\|)$ with respect to the camera view, to characterize the relationship between the relative position of i and j, and the trees' answer.

We then derive the tree response model for this last case as function $\mu'_1(t,\gamma(i,j))$, which is depicted by Fig. 7 (b). It is also computed empirically as histograms from the training samples.

Multiple Individuals in the Neighborhood. It is not trivial to extend the simplified model with at most one person in the neighborhood to the general case, because the number of neighbor locations is of the order of 100, which implies a huge number of occupancy configurations. We therefore simplify our model by assuming that only the occupied location whose sub-window intersects the most $I_c(i)$ will influence the classifier answer in i, on camera view c. We denote by $J_c^*(i)$ the occupied location from the neighborhood of i, whose corresponding sub-window covers the most $I_c(i)$

$$J_c^*(i) = \underset{j \in n_c(i), X_j = 1}{\arg\max} \|I_c(i) \cap I_c(j)\|. \quad (11)$$

This assumption makes the model tractable and has been found to hold empirically. Finally, we obtain as response model when the neighborhood is not empty, whether there is a single individual or several of them:

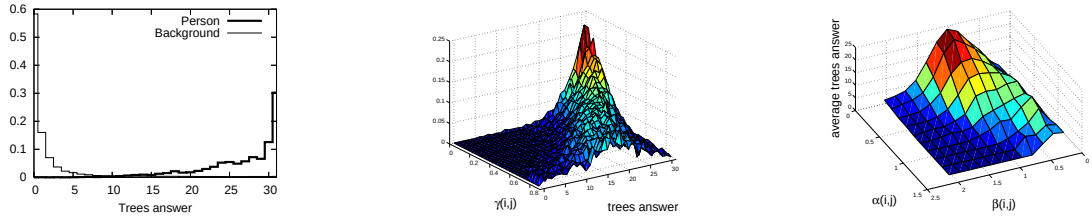
$$P(T_c(i) = t \,|\, X_i = 0, \exists j \in n_c(i), X_j = 1) = \mu'_0(t, \alpha(i, J_c^*(i)), \beta(i, J_c^*(i))) \quad (12)$$

$$P(T_c(i) = t \,|\, X_i = 1, \exists j \in n_c^+(i), X_j = 1) = \mu'_1(t, \gamma(i, J_c^*(i))) \quad (13)$$

4 RESULTS

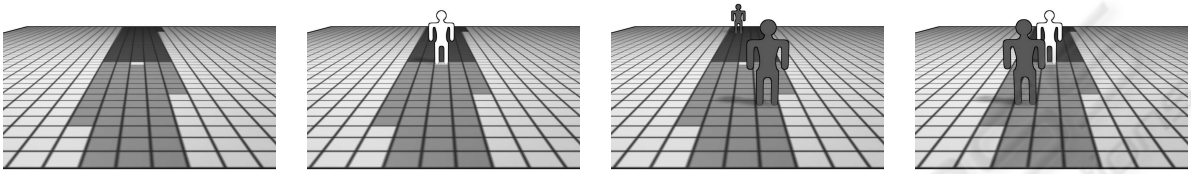
To test our approach, we acquired 30 minutes of video sequences using three outdoor cameras with overlapping fields of view. We used a 2 minutes sequence to train the system and learn the trees response model of § 3.3.2 and the remaining to test it. To demonstrate the generality of the model, we also show results in indoor sequences that were not used for training purposes. Finally, we show that our method yields meaningful results even from single views.

Baseline vs. Principled Approaches. To compare the approaches of § 3.2 and § 3.3, we randomly selected 100 frames of the outdoor sequences, manually labeled the true pedestrian locations, and compared them to both their outputs.



(a) (b) (c)

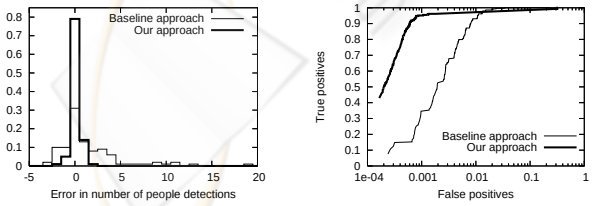
Figure 7: Tree response model. (a) shows the classifier answer distributions for a forest of 31 trees, (b) plots the distribution of the classifier answer as a function of $\gamma(i,j)$ and (c) displays the average trees' answer as a function of $\alpha(i,j)$ and $\beta(i,j)$.



(a) (b) (c) (d)

Figure 8: The images above illustrate the four cases used by the tree response model for the grid position i, colored in white. Grid positions highlighted in gray represent the neighborhood $n_c(i)$ of position i (see also Fig. 6 right, for a top view). The *visible* neighborhood $n_c^+(i)$ is shown in light gray, whereas the neighborhood $n_c^-(i)$ located beyond position i is painted in dark gray. In case (a), neither location i nor its neighborhood is occupied. In case (b), location i is occupied, but its *visible* neighborhood $n_c^+(i)$ is empty. Note that there might or might not be people in $n_c^-(i)$. In case (c), location i is empty, but there is at least one person in its neighborhood $n_c(i)$. Finally in case (d), location i is occupied, as well as at least one of the locations in $n_c^+(i)$. As in case (b), it does not matter whether $n_c^-(i)$ is occupied.

The result depicted by Fig. 9. shows that the principled approach yields much better estimates of the number of people than the baseline approach, which triggers many false positives. When setting the post-processing threshold so that both approaches have about 10% of false negatives, our approach outperforms the baseline one, by producing only about 0.06% of false positives instead of 0.81%. This result is depicted by the ROC curves of Fig. 9.b. Since our method relies on a strong model and produces very peaked occupancy probabilities, detection failures cases produce incorrect occupancy maps. This explains the crossing of the ROC curves at very high detection rates.



(a) (b)

Figure 9: Comparing the performance of the baseline and principled approaches. (a) Error distribution in the estimate of the number of people present in the scene. (b) ROC curves for the two methods. These graphs demonstrate that the principled approach truly provides a better estimate of the number of people present in the scene, and a better false positives vs. false negatives ratio.

Indoor and Outdoor Sequences. Fig. 10 depicts our results in the outdoor and indoor sequences. In both cases, people are correctly detected in spite of very real difficulties: In the outdoor images, there are strong shadows, which could create problem for methods based on background subtraction but do not affect our results. The occlusions in the indoor images are very significant but are nevertheless handled correctly, especially when one recalls that we do not enforce any form of temporal consistency and treat every time frame independently.

Thanks to the tree response model of Section 3.3.2, we can retrieve occupancy maps from the noisy classifier answers, even when using single views as shown in last raw of Fig. 10. The procedure used is the same as in the multi-view case, except that we do no longer multiply tree's answers from multiple cameras in Equation 8. Occlusions are no longer handled, as evidenced by the fact that a half-hidden person in the second image is missed. Nevertheless, the results remain meaningful.

5 CONCLUSIONS

We have shown that explicitly computing marginal probabilities of target presence given classifier responses is more reliable and accurate than simply averaging the responses across views for multi-view

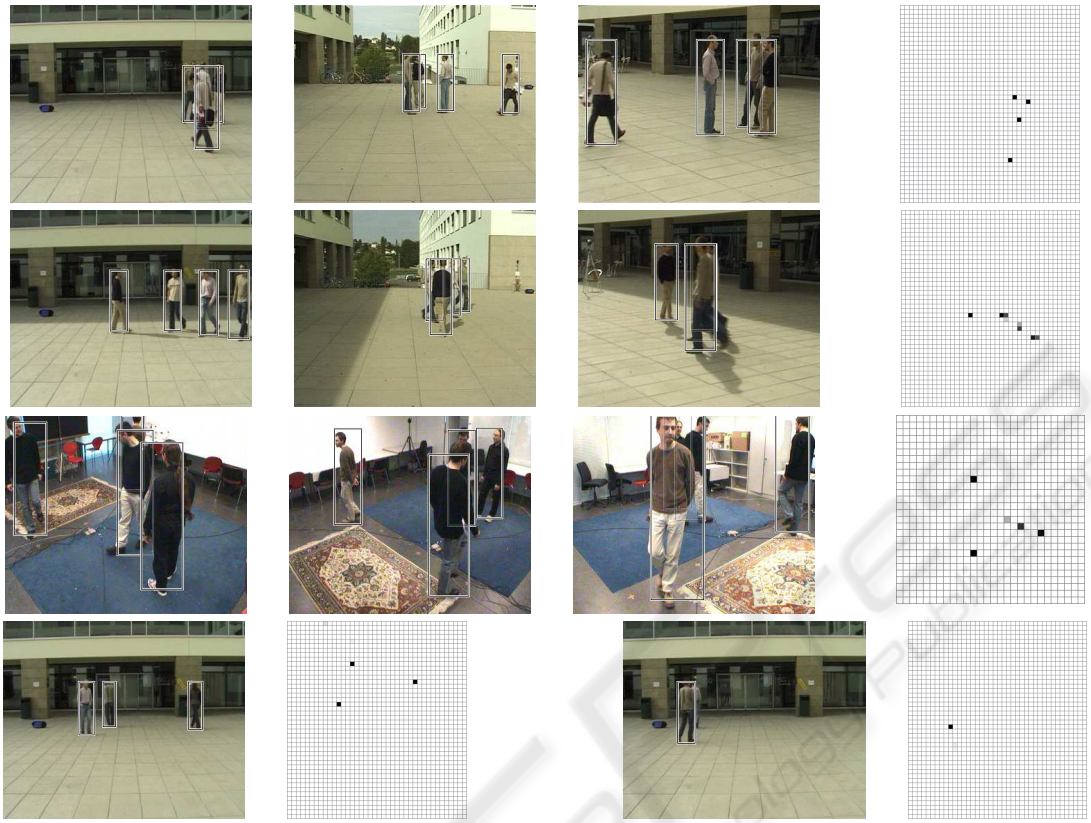


Figure 10: Results of our algorithm on real video sequences. Each one of the first three rows shows several views taken at the same time instant from different angles. Boxes are located on local maxima of the estimated probabilities of occupancy. The last column depicts the corresponding detection score map before thresholding. The last row shows two detection results obtained from single images.

people detection purposes. This is especially true in challenging situations with complex interactions between true alarms due to occlusion and very low metric accuracy in the classifier responses. Experiments show that this method allows for a reduction of one order of magnitude of false positives. As a result, we have been able to demonstrate reliable people detection at single time frames and without having to impose any temporal consistency constraints. Finally, our approach to post-processing multiple classifier outputs is generic and could be applied to other detection-by-classification problems, for which a model of the classifier response given the true detection state is available, either directly or through learning.

REFERENCES

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall, New York.

Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *CVPR*.

Elfes, A. (1989). *Occupancy Grids: A Probabilistic Framework for Robot Perception and Navigation*. PhD thesis, Carnegie Mellon University.

Fleuret, F. and Geman, D. (2002). Fast Face Detection with Precise Pose Estimation. In *CVPR*.

Khan, S. and Shah, M. (2006). A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *ECCV*.

Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *CVPR*.

Mittal, A. and Davis, L. (2003). M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. *IJCV*.

Okuma, K., Taleghani, A., de Freitas, N., Little, J., and Lowe, D. (2004). A boosted particle filter: multitarget detection and tracking. In *ECCV*.

Viola, P. and Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *CVPR*.

Viola, P., Jones, M., and D.Snow (2003). Detecting pedestrians using patterns of motion and appearance. In *ICCV*.

Zhao, T. and Nevatia, R. (2001). Car detection in low resolution aerial image. In *ICCV*.