Person Detection for Indoor Videosurveillance using Spatio-Temporal Integral Features

Adrien Descamps[1], Cyril Carincotte[2], and Bernard Gosselin[1]

[1] TCTS Lab, University of Mons, Mons, Belgium
[2] Multitel ASBL, 2 Rue Pierre et Marie Curie, Mons, Belgium

Abstract. In this paper, we address the problem of person detection in indoor videosurveillance data. We present a new method based on the state of the art integral channel features. This approach is extended to allow the use of temporal features in addition to appearance based features. The temporal features are integrated by a robust background subtraction method. Our method is then evaluated on several datasets presenting various and challenging conditions typical of videosurveillance context. The evaluation shows that additional temporal features are efficient and improve greatly the performance of the detector.

1 Introduction

Person detection in images and videos is a very active research topic in computer vision. It remains a challenging task, due to the huge variability of pedestrian appearance arising from changing pose, clothing, lighting and point of view. Inter-person occlusions, high background variability and/or small resolution images make the problem even harder in many real world scenarios.

In this paper, we address this problem in the particular context of indoor videosurveillance. Many recent works about person detection focus on automotive application [12, 9, 13], or on person detection in high resolution static images [10], but only a few of them have investigated the use of person detectors for videosurveillance and the particularities of this application [11]. The videosurveillance context has indeed some specific aspects compared to others applications : static background, relatively low resolution images, typical points of views, etc. These specific aspects impose constraints, especially the capacity to detect low resolution persons, and in densely occupied scenes. They also have advantages, the most important one being the static background, that allows use of methods like background subtraction to take into account the temporal aspect of the video data.

The method we propose integrates such temporal information into a classical appearance-based person detector. To do so, we extend integral channel features introduced in [14], and integrate a robust background subtraction method [7] in this algorithm.

We evaluate our method on various challenging data from public CAVIAR dataset [16] and a non-public real world dataset from the VANAHEIM project

[18]. Our experiments show that temporal feature can improve greatly the performances of the detector, even in conditions in which they are generally considered unreliable, and that our method achieves performance similar to state of the art, with lower detection time.

The rest of this paper is organized as follows. A brief review of related works is presented in Section 2. Section 3 presents our approach : we describe the original integral features and their extension to temporal features. Experiments are described in Section 4 and results are reported in Section 5.

2 Related work

In the literature, some early approaches for persons detection were mainly based on background subtraction (see [3]), but these methods suffer from a high sensitivity to background variability, are limited to low density of persons and are not usable with moving camera.

On the other hand, the vast majority of recent methods are based on machine learning methods and use discriminative classifier scanned over the images. One of the first method achieving good performances was the Haar cascade proposed by Viola-Jones [1], which uses simple Haar filters as feature and a cascade of adaboost classifiers. The cascade of classifiers and the use of integral image to compute Haar feature allow to achieve a very fast detector.

The Histogram of Oriented Gradient feature, described in [4], has proven very effective for person detection. The HOG feature represents the intensity distribution of the gradient depending on its orientation, and model well the shape of the person.

In [11], covariance feature was shown to be effective and, combined with a background subtraction method, leaded to a fast and effective detector with good performances in videosurveillance context. The idea of utilizing both appearance and temporal features was also developed by [15], which utilize appearance features together with short-term (frame differencing) and long-term (background subtraction) motion information.

A new generic feature described in [14], called integral channel feature, consist of sums over local rectangular regions of multiple image channels computed using linear and non-linear transformations of the input image. Combined with a soft cascade adaboost classifier [8], this feature is shown to be fast and efficient despite its simplicity, and is a generalization of other features based on integral images, like Haar or HOG.

Several studies have been conducted to standardize procedures to assess performances and compare these methods with each other [10, 12]. These studies show the overall prevalence of methods based on the HOG feature and the value of combining different features to improve performances. They also show that if the best methods reach good performance on high-quality and high-resolution images, they fail in the case of more realistic scenarios, with occlusions, variable quality images and variable background.

3 Algorithms and Method

3.1 Appearance Model

The basis of our method is the algorithm proposed in [14], which uses integral channel features based on appearance only. In this algorithm, for an input image I, a set of N channel images $C_i, i \in [1:N]$ are computed by $C_i = F_i(I)$, where $F_i(I)$ can be any operation. Integral images of these channel images are then computed, and are denoted by C_i^I. The features are defined as sum of pixels in a rectangular region of a channel image for first-order feature, or a linear combination of such sums for higher-order feature. In this work, we only use first-order feature, as higher order features were reported to have very little impact on performances ([14]). Given the integral channel images C_i^I, any feature can be computed very efficiently by four accesses and three addition operations.

The parameters of each feature are thus the rectangular region (x_j, y_j, w_j, h_j) and the channel index i_j. A set of M features is generated by choosing these parameters randomly.Despite the high number of possible features, this strategy was shown to be efficient in [14].

A soft cascade adaboost classifier is then trained using these features. The soft cascade is a variant of the cascade of classifiers, in which a rejection threshold is used after evaluation of every weak classifier, instead of using multiple distinct cascade layers. A classical adaboost classifier using decision tree as weak classifiers is first trained. Then, using [8], the rejection thresholds are determined as the lowest values allowing to reach a given detection rate on a given dataset.

The parameters of the method are thus M the number of features, the parameters of the adaboost classifier, and most importantly, the channels used. The only constraint about these channels is that the generation function F_i must be translationally invariant, $T(F_i(I)) = F_i(T(I))$, T being any translation operation. In [14], the author tested many possible channels, and retained a configuration of eleven channels, composed of three channels for the original image in the LUV color space, one channel for the gradient magnitude, and six channels for the Histogram of Oriented Gradients.

The HOG feature is composed by the normalized histogram of the gradient magnitude relative to its orientation over a subregion of the image. It can be computed with integral images as shown in [2], using one channel per histogram bin. The feature extracted from these channels are normalized a posteriori by the gradient magnitude channel, which corresponds to the L1 normalization in [4]. This feature has been shown to be very efficient to model shape of humans.

3.2 Temporal Information Extension

We propose to extend this method by using the temporal information of the video in the channels. For this, we extend the channel extraction function $C_i = F_i(I)$ to a more general one $(C_{i,t}, S_{i,t}) = F_i(I_t, S_{i,t-1})$, t being the time index. $S_{i,t}$ represents the state of the channel i at time t. This state allows the channel extraction function to use information about the previous frames of the video.

Given the new formula, we can integrate almost any video features into the method. This is illustrated in the following sections by integrating a background subtraction method.

Background subtraction We use the background subtraction algorithm described in [7], which uses a multi-layer model of background based on color and texture. In this case, the channel is a foreground probability image, providing the probability of each pixel to be part of foreground, and the state is the background model generated by the method. We have : $(C_{i,t}, B_t) = F_{i,t}(I_t, B_{t-1})$, B_t being the background model. Thus, at each frame, F_i is called with the current frame and the previous background model as input, and return a foreground probability image and an updated background model.

The foreground probability is an interesting feature which eliminates many false alarms, but is not discriminant in itself, since any moving object give similar response. The shape of the foreground is much more discriminant. We thus use, in addition to the foreground probability channel, one gradient magnitude channel and six HOG channels computed on the foreground image. These channels are computed similarly to those computed on the input image. However, unlike the classical HOG, we do not discard the information about the direction of the gradient. Indeed, this information is relevant for foreground image, because a person is supposed to appear as a high probability blob in a low probability background, and not the opposite.

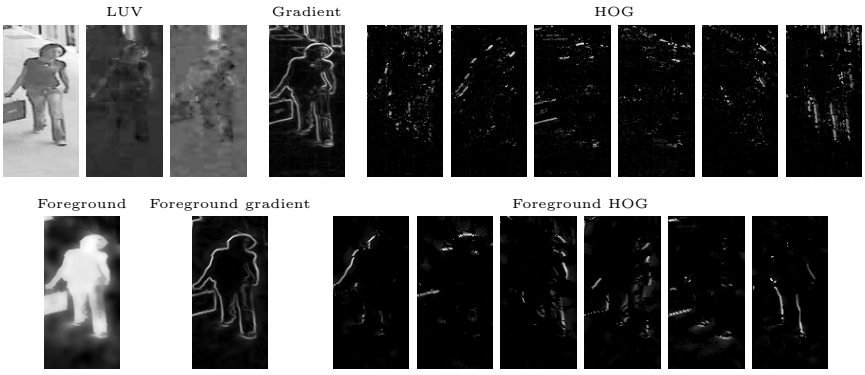
Our final method can thus use up to 18 channels : 10 for appearance and 8 for background subtraction. A sample of these channels is presented at figure 1. It can be observed that the temporal channels bring important information about the presence and the shape of a foreground object. We can also see that foreground HOG channels are less noisy and represent better the shape of the person than appearance based HOG channels. These observations must be moderated by the sensitivity of the background subtraction method to occlusion, variable background and low-contrast persons, but they demonstrate the value of using temporal features in addition to appearance.

4 Experiments

4.1 Training and Testing Data

For training and testing our method, we use several public videosurveillance datasets and a dataset from Turin metro station coming from the VANAHEIM project. For training, around 9200 positive samples are extracted from CAVIAR, PETS2009, AVSS2007 and VANAHEIM datasets. These samples come with their corresponding foreground images. Negative samples are extracted by extracting 5000 random windows from a set of 88 background images coming from various datasets, collecting 5000 additional samples by bootstrapping once on these images, and 5000 more by bootstrapping a second time on training data containing persons. The aim of the second bootstrap is to collect hard negative samples

Fig. 1. Sample images of channels



that do not appear in images containing only background, like parts of persons or groups of persons.

We evaluate our method on CAVIAR and VANAHEIM data not used during training. The CAVIAR data contains two views of a shopping center, and we use 12 short sequences of each view containing 1400 annotations of persons (with high redundancy). The VANAHEIM data contains 17 views, with 2500 annotated persons. Note that these data present a high variability and challenging condition like changing light, moving escalators, ground reflections,or high density of persons.

The high proportion of persons is seen in very low resolution in these data, and this leads us to choose a detection window size of 18x36 pixel. We can note that this size is smaller than reported in most publications : for example, a window size of 64x128 was used in [4, 10, 9, 14] (INRIA dataset [17]), and in ETHZ dataset ([6]), minimum person size is 60 pixels . Some authors [9, 13] use the Daimler dataset, with a window size of 18x36 pixels, and report the difficulty to detect persons at such low resolution.

4.2 Evaluation Methodology

For the evaluation, we consider here that only unoccluded standing people must be detected. This hypothesis is of course limiting for applications, but is justified by the fact we don't use any occlusion handling method and the difficulty of the dataset. Other persons are marked as "hard" and are optional (they do not count as false positive if detected but do no need to be detected). Performances are evaluated for four category of persons: "far", which are less than 36 pixel high, "medium", which are between 36 and 72 pixels high, "near" which are more than 72 pixel high, and "global", which includes all the persons. The performances are evaluated following [12], by applying the detector on annotated images, performing a non maximum suppression to suppress multiple detections and matching

detection bounding box with ground truth using PASCAL criterion, with an overlap threshold of 0.5. We use a simple non maximum suppression method ([14]) that suppresses the less confident of every pair of detections that overlap sufficiently according to the PASCAL criterion.

Original method, with only appearance channels (i.e. [14], referred further as Dollar), and our extension with foreground channels are evaluated. For comparison purpose, Opencv Haar cascade , Opencv LBP cascade, original HOG [4] and method based on covariance and background subtraction from [11] were also evaluated. Note that we use the same background subtraction algorithm as the covariance method, while other methods use only appearance. Except for covariance method, these methods are retrained with our dataset, and we used same training data and same parameters.

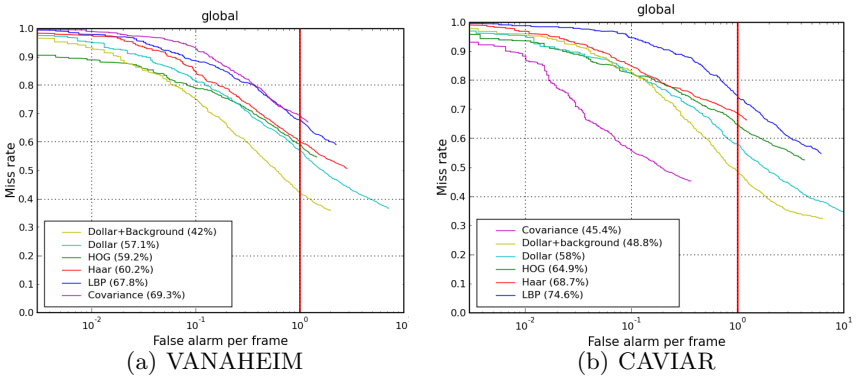
5 Results

Figure 2 shows performances of all methods on CAVIAR and VANAHEIM datasets. Note that VANAHEIM results are filtered using calibration and a ground plane hypothesis, but it doesn't influence their ranking.

The appearance methods, Haar, HOG and Dollar, give similar results, except for Dollar which outperforms slightly over methods in CAVIAR dataset. LBP method doesn't perform well on both datasets. The integration of background in Dollar improves performances greatly on both dataset and gives the best results in VANAHEIM.

The performances of the Covariance algorithm vary very much, being the best for CAVIAR, but very poor for VANAHEIM. However, this detector could not be retrained on the same dataset, due to the unavailability of the training code. This can affect greatly the results and make it difficult to draw conclusions, especially considering the CAVIAR dataset was used in the training stage [11].

Fig. 2. Global results on VANAHEIM and CAVIAR datasets



(a) VANAHEIM (b) CAVIAR

Finally, figure 3 shows performances evaluated globally on all the data, with curves for the different category of size defined above. Globally, we see the prevalence of methods which use temporal information: the Covariance method and our method with foreground. As expected, the curves show best results for near persons, and worst results for far persons. For medium and near persons, we see that foreground improves greatly performance. The gain due to foreground is especially high for medium scale, the near scale being already relatively well handled by appearance methods.

For far persons, we see that the integration background subtraction decrease performance slightly. A priori, it is for these far persons that temporal information should be the most useful, compensating the lack of appearance information. The results seem to indicate that foreground information is not discriminative enough at low scale to bring useful information.

Fig. 3. Global results

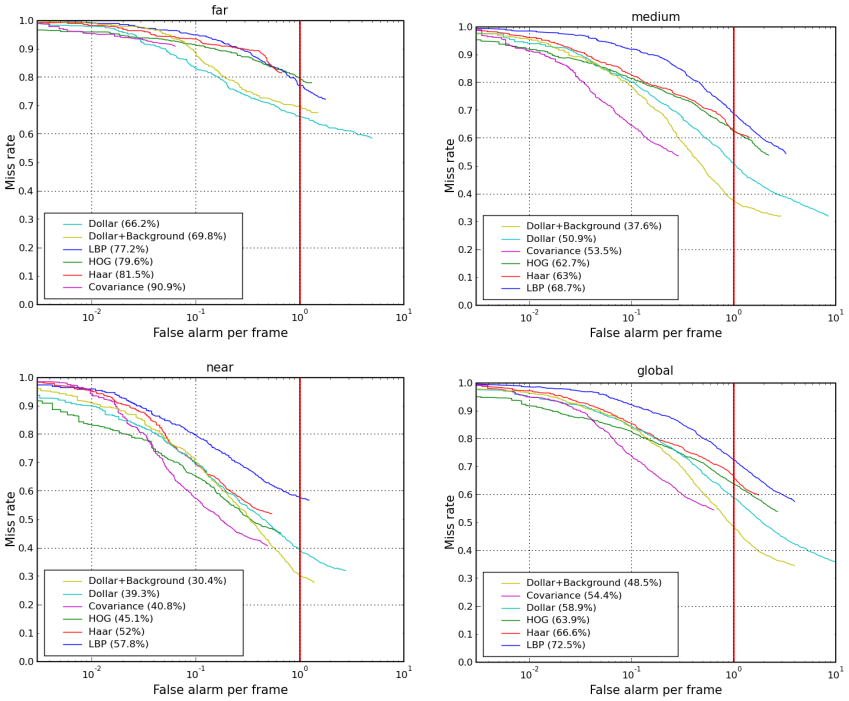


Table 1 resume the results, showing the global detection rate at one false alarm per image, and the detection time of all the methods. The detection time is reported for VANAHEIM video, at 704x288 resolution. We see that our method with background subtraction gives the best compromise between performances and detection time.

Table 1. Global detection rate at one false alarm per image.

Algorithm	VANAHEIM	CAVIAR	**Global**	Detection time (s)
Haar	39.8%	21.3%	33.4%	0.1
HOG	40.8%	35.1%	36.1%	3
LBP	32.2%	25.4%	27.5%	0.1
Covariance [1]	30.7%	**54.6**%	45.6%	2.5
Dollar	43.0%	42.1%	41.2%	0.5
Dollar+Background	**58.0**%	51.2%	**51.5**%	1.5

6 Conclusion

We presented a new pedestrian detector that integrates simply and efficiently appearance and temporal feature. The value of using temporal features for pedestrian detection in videosurveillance was demonstrated by evaluating the performance of our method on various and challenging data. However, performances achieved by the best detector are still far from those needed for real-world videosurveillance applications.

Many improvements of our method are possible. The first one is the addition of a tracking method to the detector, that should eliminate some false alarms and miss detections in persons trajectories. The features have still to be further investigated to find additional efficient temporal features, e.g. based on optical flow. Finally, the inter-person occlusions are an important problem in many cases in videosurveillance, and modelling explicitly these occlusions could help to improve greatly performances in such cases.

7 Acknowledgement

The research leading to these results has received funding from the European Communitys Seventh Framework Programme FP7/2007-2013 - Challenge 2-Cognitive Systems, Interaction, Robotics - under grant agreement n 248907-VANAHEIM.

References

1. P. Viola and M. Jones.: Rapid object detection using a boosted cascade of simple features. Computer Vision and Pattern Recognition, 2001.
2. F. Porikli.: Integral histogram: A fast way to extract histograms in cartesian spaces. Computer Vision and Pattern Recognition, 2005.

[1] For covariance method, the curve doesn't reach the rate of one false alarm per image for CAVIAR and global datasets. The detection rate is thus reported at 0.3 and 0.7 false alarm per image respectively.

3. Neeti A. Ogale.: A survey of techniques for human detection from video. Master Thesis, University of Maryland, 2006.
4. N. Dalal. Finding people in images and videos. PhD Thesis, Institut National Polytechnique de Grenoble, 2006.
5. Q. Zhu, S. Avidan, M.C. Yeh, and K.T. Cheng.: Fast human detection using a cascade of histograms of oriented gradients. Computer Vision and Pattern Recognition, 2006.
6. A. Ess, B. Leibe, and L. Van Gool.: Depth and appearance for mobile scene analysis. International Conference on Computer Vision, 2007.
7. J. Yao and J.M. Odobez.: Multi-layer background subtraction based on color and texture. Computer Vision and Pattern Recognition, 2007.
8. C. Zhang and P. Viola.: Multiple-instance pruning for learning efficient cascade detectors. Neural Information Processing Systems, 2007.
9. O. Tuzel, F. Porikli, and P. Meer.: Pedestrian detection via classication on riemannian manifolds. Pattern Analysis and Machine Intelligence, 2008.
10. C. Wojek and B. Schiele.: A performance evaluation of single and multi-feature people detection. Lecture Notes in Computer Science, 2008.
11. J. Yao and J.M. Odobez.: Fast human detection from videos using covariance features. Computer Vision Visual Surveillance Workshop, 2008.
12. P. Dollar, C. Wojek, B. Schiele, and P. Perona. : Pedestrian detection: A benchmark. Computer Vision and Pattern Recognition, 2009.
13. M. Enzweiler and D.M. Gavrila.: Monocular pedestrian detection: Survey and experiments. Pattern Analysis and Machine Intelligence, 2009.
14. P. Dollar, Z. Tu, P. Perona, and S. Belongie.: Integral Channel Features. British Machine Vision Conference, 2009.
15. J. Zhang and S. Gong.: People Detection in Low-Resolution Video with Non-Stationary Background. Image and Vision Computing, 2009.
16. CAVIAR dataset : http://homepages.inf.ed.ac.uk/rbf/caviar/.
17. INRIA dataset : http://pascal.inrialpes.fr/data/human/.
18. VANAHEIM project : http://www.vanaheim-project.eu/, 2010-2013.