People detection in complex scene using a cascade of Boosted classifiers based on Haar-like-features

M. Siala[1], N. Khlifa[1], F. Bremond[2], K. Hamrouni[1]
1. Research Unit in Signal Processing, Image Processing and Pattern Recognition (ENIT, Tunisia)
2. PULSAR Team (INRIA, Sophia Antipolis, Nice)

***Abstract*—Pedestrian detection in a real scene is an interesting application for video surveillance systems. This paper presents our contribution to improve the work of Viola and Jones, originally designed to detect faces. This work uses a cascade of classifiers based on Adaboost using Haar features. It improves the learning step by including a decision tree presenting the different poses and possible occlusions. The method has been tested on real and complex sequences and has given a good detection despite occlusions and poses variation.**

I. INTRODUCTION

In this paper, we focus on the problem of detecting people in video data, such a system could be used in surveillance systems, driver assistance systems, and image indexing. Detecting people in images is more challenging than detecting many other objects due to several reasons. The main challenges of people detection in video sequences are that there is a large variation in the appearance due to changes in clothing and cameras positions. The most difficulty is that in crowded scene, there is a large amount of occlusion among people which makes the task of people detection very difficult.

The cascade of classifiers based on Adaboost is a good and robust method for characterization and detection, but it presents some limitation in complexes scenes. That's why we propose to study this method and try to improve its performance to detect people in complex environment.

The paper is organized as follows: section 2 summarizes some important work, the next section presents briefly the cascade of Adaboost method and explains the proposition for improving its training phase. The last section gives some results..

II. RELATED WORK

There are several techniques that have been proposed in the literature addressing the problem of people detection. Here, we will only present a few of the more recent ones.

Papageorgiou and al. have successfully employed example-based learning techniques to detect people in complex static scenes without assuming any a priori scene structure or using any motion information. Their system detects the full body of a person. Haar wavelets [4] are used to represent the images and Support Vector Machine (SVM) is used to classify the patterns [5], in this paper The system derives much of its power from a representation that describes an object class in terms of an over-complete dictionary of local, oriented, multi-scale intensity differences between adjacent regions, efficiently computable as a Haar wavelet transform. This example-based learning approach implicitly derives a model of an object class by training a support vector machine classifier using a large set of positive and negative examples. Authors present results on face, people, and car detection tasks using the same architecture. In addition, they quantify how the representation affects detection performance by considering several alternate representations including pixels and principal components. We also describe a real-time application of their person detection system as part of a driver assistance system.

Regarding person detectors that incorporate motion descriptors, Haar feature [2], first introduced by Viola and Jones for face detection, have been also used for people detection by Viola et al. [1] and an extension of these have also been proposed by Lienhart et al. [3].

Recently, Dalal and Triggs have further developed this idea of histogram of gradient and have achieved excellent recognition rate of human detection in images [6], They study the question of feature sets for robust visual object recognition, adopting linear SVM based human detection as a test case. they show that grids of Histograms of Oriented Gradient (HOG) descriptors significantly outperform existing feature sets for human detection. They study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results.

Leibe et al. [7] and [9] developed an effective static-image pedestrian detector for crowded scenes by coding local image patches against a learned codebook and combining the resulting bottom up labels with top-down refinement. The approach considers object categorization and figure-ground segmentation as two interleaved processes that closely collaborate towards a common goal. The approach is a learned representation for object shape that can combine the information observed on different training examples in a probabilistic extension of the Generalized Hough Transform

Mikolajczyk et al. [8] use position-orientation histograms of binary image edges as image features, combining seven part detectors to build a static-image detector that is robust to occlusions.

In [10], Yao et al. present a fast method to detect humans from videos captured in surveillance applications. It is based on a cascade of LogitBoost classifiers relying on features mapped from the Riemanian manifold of region covariance matrices computed from input image features.

Indeed, methods cited above do not address occlusion situations. In addition, the cascade of AdaBoost method presents difficulties for the people detection in complex scenes. We will try then to improve its detection rate in real scenes.

III. The Ameliorated/ Improved Cascade of Adaboost Method

Very often in crowded scenes, people are only partially visible to the camera. Hence approaches, that attempt to detect full body, fail in most cases. We have adopted an approach that uses the coarse-to-fine strategy to divide the entire body space into smaller and smaller subspaces to tackle this problem. We learn several body poses of humans using AdaBoost separately, and obtain detectors for each of these body poses. Annotated data of these body poses are fed, during training, separately to AdaBoost algorithm that use Haar features to generate reliable classifiers for the corresponding body poses as in Fig. 1. During the training phase, we typically tune each classifier to obtain a high detection rate even at the cost of a higher false alarm rate. However, the proposed algorithm is able to reliably detect humans and reject false alarms despite the higher false alarm rates of the initial classifiers.

In the next subsection, we will introduce and define the adaboost algorithm, before presenting our improvement proposal.

Fig. 1. Different poses of a person

A. AdaBoost

AdaBoost, short for Adaptive Boosting, is a machine learning algorithm [11], formulated by Y. Freund and R.Schapire. It can be used in conjunction with many other learning algorithms to improve their performance. AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. Otherwise, it is less susceptible to the over fitting problem than most learning algorithm.

A variant of AdaBoost is used both to select a small set of features and train the classifier. In its original form, the AdaBoost learning algorithm is used to boost the classification performance of a simple (sometimes called weak) learning algorithm.

The main idea of boosting algorithms is to combine many simple and moderately accurate hypotheses (called weak classifiers) into a single, highly accurate classifier for the task at hand. The weak classifiers are trained sequentially and, conceptually, each of them is trained mostly on the examples which were most difficult to classify, by the preceding weak classifiers.

The boosting algorithm takes as input a training set of m examples $S = (x1, y1), \ldots (xm, ym)$ where each instance xi is a vector of attribute values that belong to a domain or instance space X, and each label yi is the class label associated with xi that belongs to a finite label space Y which $Y = \{0, 1\}$. In addition, the boosting algorithm has access to another unspecified learning algorithm, called the weak learning algorithm, which is denoted generically as weak learner. The boosting algorithm calls the weak learner repeatedly in a series of rounds. Adaboost manipulates the training examples to generate multiple hypotheses. Adaboost maintains a probability distribution $p(x)$, or weight function over the training examples. In each iteration t, it weights the training samples the probability distribution $p(x)$. The learning algorithm is then applied to produce a classifier h. The error rate ε of this classifier on the training examples (weighted according to $p(x)$) is computed and used to adjust the probability distribution on the training examples (note that the probability distribution is obtained by normalizing a set of weights $w(i)$ over the training examples. The effect of the change in weights is to place more weight on training examples that were misclassified by h and less weight on examples that were correctly classified in the last stage. In subsequent iterations, therefore, Adaboost tends to construct progressively more difficult learning problems. The final classifier, *hfinal*, is constructed by a weighted vote of the individual classifiers $h1, h2, \ldots$. Each classifier is weighted according to its accuracy for the distribution p that it was trained on.

B. Improvement of the training phase

The boosted cascade method was used by Viola and Jones originally for face detection, there results were good. They improve their works for pedestrian detection but they used a simple database (images present from one to two persons).

We propose to improve the results of detection of the Viola and Jones work, by changing the learning method by adopting the principle of Huang et al.[11], and this to detect people in complex scene.

In [13], Huang et al. have proposed a learning method to

detect faces. They proposed a WFS tree (Width-First-Search), which is based on a decision tree that divides the entire face space into smaller and smaller subspace.

In the present work, we propose to generalize this learning method to detect human body. We have to divide the human body configuration in various poses while including the occlusion case.

An example of our tree is shown in Fig. 2.

During training phase, we have to train each body poses separately front and rare pose (full left and right profile pose, half left and right profile pose).

People being non-rigid objects in nature have different shapes and appearances based on the pose and articulation of the person.

There is a large variation in the appearances due to varied clothing, natural variation in the height/shape of the person and variation in the camera views. Large amount of occlusions in the scene complicate the process, that's why we have trained each case of occlusion (person/person or person/object) with different poses of humans separately (front and rare occluded pose, full left and right profile occluded pose, half left and right profile occluded pose).

In order to overcome the above problems, we have done an analysis on the performance of each detector (by observing its detection and false alarm rates during detection process) for a given body poses and chose the most reliable ones. We identified 6 distinct poses of a person and each pose was trained separately so that a person in any possible pose will be detected. We have obtained detectors for each pose trained pose. From these detectors, we obtained 6 different detectors for the 6 occluded body poses. So incorporating all the distinct body poses (simple and occluded), we finally obtain 12 pose detectors that are used in the algorithm.

Each pose detector is tuned to give a very high detection rate with few false alarms allowed. The advantage of this requirement is that the classifier cascade is reasonably small and therefore fairly fast.

During the detection phase, we should combine the various body pose (simple and occluded) detectors to get a unified detection of a person.

IV. RESULTS

To evaluate the method performance, we have used the NIST video database which has been constructed for TrecVid 2008 challenge from the Gatwick airport of London [12].

Results on samples images from camera scene are shown in Fig. 3. The contrast of the camera videos is poor and the resolution is low (720 x 576 pixels). We can observe that the algorithm is able to detect people in spite of many such difficulties. There is a wide variation in clothing and baggage they are carrying. People are walking at different locations in the scene with different poses. We can also observe that the people in the scene are crowded and occluded. However, the proposed algorithm is able to detect people in such adverse conditions; it is able also to detect people in all possible poses.

We have tested the performance of our algorithm in three other video sequences obtained from the same database as shown in Fig. 4. The algorithm has detected most people in the scene despite their small size and the bad luminosity.

In training, we use more than 1800 images including varied appearance of people in all poses and their clothes, as positive samples and we use about 2200 background images from varied scene as negative samples in the algorithm. We use specific data samples to take advantage of fix cameras. Such number of data samples is fed into AdaBoost algorithm in order to obtain efficient and robust detectors.

During the training phase, the cascade parameters (number of layers) and threshold values (false positive rate (0.6) and false negative rate (0.1) for the strong classifiers) are tuned to give us a higher detection rate.

Fig. 3. Example of Detection results

In the test phase, we have annotated about 270 images corresponding to 1540 people visible in these images. 1321 people have been properly detected by our method, which gives a good detection rate of 80%. It follows that the number of missing detections for our method is fairly low, the rate of false positives remains little bit high, around 60%. It will be evident from the literature that the above mentioned detection rate are high for a complex real crowded scene, when the image quality and its resolution is poor. These rates have been obtained by taking 90% of samples for training and 10% for testing. The results mentioned above are summarized in table 1.

TABLE I
DETECTION'S QUANTIFICATION

works	proposed Method	Viola and al Method
Number of Images	270	270
Number of persons	1540	1540
TP (good detection)	85.77%	30.45%
FP (false positif)	60% 2000/3321	60.65% 723/1192
missed Detection	14.29%	69.54%

I. CONCLUSION

In this paper, people detection, is performed using AdaBoost algorithm, which allows selecting discriminative features and combining many weak classifiers to obtain a strong one.

We have improved the training phase of Viola and Jones work by adopting the Huang and al. work. We have divided the human body in various poses while including the occlusion case and we have trained each body poses separately.

We conclude that the improved algorithm is robust in nature and works well in various data conditions. It gives a good detection rate despite occlusions, poor people resolution and poses and scale variation.

As future works, we propose to use a large database and the leave-one-out method in the training phase. We can also, try to add body part detection (face, foot, shoulder) to improve the people detection rate, and use 3D windows of human size and moving regions to speed up the process. Finally, we can add tracking to improve the global detection rate.

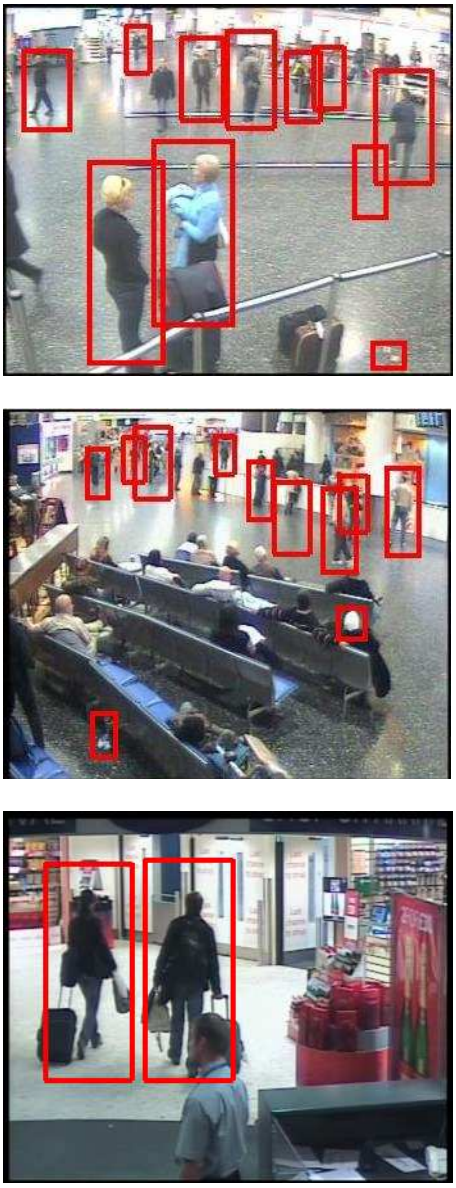

Fig. 4. Detection results in other video sequences

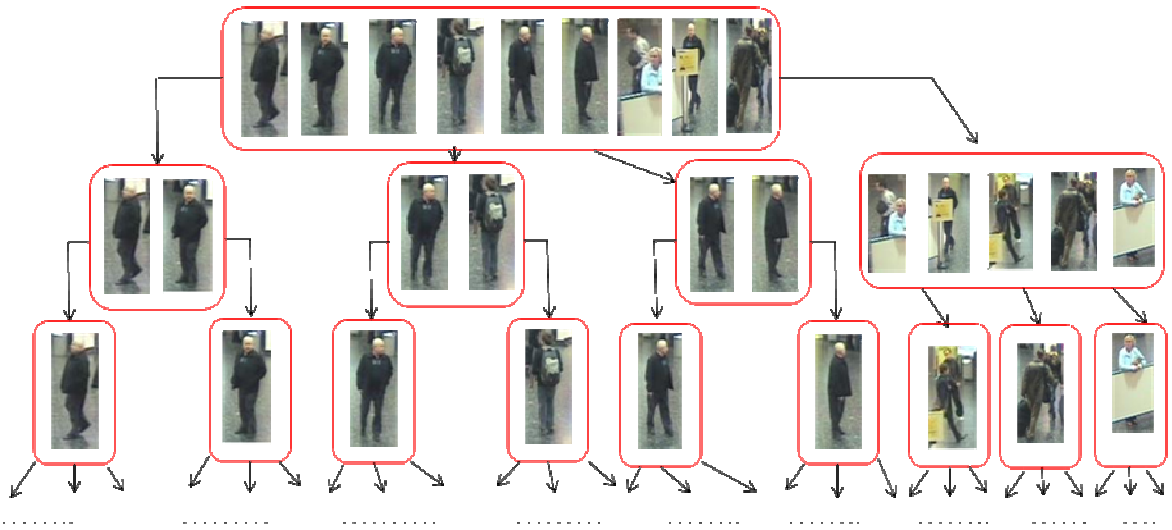

Fig. 2. The proposed decision tree used to define the body poses

REFERENCES

[1] P. Viola, M.J. Jones et D. Snow, "*Detecting Pedestrians Using Patterns of Motion and Appearance*", Proceedings of the Ninth IEEE International Conference on Computer Vision 2-Volume Set, ICCV 2003.

[2] P. Viola et Michael J. Jones. "*Robust real-time face detection*", Int. J.Comput. Vision, 57(2):137–154, 2004.

[3] R. Lienhart, J. Maydt, "*An Extended Set of Haar-like Features for Rapid Object Detection*", IEEE ICIP 2002, Vol. 1, pp. 900-903, Sep. 2002.

[4] S. Mallat, "*A Theory for Multiresolution Signal Decomposition: The Wavelet Representation*", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674-693, July 1989.

[5] C. Papageorgiou et T. Poggio, "*A trainable system for object detection*". International Journal of Computer Vision, 38(1):15–33, 2000.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[7] B.Leibe, E.Seemann et B.Schiele, "*Robust object detection with interleaved categorization and segmentation*", IJCV Special Issue On Learning for Vision and Vision for Learning, august 2007.

[8] K. Mikolajczyk, C. Schmid, and A. Zisserman, "*Human detection based on a probabilistic assembly of robust part detectors*", In Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic, volume I, pages 69–81, 2004.

[9] B. Leibe, E. Seemann, and B. Schiele, "*Pedestrian detection in crowded scenes*", In Proceedings of the Conference on Computer Vision and Pattern Recognition, San Diego, California,USA, pages 876–885, June 2005.

[10] J. Yao and J.M. Odobez, "*Fast Human Detection from Videos Using Covariance Features",* in 8th European Conference on Computer Vision Visual Surveillance workshop (ECCV-VS), Marseille, Oct. 2008.

[11] C.Huang, H.Ai, Y.Li et S.Lao, "*Vector Boosting for Rotation Invariant Multi-View Face Detection*", International Conference on Computer Vision (ICCV), 2005.

[12] A. F. Smeaton, P. Over and W. Kraaij, "*Evaluation campaigns and TRECVid"*, In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330.