Fast and viewpoint robust human detection in uncluttered environments

Paul Blondel, Alex Potelle, Claude Pégard, Rogelio Lozano

Fast and viewpoint robust human detection in uncluttered environments

Paul Blondel [#1], Alex Potelle [#1], Claude Pégard [#1], Rogelio Lozano [#2]

[#1] *Université Picardie Jules-Vernes, 80000 Amiens, France*
[#2] *Université de Technologie Compiègne, 60200 Compiègne, France*

[1] firstname.lastname@u-picardie.fr [2] rogelio.lozano@hds.utc.fr

***Abstract*—Human detection is a very popular field of computer vision. Few works propose a solution for detecting people whatever the camera's viewpoint such as for UAV applications. In this context even state-of-the-art detectors can fail to detect people. We found that the Integral Channel Features detector (ICF) is inoperant in such a context. In this paper, we propose an approach to still benefit from the assets of the ICF while considerably extending the angular robustness during the detection. The main contributions of this work are: 1) a new framework based on the Cluster Boosting Tree and the ICF detector for viewpoint robust human detection, 2) a new training dataset for taking into account the human shape modifications occuring when the pitch angle of the camera changes. We showed that our detector (the PRD) is superior to the ICF for detecting people from complex viewpoints in uncluttered environments and that the computation time of the detector is real-time compatible.**

***Index Terms*—human detection, machine learning, multi-viewpoint, viewpoint robust, supervised training**

I. Introduction

The performance of object detectors can be impacted by changes in the camera's viewpoint. These algorithms can be inoperative if they are not robust enough. Most of the human detection algorithms in the literature suffer from this problem because they are designed to work with a very specific view: the pedestrian view. For numerous applications the viewpoint can be complex and/or can change over time because the camera is moving: the video surveillance, sports event filming, aerial filming, etc. Theorically there is an infinite number of possible camera's viewpoints. During the detection stage the detector should be capable of dealing with the maximum of these cases.

A. Human detection algorithms

For a more generalized use of the detection algorithm: no assumptions are made about the pose of the camera, or about the movement of people. Thus, detection methods based on background substraction are not well suited. The detection process should only be based on the visual information contained in one frame. Monolithic and part-based detectors fit this condition.

Fig. 1. a) results obtained with the ICF detector [1] b) and with the PRD: the detector is less sensitive to changes of shape and angle.

1) Monolithic detectors: Monolithic detectors search for monolithic parts of the image looking like people. Papageroriou et al [2] use wavelet descriptors, a sliding-window method to exhaustively scan the image and a SVM classifier. Many of current detectors are based on this approach. Viola et al [3] use integral images and a cascade classifier to speed up the computation of the Haar-like wavelet features and reach real-time performance for face detection. The Histogram of Oriented Gradients (HOG) detector of Dalal and Triggs [4] is an efficient human detector using a variant of the very well-known and quite efficient SIFT descriptor [5].

Dollár et al [1] proposed a mix between Viola et al's detector and the HOG detector: the ICF detector. It computes simple rectangular features on integral images of different channels. The classification is performed using a fast soft-cascade classifier.

2) Part-based detectors: Part-based detectors consider the human body as a set of parts. Felzsenzwald et al [6] propose a method to detect people by fragments and re-build a human model by using a pictural structure reprensentation. Each part of the human model is separately learned. These detectors are slower than monolithic detectors and are not well adapted to detect far-off situated people.

B. The multi-viewpoint context

Let's first define the camera's viewpoint: it is the view obtained through this camera and for a specific configuration of its roll, pitch and yaw angles (Fig.2 a). The camera has six degrees of freedom.

In addition to the perspective effect, the impact of the

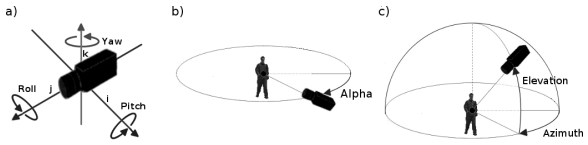

Fig. 2. a) camera angles, b) camera poses in a pedestrian context, c) camera poses in a general context.

Fig. 3. Examples of people images with different roll and pitch angles.

camera angles on the human appearance are the followings: the roll angle tends to rotate the shape of people and the pitch angle tends to change the shape of people (Fig.3).

When the visual changes are too important, they cannot be managed by a detector designed to detect people in a pedestrian view (Fig.2). This lack of robustness can be due to the combination of several factors: the nature of the visual descriptor, an unadapted training dataset, and/or the scanning process itself.

1) The nature of the visual descriptor: Some decriptors are natively dependant on orientation, such as: Haar-like [3] and HOG [4] descriptors. A rotation invariant descriptor should have both its shape and its metric invariant to rotation. However, in our case, rotation invariant descriptors are not a solution because we have to deal with the pitch angle of the camera which causes changes of people appearance (Fig.2 and Fig.3).

2) The training dataset: Most human training datasets are not adapted to deal with changing viewpoints. To the best of our knowledge only one training dataset is designed to permit an improvement of the detection capabilities for changing viewpoints: the GMVRT-v1 dataset[1]. Nevertheless, this dataset only takes into account changes implied by the pitching of the camera. The ideal dataset should allow the training algorithm to face the maximum of viewpoints during the training, i.e: with rolling and pitching combined.

3) The scanning process: This part is often designed to work for a single view: the pedestrian view. Detectors using the sliding window approach often use a vertical one-half ratio detection window to scan images for human candidates in different places, and for different depths [3][4][1]. This technique is obviously not appropriate when there is an important roll of the camera, because, in this case, human shapes are not vertical but rotated.

C. Content of the paper

The goal of this work is to reach fast and viewpoint robust detection of upright human beings for applications in uncluttered environments. This work is mainly about a new framework designed to combine the assets of the ICF detector to the robustness of a multi-view classifier.

[1]http://mis.u-picardie.fr/∼p-blondel/papers/data

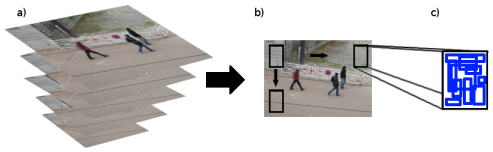

Fig. 4. a) image pyramid, b) scanning with the sliding window, c) computation of the visual features (in blue).

Section II is about the ICF detector and the training framework. The second part of this section is about the adaptations proposed to reach fast and viewpoint robust human detection. Section III is about the tests conducted on a test dataset containing images with complex viewpoints. The results of the tests are also discussed in this section.

II. Proposed approach

There are different approaches for improving the robustness to viewpoints with a supervised monolithic detector: training a combination of binary classifiers, training a multi-class classifier or training a view-aware binary classifier.

However, detecting human people whatever the viewpoint is a pure binary problem: only the presence or the non-presence of human people is of interest. Considering one class considerably simplify the labelization problem as well as the tuning of the parameters.

A. ICF detector

Among all the monolithic supervised pedestrian detectors the ICF detector of Dollár and al [1] is a good candidate to approach a fast and competitive detection [7]. This part describes the two phases of the detector.

1) The detection phase: The input image is exhaustively scanned by a sliding detection window (Fig.4 c). To detect persons at different distances all the levels of an image pyramid are scanned as showed in Fig.4 a.

The visual features are computed on integral images of ten different channels, which are: L, U, V, gradient magnitude and six "HOG channels" [7]. A feature is simply the sum of the pixels contained within a rectangle and associated to one of the channels. A person is detected in a window if the set of visual features matches the human model.

A coarse-to-fine approach is adopted to speed the classification: the soft-cascade. Dollár et al proposed to approximate the features between pyramid levels to speed up the detection: the Fastest Pedestrian Detector In the West (FPDW) [8], and using Aggregated Channel Features (ACF) [9]. These techniques can also be used with our PRD for greater speed performance.

2) The training phase: During the training phase AdaBoost [10] is used to select the best discriminant features on positive and negative training images. The final classifier is a combination of weak-classifiers (depth-2 decision trees).

B. The new framework

Wu et al [11] show that AdaBoost can fail to find an optimal solution when the positive training dataset contains important variations of shape. They propose a method to cope with

this problem by learning the different aspects of the same object class: the Cluster Boosting Tree (CBT). The idea is to clusterize (k-means) the training positive dataset during the training process to reach an optimal solution for each cluster. A binary tree structure of the training dataset is generated by the clusterizations. This allows both sharing and optimized choices of features [11].

The three following sections are about the adaptations of the ICF detector and the CBT training algorithm for viewpoint robust human detection. We named the final solution the pitch and roll-trained detector (PRD).

1) ICF adaptations: The features are computed in a circle with a radius of 64 pixels. Dollár and al recommend to use 30.000 random feature condidates [1]. As the surface of the circle is about 1.5 bigger than the surface of the classic detection window, 45.000 features candidates are generated in order to keep a relatively similar density of candidate features. The classification is performed by a depth-first search tree traversal of the classifier (trained with CBT).

2) CBT adaptations: Alg.1 presents our modified version of the CBT, see [11] for the original implementation. Our version is lighter and proved to be more efficient for our needs. Lines 7 and 8 (Alg.1) are typical AdaBoost procedures, except that S_+^k is a subset of S_+^0 for k>0. Line 9 is the condition to trigger the clusterization: the classification power (h(t,k).Z) of the three latest trained weak-classifiers are compared to θ_Z. Lines 12 and 13 are retraining procedures of the previously trained weak-classifiers. We found it better to set θ_Z to 0.98 and to authorize as many clusterizations as possible.

Fig. 5. Examples of GMVRT-v2 images.

3) Adapted training data: The GMVRT-v1 dataset was a first attempt to deal with the pitch angle, but the lack of samples forced us to re-build a new training dataset: the GMVRT-v2[1] (Fig.5). This new dataset contains 3846 images of people taken at different pitch angles of the camera (extracted from one hundread aerial movies). Some INRIA positive training images were added to complete the dataset.

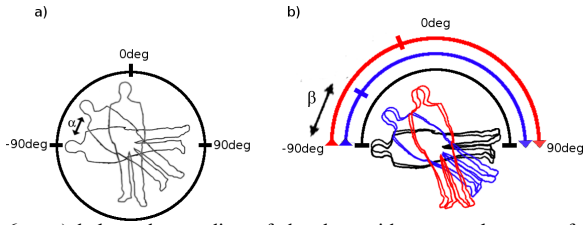

Fig. 6. a) balanced spreading of the data with an angular step of αdeg, b) multiple balanced spreading shifted by angular offsets multiple of β (one color = one spreading).

Subsets of images are rotated every 5 degrees during the training and between -90 to 90 degrees (α = 5degrees, Fig.6.a). We named it an angular spreading. Spreading the data reduces the learning robustness. We overcame this problem by repeating the same operation several times with different angular offsets (β = angular offset, Fig.6.b). The first spreading is from -90deg to 90deg, the second spreading is from -90+βdeg to βdeg, etc. It guarantees more positive samples by degree.

input : GMVRT-v2 training dataset[1]
output: viewpoint robust classifier

1 *several angular spreading of the data*;
2 *extracting all candidate features for all the data*;
3 $c \leftarrow 1$;
4 **for** $k \leftarrow 0$ **to** c **do**
5 *reset default weights of* S_+^k *and* S_-;
6 **for** $t \leftarrow tinit(k)$ **to** T **do**
7 *build best weak-classifier h(k,t)*;
8 *update weights of* S_+^k *and* S_-;
9 **if** *h(k,t).Z* $> \theta_Z$ *and h(k,t-1).Z* $> \theta_Z$ *and h(k,t-2).Z* $> \theta_Z$ **then**
10 *split* S_+^k *into* S_+^k *and* S_+^{c+1} ;
11 *h(c+1,t') = h(k,t'),* $\forall t' \in [0, t]$;
12 *retrain weak-classifiers h(k,t'),* $\forall t' \in [0, t]$ *with* S_+^k *and* S_-;
13 *retrain weak-classifiers h(c+1,t'),* $\forall t' \in [0, t]$ *with* S_+^{c+1} *and* S_-;
14 *tinit(c+1) = t*;
15 *c++*;
16 **end**
17 **end**
18 **end**
19 $\forall k \in [0, c]$ *compute the soft-cascade for channel k*;

Algorithm 1: Our CBT implementation. c: number of clusterizations, k: index of the cluster (or channel), T: maximum number of weak-classifiers, h(k,t): weak-classifier number t of channel k, tinit(k): starting index for cluster k, θ_Z: clustering critera, S_+^k: cluster k of positive image, S_-: all the negative images.

III. Tests

In the following section different aspects of the detector have been tested and compared: the general performance, the computation time and the angular robustness. The general performance and the computation time have been tested on a dataset of 210 images taken from complex camera's viewpoints[1]. Each image contains from 3 to 5 people. The angular robustness has been tested on two other datasets of 180 images each: a dataset for testing the rolling robustness and a dataset for testing the pitching robustness[1]. We named our final solution: the pitch and roll-trained detector (PRD).

The color code is the following: in red, the original Integral Channel Features (ICF) detector, in green, a detector trained with the GMVRT-v2 dataset for pitch robustness (pitch-trained detector or PD), in blue, a detector trained with the INRIA dataset but for different roll angles (roll-trained detector, RD) and in purple, the pitch and roll-trained detector (PRD).

A. The general performance (Fig.7)

The ICF detector fails to succeed in most cases (Fig.7). There is a slight improvement of the detection performance

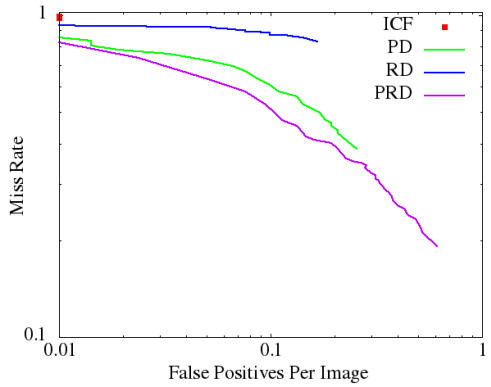

Fig. 7. ROC curves for the four detectors. The two best performance are obtained with the PD and the PRD detector. The ICF detector clearly fails on this dataset.

when the roll of the camera is taken into account during the training (RD). The improvement is bigger when the pitch of the camera is considered instead (PD). The best performance are obtained when both the pitch and the roll of the camera are considered (PRD).

B. The angular robustness (Fig.8 and 9)

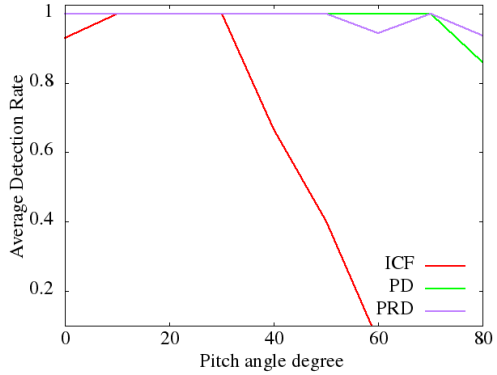

Fig. 8. The PD and the PRD are quite robust to pitch angle variations unlike the ICF.

1) The pitch angle robustness: The performance of the ICF begins to fall down from about 35 degrees (Fig.8). The average detection rate is null at 60 degrees. Conversely, the PD and the PRD have relatively similar average rates whatever the pitch angle between 0 and 80 degrees included.

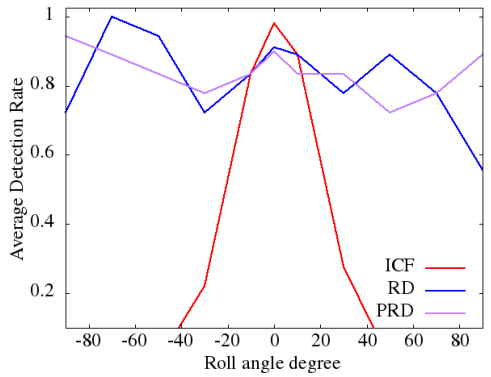

Fig. 9. The RD and the PRD are more robust to roll angle variations than the ICF. The PRD seems to be more stable though.

2) The roll angle robustness: The average rate of the ICF is null before -40 degrees and after 40 degrees. The average rates can be considered acceptable for human detection between -20 and 20 degrees. The RD and the PRD have relatively stable average detection rates from -90 to 90 degrees. However, the PRD seems more stable than the RD.

C. The computation time (Tab.1)

	ICF	**PD**	**PRD**
w/o FPDW	T	T	1.75×T
w/ FPDW	0.35×T	0.38×T	1.05×T

TABLE I
COMPUTATION TIME.

The average speed of the RD detector is not tested in this part due to its poor detection performance. The PRD is 1.75 times slower than the ICF and the PD. This slow-down is due to the number of weak-classifiers of the PRD. The tested PRD has six times more weak-classifiers than the PD or the ICF. The FPDW optimizations allow the PRD to approximatively reach the average speed of the ICF, which is one of the fastest pedestrian detector of the state-of-the-art [7].

IV. CONCLUSION

In this paper, we proposed a new detector particularly adapted to moving cameras where the viewpoint is likely to be complex and changes over time: the pitch and roll-trained detector (PRD). We showed that the PRD outperforms the Integral Channel Features (ICF) detector for complex camera's viewpoints. The main contributions of this work are: 1) a new framework based on the Cluster Boosting Tree and the ICF detector for viewpoint robust human detection, 2) a new training dataset for taking into account the human shape modifications occuring when the pitch angle of the camera changes. The next objective is to reinforce the detector with other features or signals to extend the capabilities of the detector to cluttered and more complex scenes such as urban scenes.

REFERENCES

[1] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features," in *Proceedings of the British Machine Vision Conference*, 2009.
[2] C. Papageoriou and T. Poggio, "A Trainable System for Object Detection," *International Journal of Computer Vision*, 2000.
[3] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Computer Vision and Pattern Recognition*, 2001.
[4] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Conference on Computer Vision and Pattern Recognition*, 2005.
[5] D. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision - Volume 2*, 1999.
[6] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision*, 2005.
[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Conference on Computer Vision and Pattern Recognition*, 2009.
[8] P. Dollár, B. S., and P. Perona, "The Fastest Pedestrian Detector in the West," in *British Machine Vision Conference*, 2010.
[9] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection ," *Transactions on pattern analysis and machine intelligence (TPAMI)*, pp. 1–14, 2014.
[10] R. E.Schapire and S. Yoram, "Improved Boosting Algorithms Using Confidence-rated Predictions," *Machine Learning*, 1999.
[11] B. Wu and R. Nevatia, "Cluster Boosted Tree Classifier for Multi-View, Multi-Pose Object Detection," in *Computer Vision and Pattern Recognition*, 2007.