Human Detection in Digital Videos Using Motion Features Extractors

Rodrigo F. S. C. Oliveira
Universidade de Pernambuco (UPE), Recife, Brazil
Email: rfsco@ecomp.poli.br

Carmelo J. A. Bastos-Filho
Universidade de Pernambuco (UPE), Recife, Brazil
Email: carmelofilho@ieee.org

***Abstract*—Human detection in digital videos is challenging since the human appearance may widely vary. Several algorithms to detect humans in digital images have been recently developed, such as the Aggregated Chanel Features (ACF). Most of them are based on features related to the shape. These algorithms give the best results regarding accuracy but generate many false alarms. In this paper, we propose to use motion features in the ACF to accurately detect humans in digital videos. Three motion feature channels are assessed: MBH, IMHcd and WSTD. The IMHcd presented the best results within the ACF. We demonstrate that our proposal returns more accurate results than the original ACF and presents a reduction in false positive detection rate.**

I. INTRODUCTION

The detection of humans in digital images and videos is a repetitive task and appears in many applications, such as for autonomous robots, pedestrian protection, and surveillance.

Human detection in digital images can be defined as Given an image (or a sequence of images), the algorithm has to determine the locations of these people in the analyzed images, by identifying all the bounding boxes that contain humans. This task is challenging since the human appearance can present many variants on their characteristics, such as height, body shape, color, and texture. Moreover, the characteristics of urban scenes make this task even more challenging, such as complex background, the position of the camera, uncontrolled and unbalanced illumination and occlusion.

Many algorithms for Human detection have been developed in the last years, including approaches based on sliding window detection, Haar wavelets Features, and Support Vector Machines, shape features called Histogram of Oriented Gradient (HOG), color features and texture features.

Aggregated Channel Features (ACF) was proposed by Dollar et al. [1] and it is based on the shape analysis. ACF is considered as one of the state-of-art algorithms. ACF uses a multi-scale sliding window approach aiming to associate it with an output window that can be a candidate for an image of a human being. Then, is evaluates several features channels and sums and smooths every block of 4×4 pixels in these channels. This process results in lower resolution channels with the same size and different amount of layers, in which features are single pixel lookups in the aggregated channels. These aggregated channels are transformed into feature vectors that are used to train a classifier. ACF uses ten features channels: the normalized gradient magnitude (1 layer), the histogram of oriented gradients (6 layers), and the LUV color channels (3 layers). ACF uses Boosting to train and combine decision trees over vectors of features produced as explained before. AdaBoost is used to train and combine 2048 depth-two trees over the $128 \cdot 64 \cdot 10/16 = 5120$ candidate features in each 128×64 window. ACF can achieve good accuracy rates, but produces a reasonable amount of false positives (e.g. background pieces misclassified as human).

II. NEW ACF EXTENSION

We propose an extension of the ACF for human detection in monocular images. Our proposal uses the extraction of characteristics using motion information combined with other features and mechanisms already deployed in the ACF. The following motion features extractors are tested in this paper: *Motion Boundary Histogram* (MBH) [2], *Internal Motion Histogram Central Difference* (IMHcd) [2] and *Weak Stabilized Temporal Difference* (WSTD) [4]. Besides, a shape feature extractor called Entropy of Histogram Oriented Gradient (EHOG) is proposed.

MBH produces channel features containing twelve layers and computes the optical flow between two consecutive frames, uses the *Lucas-Kanade* algorithm to estimate the optical flow and apply the same procedure used to compute HOG features. IMHcd produces a six layers channel and can filter out camera-centric motion and object-centric motion, whereas it captures part-centric motion such as movements of articulated body parts. This features extractor aims to capture movements of legs, arms and other articulated parts of humans from a sequence of images. WSTD is a hybrid motion features extractor that uses optical flow and the subtraction of frames. This features extractor can filter out camera-centric motion and object-centric motion while captures part-centric motion. Therefore, these features can capture movements of arms, legs, and other articulated parts.

EHOG consists of a feature channel that has one layer. It may be considered an extension of HOG features where the *Shannon* entropy is used to measure the histogram uniformity. This feature can distinguish between regions with similar shapes and areas with diverse forms, for instance, background and humans. The first step is to calculate histograms of gradient orientation for every $c \times c$ pixels cell in the image. These histograms are computed in the same way of HOG features, resulting in a map of cells, in which each cell has one histogram with six bins, and each bin represents an

angle interval. These histograms are normalized dividing each bin frequency by the overall incidence. This normalization makes each histogram a kind of random variable where the angles ranges are the outcomes and its frequencies are the distributions. The Shannon entropy is computed for each one of these six random variables.

III. Experimental Results

We performed some evaluations using the Caltech Pedestrian data set [3], the most used benchmark for human detection. We used the subset called "reasonable". It has 10 hours of real scenes recorded by a camera mounted on the front of a car driven through urban areas with a regular traffic of pedestrians and automobiles.

We deployed a 10-fold cross validation for the experiments. The detection samples used as positive and negative examples are randomly selected from the detection windows extracted from the images of the training or test collection. The ACF and motion feature extractors parameters are set up as suggested by their original papers. For more information about these parameters, please refer to [1], [2], [4].

We used Miss rate (MR) and False Positive Per Image (FPPI) as the performance metrics [3]. MR means the number of humans in an image that was not detected divided by the total number of people present in the pictures. On the other hand, FPPI is the average number of false alarms generated by the detector divided by the number of non-humans wrongly classified as humans (false positives). We used the methodology described in [3] to eliminate the ambiguity. By this rule, a detected bounding box and a ground truth bounding box matches if their overlap area exceeds the detection threshold, set to 0.5 in our paper.

The first goal is to determine which extractor of motion features is the most indicated for the human detection together with the ACF framework. From preliminary results, we observed that all detectors presented a similar performance for high detection threshold values. However, IMHcd gives the lowest MR for an FPPI equal to 10^0. Fig. 1 shows the boxplots of the MR for the four combinations of motion features extractors with the original ACF features. This graph demonstrates that IMHcd reaches the best performance considering the log-average MR along the 10-fold cross validation process. Therefore, IMHcd is the most accurate motion features extractor among the assessed features.

Fig. 2a shows a comparison between the ACF + IMHcd and ACF. One can observe that the ACF + IMHcd is more accurate than the original ACF as a whole. This improvement is more evident for higher FPPI values. Another improvement reached by the proposed new ACF extension is the reduction of false alarms emitted compared to the original ACF.

We present in Fig. 2b the FPPI values for MR reference level equal to 10%. Comparing the boxplots, one can observe the new ACF extension offers lower values of FPPI than the original ACF. This boxplot gives the performance for a low detection threshold setting in which a higher amount of false alarms is often emitted. However, the new ACF extension can reduce the false alarms emission considerably. Therefore, the new ACF extension produces less false alarms than the original ACF.

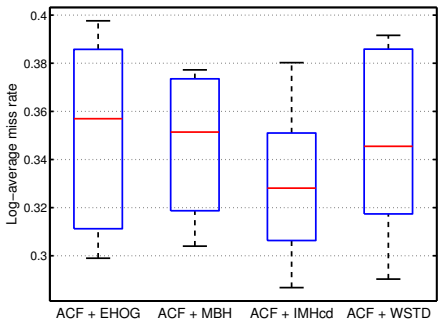
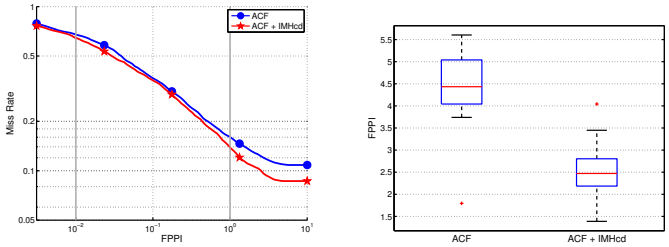


Fig. 1. Box plot of log-average MR for human detectors generated by combining motion features extractors with the original ACF features.



(a) (b)

Fig. 2. (a) Curve MR x FPPI for the ACF and the ACF + IMHcd and (b) Box plots of FPPI values on MR reference level equal to 10%.

IV. Conclusions

This paper proposes a human detector that it is an extension of the ACF detector [1]. We showed that the use of the IMHcd motion features extractor improves the detection accuracy and mitigates the false alarms emission. Emission of false alarms is usually a critical point for human detector employment in most of the real-world applications. Specifically, the new ACF extension is more accurate than the original ACF, mainly in low detection threshold settings. Our proposal also reaches lower FPPI values for individual miss rate reference values compared to the original ACF, primarily for small values of MR.

References

[1] Appel, R., Perona, P., Belongie, S.: Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **99**(PrePrints), 1 (2014)

[2] Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Proceedings of the 9th European Conference on Computer Vision - Volume Part II, ECCV'06, pp. 428–441. Springer-Verlag, Berlin, Heidelberg (2006)

[3] Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(4), 743–761 (2012)

[4] Park, D., Zitnick, C.L., Ramanan, D., Dollar, P.: Exploring weak stabilization for motion feature extraction. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13, pp. 2882–2889. IEEE Computer Society, Washington, DC, USA (2013)