Gradient Histogram Background Modeling for People Detection in Stationary Camera Environments

Víctor Borjas
University of Barcelona
Computer Vision Center
victor.borjas@ub.edu

Jordi Vitrià
University of Barcelona
Computer Vision Center
jordi.vitria@ub.edu

Petia Radeva
University of Barcelona
Computer Vision Center
petia.radeva@ub.edu

Abstract

One of the big challenges of today person detectors is the decreasing of the false positive rate. In this paper, we propose a novel framework to customize person detectors in static camera scenarios in order to reduce this rate. This scheme includes background modeling for subtraction based on gradient histograms and Mean-Shift clustering. Our experiments show that the detection improved compared to using only the output from the pedestrian detector reducing 87% of the false positives and therefore the overall precision of the detection was increased significantly.

1 Introduction

Within the marketing business, several analysis regarding concurrence of people in a specific hall, area, or zone are of great importance to define the value of marketing spots, effectiveness of merchandising programs, better store locations, and many others. People detection is therefore critical for these studies.

People detection in images and videos has been a widely researched field. There has been significant progress in the last decade [2, 5, 6] just to mention a few. The accuracy of these algorithms is satisfactory. Most of them are trained in a generic way that tries to deal with all the possibilities of person and background. This approach is very useful as a general tool for detection, adding up the near real-time performance those algorithms have lately arrived to [1]. However, this approach requires a huge training set to cover a very large variety of viewpoints, resolutions, lighting conditions, blur effects and many other variations. This condition leads to the drop in accuracy in specific video sequences. Let us take surveillance as an example. Variations in viewpoints, resolutions and background are considerably reduced and make easier to train a detector for this specific scene.

Recently, many efforts have been made that aim to train specific scenario detectors [7, 8, 9]. In [10], the authors developed an automatic framework that trains a generic detector for a specific scenario using the technique of Transfer Learning.

On the other hand, background subtraction is widely used for detecting moving objects from static cameras. The main idea is that detecting moving areas on the current frame comes from the difference from a reference frame (also called "Background frame" or "Background Model"). Many different methods have been proposed including Gaussian average models [12], Temporal median filters [13], Mixture of Gaussians [14], etc.

Our proposal presents a novel approach to model background and joins both, people detection and background subtraction, to reduce the amount of false positives. This new scheme has as advantages the ease of implementation, the lack of a retraining step of the detector, and the modular design of the algorithm. All of these characteristics help in the reduction of time in the stages of development and installation of real life applications.

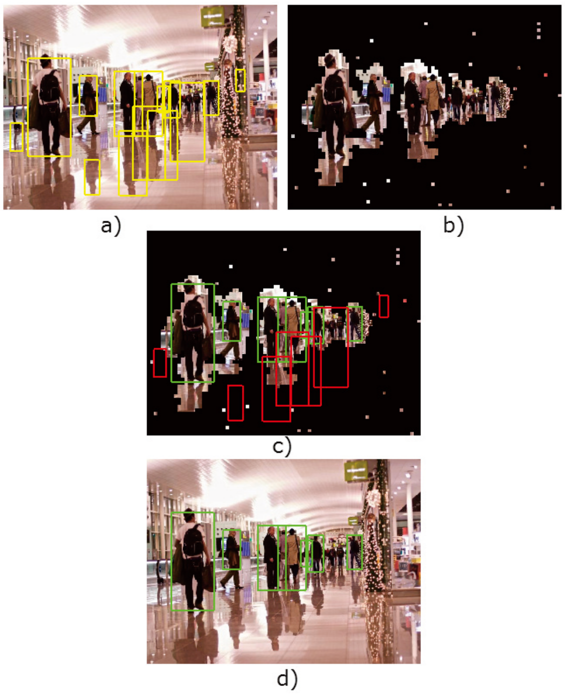

Figure 1. a) People Detection, b) Background Subtraction, c) Intersection between detections and foreground, d) Final result

2 Our Approach

Given the objective of improving a person detector within a stationary camera scenario, we propose a parallel post-processing scheme that applies background per-cell gradient histogram modeling to define possible foreground areas and therefore reject false positives from the generic person detector.

2.1 People Detection

Up to our knowledge, Dollár et al. in [1] developed one of the best state-of-the-art real-time people detector:

Fastest Pedestrian Detector in the West: [1] A multi-scale pedestrian detector based on [4] uses a novel re-scaling technique to construct the image pyramid nearly on real time. The key insight is that the feature responses, in this case gradient histograms [2], computed at a single scale can be used to approximate feature responses at nearby scales. This approximation accelerates the detection 10-100 times with only a minor 1-2% of loss in accuracy.

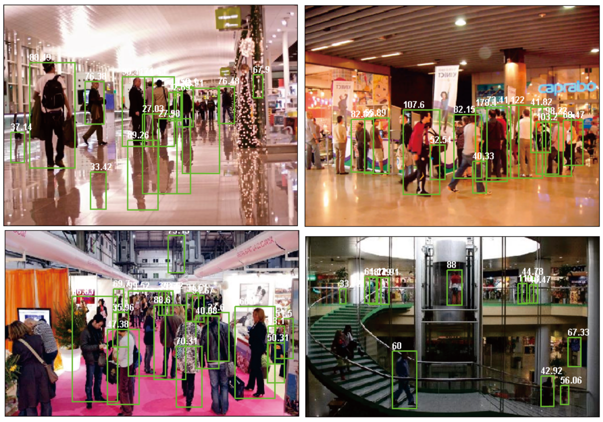

Figure 2. Examples of person detector [1]

2.2 HoG Cells

As mentioned before, to model the background, researchers commonly use per-pixel frameworks and perform different types of parametric or non-parametric algorithms [12, 13, 14, 15, 16]. These techniques have trouble dealing with high resolution images since the amount of computational work increases considerably as image resolution raises. Together with that, per-pixel modeling is affected by changes in illumination in the scene. Therefore, we propose to work on a higher level than pixels. Each frame is divided into squared cells, from which the gradient histogram is extracted. For each cell, we compute the gradient histogram as in [2], with 9 bins for angle values. This idea has the advantage of reducing the amount of computation, and also inherits the property of gradient histograms that are illumination invariant.

The size of the cells can be defined depending on the specific application. This parameter has to be coherent with the size of the detections the application is built to find.

2.3 HoG Cell Mean-Shift Clustering for Background Modeling

Similarly to [11], where the author uses a Mean-Shift clustering to model the background, we build the model from the output of a per-cell Mean-Shift clustering over a number N of recent or previous frames. The hypothesis is that the background cells will always correspond to the biggest nc clusters. Figure 3 shows an example of a clustering result of a cell in a video sequence. The nc largest clusters represent the structures that are more constant along time, which means they belong to background. In order to define this, we set a time-window represented by a certain number of frames.

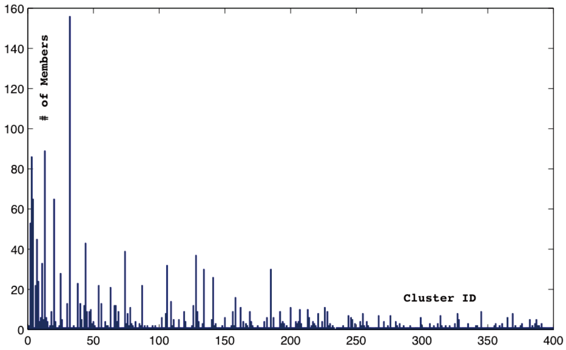

Figure 3. Example of cluster frequency. Larger clusters correspond to more constant structures over time, therefore classified as background.

This scheme has the advantage of working with a very simple clustering algorithm and only consists of two parameters: the mean-shift bandwidth and the time-window (number of frames considered to model the background).

2.4 Background Model Update

Updating the model is critical in this application since the background may have slow and almost unnoticeable changes along time. These variations can come in the way of day-night light changes, indoor lighting, moving lifts or doors, furniture changes, etc.

To cope with this, every n number of frames (or time), parallel to the main detection framework, the algorithm will recompute the clustering of every cell and define the clusters which represent background.

2.5 Background and Detection Intersections

In every new frame from the sequence, we select the person detection (from FPDW [1]) that corresponds to the target Recall of the application, and in which the goal is to improve the Precision. Secondarily, for each cell in the frame, we compute their gradient histogram and look up in the model if this cell belongs to any of the previously learned background gradient histogram clusters. Using this last output, we build a foreground mask. Algorithm 1 shows this procedure, and Algorithm 2 shows the procedure of updating the model.

Finally, to evaluate the person bounding boxes, we searched for the intersection of both outputs and threshold the detections that are lower than certain intersection percentage. This operation will reject detections activated within a background zone on the image, reducing significantly the amount of false positives from the detector.

3 Experiments

Our database consists on five videos, each containing 2700 frames, of different stationary indoor scenes. We ran several experiments changing the parameters *bandwidth* for Mean-Shift clustering and *timeThr*

Algorithm 1 New Frame Computation

```
for Every new frame do
    —Run FPDW to get detections
    —Reject detections which have a bounding box height smaller than $h$ specified minimum
    —Compute HoG in Cells
    —Add frame to history (set of frames determined by the time-window)
    —Compute distance of current frame cells from BG clusters
    —Threshold distance greater than a threshold from any BG cluster and define current frame background $cfBG$
    —Compute intersection of $cfBG$ and detections
    —Reject detections with interesection less than $minInt$ percentage
end for
```

Algorithm 2 Background Update

```
for all cells in history do
    —Compute Mean-Shift clusters with bandwidth parameter
    —Threshold clusters with number of members less than $timeThr$
end for
Output BG model from clusters
```

which corresponds to the amount of members a cluster should have during the time-window in order to be labeled as background. To generate the background model, we used the first 1000 frames from each sequence (update parameter $x = 1000$) and evaluate the performance on the rest 1700 frames. The minimum detection height was set to $h = 50$.

As evaluation technique, we obtained the *Precision*, *Recall* and $F\!-\!score$ from the entire set of frames in the videos. For this evaluation, the parameters BW and $timeThre$ used were varied along from $BW = 0.05$ to $BW = 0.35$, and from $timeThr = 50$ to $timeThr = 1000$. Figures 4, 5, 6, 7 and 8 show the results of the top 7 configurations, over the minimum percentage of intersection $minInt$ used to reject detections.

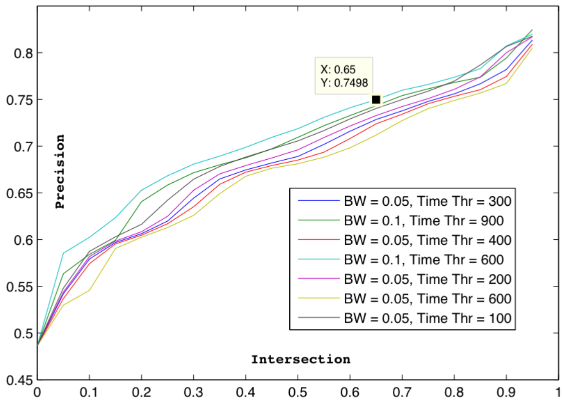

Figure 4. Precision vs Intersection. Tip at configuration with maximum F-score. See Fig. 6

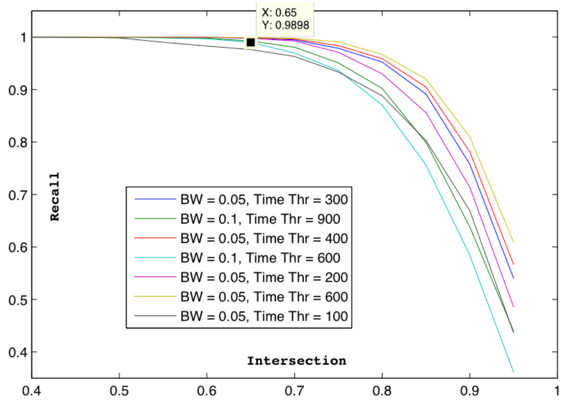

Figure 5. Recall vs Intersection. Tip at configuration with maximum F-score. (see Fig. 6)

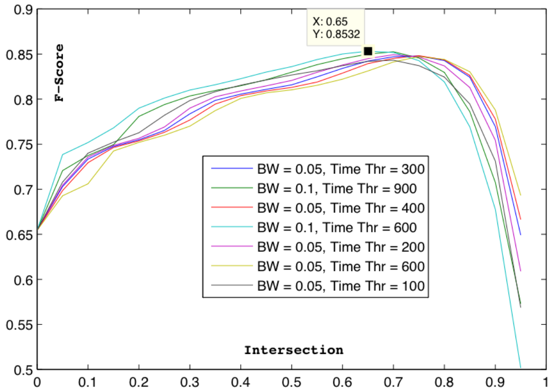

Figure 6. F-Score vs Intersection. Tip at configuration with maximum value.

Results show that the best configuration of BW and $timeThr$ leads to a precision of 0.75, a recall of 0.99 and therefore a F-Score of 0.85.

4 Discussion and Conclusions

We presented a post-processing framework for people detection within stationary camera applications. Since this algorithm works as a post-processing stage from the people detector, the final miss rate is the union of both performances. This makes applicable the proposed framework on person detection where we can choose the necessary recall since, applying it, we improve only the precision of the method. Still, the results show that the detection improved compared to using only the detector since it reduces the amount of false positives (from 16000 to 2000 in out Database) and therefore we obtained the optimal precision and recall of 0.75 and 0.99, respectively.

Using gradient histograms for background modeling instead of building it pixel-wise, showed more robustness even when dealing with short amount of information. Capturing the structure of the background instead of the intensity, gives better results for foreground-background classification.

However, this algorithm is not capable of reducing the amount of false positives that are fired inside the foreground cells. This situation can be attacked by additionally performing an occlusion reasoning which will throw better results.

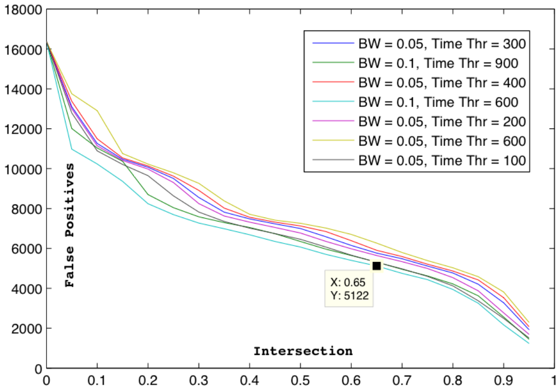

Figure 7. False Positives vs Intersection. Tip at configuration with maximum F-score (see Fig. 6)

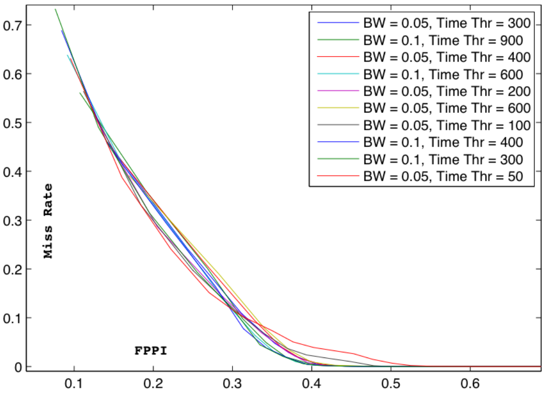

Figure 8. False Positives Per Image vs Miss Rate

There are several conditions that are dependent on the real application addressed. First of all, the size of the expected detections or in other words, the distance from which we want to detect persons. This condition is very important to define the correct detector or the right parameters of it, as well as the minimum height we want to threshold. Another important characteristic to have in mind regards the background. The variability of it has to be taken into account when defining the time window in which we want to work. Parameter *timeThr* addresses this. There are two types of variability. First, the short term which is for example trees or lifts and other stuff that moves continuously along the sequence. As for a long term variability an example is day light. Varying both *timeThr* and x (frames used for updating), the algorithm will attack both long and short term changes.

As future work, we contemplate the possibility of building a complete HoG model of background and including it on the detection itself. In addition, as mentioned before, an occlusion detection to reduce false positives in foreground areas will also improve the performance. Another way of improving the performance can be to use different updating scheme. Different clustering algorithms may be also evaluated, for better BG modeling. Finally, another idea will be to use Conditional Random Fields for foreground false positive elimination.

References

[1] P. Dollár and S. Belongie and P. Perona. "The Fastest Pedestrian Detector in the West," *In Proc. BMVC, 2010.*

[2] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection," *In Proc. CVPR, 2005.*

[3] Fukunaga, K. and Hostetler, L. "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory, 1975.*

[4] P. Dollár and Z. Tu and P. Perona and S. Belongie. "Integral Channel Features" *In Proc. BMVC, 2009.*

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. "Object Detection with Discriminatively Trained Part-Based Models," *In PAMI, 2010*, vol 32, num 9, pp. 1627-1645.

[6] L. Bourdev and J. Malik. "Poselets: Body part detectors trained using 3d human pose annotations," *In Proc. ICCV, 2009.*

[7] Levin, Viola, and Freund. "Unsupervised Improvement of Visual Detectors using Co-Training," *In Proc. ICCV, 2003.*

[8] C. Rosenberg, M. Hebert, and H. Schneiderman. "Semi-supervised self-training of object detection models," *In Proc. of IEEE Workshop on Application of Computer Vision, 2005.*

[9] P. Roth, S. Sternig, H. Grabner, and H. Bischof. "Classifier grids for robust adaptive object detection," *In Proc. CVPR, 2009.*

[10] Meng Wang, Wei Li and Xiaogang Wang. "Transferring a Generic Pedestrian Detector Towards Specific Scenes," *In Proc. CVPR, 2012.*

[11] M. Piccurdi and Z. Jan. "Mean-Shift Background Image Modelling," *In Proc. ICIP, 2004*

[12] C. Wren, A. Azarbayejani, T. Darrell and A.P. Pentland. "Pfinder: real-time tracking of the human body," *In PAMI, 1997*, vol 19, no 7, pp. 780-785.

[13] R. Cucchiara, C. Grana, M. Piccardi and A. Prati. "Detecting moving objects, ghosts, and shadows in video streams," *In PAMI, 2003*, vol. 25, n. 10, pp. 1337-1342.

[14] C. Stauffer and W.E.L. Grimson. "Adaptive background mixture models for real-time tracking," *In Proc. CVPR, 1999*

[15] P. Wayne Power and J.A. Schoonees. "Understanding background mixtyre models for foreground segmentation," *In Proc. IVCNZ, 1999*

[16] A. Elgammal, D. Harwood and L.S. Davis. "Non-parametric model for background subtraction," *In Proc. ECCV, 2000*