A Region Driven and Contextualized Pedestrian Detector

Thierry Chesnais[1], Thierry Chateau[2], Nicolas Allezard[1], Yoann Dhome[1], Boris Meden[1], Mohamed Tamaazousti[1] and Adrien Chan-Hon-Tong[1]

[1]*CEA, LIST, Vision and Content Engineering Laboratory, Point Courrier 94, F-91191 Gif-sur-Yvette, France*

[2]*Institut Pascal, UMR 6602 CNRS, Blaise Pascal University, Campus des Cézeaux, 63170 Aubiere, France*

Keywords: Videosurveillance, Object Detection, Pedestrian Detection, Semi-supervised Learning, Oracle.

Abstract: This paper tackles the real-time pedestrian detection problem using a stationary calibrated camera. Problems frequently encountered are: a generic classifier can not be adjusted to each situation and the perspective deformations of the camera can profoundly change the appearance of a person. To avoid these drawbacks we contextualized a detector with information coming directly from the scene. Our method comprises three distinct parts. First an oracle gathers examples from the scene. Then, the scene is split in different regions and one classifier is trained for each one. Finally each detector are automatically tuned to achieve the best performances. Designed for making camera network installation procedure easier, our method is completely automatic and does not need any knowledge about the scene.

1 INTRODUCTION

Recently several applications, like videosurveillance, have promoted the development of the pedestrian detection algorithms. Classical approaches to detect objects are based on machine learning. Training a detector consists in extracting the best discriminative features between pedestrian and background from a labeled training dataset. Then the detector compares the selected features of a new image with these of the database to predict the presence of a pedestrian. But the appearance of pedestrians varies a lot in terms of size, angle and posture, depending on the viewpoint of the camera. These large variations disrupt the detector.

In the case of a videosurveillance system, most of the characteristics of the scene are known and are stable for a long time. Taken into account information coming from the scene to contextualize a detector could simplify the pedestrian detection problem.

In this article we focus on a videosurveillance problem using a calibrated static camera. We demonstrate that contextualizing and restraining a detector inside some predefined regions improves local and global performances of the system. Our automatic method takes into account the perspective and the pedestrian density of the scene to build these regions.

This paper is organized as follows. The section 2 introduces some approaches recently proposed to mitigate the problems mentioned above: generic classifier and perspective deformations. The sections 3 and 4 present our method to contextualize a detector using the geometric information of the scene. Finally evaluations of our approach are given in section 5.

2 RELATED WORK

Different strategies exist to build a detector. Classical methods consist in computing a global classifier, that is used to entirely scan an image. This approach is commonly used in the case of a generic learning algorithm when the context of the classifier is unknown. A representative dataset is difficult to gather, especially if the classifier had to work in a broad range of applications. This kind of classifiers is not very specific and can fail if the scene characteristics are different from these encountered during the learning.

Secondly, it is difficult to design a pedestrian detector, because of i) the perspective issues and ii) as the human body is articulated, it yields a lot of different configurations. Since (Dalal and Triggs, 2005), there has been a lot of work trying to make the model learned even more complex to take these deformations into account. The most successful approach is probably the Deformable Part Model from (Felzenszwalb et al., 2008). Instead of considering a global and complex classifier to recognize a pedestrian in ev-

Chesnais T., Chateau T., Allezard N., Dhome Y., Meden B., Tamaazousti M. and Chan-Hon-Tong A..
A Region Driven and Contextualized Pedestrian Detector.
DOI: 10.5220/0004292607960799
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 796-799
ISBN: 978-989-8565-47-1

ery posture seen from any angle, the complexity is externalized and several specific classifiers are used.

Local methods, like (Grabner et al., 2007), exist to reduce the complexity of the learning problem in the case of a static camera. Instead of computing a single complex detector, a lot of simpler classifiers are learned. As these methods are extremely local they could be sensitive to noise in the video or global illumination changes. Moreover some areas are almost empty, leading to some poorly trained detectors.

Finally some intermediary methods have been published. (Park et al., 2010) suggest to build two different classifiers. One for small pedestrians with a few details and one more complex for these with a sufficient definition. Mixing both classifiers allows to improve the global detector performances in case of scale variations due to the perspective.

In videosurveillance, the camera is static and continually observes the same place. So, after studying the scene during a while, it is possible to predict the majority of the events encountered by the system and their positions in the image. Taking into account this information could increase performances.

3 ORACLE

Our goal is to contextualize a pedestrian detector. Two steps are necessary to achieve it. First we need to collect some pedestrian examples coming from the scene in order to train a classifier. This function is provided by an oracle which automatically annotates the video and finds useful pedestrian and background examples. An oracle has a low recall and a high precision. The oracle uses a combination of elementary algorithms: a generic pedestrian detector trained on an independent database and a background subtraction algorithm.

3.1 Positive Examples

To extract some pedestrian examples, we need to merge signals provide by these two algorithms. Our approach is inspired by a previous work of (Rodriguez et al., 2011). The generic classifier provides a vector $s_c = (s_i)_{1 \leq i \leq n}$ containing the classification score of the n boxes in the image. Each score s_i is normalized between 0 and 1. The background subtraction also provides a binary segmentation, I_{Bkg}, of the image. The oracle output is a Boolean vector $x \in \{0,1\}^n$. The i^{th} component of x indicates if the i^{th} box is a positive detection.

We are considering the merging step as a minimization of an energy E, formulated in the equation 1:

$$E = -s_c^T x + x^T W x + \alpha ||I_{Bkg} - I_{Model}(x)||^2 \quad (1)$$

The matrix W avoids that two boxes could both be selected if they are too close. W is a symmetric matrix $n \times n$ with $W_{i,j} = \infty$ if the two boxes i and j are similar and 0 else. Notice that if we omit the last term in the equation we could recognize a classical non maximum suppression method. This last term controls the consistency of a detection with the background subtraction. To obtain the binary model $I_{Model}(x)$ we model detections in x by a full box. The parameter α adjusts the importance between the classification and the background subtraction segmentation. In our experiment, α is fixed to 1.

3.2 Negative Examples

The oracle gives a set of pedestrian positions. In order to train a classifier we still need a set of background examples. Our strategy consists in choosing random boxes, non overlapping with positive detections. As the oracle has a low recall, it does not detect all pedestrians. Then some false negative detections will be incorporated in the negative base. However this is rather unlikely because statistically there are more negative examples in an image than positive ones.

4 REGION DRIVEN PEDESTRIAN DETECTION

The second step of the contextualization consists in incorporating some information on the geometry of the scene. Spatially contextualizing a detector is a way to limit the influence of the perspective in the image. The contextualization can be useful during the detection phase because it can avoid scanning empty areas (section 4.1). Moreover the contextualization improves the training step (see 4.2). During this phase, several regions are defined in the scene and a classifier is trained for each region. Thus a spatially driven classifier is not forced to recognize all the possible poses of a pedestrian but only a coherent subset. Creating groups with pedestrians sharing common properties simplifies the training problem.

4.1 Empty Areas

In a scene, we frequently observe some inaccessible areas like walls. Scanning these regions can only increase the false positive rate and slow down the scan.

To find these empty areas we first segment the image in order to reveal its structure. In this optic, we

use a superpixel algorithm (Achanta et al., 2012), that spatially splits, in the color domain, an image into areas. We only retain areas containing at least one detection (defined by its 3D feet position) provided by the oracle. The other ones are not scanned during the detection step.

4.2 Building Regions to Drive a Detector

Regions ensure a good locality of the examples for one classifier. This locality allows to learn pedestrians with similar scale and lean characteristics.

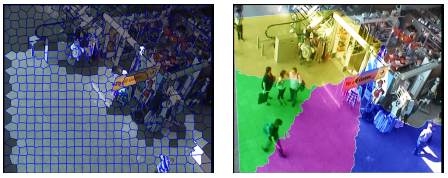
To build the different regions, we still exploit a superpixels algorithm. We use a k-means algorithm working in a four parameters space: (x, y, s_p, s_n), with x and y corresponding to a 3D position of a pedestrian previously detected by the oracle in the ground plane. $s_p(x,y)$ is the score given by the generic classifier for it and $s_n(x,y)$ is the score of the background element at the position. We obtain $s_n(x,y)$ by averaging the classification scores at this position on the complete sequence. We use the same assumption as for a background subtraction algorithm: a pedestrian is seldom immobile and its influence on the average is low.

5 EVALUATIONS AND DISCUSSIONS

In this section we compare performances of a generic detectors with two contextualized detectors: the first one is trained on data coming from the whole image where as the second one is also spatially driven. This comparison comprises two parts: the first one presents the performances by region whereas the second one highlights the global performances. All these comparisons show the competitiveness of our system. We test our algorithm on a sequence coming from the European project ITEA2 ViCoMo.

5.1 Results by Region

In this part, we demonstrate the interest of the spatial contextualization. First we use our oracle to detect some pedestrian examples. The figure 2 shows some examples coming from the contextualized training dataset. For a given region, pedestrians share similar appearance characteristics like scale. Errors are often recurrent. Concerning the negative examples some stationary pedestrians are not detected. For the positive examples the oracle is not always able to correctly center a person, especially when two people are closed. On the groundtruth sequence, our oracle reaches a recall of 0.66 and a precision of 0.97.



(a) Empty areas (dark) (b) Regions

Figure 1: Empty areas and the four regions built by our method.

Figure 2: Positive (continuous line) and negative (dotted line) examples gathered from each region. Examples with label errors are surrounded by a red box.

Our algorithm defines 4 regions, shown on the figure 1.b. For each one, a classifier is trained with some examples coming from the specific region. We compare their performances with a global classifier, trained with examples coming from the entire image. Results are shown on the figure 3.

The region with the higher deformation due to the perspective has the id 0 and is located at the right bottom corner. Our contextualized and spatially constrained detector (blue curves) performs better than the global contextualized detector (green curves). In the other parts of the image, both classifiers achieve similar performances. This can be explained by the fact that the perspective has less influence in these regions. The shape of the curves for the last region (id 3) is due to the fact that some false positive detections have a higher score than true positive ones.

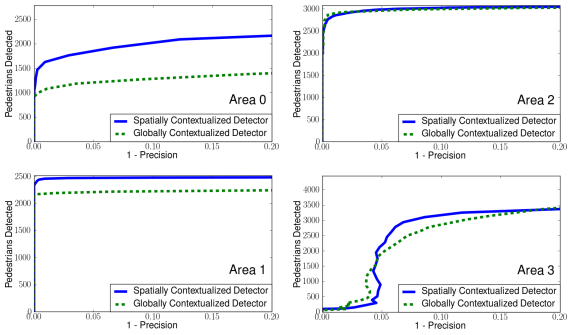


Figure 3: Precision-recall curves for each region.

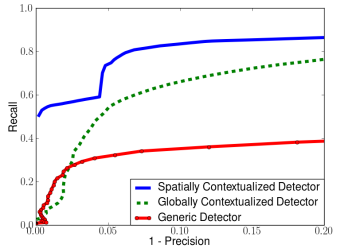


Figure 4: Global precision-recall curves.

5.2 Global Results

In the last paragraph we present the results of our system for each region separately. In this section we are interested in its global performance. We compute precision-recall curves on the whole image.

Our results are shown on the figure 4. We can see that both contextualized detectors (blue and green curves) are better than the generic one (red curve). As expected, the contextualized and region driven detector performs better than the solely contextualized detector. On the blue curve, the distortion is easily explained by the performances of the classifier in the fourth region (id 3).

All these experiments tend to prove that a global classifier, even if it is contextualized, is not optimal in our application. A region driven classifier can achieve better performances.

5.3 Automatic Detection Threshold Estimation

Once all classifiers have been trained, we need to tune them to achieve the best performances. In a video-surveillance system, as the scene is stable, one false positive detection is fairly sure to come back in the following frames. So it is essential to filter out the false positive detections. This mostly consists in adjusting each detection threshold independently.

The optimal threshold is the one maximizing the best F-measure, which is a trade-off between the precision and the recall. In the case of a generic learning algorithm, we simply use an annotated testing dataset to estimate this threshold. In a contextualized approach, we do not have such a dataset like a groundtruth. So we reuse the oracle to collect new examples to create an estimated groundtruth. This time we replace the generic classifier of the oracle by a contextualized one. With this estimated groundtruth, it is possible to compute a recall and a precision for a given threshold and to deduce the F-measure. A 1-D maximization is done to find the best threshold.

Table 1: Comparison between the detectors (global and spatial) performances at their optimal thresholds θ_{opt}^{pr} and at their estimated ones θ_{opt}^{auto} (T: threshold, P: precision, R: recall, F: F-measure).

	Global		Region 0		Region 1		Region 2		Region 3	
	θ_{opt}^{pr}	θ_{opt}^{auto}	θ_{opt}^{pr}	θ_{opt}^{auto}	θ_{opt}^{pr}	θ_{opt}^{auto}	θ_{opt}^{pr}	θ_{opt}^{auto}	θ_{opt}^{pr}	θ_{opt}^{auto}
T	24.6	21.9	11.5	16.2	1.4	17.9	3.1	32.5	28.3	27.6
P	0.87	0.73	0.88	0.99	0.99	1.0	0.97	1.0	0.88	0.86
R	0.69	0.72	0.75	0.60	0.98	0.89	0.96	0.88	0.73	0.74
F	0.77	0.72	0.81	0.74	0.98	0.94	0.96	0.94	0.80	0.80

To evaluate the accuracy of the estimated thresholds, θ_{opt}^{auto}, we compare them to the optimal thresholds θ_{opt}^{pr}. Results are shown on the table 1. As the contextualized oracle is not perfect, there are some mislabeled examples in the estimated groundtruth. So the estimated thresholds can be slightly different from the optimal ones. But usually they achieve similar F-measures.

6 CONCLUSIONS

In this article, we propose a system to automatically build a contextualized pedestrian detector for video-surveillance applications. First, an oracle with a high precision gathers scene specific pedestrian examples. This dataset and the geometry of the scene are then used to design 3D regions where pedestrians share similar appearance characteristics. The idea is to externalize the classifier complexity. Finally one detector, composed by the classifiers trained for each region and set to their optimal working points, is run.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). SLIC Superpixels Compared to State-of-the-art Superpixel Methods. *PAMI*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*.

Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *CVPR*.

Grabner, H., Roth, P. M., and Bischof, H. (2007). Is pedestrian detection really a hard task? In *PETS*.

Park, D., Ramanan, D., and Fowlkes, C. (2010). Multiresolution models for object detection. In *ECCV*.

Rodriguez, M., Sivic, J., Laptev, I., and Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. In *ICCV*.