Local Response Context Applied to Pedestrian Detection

William Robson Schwartz[1], Larry S. Davis[2], and Helio Pedrini[1]

[1] University of Campinas, Institute of Computing, Campinas, SP, 13084-971, Brazil
[2] University of Maryland, Dept. of Computer Science, College Park, MD, 20742, USA
{schwartz,helio}@ic.unicamp.br, lsd@umiacs.umd.edu

Abstract. Appearing as an important task in computer vision, pedestrian detection has been widely investigated in the recent years. To design a robust detector, we propose a feature descriptor called Local Response Context (LRC). This descriptor captures discriminative information regarding the surrounding of the person's location by sampling the response map obtained by a generic sliding window detector. A partial least squares regression model using LRC descriptors is learned and employed as a second classification stage (after the execution of the generic detector to obtain the response map). Experiments based on the ETHZ pedestrian dataset show that the proposed approach improves significantly the results achieved by the generic detector alone and is comparable to the state-of-the-art methods.

Keywords: pedestrian detection, local response context, partial least squares regression.

1 Introduction

Pedestrian detection is of fundamental importance in computer vision due to the use of people's location for tasks such as person recognition, tracking, pose estimation, and action recognition. To reduce the amount of noise (false detections) input to these tasks, it is important to maintain a low miss-detection rate while reducing as much as possible the number of false alarms, which can only be achieved with the use of robust detection algorithms.

The main challenges faced to locate people in images are related to pose variation, illumination changes, blur, and partial occlusions. To deal with such conditions, most pedestrian detectors are either holistic or part-based [14]. While the latter, which employs a generative process to combine detected parts to a prior human model, is more suitable to handle conditions such as pose variation and partial occlusions, the former is able to collect more discriminative information by performing a statistical analysis to combine a set of low-level features within a detection window due to the larger size of the whole body, compared to the size of the parts.

To be able to locate all humans in an image, a holistic detector employs an image sweeping based on a sliding window which considers multiple scales and small strides. A consequence resulting of this approach is the existence of multiple

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 181–188, 2011.

decreasing responses around the person's location. These multiple responses are normally removed by non-maximum suppression.

The evaluation of the response provided by a single detection window (as done when non-maximum suppression is applied) may generate ambiguities since that peak response can be caused by false alarms such as trees or poles, which present shapes similar to pedestrians. Nevertheless, the analysis of the spatial distribution of responses around detection windows (context) might reduce or even remove such ambiguities because the behavior of detector responses may vary according to the type of object.

The contribution of the context around the person's location for detection has been observed in the work of Dalal and Triggs [2] with the addition of a number of background pixels on the four sides of the detection window. Therefore, one way of incorporating more context is to increase even more the detection window size to add more background information. However, there is the consequence that the feature space becomes extremely high dimensional since robust detection is better achieved through feature combination [11].

The addition of more descriptors to capture extra background information, besides increasing the feature space dimensionality, does not incorporate information regarding the object being detected (pedestrians) because descriptors are general and only add such information after a learning process. On the other hand, if detection responses are considered, some information regarding the problems is already incorporated since the responses depend on the object class, and the dimensionality of the feature space for the detector is not changed.

This work proposes the use of local response context to improve pedestrian detection. The process works as follows. After the execution of a holistic detector and the composition of the response map for an image, responses around each detection window are sampled to compose a feature vector, referred as to *Local Response Context* (LRC). In the training phase, feature vectors located around detection windows containing pedestrian and detection windows with background are used to learn a regression model. Therefore, the responses are used as a new set of descriptors. Finally, during the classification, LRC feature vectors are projected onto the model and classified as pedestrian or background.

2 Related Work

Dalal and Triggs [2] proposed the use of histogram of oriented gradient (HOG) as feature descriptor for human detection, whose results outperformed other features. Zhu et al. [20] presented a method that significantly speeds up human detection by combining HOG descriptors with a cascade of rejectors. Variable size blocks are used in their method, such that larger blocks allow rejection of the majority of detection windows in the early few stages of the cascade. Zhang et al. [19] described a multiple resolution framework for object detection based on HOG to reduce computational cost, where lower resolution features are firstly used to reject most of the negative windows, then expensive higher resolution features are used to obtain more precise detection. Begard et al. [1] developed two learning algorithms for real-time pedestrian detection using different implementations of AdaBoost to optimize the use of the local descriptors.

A human detection method using covariance matrices as feature descriptors and a learning algorithm based on a Riemannian manifold was presented by Tuzel et al. [15]. Their method produced superior results when compared with the methods proposed by Dalal and Triggs [2] and Zhu et al. [20]. Mu et al. [10] developed a human detection method based on two variants of local binary patterns (LBP), comparing the results against the use of covariance matrix and HOG descriptors. Wu and Nevatia [17] proposed a cascade-based framework to integrate heterogeneous features for object detection, such as edgelet, HOG and covariance descriptors. Maji et al. [8] proposed features based on a multi-level version of HOG and histogram intersection kernel support vector machines (IKSVM) to obtain a good balance between performance and accuracy in pedestrian detection.

Part-based representations have also been used for human detection. Shet and Davis [13] employed a logical reasoning approach to utilizing contextual information and knowledge about human interactions, extending the output of different low-level detectors for human detection. Lin and Davis [7] proposed a pose-invariant feature descriptor for human detection and pose segmentation. Tran and Forsyth [14] developed a two-step pedestrian detection strategy, where the configuration of the best person within each detection window is firstly estimated, then features are extracted for each part resulting from this estimation and passed to a support vector machine classifier to make the final decision.

Context information has been used to increase accuracy of the human detection process. Gualdi et al. [6] exploited context information through a relevance feedback strategy, which enhances the pedestrian detection step by using training on positive and negative samples, and a weak scene calibration, which estimates the scene perspective to discard outliers. Statistical relationship between objects and people, modeled by Markov logic networks, was used by Wu et al. [18] to incorporate user activities as context information for object recognition. Morency [9] described co-occurrence graphs for modeling relations between visual head gestures and contextual cues, such as spoken words or pauses, to select relevant contextual features in multiparty interactions.

3 Proposed Method

In this section, we describe the proposed method for incorporating local response context for pedestrian detection. Since the addition of feature descriptors extracted from surrounding regions of a detection window would result in an extremely high dimensional space (millions of descriptors to describe a single detection window), we use detection responses of a local neighborhood to build a new classifier to improve the discrimination between human and non-human samples.

A holistic sliding-window detector (referred as to *generic detector*) extracts feature descriptors for each detection window, then presents the resulting feature vector to a classification method, which results in a response value used as

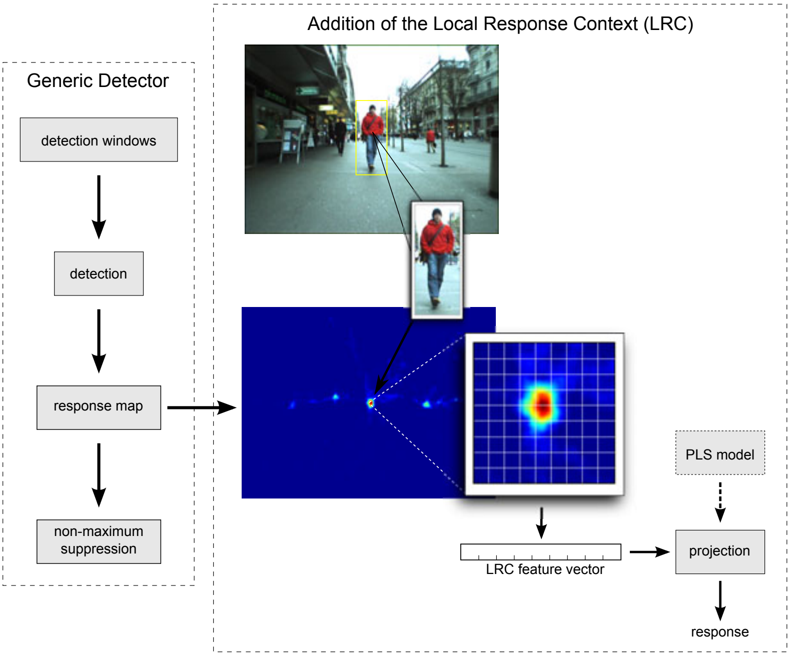

Fig. 1. Proposed method. The left-hand side shows the detection process performed by a generic pedestrian detector. The right-hand side shows the incremented detection process with the addition of the local response context descriptor. Using the resulting response map, LRC descriptors are extracted for each detection window and projected onto a PLS model, resulting in a more accurate classification between humans and non-humans.

confidence to separate humans from background. A response map, $R(x, y)$ with the image size, is built using the resulting set of responses (a response is placed at the centroid of the location of its corresponding detection window). Finally, a non-maximum suppression procedure is executed to maintain only detection windows with the highest responses. This process is illustrated on the left-hand side of Figure 1.

In contrast to the generic detector, which applies the non-maximum suppression after building the response map, we use this map to sample responses in the neighborhood of each detection window to extract the *local response context* descriptor, which will be used for a second and more accurate detection process, as illustrated on the right-hand side of Figure 1. The complete procedure is composed of feature extraction, training and classification.

The feature extraction works as follows. Let d_i be a detection window with centroid located at (x, y) and its local neighborhood defined by a square with left-most corner at $(x - \Delta, y - \Delta)$ and right-most corner $(x + \Delta, y + \Delta)$, where Δ is a value defined experimentally. The responses $R(x', y')$ inside this region are sampled and linearized to compose the feature vector $\boldsymbol{v}_{\mathrm{LRC}}$, used to describe d_i during its classification.

For the training, a generic pedestrian detector is executed for a previously labeled image sequence, resulting in a set of response maps. For each detection window, a feature vector based on LRC is extracted. Then, the detection windows are separated according to their classes (humans or non-humans) given by the ground-truth locations. Finally, a Partial Least Squares (PLS) model [12] is built to classify samples using labels +1 for human and −1 for non-human.

Partial least squares is a method for modeling relations between sets of observed variables in a latent space. It constructs new predictors as linear combinations of the original variables summarized in a matrix $\boldsymbol{X}$ of descriptor variables (matrix with feature vectors) and a vector $\boldsymbol{y}$ of responses (class labels). PLS decomposes the input variables as

$$\begin{aligned}\boldsymbol{X} &= \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}\\ \boldsymbol{y} &= \boldsymbol{U}\boldsymbol{q}^T + \boldsymbol{f}\end{aligned}$$

where $\boldsymbol{T}$ and $\boldsymbol{U}$ are $n \times p$ matrices containing p extracted latent vectors, the $(m \times p)$ matrix $\boldsymbol{P}$ and the $(1 \times p)$ vector $\boldsymbol{q}$ represent the loadings and the $n \times m$ matrix $\boldsymbol{E}$ and the $n \times 1$ vector $\boldsymbol{f}$ are the residuals. The PLS method, using the nonlinear iterative partial least squares (NIPALS) algorithm [16], constructs a matrix of weights $\boldsymbol{W}$ indicating the importance of each descriptor. Using these weights, the regression coefficients $\boldsymbol{\beta}_{m\times 1}$ can be estimated by

$$\boldsymbol{\beta} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\boldsymbol{T}^T\boldsymbol{y}. \tag{1}$$

Then, the regression response, y_v, for a feature vector $\boldsymbol{v}_{\mathrm{LRC}}$ is obtained by

$$y_v = \overline{y} + \boldsymbol{\beta}^T\boldsymbol{v}_{\mathrm{LRC}} \tag{2}$$

where $\overline{y}$ is the sample mean of $\boldsymbol{y}$.

Once the PLS model has been estimated in the training process, it is stored to be used during the classification, when test sequences are presented. The classification is illustrated on the right-hand side of Figure 1 and works as follows. First, for each image, a generic pedestrian detector is executed to obtain the response map. Then, the feature vector $\boldsymbol{v}_{\mathrm{LRC}}$ is extracted for a detection window d_i and projected onto the PLS model using Equation 2. The higher the response, the more likely is that d_i contains a human (due to the class labeling used). Finally, using the response map generated by this classification process, the non-maximum suppression can be performed to locate pedestrians individually.

4 Experimental Results

This section presents and compares results obtained with local response context. First, we present a brief summary of the human detector used to obtain the response maps. Then, we describe the parameter choice considered and, finally, we compare detection results achieved when LRC is incorporated to results obtained by state-of-the-art approaches.

The PLS Human Detector. As generic detector, used to obtain responses for each detection window to build the response map, we employed the human detection method proposed by Schwartz et al. [11], which is available for download. This is a holistic detector based on a combination of descriptors focusing on shape (histograms of oriented gradients), texture (co-occurrence matrices), and color (color frequency) that uses PLS to reduce the dimensionality of the feature space and provide discriminability between the two classes.

The detection is performed as follows. Each detection window is decomposed into overlapping blocks and feature descriptors are extracted and concatenated in a feature vector. This feature vector is then projected onto a PLS model and the resulting latent variables are classified as either a human or non-human by a quadratic classifier. Finally, a response map is output for each image.

Experimental Setup. To evaluate our method, we use the ETHZ pedestrian dataset [3], which is composed of three video sequences collected from a moving platform. These sequences contain frames of size 640 × 480 pixels. For all the experiments, the detection is performed over 16 scales to consider humans with heights between 60 and 500 pixels, with strides of 4 pixels in the x-axis and 8 pixels in the y-axis. This setup results in $64,292$ detection windows per frame.

To learn the PLS model with the local response context, the initial 50 frames of the training sequence #0 from the ETHZ pedestrian dataset is used. From these frames, 163 human exemplars and 750 counter-examples (15 per frame chosen in decreasing order according to the response of the corresponding detection windows) are sampled. The PLS model is built considering a latent space of 10 dimensions. In addition, $\varDelta = 19$ is used for the neighborhood, resulting in a feature vector with 1521 descriptors. These parameters were chosen empirically during the training and kept fixed during the classification stage.

Comparisons. Figure 2 shows curves comparing the proposed method (referred as to *local response context*) in the three ETHZ sequences to other methods of the literature. In these plots, the x-axis shows the number of false positive per image (FPPI) and the y-axis shows the recall, which is the fraction of detected pedestrian samples among all pedestrians in the video sequence.

Curves shown in Figure 2 compare our method to four state-of-the-art approaches in the literature. The most important comparison is to the PLS human detector [11] since the proposed method uses the response map generated by it. Therefore, any gain in performance compared to that method is due to the addition of the LRC. The other approaches in the comparison were proposed by Ess et al. [3,4,5]. These methods employ not only low-level descriptors, as in the proposed method, but also scene information such as depth maps, ground-plane estimation, occlusion reasoning, and tracking to detect pedestrians.

Discussion. According to the results displayed in Figure 2, there are significant improvements on the recall when compared to the use of the the PLS human detector [11], which shows a clear contribution of the LRC. In addition, even though the other methods in the comparison use extra information, the proposed method presents very similar or better results in all video sequences. Therefore,

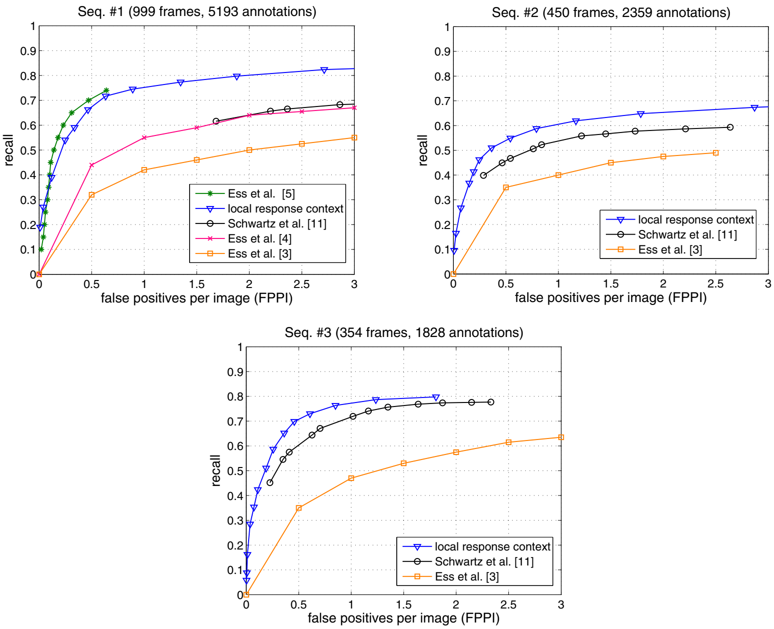

Fig. 2. Comparisons using three video sequences of the ETHZ pedestrian dataset. The proposed method is referred as to local response context.

the use of extra information such as ground-plane estimation and tracking might be exploited to achieve further improvements in the future.

One of the advantages of using LRC descriptors instead of considering low-level descriptors extracted from the neighborhood is the fairly low dimensionality of the feature vectors (1521 descriptors for a neighborhood with 19 pixels). If the detection window of the PLS human detector [11] were increased to consider a local neighborhood, the number of descriptors in the feature vector would be easily higher than one million, which might prevent the method from running due to the extremely high memory consumption and computation.

5 Conclusions

This work presented a pedestrian detection approach based on the local response context. This method uses response maps computed by a generic pedestrian detector (PLS Human Detector for this work) to extract feature descriptors that are used to build a PLS model employed to classify detection windows as humans or non-humans. Experimental results presented improvements on detection rates on the ETHZ dataset when compared to the PLS Human Detector. In addition, the proposed detector achieved results comparable to state-of-the-art methods.

Acknowledgements. The authors are thankful to FAPESP, CNPq and CAPES for the financial support. Larry Davis acknowledges the support of the Office of

Naval Research under a subcontract from Carnegie Mellon University on the MURI grant "Rich Representations With Exposed Semantics for Deep Visual Analysis."

References

1. Begard, J., Allezard, N., Sayd, P.: Real-time human detection in urban scenes: Local descriptors and classifiers selection with adaboost-like algorithms. In: CVPR Workshops (2008)
2. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR (2005)
3. Ess, A., Leibe, B., Gool, L.V.: Depth and Appearance for Mobile Scene Analysis. In: ICCV (2007)
4. Ess, A., Leibe, B., Schindler, K., van Gool, L.: A Mobile Vision System for Robust Multi-Person Tracking. In: CVPR (2008)
5. Ess, A., Leibe, B., Schindler, K., Gool, L.V.: Moving Obstacle Detection in Highly Dynamic Scenes. In: ICRA (2009)
6. Gualdi, G., Prati, A., Cucchiara, R.: Contextual Information and Covariance Descriptors for People Surveillance: An Application for Safety of Construction Workers. EURASIP Journal on Image and Video Processing (2011)
7. Lin, Z., Davis, L.S.: A pose-invariant descriptor for human detection and segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 423–436. Springer, Heidelberg (2008)
8. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR (2008)
9. Morency, L.P.: Co-occurrence graphs: contextual representation for head gesture recognition during multi-party interactions. In: WUCVP (2009)
10. Mu, Y., Yan, S., Liu, Y., Huang, T., Zhou, B.: Discriminative local binary patterns for human detection in personal album. In: CVPR (2008)
11. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human Detection Using Partial Least Squares Analysis. In: ICCV (2009)
12. Schwartz, W., Guo, H., Davis, L.: A Robust and Scalable Approach to Face Identification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 476–489. Springer, Heidelberg (2010)
13. Shet, V., Neuman, J., Ramesh, V., Davis, L.: Bilattice-based logical reasoning for human detection. In: CVPR (2007)
14. Tran, D., Forsyth, D.: Configuration estimates improve pedestrian finding. In: NIPS (2007)
15. Tuzel, O., Porikli, F., Meer, P.: Human Detection via Classification on Riemannian Manifolds. In: CVPR (2007)
16. Wold, H.: Partial least squares. In: Kotz, S., Johnson, N. (eds.) Encyclopedia of Statistical Sciences, vol. 6, pp. 581–591. Wiley, New York (1985)
17. Wu, B., Nevatia, R.: Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. In: CVPR (2008)
18. Wu, C., Aghajan, H.: Using context with statistical relational models: object recognition from observing user activity in home environment. In: WUCVP (2009)
19. Zhang, W., Zelinsky, G., Samaras, D.: Real-time accurate object detection using multiple resolutions. In: ICCV (2007)
20. Zhu, Q., Yeh, M.C., Cheng, K.T., Avidan, S.: Fast human detection using a cascade of histograms of oriented gradients. In: CVPR (2006)