Context Awareness for Object Detection[1)]

Roland Perko and Aleš Leonardis
University of Ljubljana, Slovenia
{roland.perko, ales.leonardis}@fri.uni-lj.si

Abstract:
A wide range of algorithms have been proposed to detect objects in still images. However, most of the current approaches are purely based on local appearance and ignore the context in which these objects are embedded. This paper proposes a general approach to extract, learn and use contextual information from images to increase the performance of classical object detection methods. The important properties of the proposed approach are that it can be combined with any existing object detection method and it provides a general framework not limited to one specific object category.

1 Introduction

Object detection is a classical discipline in computer vision and is used in a large field of applications. Many concepts have been developed (see e.g. [5]) where, independent of the particular representation model used (bag-of-words model [6], part-based model [7] or discriminative model [8]), the employed object detector has been based on local appearance only. This is done by scanning the whole image pixel-wise and calculating the probability for the presence of an object of interest only within a surrounding window of a certain size without using contextual information, e.g. [4, 13]. However, there is strong evidence from psychology, e.g. [1], that context plays a crucial role in scene interpretation and understanding. Humans can identify objects even when the local appearance is too weak for a unique decision. This is done by using contextual information and by applying a reasoning to identify the object of interest. An example is shown in Figure 1, where most people will have little trouble to recognize the marked objects in the image. However, shown in isolation an indisputable recognition of these patches is not easily possible. In this paper we present a framework on how to extract and learn contextual information from images, which is then used to boost the performance of state-of-the-art object detection methods. The principle idea is to extract context probability maps from images and learn their configuration for certain object classes (e.g. for pedestrians). The learned model is then used to calculate a context confidence map

[1)] This research has been supported in part by the following funds: Research program Computer Vision P2-0214 (RS) and EU FP6-511051-2 project MOBVIS.

for an arbitrary image. This data is then used in a cascade to filter out incorrectly detected objects.
The remainder of this paper is organized as follows: Section 2 provides an overview of related work. We turn our attention to our approach in Section 3. Experimental setup is described in Section 4 and results are reported in Section 5. Conclusions are drawn in Section 6.

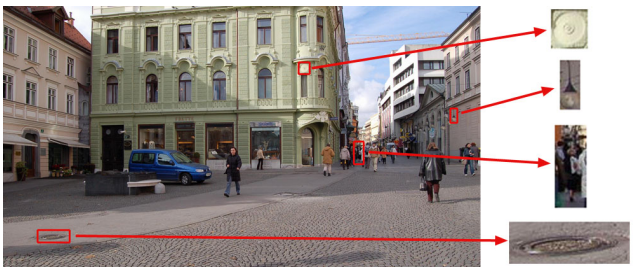


Figure 1: The object hypothesis formed from local appearance is rather weak for unique object recognition. Using the surroundings of the patches significantly aids recognition.

2 Related work

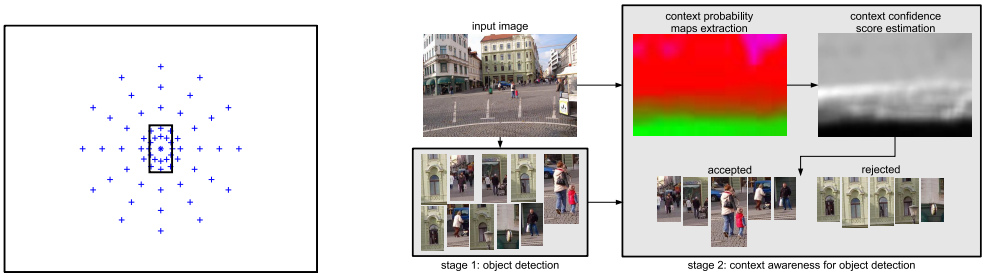
Extensive study on context for computer vision was done by Oliva and Torralba [11, 14]. The main idea is to categorize scenes based on the properties of the power spectrum of images. Out of the spectrum, semantic categories are extracted in order to grasp the gist of the scene. The image is classified as, e.g., an urban environment, a coastline, a landscape, a room, etc. As the category of an image is determined, the average position of objects of interest within the image (e.g., a pedestrian, a car, etc.) is learned from a large database (the LabelMe image database is often used [12]). This coarse position could then be used as a prior to limit the search space for object detection.
Another definition of spatial context is given by Hoiem *et al.* [9], where the idea is to extract a geometric context from a single image. The image is classified into three main classes, namely "ground", "vertical" and "sky". This classification is done by using many features, including texture, shape and color information in combination with geometrical cues. Then, from a huge labeled database, a classifier is trained using AdaBoost based on weak decision tree classifiers. Using these geometrical context classes as a prior, Hoiem *et al.* extended classical object detection into 3D space by calculating a coarse viewpoint prior [10]. The knowledge of the viewpoint limits the search space for object detection (e.g. cars should not occur above the horizon). In addition, the possible sizes of the objects of interest are limited given the geometric relationship between the camera and the scene.
Bileschi [2] classifies an image into seven pre-defined semantic classes. Four are texture based (building, road, sky, tree), where the remaining ones are defined by shape (cars, bicycles, pedestrians). These classes are learned from different sets of "standard model features" (also known as HMAX). Bileschi then defines the context by using low-level vision features from the

Blobworld system [3] (three color and three texture based features), in addition 10 absolute image positions are encoded followed by four binary sematic features, representing the four extracted classes (building, road, sky, tree). To extract a context vector for one given position in the image, the data is sampled relatively to the object center for 5 radii and 8 orientations, which results in an 800 dimensional feature vector. However, when using this type of contextual information for object detection in addition to a standard appearance-based approach, the gain in the detection rate is negligible. This issue is also confirmed by [15]. Another outcome of the extensive studies by Bileschi is that using global position features (also used by Torralba and Hoiem) indeed helps to improve the detection rate, however this is due to the input image data being biased (e.g. cars are more likely to be in the lower half of the image, as all the input images are aligned that the horizon is centered in the image in Bileschi's image database).
One major drawback of all listed methods is that the positions of the objects of interest are learned from a labeled database comprising of images shot in a limited set of predictable compositions. In fact, when acquiring images to label objects, it is very likely that the objects of interest will be placed in the center of the image or at least not positioned close to the image borders. That is why the relative object position from the LabelMe database, for example, is biased and therefore this position prior only holds for average standard images, but not for arbitrary rotated or tilted images. Our approach avoids this issue by first, providing a general framework not limited to one specific definition of context and second, learning contextual information instead the object positions.

3 Our approach

In this work we want to boost the performance of any object detection algorithm by filtering out incorrect detections based on context in a cascade. In the first stage of the cascade the object detection algorithm gives bounding boxes of potential objects of interest. In the second stage, the context of these rectangles is explored to reject objects that are at unrealistic positions in terms of context. The object detection algorithm is fully separated from context extraction and filtering. We assume that contextual information can be stored in maps containing probabilities for certain semantic classes. These context probability maps are directly extracted from an image and examples could be vegetation, sky, cars, etc. Logically, these classes should be directly related to the objects to be detected. The properties of all maps are extracted for one specific object category (e.g. pedestrians). This is done by sampling the data of the maps relatively to the object centers for a certain number of radii and orientations (as is done by Bileschi and visualized in Figure 2(a)). The receptive field, that is, those pixels in the image which influence that feature, is chosen to be quite large, so as to capture a more global context. Each object is used to construct a feature vector containing the contextual information. These positive feature vectors together with negative feature vectors, extracted from randomly drawn image patches of images not containing the specific object category, are

passed to a strongly supervised learning algorithm. The learned model should be capable of discriminating between realistic or unrealistic context for object detection. Using this model a context confidence score is estimated for each position in the image. In the final step the object recognition scores achieved from an object detection algorithm are combined with the context confidence scores, to filter out incorrect detections. The main concept of our method for performing context aware object detection is shown in Figure 2(b). To demonstrate that



(a) (b)

Figure 2: (a) An illustration of the 60 relative sampling positions, plotted as '+' signs, relative to the object center marked as a star. The thick black rectangle represents the average size of pedestrians. (b) Novel concept of context awareness for object detection for the example of pedestrian detection: Context probability maps are extracted from the input image, from which a context confidence score is estimated for one specific object category. In parallel a standard object recognition method is used to detect this object category. In the final step both results are fused improving object detection accuracy.

this concept is feasible, we visualize the average object of interest, the average magnitude and the average context in Figure 3(a), where we choose pedestrians as objects of interest. The employed context probability maps are the three layers from Hoiem's approach [9]. The average pedestrian is standing on the ground, the body is in the vertical context class and is not located in the sky. Since this is not a random configuration, it could be learned, given positive and negative examples.

4 Experimental setup

For testing and evaluating our method, we use pedestrians as objects of interest. However, this is an arbitrary choice and other classes could be used as well. To detect pedestrians in images, we tested methods of Dalal and Triggs [4] and Seemann *et al.* [13]. Both detectors are shape based methods, where the former uses histograms of gradients and the latter uses implicit shape models. We determined that Dalal's approach produces fewer false positives, therefore it is used in our experimental setup. For defining the context probability maps we use Hoiem's definition of context, as shown in Figure 3(b). As with the objects, our method is not limited to this specific kind of context. On the contrary, any type of probability map could be used in this framework. We could take other object recognition cues and use their output probability map as a input for our algorithm (e.g. the existence of a pedestrian crossing could

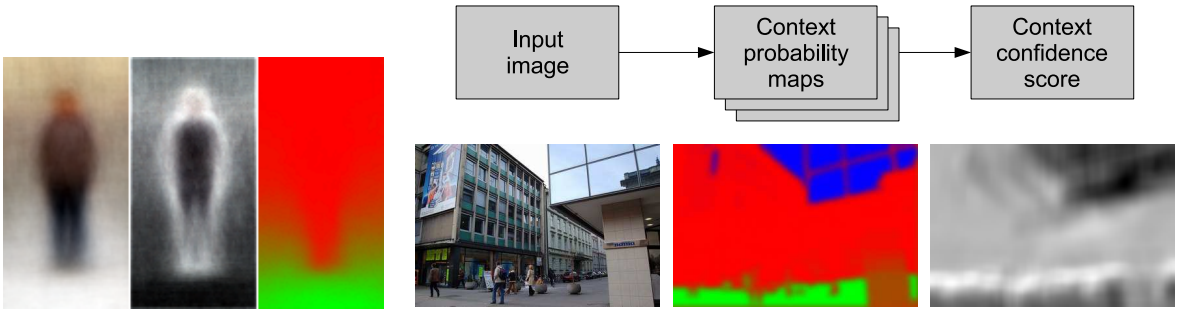


Figure 3: (a) Average pedestrian out of 2175 detected pedestrians from LPP-34 image database. Shown are the average pedestrian, the magnitude image and the average Hoiem's geometrical context. The context is color coded: ground (green), vertical (red) and sky (blue). (b) Novel concept of calculating a prior for object detection (in this case for pedestrians) using only contextual information. Starting from an input image context probability maps are extracted (color coded and histogram equalized), which are used to calculate a context confidence score for the presence of pedestrians. Bright regions indicate likely positions where pedestrians could occur in the image, whereas dark regions indicate unlikely positions. Best viewed in color.

be helpful for pedestrian detection).

Extracting context: Using Dalal's pedestrian detector we extract pedestrians from the given image database and label the extracted objects as true or false positives using the ground truth information. Most of the false positives are detected because the shape of the objects is characteristic for human silhouettes (e.g. rounded objects are misinterpreted as a human shoulder, etc..). For the remaining positive examples, geometrical context vectors are formed, containing the class probabilities of the three context classes extracted with Hoiem's method. The context probability maps are downsampled to a width of 80 pixels and smoothed with a 5×5 pixel average filter. This specific width of 80 pixels was inspired by [2] and is meant as a tradeoff between capturing the gross contextual information and being computationally efficient. By using 12 orientations and 5 radii ($r \in [3, 5, 10, 15, 20]$ percentage of the image diagonal) for each pedestrian, a 180 dimensional context feature vector is extracted. We also tested various other combinations of orientation counts and radii. We found that small variations do not have strong influence on the results and that the given combination gives best results. Note, that the receptive field used is quite large. In fact, one quarter of the input image contributes to the context feature extraction for each object.

Learning context: To be able to learn this context, negative examples are drawn from images containing no pedestrians, at random positions. Having positive and negative examples, a linear support vector machine (SVM) is trained, which should be able to separate true context feature vectors from false ones. The positive and negative examples are divided randomly into two sets, where the first is used for training and the second one for testing. To verify the robustness of the SVM learning this cross-validation is repeated several times. In this test the classification rate is very stable (changes less than 1% over 100 iterations).

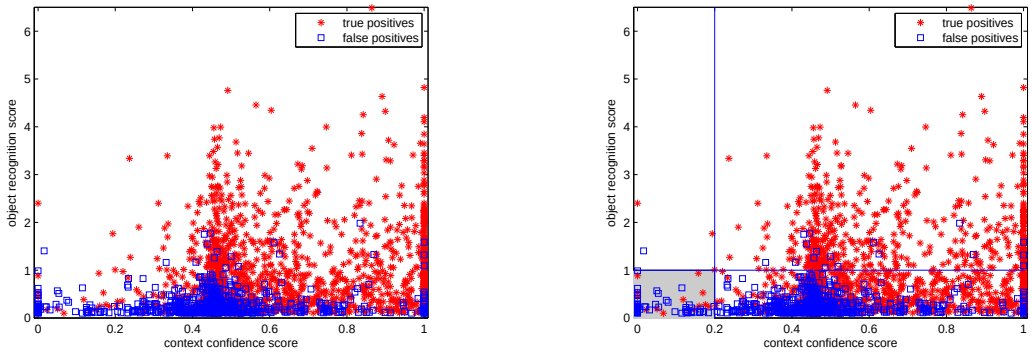
Using context: Then, using the learned SVM model a probability map is calculated for

arbitrary test images, representing the probability that at the current pixel position a pedestrian is present, using only context information (called context confidence score). This is done by converting the SVM output score into a probability score, where several parameters are needed. This is done by first zero-meaning the data, then setting the standard deviation to $1/\sqrt{2}$, clipping values below -1 and above $+1$ and finally scaling the data to $[0, 1]$. The basic reason for the clipping is to get rid of outliers. Figure 3(b) illustrates the concept and shows the result on one LPP-34 image.

5 Results

The novel concept described in the previous sections is tested on the MOBVIS LPP-34 image database, comprised of 612 images of Ljubljana's center with a resolution of 3008×2000 pixels. For our experiments the images are downsampled to 1504×1000 pixels. The images contain standard urban scenes, therefore this specific context is learned. It is very unrealistic to expect that the same learned model would also be able to provide useful information when, e.g., using non-urban images. All pedestrians which can be detected using Dalal's algorithm, that is, pedestrians with a height from 78 to 500 pixels, are manually labeled in the image database to have ground truth bounding boxes. Overall the ground truth consists of 2060 bounding boxes covering pedestrians. Using Dalal's pedestrian detector we extract 2175 objects. The detected pedestrians have an average size of 109×215 pixels with a standard deviation of 53 pixels in x direction and 105 in y direction. 1500 (69%) detected objects are pedestrians and 675 (31%) are incorrectly detected objects while 560 pedestrians are not detected. A correct detection (true positives) should cover the whole pedestrian including one quarter of its width as border according to Dalal's definition (this means that a object of the size of e.g. 64×128 pixels detected as a pedestrian should contain the pedestrian in the center surrounded by a margin of 16 pixels). Therefore, partially detected pedestrians are declared as an incorrect detection (false positives). Note that Dalal's detector is among the best object detectors even though many false positives are generated and many objects are missed. For each detected object the 180 dimensional context feature vector is extracted, defining the true context. Negative examples are drawn from manually selected images containing no pedestrians. For this experiment 5000 objects are cropped at random positions, where the sizes adhere to the statistics from the true pedestrian samples. A linear support vector machine (SVM) is subsequently trained with these positive and negative examples. The accuracy of the training process is actually not very high (approximately 81% of the data is classified correctly). This comes from the fact that the data is not separable (many pedestrians are completely contained within the vertical context class, which is also true for arbitrary negative examples). Since the geometric context is only used to calculate a position prior to limit the search space for finding potential matches, it is important to filter out unlikely areas and emphasize likely ones. With the learned SVM a probability map is calculated for arbitrary test images, representing the probability that at the current pixel position a pedestrian is present, using contextual

information only. Figure 4 shows the plot of object recognition score versus the context confidence score. It is obvious that with the introduction of this new "dimension" we can filter out some false positives, which cannot be filtered just by using the appearance-based object recognition score. On the other hand objects sharing the mean context confidence score (0.4 - 0.6) are not distinguishable in true and false positives. So only objects which are heavily out of context are detected by this filtering procedure. Figures 5(a) show examples of correctly filtered false positives, while Figure 5(b) shows incorrectly filtered true positives. In this case we reject 68 false positives while losing only 18 true positives. In other words, 10% of incorrect objects were detected by losing 1.2% of the correct ones. These results are a proof of concept and they should be improved to be useful in practice.



(a) (b)

Figure 4: Results: (a) visualization of object recognition score (ORS) versus context confidence score (CCS). Two observations are made: Detected objects with an ORS higher than 1.0 are very likely to be true positives (only 3% of false positives have a higher ORS than 1.0) and objects with a CCS with less than 0.2 are very likely to be false positives (99% of true positives have a higher CCS than 0.2). (b) visualization of the filtered out objects (within the gray rectangle).

(a) (b)

Figure 5: Results: (a) filtered out false positives by using context (24 out of 68 examples are shown). (b) filtered out true positives by using context (8 out of 18 examples).

6 Conclusion and future work

Using our approach on *context awareness for object detection* we are able to filter out incorrectly detected objects which are out of context. This means that we can decrease the number

of incorrect detections. For the image database we used 10% of wrongly detected objects were rejected by only losing 1.2% correct detections. However, in the future we also want to increase the number of correct detections that are not recognized by our current object detection algorithm. Therefore, we will need the object recognition score and the estimated size of the object for each pixel position in the image. Only the bounding box and the object recognition score for the detected object, such as what we retrieve with Dalal's algorithm, is not enough. With this information we will be able to combine the object recognition score and the context confidence score as before, but in this case, instead of rejecting false detections, we will add correct detections. In addition the research will include results from different object detection algorithms and will explore novel context probability maps, e.g. using texture as a contextual cue.

References

[1] I. Biederman. *On the semantics of a glance at a scene*, chapter 8, pages 213–263. Perceptual Organization. Lawrence Erlbaum, 1981.

[2] S. Bileschi. *StreetScenes: Towards Scene Understanding in Still Images*. PhD thesis, Massachusetts Institute of Technology, May 2006.

[3] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *VISUAL*, Amsterdam, June 1999.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *CVPR*, volume 2, pages 886–893, San Diego, June 2005.

[5] L. Fei-Fei, R. Fergus, and A. Torralba. Recognizing and learning object categories. `http://people.csail.mit.edu/torralba/iccv2005`, Tutorial presented at ICCV 2005, October 2005.

[6] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *CVPR*, San Diego, June 2005.

[7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJRS*, 61(1):55–79, January 2005.

[8] J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Dept. of Statistics, Stanford University, Technical Report, August 1998.

[9] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, October 2005.

[10] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, New York, June 2006.

[11] A. Oliva, A. Torralba, A. Guerin-Dugue, and J. Herault. Global semantic classification of scenes using power spectrum templates. In *The Challenge of Image Retrieval*, 1999.

[12] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. Technical Report AIM-2005-025, MIT AI Lab Memo, September 2005.

[13] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *CVPR*, New York, June 2006.

[14] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):153–167, July 2003.

[15] L. Wolf and S. Bileschi. A critical view of context. In *CVPR*, New York, June 2006.