Learning visual context for object detection *

Roland Perko and Aleš Leonardis
University of Ljubljana, Faculty of Computer and Information Science
Tržaška 25, SI-1001 Ljubljana, Slovenia
{*roland.perko,ales.leonardis*}*@fri.uni-lj.si*

Učenje vizualnega konteksta za detekcijo objektov

Kontekst ima pomembno vlogo pri splošnem zaznavanju prizorov, saj zagotavlja dodatno informacijo o možnih lokacijah objektov v slikah. Detektorji objektov, ki se uporabljajo v računalniškem vidu, tovrstne informacijo običajno ne izkoristijo. V članku bomo zato predstavili koncept, kako se lahko kontekstualne informacije naučimo iz primerov slik prizorov. To informacijo bomo uporabili za izračun kontekstnega polja, ki predstavlja apriorno informacijo za detekcijo objektov glede na možne lokacije. Detekcija objektov, ki temelji na lokalnem videzu, je potem selektivno uporabljena le na nekaterih delih slike. Predlagano metodo smo preizkusili na primerih detekcije pešcev, avtomobilov, in oken, pri čemer smo uporabili zahtevne podatkovne zbirke slik urbanih okolij. Rezultati so pokazali, da kontekstualna informacija dopolnjuje lokalno informacijo na podlagi videza, ter tako zmanjša kompleksnost iskanja in poveča robustnost detekcije predmetov. Prednost predlagane metode je tudi v tem, da je učenje kontekstualnih konfiguracij za različne kategorije objektov neodvisno od specifičnih modelov za posamezne naloge.

1 Introduction

In the real world there exists a strong relationship between the environment and the objects that can be found within it. Experiments with scene perception, interpretation and understanding have shown that human's visual system extensively use these relationships to make object detection and recognition possible [1]. Humans can identify a given object in a scene, even if they would not normally identify the same object when it is presented in isolation. The limitation of a local appearance being too vague is resolved by using contextual information and by applying a reasoning mechanism to identify the object of interest. An example is shown in Fig. 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 achieved. In general context plays a useful role in object detection in at least two ways. First, it helps detection when local intrinsic information about the object is insufficient. Second, even when local appearance based object detection is possible, the search space can be cut down by setting the attention to image regions where the occurrence of the objects of interest is most likely. For example, when searching for manholes in Fig. 1 the search can be constrained to the ground plane.

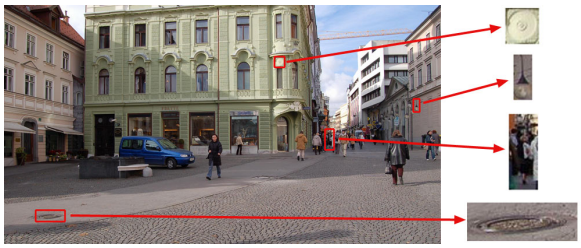

Figure 1: The object hypothesis formed from local appearance is rather weak for unique object recognition. Using the surroundings of the patches significantly aids recognition.

Even though object detection is a well established discipline in computer vision and is used in a large number of applications, contextual information is typically ignored. Many concepts for object detection have been developed where the employed object detector is based on local appearance alone (see e.g. [4] for a review). In this paper, we present a concept of how to extract and learn contextual information from examples. This context is used to calculate a context confidence map, that represents a prior for object detection. The configuration of the contextual features is learned from given examples with machine learning approaches. Therefore, no task specific models are required and the approach is not limited to one specific object category. We show that our approach is general by learning visual contextual features for pedestrian, car and window detection without tweak-

* 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.

ing the underlying model. An illustration of the concept is given in Fig. 2.

The remainder of the paper is organized as follows: Sec. 2 provides an overview of related work. Our approach is presented in Sec. 3. Experimental results are reported in Sec. 4 and the conclusion is given Sec. 5.

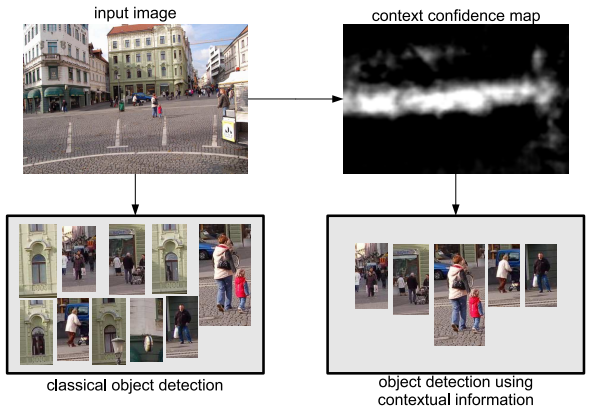

Figure 2: Illustration of using visual context for object detection shown on the example of pedestrian detection. Classical detection approaches search for the objects in the whole image and therefore are more prone to making incorrect detections (left). Our approach derives a context confidence map and then the object detector is applied only in regions where the presence of the objects is most likely. Therefore, the number of incorrect detections is decreased (right).

2 Related work

Extensive study on visual context for computer vision was done by Torralba and Oliva [7]. 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 so called *gist* of the scene. The image is classified as, e.g., an urban environment, a coastline, a landscape, a room. As the category of an image is determined, the average position of objects of interest within the image is learned from a large database. This coarse position can then be used as a prior to limit the search space for object detection. In contrary to the work by Torralba *et al.*, we use very different features and also completely different feature descriptions. In addition, we provide an evaluation for combining our contextual priming method with a state-of-the-art object detector.

A definition of spatial context is given by Hoiem *et al.* [5], 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*. Using these geometrical context classes as a prior, Hoiem *et al.* extended classical object detection into 3D space by calculating a coarse viewpoint prior [6]. 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. We use the geometrical context provided by Hoiem *et al.* as a contextual cue (beside other cues). The main difference is that we learn the configuration of this contextual information directly and not only to calculate a horizon estimate.

Bileschi [2] classifies an image into seven predefined semantic classes. Four classes (building, road, sky, tree) are texture based, where the remaining ones (cars, bicycles, pedestrians) are defined by shape. These classes are learned from different sets of *standard model features*. Bileschi then defines the context by using low-level visual features from the *Blobworld* system [3], 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. 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. Another outcome of the extensive studies by Bileschi is that using global position features (also used by Torralba and Hoiem) indeed help to improve the detection rate, due to the input image data being biased. For example, cars are more likely to be in the lower half of the image. This is because the horizon being in the center of each image, in Bileschi's image database. Our approach does not require such global position priors. In addition, we encode our contextual features as probabilities rather than binary values.

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 image databases are often 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 firstly, providing a general framework not limited to one specific definition of context and secondly, learning contextual information instead the object positions.

3 Our approach

The used approach to extract visual contextual features from images is based upon our work in [8]. It is assumed that contextual information can be stored in probability maps, which can encode high-level semantic or low-level image features. In this work, two complementary types of features are used to form contextual information: geometrical features and texture features. The former are the three semantic classes from Hoiem's approach [5] which give the probabilities that the current pixel belongs to the ground, the vertical class (buildings, trees, etc.) or the sky. The latter describe texture features as defined within the *Blobworld* system [3], which capture information on the local structure and the local gradient magnitude.

Feature vectors are extracted for an object of interest by sampling these contextual feature maps relatively to the object's center for 5 radii and 12 orientations. Therefore, an 180 dimensional feature vector (60 samples and 3 context probability maps) sparsely represents the contextual information surrounding the current object of interest. Fig. 3 illustrates the workflow of how to extract contextual features from an image.

In the learning phase, for manually selected objects in the images such feature vectors are extracted (used objects of interest are pedestrians, cars and windows). These positive feature vectors together with negative feature vectors are passed to a strongly supervised learning algorithm - a support vector machine (SVM). Negative contextual feature vectors are extracted from randomly drawn image patches of images not containing the specific object category.

The learned contextual SVM model is used to calculate a context confidence score for each position in the image. This score gives the likelihood of the presence of the object of interest at this spatial position (see Fig. 2). It is calculated by extracting the corresponding contextual feature vector for each position and supplying it to the learned contextual model. Typically, the output of the machine learning algorithm is not probabilistic, so it has to be converted into the domain $[0, 1]$ to be probabilistic. We use robust statistics to fit a Gaussian distribution into the SVM outputs.

Overall, we are able to calculate a probability map, called context confidence map, for the presence of previously learned object categories (see Fig. 2).

4 Experimental results

The novel concept is tested on the LPP-34 image database, comprised of 612 images of Ljubljana's center with a resolution of 3008×2000 pixels each.

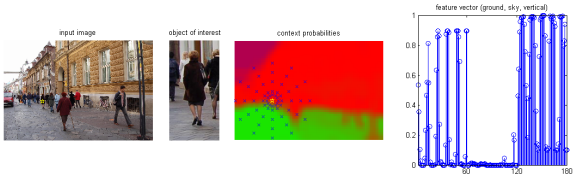

Figure 3: Workflow from an object of interest in an image to the corresponding contextual feature vector: input image with object of interest marked by a star; detailed view of the object of interest; geometrical context probability maps and the positions where the values are sampled to form the contextual feature vector; feature vector holding a sparse representation of the geometrical context of the object of interest.

For our experiments, the images are downsampled to 1504×1000 pixels. We manually labeled 3803 pedestrians, 803 cars and 3601 windows in the images. Each object is defined by the corresponding bounding box. Two sets of experiments were conducted to demonstrate the approach.

First, we learned three contextual SVM models for the given object categories. As a result we visualize two input images together with the calculated context confidence maps (see Fig. 4). Note, that no local information was used to extract these priors. As the existence of an object also influences the visual contextual features, in the case of the object category *cars* only regions where there are actually cars in the image get high scores not e.g. the whole road. In Fig. 4 we see, that pedestrians should appear on pavements, that is an image region where the ground plane (road) intersects with vertical structures (buildings). Cars occur on the road, which is next to the possible presence of pedestrians. Windows are basically above the road on facades of buildings. These different image regions are detected using our definition of visual context. In the object detection step it is sufficient to search in areas with high context confidence scores only. In case of pedestrian detection using textural context only 14% of the image pixels had to be searched on average, which yield a speed-up of factor 7.1 in the object detection step.

Second, we applied the local appearance based pedestrian detector by Seemann *et al.* [9] on our database[1]. The detection rate is plotted versus the false positives per image (FPPI) in Fig. 5. We calculated the detection rate for the local appearance detector alone, for the combination of local appearance with different contextual cues and finally for the combination of local appearance with all contextual cues. As the detector and all cues provides a probabilistic output, they can be fused by multiplying the

[1]We thank Edgar Seemann for providing the detections on our database.

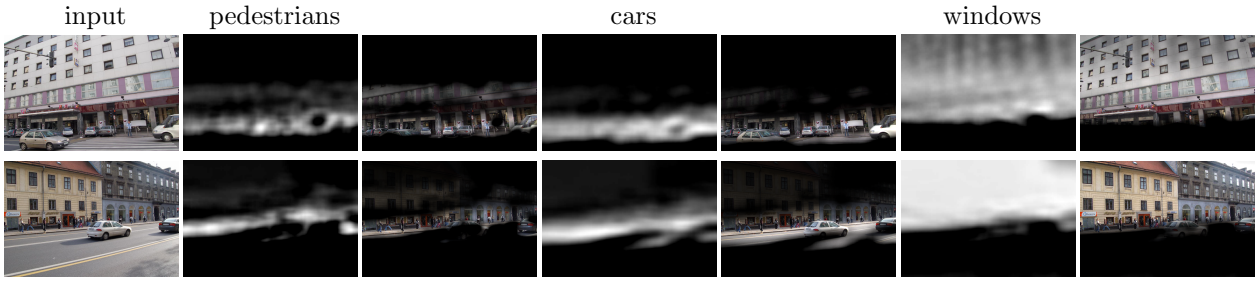

Figure 4: Context confidence maps for three different object categories shown for two images of our database. Shown are the context confidence maps for pedestrians, cars and windows and the input image multiplied by these maps. Bright areas indicate locations where the probability for the presence of the given object category is high.

individual scores (assuming that the underlying features are statistically independent). These curves reveal that geometry is the weakest contextual cue and the strongest cue to aid local detection is texture. The fusion of both contextual cues with the local appearance based method provides the best detection rate. In the range from 1 FFPI to 5 FPPI the detection rate is boosted by 22.8% on average when using visual context in addition to local appearance.

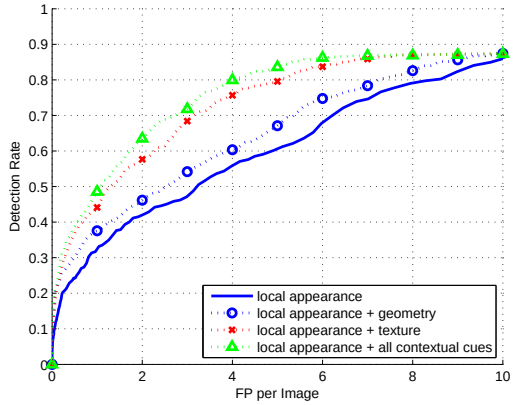

Figure 5: Detection rate curves using local appearance in combination with contextual cues are shown for pedestrian detection.

5 Conclusion

We presented a concept of how to extract and learn visual context from examples. This context was used to calculate a context confidence map, that represents a prior for object detection for a single image. Local appearance based object detection was then applied on selected parts of the image only. We have shown that our concept is very general and we do not have to design task specific models for different object categories. Results also revealed that context aware object detection provides speed-up and increases the robustness of the detections.

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. M. 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*, pages 509–516, Amsterdam, June 1999.

[4] L. Fei-Fei, R. Fergus, and A. Torralba. Recognizing and learning object categories. `http://people.csail.mit.edu/torralba/shortCourseRLOC`, Tutorial presented at CVPR, June 2007.

[5] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, volume 1, pages 654–661, October 2005.

[6] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, volume 2, pages 2137–2144, June 2006.

[7] A. Oliva, A. Torralba, A. Guerin-Dugue, and J. Herault. Global semantic classification of scenes using power spectrum templates. In *CIR*, 1999.

[8] R. Perko and A. Leonardis. Context awareness for object detection. *OAGM/AAPR*, pages 65–72, May 2007.

[9] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *CVPR*, volume 2, pages 1582–1588, June 2006.