LEARNING CONTEXTUAL RULES FOR PRIMING OBJECT CATEGORIES IN IMAGES

Roland Perko[1,2], *Lucas Paletta*[2], *Aleš Leonardis*[1]

[1] University of Ljubljana, Slovenia; [2] Joanneum Research, Austria

ABSTRACT

In this paper we introduce and exploit the concept of contextual rules in the field of object detection. These rules are defined as associations between different object likelihood maps and are learned from given examples. The contextual rules can be used to prime regions where a target object category occurs in an image given areas of other object categories. The principal idea is to locate several basic object categories in an image and then use this information to infer object likelihood maps for other object categories. The proposed framework itself is general and not limited to specific object categories. For demonstrating our approach, we use likely occurrences of pedestrians and windows in urban scenes, extracted by a technique employing visual context, and use them to prime for shop logos.

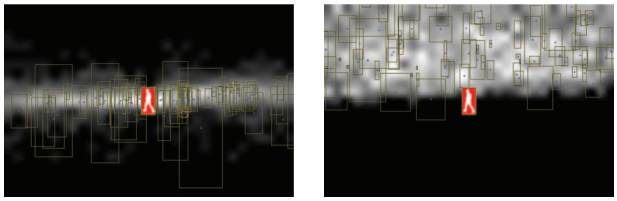
***Index Terms*—** Visual context, machine learning, object detection, scene understanding.

1. INTRODUCTION

In the real world, objects tend to co-vary with other objects and particular environments, providing a rich collection of contextual associations to be exploited by the visual system [1]. As it is known from the literature on visual cognition [2], cognitive neuroscience [3] and computer vision [4, 5], the human and animal visual systems use relationships between the surrounding and the objects to improve their ability of categorization. In particular, this *visual context* provides important cues about an object's presence, position and scale within the observed scene or image. In addition, the spatial relations between objects and between object categories are exploited by the visual system, as objects mostly do not occur in isolation. According to [1], a natural way of representing the context of an object is in terms of its relationship to other objects. That is, if one object in a scene or image is known, the locations and sizes of other target objects are strongly constraint.

To clarify the important role of context we conducted a proof-of-concept experiment. We fixed the location of one pedestrian in the image center and marked the relative occurrences of other pedestrians and windows using the urban test image data set of [6]. In Fig. 1 the probability distributions for the two object categories are visualized. It is clearly visible that the occurrence of different objects is not independent and that this dependency could and should be exploited for the task of object detection.

However, in the state-of-the-art literature of context based object detection the *context* is primarily determined with respect to a single object category. In this paper, we significantly advance the usage of visual context by proposing a concept of how to learn contextual associations, i.e. contextual rules, from previously extracted semantic scene information to prime for an object category of interest. Literally, we prime from the spatial locations of detected or given object categories to a target object category location, thus improving object detection accuracy. Using this novel concept, we do not have to define a new context for every object category. On the contrary, existing context frameworks are exploited where their spatial relation already provides a new context. Throughout this work we will use likely occurrences of *pedestrians* and *windows* in urban scenes as input object categories and use them to prime for *shop logos*. However, the presented approach is general and not limited to these specific categories. Our argument for using this specific object category is that we humans would define the possible location of shop logos (thereafter called *logo zone*), as above eye-level and below the first rows of windows on the building's facade. Logos are in general not defined by a fixed given shape, as it is the case with other categories in urban scenes, like cars, traffic signs, or pedestrians. A shop logo is rather defined by its relative position within the observed scene and of course by its visual protruding appearance. We try to exploit the spatial relation to other objects, rather than papers focusing on classical visual attention concepts for object detection, see e.g. [7].



(a) (b)

Fig. 1. Conditional dependencies of objects. The probability distribution of (a) pedestrians and (b) windows is visualized, under the condition of the presence of one pedestrian in the center of the image. A couple of object locations are shown by their corresponding bounding boxes. These results were obtained using a database of several hundred images of urban scenes. Best viewed in color.

Our basic idea for applying contextual rules is sketched in Fig. 2. From an input image the possible occurrences of pedestrians and windows are extracted, integrating visual contextual information alone. In the next step the spatial relation of these priors, also called foci of attention or object likelihood maps, is exploited and used to prime the location of shop logos.

The proposed framework is very general and enables learning of different contextual associations. The relations between objects can be used, or between an object and an object category (cf. Fig. 1), or even between object categories. In this paper we only aim to exploit the relationship of different object categories and not of objects within one category. The reason for this is, that in many urban images only a few objects of the same type may exist, so that the

The work is partially supported by the European Commission funded project MOBVIS under grant number FP6-511051, by the FWF Austrian National Research Network on Cognitive Vision under sub-project S9104-N13, and by the J2-2221-1539 and P2-0214 Slovenian Research Agency projects.

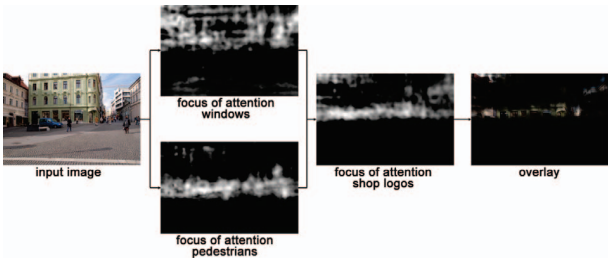


Fig. 2. The basic idea of object priming based on contextual rules. Given an image, foci of attention for specific object categories are extracted. Then, these *regions* are used to calculate probable regions for a target object category. The resulting image shows the primed regions, in this case for shop logos, overlayed on the input image.

possible priming within object categories is inefficient or even infeasible. In addition, we will show, that we can also prime from object categories even in case when no instance of this category is existing in the image.

Our contribution. We describe how to transfer the concept of *contextual rules* into a working framework for priming object categories of interest in a single image. We incorporate a level of semantic abstractions, which is learned and exploited by a spatial sampling procedure. The proposed framework is general and could be combined with any existing visual context extraction approach and with any object detection method.

The concept is based on hierarchically using prior extracted knowledge to derive novel information, where our ambition is to leverage context for object detection towards the ultimate goal of automatic scene understanding. Ideally, only a few basic object categories have to be detected in an image, to automatically locate several other categories using the presented sketch. Our method allows to arbitrarily combine prior knowledge to learn novel or hitherto unknown contextual relations, which was not possible with existing methods. As a proof of concept we report the results for the object category *shop logos* in urban scenes. However, the actual combination of a local appearance-based object detector with the novel extracted context is not within the focus of the presented work.

2. RELATED WORK

In recent years computer vision researchers have developed powerful concepts for appearance-based object detection. Famous examples are the algorithm in [8] for face detection based on a cascade of boosted weak classifiers, [9] for pedestrian detection based on histograms of gradients and [10] for generic object detection based on implicit shape models. Even though these methods are based on different methodologies they have at least one common aspect, i.e. ignoring prior scene knowledge. It is implicitly assumed that the probability of an object occurrence is uniform over the whole image (which is not the case, cf. Fig. 1). Lately, research on general scene perception and categorization provided the novel concept of using *visual context* to guide object detection. The most influential work to be mentioned is the one in [11], where so called *gist* features are defined. They are used to rapidly categorize the content of an image into coarse semantic scene types. The knowledge about the scene content is hence used to prime for the possible type of objects, their location and their size in an image. Another example is [12], where the relation of objects, together with the geometry of the scene are combined into a framework using inference. The authors of [5] reported on a method of context aware object detection, where priors are extracted from geometrical and texture features. These priors are then used to limit the search space for object detection while simultaneously increasing detection performance. In [13] priming from objects to objects is performed. In [14] object detection is facilitated by performing inference between objects and their surrounding. To the best of our knowledge, there is no paper on learning contextual rules, that are used to prime for a target object category given other object categories. Therefore, this basic idea is exploited in this work.

3. OUR APPROACH

Our approach builds upon the works in [15, 5], which enable the extraction of priors for specific object categories from a single image, visualized in Fig. 3(a). As shown in Fig. 3(b) we add another layer to the framework. The intermediate layer holds object likelihood maps for pedestrians and windows. These maps are derived directly from image features, which are based on geometry [12] or texture [16]. As these maps hold semantic information they are called *semantic abstractions*. In the consecutive step, the spatial configuration of these object likelihood maps is learned and exploited via contextual rules. These rules take the semantic abstractions as input and derive another object likelihood map for a novel object category (i.e. shop logos in our example).

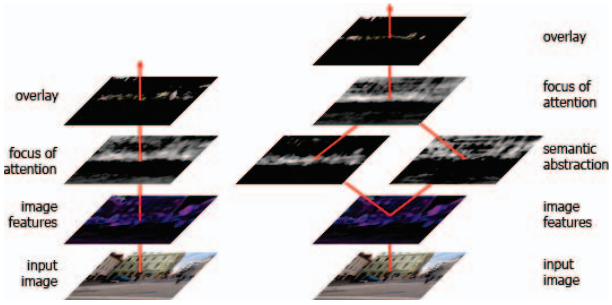
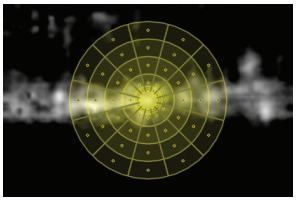


Fig. 3. Visual context extraction of the object category *shop logos* in urban scenes. (a) Approach for extracting visual context as in [5]. (b) Novel method exploiting contextual rules, i.e. contextual associations, formed using semantic abstractions. In this specific example the semantic abstractions are foci of attention for pedestrian and window occurrence in the image.

Semantic abstraction. To get from images to the locations where objects of a category of interest are likely to occur, the algorithm of [5] is used. This method can be triggered by different input image features, where geometry and texture features are used for evaluation. The output of this algorithm is a probability map, the focus of attention, for each pixel in the input image (spatially reduced for speedup) holding the information on how likely an object may occur at the current location. For our tests we run the method twice to get priors for two pre-trained classes, namely pedestrians and windows. These probability maps, are encoding semantic scene information and are the only information used to learn the contextual rules.

Learning contextual rules. As stated we define the contextual rules as the spatial relationship between object categories. Therefore, the representation should be capable of holding semantic information like “is left of”, “is in between of”, or “is above of”. Obviously, in the case of shop logos, the contextual rules should learn information like “shop logos are located above pedestrian heads and below the windowed facade of buildings”. To extract the data needed to

construct such rules, we use a spatial sampling, like in [15]. We therefore extract samples from the semantic abstractions, i.e. probability maps, for a given number of radii and orientations, with a rather big receptive field, i.e. one quarter of all pixels in the input image. We stick to 5 radii and 12 orientations as in [5], resulting in a 60-dimensional feature vector for each semantic abstraction. The sampling process is illustrated in Fig. 4.



$$\longrightarrow \quad f = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{60} \end{bmatrix} \in [0,1]^{60}$$

Fig. 4. Gathering semantic abstractions from an image. As input data a focus of attention is used, which holds the probabilities of how likely an object of interest occurs at this image location. Then the feature vector for one given location in the image is defined in a radial manner. In each of the visualized regions (illustration left) the central probability value in the domain $[0, 1]$ is extracted. The feature vector f is formed by clockwise sampling while increasing the radius, yielding a 60-dimensional vector. Such a feature vector contains the spatial relation of the sematic abstraction given for one object category.

We do not want to bring any explicit modeling into the system, keeping the approach general for any kind of object categories. Therefore, we just rely on positive and negative instances of the target object category. While positive samples are randomly drawn from manually labeled logo zone regions, negative samples are randomly drawn outside the logo zone. At every location of these supplied samples one feature vector is extracted holding the semantic abstractions (cf. Fig. 4) (i.e. an $n \times 60$-dimensional vector, with n being the number of input objects categories used). These labeled feature vectors are passed to a strongly supervised learning algorithm, where we use a linear support vector machine (SVM). The so learned contextual model represents our contextual rules and is able to "predict" regions where the occurrence of our target object category is likely.

Applying contextual rules. Once we have the contextual rules, i.e. an SVM model, learned for our target category, they can be applied to any new image. First, the foci of attention are extracted for the basic object categories. Second, new feature vectors are extracted for every position in the image holding the semantic abstractions. This procedure is the same as in the training step (cf. Fig. 4). Third, these feature vectors are supplied to the learned SVM yielding a score for each position in the image. Rather than mapping the SVM scores to probabilities [17], we use robust statistics for mapping the score into the domain $[0, 1]$, assuming that the data is approximately Gaussian distributed. Overall, we get a context score for each image location which represents the confidence that a target object occurs at this specific location. Large confidence values are typically clustered and define regions of high probability for the existence of a target object. These probable object locations are primed using spatial contextual rules, holding the relation of other object categories within the image.

Contextual rules and object detection. Once the contextual confidences are determined, there are several ways to combine them with an object detection task. One classical approach is to use the extracted *context* in a cascade just to limit the search space for object detection, yielding speedup [15]. Another method is to set up a probabilistic framework [18] which combines contextual and object detection scores. By re-ranking the appearance-based detections, the object detection accuracy is improved.

4. EXPERIMENTAL RESULTS

Data set and ground truth. We use the Ljubljana urban image data set[1] [6] containing 612 images of Ljubljana's center with a resolution of 1504×1000 pixels each available at the given website. The images are captured at three distinct tilt angles, which is a very challenging aspect, since the shop logos are in general not just in the center of the image. To get the ground truth for potential logo zone regions we manually labeled these regions in all images, where some examples are given in Fig. 5(a-b). Our logo zones are indicated as white regions in Fig. 5(b). Then, from the whole set of images 10000 positive and the same number of negative examples are randomly drawn and supplied to the SVM learning approach.

Learning contextual rules. In our implementation the feature vectors have 60 dimensions for each semantic abstractions. Since, two abstractions are used (for pedestrians and for windows), the combined feature vectors are 120-dimensional. For learning, the set of positive and negative feature vectors were randomly split into two sets, one used for training, the other one for testing. The classification accuracy for texture-based features was 78% and for geometry-based features also 78%, each with a standard deviation of approximately 0.3% for 100 different random splits. However, the confusion matrices reveal a difference between the two set of features. The geometry-based learning was able to classify 97% of positives samples correctly, but only 57% of the negative ones. For texture-based features the numbers are 79% and 76%. The explanation is, that in the case of using the geometry features, the class of window priors is highly correlated with the extracted class of buildings (i.e. one layer of the semantic classification [12]). As shop logos are located at facades of buildings, positive examples can be well classified. However, negative examples which are on a buildings facade cannot be distinguished from positive samples and are therefore in general classified incorrectly.

Evaluation framework. The learned contextual rules are applied on the test images. Then, the confidence values are calculated inside and outside the manually labeled logo zones. Obviously, if the rules are representing the correct contextual associations, the average confidences should be significantly higher within the logo zones than outside. We made two kind of experiments, where we just changed the underlying low-level image features in the calculation of the semantic abstractions. Therefore, we are able to compare the impact of texture to geometry cues within this framework.

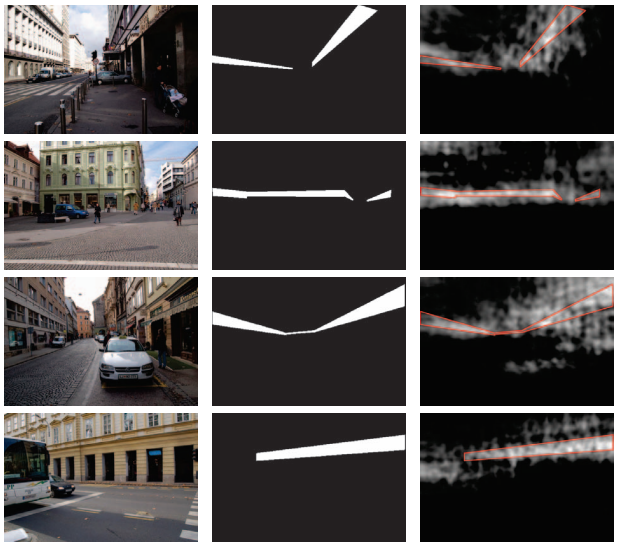
Results using contextual rules. The results are summarized in Table 1. The mean context confidence values together with the standard deviations are given for the two underlying image features, for the regions inside and outside the logo zone (cf. Fig. 5(b) for examples). In both experiments the context confidence values inside the logo zone are significantly higher than outside with a factor of 3.16 for texture and 2.16 for geometry. These factors lead to the conclusion, that in fact the cue based on texture provides a more reliable prior than the one using geometry. This matches the observations in [5]. Furthermore, it can be assumed that a combination of geometry and texture would even lead to better results, as these fea-

[1] http://vicos.fri.uni-lj.si/luis34/

tures are - at least partly - complementary. Overall, the algorithm is able to prime for our target object category given the two input categories. In Fig. 5(c) the confidence maps for the images in Fig. 5(a) are shown, overlayed by the ground truth logo zones. We want especially emphasize the results shown in the last row of Fig. 5. Even though there are no pedestrians in the image, the proposed concept is still working. This is simply to the fact, that the concept is based on object categories extracted using visual contextual information alone, not relying on instances of objects. In the concrete example the occurrence of pedestrians is likely close to the building, providing the correct context for shop logo detection.

		context confidence		factor of distinctiveness between these regions
		inside the logo zone	outside the logo zone	
texture	μ	0.45	0.14	3.16
	σ	0.22	0.06	
geometry	μ	0.33	0.15	2.16
	σ	0.06	0.03	

Table 1. Quantitative results of applying the concept of contextual rules for the task of shop logo priming. Mean values and standard deviations of the extracted context confidences are given for regions inside and outside the manually labeled logo zones, where the semantic abstractions are once extracted using texture-based features and once using geometry-based cues. The given factors indicate the distinctiveness between the logos zone and the rest of the images.



(a) (b) (c)

Fig. 5. Examples of input data and result. (a) Test images and (b) the manually defined logo zones in the training set. Note, that logo zones are labeled rather than individual instances of shop logos. (c) The extracted contextual confidence maps overlayed by the ground truth logo zones, here shown using texture features. From visual inspection it can be seen, that the algorithm is able to prime the locations where shop logos are existing. However, also some other image regions are in the focus of attention.

5. CONCLUSION AND OUTLOOK

In this paper we introduced a concept of contextual rules in the field of object detection. These rules were defined as contextual associations between context classes of object categories and were learned from given examples. As results these rules were used to prime regions where a target object category occurs in an image given regions of other object categories. The framework itself is general and not limited to specific object categories. For demonstrating the approach we took the task of detecting shop logos in urban scenes, where the possible occurrence of pedestrians and windows in the scene are used to define the contextual rules. It turned out that this concept is able to prime for the new category, where in the case of using texture features, the logo zone regions got larger context confidences scores, than the rest of the images (larger by a factor of 3.16).

The next steps are first, to arrange a large scale experiment, by e.g. using 10 input categories and prime for 10 output categories. Second, to fully employ object to object, object to object category and category to category relations. And third, to combine a classical appearance-based object detection method with the achieved results of the presented contextual priming.

6. REFERENCES

[1] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.

[2] S. E. Palmer, "The effects of contextual scenes on the identification of objects," *Memory and Cognition*, vol. 3, pp. 519–526, 1975.

[3] M. Bar, "Visual objects in context," *Nature Reviews, Neuroscience*, vol. 5, pp. 617–629, 2004.

[4] A. Torralba, A. Oliva, M. Castelhano, and J. M. Henderson, "Contextual guidance of attention in natural scenes: The role of global features on object search," *Psychological Review*, vol. 113, pp. 766–786, 2006.

[5] R. Perko and A. Leonardis, "Context driven focus of attention for object detection," in *WAPCV*, vol. 4840, chapter 14, pp. 216–233. 2007.

[6] D. Omerčević, O. Drbohlav, and A. Leonardis, "High-dimensional feature matching: Employing the concept of meaningful nearest neighbors," in *ICCV*, 2007.

[7] L. Paletta and C. Greindl, "Context based object detection from video," in *ICVS*, 2003, pp. 502–512.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, vol. 2, pp. 886–893.

[10] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *IJCV*, vol. 77, no. 1-3, pp. 259–289, 2008.

[11] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[12] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *ICCV*, 2005, vol. 1, pp. 654–661.

[13] K. Murphy, A. Torralba, and W. T. Freeman, "Using the forest to see the trees: A graphical model relating features, objects and scenes," in *NIPS*, 2003, vol. 16.

[14] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *ECCV*, 2008, pp. 30–43.

[15] S. M. Bileschi, *StreetScenes: Towards Scene Understanding in Still Images*, Ph.D. thesis, Massachusetts Institute of Technology, 2006.

[16] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *PAMI*, vol. 24, no. 8, pp. 1026–1038, 2002.

[17] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[18] A. Torralba and P. Sinha, "Statistical context priming for object detection," in *ICCV*, 2001, vol. 1, pp. 763–770.