Context Driven Focus of Attention for Object Detection

Roland Perko and Aleš Leonardis

University of Ljubljana, Slovenia
{roland.perko,ales.leonardis}@fri.uni-lj.si

Abstract. Context plays an important role in general scene perception. In particular, it can provide cues about an object's location within an image. In computer vision, object detectors typically ignore this information. We tackle this problem by presenting a concept of how to extract and learn contextual information from examples. This context is then used to calculate a focus of attention, that represents a prior for object detection. State-of-the-art local appearance-based object detection methods are then applied on selected parts of the image only. We demonstrate the performance of this approach on the task of pedestrian detection in urban scenes using a demanding image database. Results show that context awareness provides complementary information over pure local appearance-based processing. In addition, it cuts down the search complexity and increases the robustness of object detection.

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 the human visual system extensively uses these relationships to make object detection and recognition more reliable [1,2,3]. In the proper context, humans can identify a given object in a scene, even if they would not normally recognize 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 have little trouble in recognizing 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 attending 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.

Even though object detection is a well established discipline in computer vision and is used in a large number of applications, contextual information is

L. Paletta and E. Rome (Eds.): WAPCV 2007, LNAI 4840, pp. 216–233, 2007.

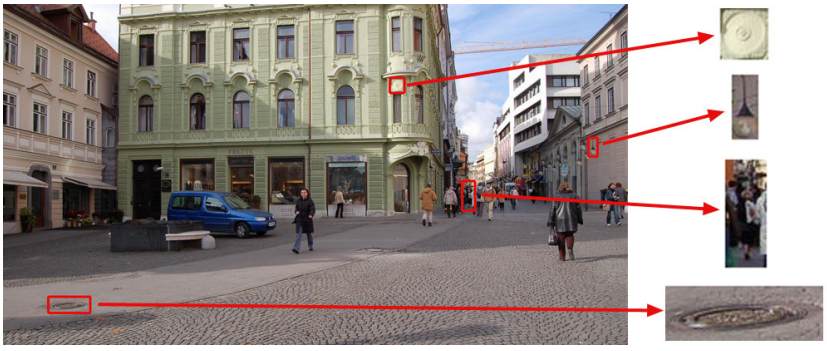


Fig. 1. The object hypothesis formed from local appearance is rather weak for a unique object recognition. Using the surroundings of the patches significantly aids recognition.

typically ignored. Many concepts for object detection have been developed where, independent of the particular representation model used, the employed object detector is based on local appearance alone (see e.g. [4] for a review). Standard representation models are bag-of-features models [5,6,7], part-based models [8,9] or discriminative models [10,11]. One could argue that part-based models use some sort of contextual information as not only the visual appearance of the parts are used, but also their locations. However, this context is very locally defined and differs from the context definition in the paper, which is a more global measurement, typically involving the background.

In this paper, we present a concept of how to extract and learn contextual information from examples. This context is used to determine a focus of attention, that represents a prior for object detection. State-of-the-art local appearance-based object detection methods are then applied on selected parts of the image only. We also explore which kinds of contextual features are useful for this problem setting. 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 demonstrate the performance on the task of pedestrian detection in urban scenes using a demanding image database. Results show that context awareness provides complementary information over pure local appearance-based processing. In addition, it cuts down the search complexity and increases the robustness of object detection. An illustration of the overall concept is given in Fig. 2.

The presented work can be seen as an extension to our previous paper [12]. Here we emphasize the concept of the focus of attention which is the main novelty. In addition, new contextual features are proposed and evaluated using two object detection algorithms.

In the field of cognitive attention for object detection, researchers usually decouple the contextual information processing from object detection. These can be performed in a cascade or calculated in parallel and fused in the final

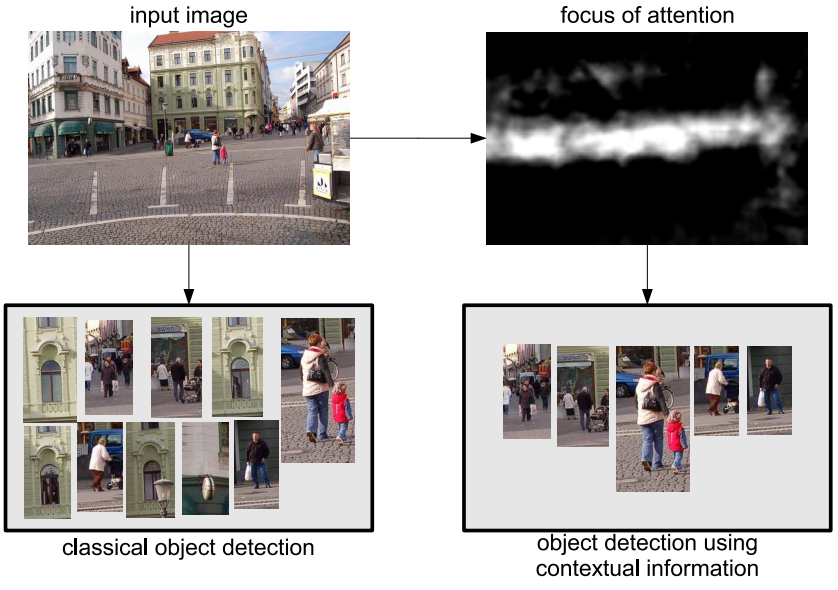


Fig. 2. Illustration of the *context driven focus of attention for object detection* concept shown on the example of pedestrian detection. Classical detection approaches search for the objects in the whole image and therefore are more prone to making incorrect detections (left). Our approach derives a focus of attention by using contextual information and then the object detector is applied only in regions where pedestrians are most likely to be found. Therefore, the number of incorrect detections is reduced (right).

stage. Using such an approach, any object detector can be combined with the contextual processing. Therefore, the related work is focused on how context can improve object detection rather than on object detection approaches.

The remainder of the paper is organized as follows: Sec. 2 provides an overview of related work. Our approach is presented in Sec. 3. Implementation details are described in Sec. 4. Experimental results are reported in Sec. 5 and the conclusion is given Sec. 6.

2 Related Work

Extensive study on context for computer vision was performed by Torralba and Oliva [13,14,15]. The main idea is to categorize scenes based on the properties of the power spectrum of images. Semantic categories are extracted from the spectrum in order to grasp the so called *gist* of the scene. The images are 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 (e.g., a pedestrian, a car) is learned from a large database. The LabelMe image database [16] is often used for such purposes. This coarse

position can then be used as a prior to limit the search space for object detection. In contrast 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 state-of-the-art object detectors.

A definition of spatial context is given by Hoiem *et al.* [17], where the idea is to extract a geometric context from a single image. The image is segmented into three main classes, namely *ground*, *vertical* and *sky*. This segmentation is done by using several features, including texture, shape and color information in combination with geometrical cues. A classifier is trained using AdaBoost, based on weak decision tree classifiers, from a large labeled database. Using these geometrical context classes as a prior, Hoiem *et al.* extended classical object detection into 3D space by calculating a coarse viewpoint prior [18]. The knowledge of the viewpoint limits the search space for object detection, e.g. cars should not appear above the horizon. In addition, 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 (alongside 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 [19] classifies an image into four pre-defined semantic classes. These classes indicate the presence of buildings, roads, skies, and trees, which are identified using their texture properties. These classes are learned from different sets of *standard model features* (also known as HMAX [20]). Bileschi then defines the context by using low-level visual features from the Blobworld system [21] (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 relative 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 is also confirmed in [22]. 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, due to the input image data being biased. In Bileschi's image database for example, cars are more likely to be in the lower half of the image. This is because the horizon is in the center of each image. 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 positioned 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 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

In this Section we describe all necessary steps to achieve context aware object detection. We start with the mathematical formulation (Sec. 3.1). We then state what the contextual features are and how they are calculated (Sec. 3.2). Next, we show how these features can form feature vectors, describing the surrounding area of a given position in the image. These feature vectors are used for learning in the training step (Sec. 3.3) and for calculating the final focus of attention in the testing step (Sec. 3.4). To show that our concept is very general we also add an additional contextual cue, the viewpoint prior, derived from a horizon estimate (Sec. 3.5). An illustration of our approach is given in Fig. 3.

3.1 Mathematical Formulation[1]

In general, the problem of object detection requires evaluation of the conditional probability density function (pdf) $p(\mathbf{v}|\mathbf{x}, o)$. This is the probability of observing

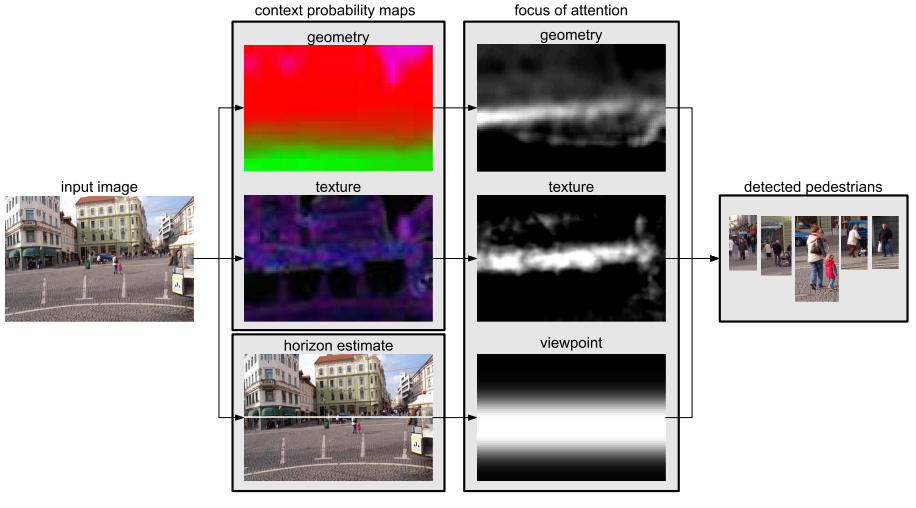


Fig. 3. Concept of context driven focus of attention for object detection on the example of pedestrian detection. Context probability maps based on geometrical and textural features are extracted and are used to calculate a focus of attention. In addition, a horizon estimate contributes a viewpoint prior. The local appearance-based object detector searches one of the individual focus of attention areas or a fusion of individual focus of attention areas, instead of searching the whole image. Best viewed in color.

[1] The mathematical formulations are adopted from [14].

a set of features (image measurements) $\mathbf{v}$, given an object o at a fixed location $\mathbf{x}$. It is also called the *likelihood* of an object o at location $\mathbf{x}$ given the observed set of features $\mathbf{v}$. In order to reduce the dimensionality of the vector $\mathbf{v}$, only a local neighborhood is used to calculate the object's presence (e.g. [23]). The local pdf $p(\mathbf{v}_L|\mathbf{x}, o)$ where $\mathbf{v}_L$ is a set of local image measurements formalizes the main principle of classical object detection. It states that the only features relevant for detection of an object are the features that potentially belong to the object and not to the background. In this case $\mathbf{v} \simeq \mathbf{v}_L$ and $p(\mathbf{v}|\mathbf{x}, o) \simeq p(\mathbf{v}_L|\mathbf{x}, o)$ holds. To model the contextual information, which provides the relationship between the background and the objects, we formulate a second pdf $p(\mathbf{v}_C|\mathbf{x}, o)$ based on contextual features $\mathbf{v}_C$ only. The whole feature set $\mathbf{v} = \{\mathbf{v}_L, \mathbf{v}_C\}$ is split into local and contextual features. They are extracted from complementary spatial locations in the image and are assumed to be independent. The joint conditional pdf of $\mathbf{v}_L$ and $\mathbf{v}_C$ is defined as

$$p(\mathbf{v}_L, \mathbf{v}_C|\mathbf{x}, o) = p(\mathbf{v}_L|\mathbf{v}_C, \mathbf{x}, o) \cdot p(\mathbf{v}_C|\mathbf{x}, o)\,. \tag{1}$$

Since $\mathbf{v}_L$ is assumed to be independent of $\mathbf{v}_C$, given $\mathbf{x}$ and o,

$$p(\mathbf{v}_L|\mathbf{v}_C, \mathbf{x}, o) = p(\mathbf{v}_L|\mathbf{x}, o) \tag{2}$$

and therefore the joint pdf further simplifies to

$$p(\mathbf{v}|\mathbf{x}, o) = p(\mathbf{v}_L|\mathbf{x}, o) \cdot p(\mathbf{v_C}|\mathbf{x}, o) = p_L \cdot p_C\,. \tag{3}$$

In the remainder of the paper we abbreviate $p(\mathbf{v}_L|\mathbf{x}, o)$ as p_L and $p(\mathbf{v}_C|\mathbf{x}, o)$ as p_C. From this formulation we see that the pdf based on local appearance p_L can be independently calculated from the pdf based on contextual information p_C and fused by multiplication in the final step.

In general the pdf p_L based on local appearance only provides local evidence and should have narrow maxima providing high confidence. The evaluation of this function requires exhaustive spatial and multiscale search and is therefore computationally expensive.

The function p_C using contextual information provides image regions where the presence of the object of interest is likely (it is assumed that this function is smooth and will not have narrow peaks). Therefore, it acts as a *focus of attention* or *prior* for exhaustive local search.

3.2 Extracting Contextual Features

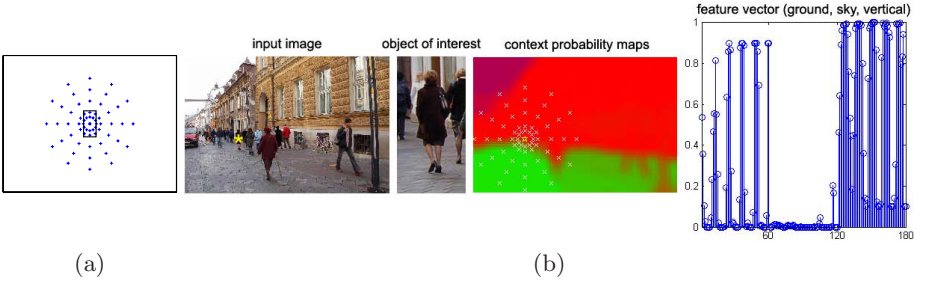
We assume that contextual information can be stored in probability maps, which are images themselves. These maps are stored in a lower resolution compared to the input image since the contextual information is aggregated over a large part of the image. The maps can encode high-level semantic features or low-level image features. Examples for semantic maps could be vegetation, sky, cars, etc., whereas low-level features could be information of gradients, texture descriptors, shape descriptors, color descriptors, etc. In this work, two complementary types of features are used to form contextual information: geometrical features and texture features.

Geometrical Features: The employed context probability maps are the three semantic classes from Hoiem's approach [17] which give the probabilities that the current pixel belongs to the ground, the vertical class (buildings, trees, etc.) or the sky. Therefore, the contextual features consist of a three layer image holding the probabilities of the three semantic classes. An example is shown in Fig. 3 where these three layers are color coded (ground (green), vertical (red), sky (blue)).

Texture Features: For describing texture, three features proposed within the Blobworld system [21] are used, which capture information about the local structure and the gradient magnitude. They are polarity, anisotropy and texture contrast, which measure the likelihood of the local gradient to switch direction, the relative strength of the gradient in orthogonal directions and the roughness of the region. These features are extracted from the second moment matrix (also called structure tensor) over a multiscale search. This matrix could be calculated in a very efficient way and is well know in the field of *point of interest detection* (e.g. [24]). An example is shown in Fig. 3 where these three layers are color coded (anisotropy (red), polarity (green), texture contrast (blue)).

3.3 Extracting and Learning Contextual Features Vectors

The next step is to extract feature vectors from the previously calculated context feature maps for a given position in the image. In the training step, positive feature vectors are collected, describing the context of the objects of interest. Therefore, bounding boxes of example objects are used that are typically manually selected and establish the ground truth. A feature vector is extracted by sampling the data of the maps relative to the object centers for a certain number of radii and orientations (as is done in [19]). Fig. 4(a) illustrates this concept,



(a) (b)

Fig. 4. (a) An illustration of relative sampling positions (for 5 radii and 12 orientations), plotted as '+' signs, relative to the object center marked as a star. The thick black rectangle represents the average size of pedestrians. (b) 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. Best viewed in color.

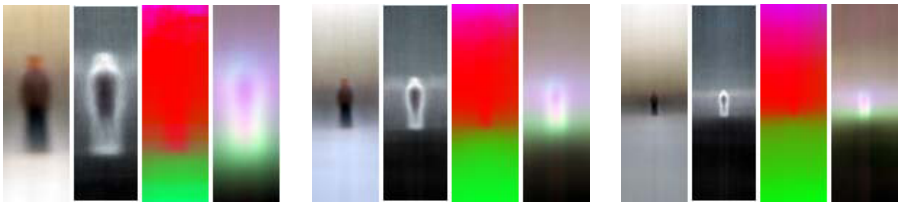
where the probability maps are sampled at 60 positions (5 radii and 12 orientations) relative to the objects' centers. These 60 values are stacked into a single feature vector for each layer of the probability maps. These vectors are then concatenated and form the final feature vector. This vector is a sparse representation of the contextual information surrounding the current object of interest. The receptive field, that is, those pixels in the image which influence the feature, is chosen to be quite large, so as to capture a more global context. This yields a low-dimensional context representation, i.e., a 180-dimensional vector in the used implementation per object. In comparison, the object detector based on local appearance in [25] uses a 3780-dimensional feature vector.

Such a contextual feature vector can be extracted for each object in a training set. These positive feature vectors together with negative feature vectors are passed to a strongly supervised learning algorithm, e.g. 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 model should be capable of discriminating between realistic or unrealistic context for object detection. Fig. 4 illustrates the workflow of how to extract a contextual feature vector from an image, where one quarter of the image information is used to describe the contextual information.

3.4 Using Learned Contextual Model

To extract the focus of attention for each position in the image the corresponding contextual feature vector is extracted and supplied to the learned contextual model. In this way, for each position in the image a context confidence score is calculated. Typically, the output of the machine learning algorithm is not probabilistic so it has to be normalized to the domain $[0, 1]$. Then the output is a probabilistic context confidence map, which gives the likelihood of the presence of the object at this spatial position (see Fig. 3). Very unlikely positions could be rejected by applying a threshold on the context confidence maps. In our implementation two different thresholds are used. A *conservative* one that rejects only objects with very low context confidence scores and a *liberal* one rejecting objects in a less restrictive way. This modified version of the context confidence map p_C is called *the focus of attention*.

To demonstrate this concept, we visualize the average object of interest, the average magnitude, the average geometrical context and the average textural context in Fig. 5 for different scales, where we choose pedestrians as objects of interest. The employed context probability maps are the ones described in Sec. 3.2. The three geometrical and textural features are color coded (ground (green), vertical (red), sky (blue); anisotropy (red), polarity (green) and texture contrast (blue)). It is obvious that the average pedestrian's contextual arrangement is well defined. Pedestrians are standing on the ground; the body is in the vertical context class and is not located in the sky; the pedestrians themselves are highly textured; areas above are strongly textured and areas below are more or less homogeneous. Since this is not a random configuration, it could be learned, given positive and negative examples.



(a) (b) (c)

Fig. 5. An average pedestrian out of 3803 manually selected pedestrians from Ljubljana urban image data set. The average pedestrian, the magnitude image, the average geometrical context and the average textural context are shown for three different scales (a) 1.0 (b) 2.0 (c) 4.0. Best viewed in color.

3.5 Additional Contextual Priors

To demonstrate the generality of the proposed approach we also add an additional contextual cue, the viewpoint prior, derived from a horizon estimate. We calculate the position of the horizon in the image, which is described by the tilt angle of the camera's orientation (this is feasible if the image was captured without a significant in-plane rotation). In the current implementation we relate the image of interest by means of wide baseline stereo matching to geo-referenced reference images and calculate the camera's orientation via triangulation[2]. We use algorithms similar to the one proposed in [26]. Using this horizon estimate a Gaussian-like pdf is calculated (see Fig. 3). In contrast to the geometrical and texture feature, this pdf is modeled for the specific object category. In general, any pdf can be added into our framework.

4 Implementation Details

To test our approach, the proposed framework was implemented and details are given in this Section. We applied pedestrian detectors from Seemann *et al.* [27] and Dalal and Triggs [25] to get the local pdf p_L. Both detectors are shape-based methods, where the former uses implicit shape models and the latter uses histograms of gradients. Since only the shape is used for object detection, these approaches are not limited to pedestrian detection. All classes of objects that share a common and unique shape can be detected (e.g. cars, when front and side view is trained). For the first detector, the results were provided directly by Seemann after we shared our database. For the second detector, we used Dalal's

[2] Calculating a viewpoint prior in the proposed way is only feasible if this information is also extracted for other reasons, as is the case in the *Vision Technologies and Intelligent Maps for Mobile Attentive Interfaces in Urban Scenarios* project. At this point we just want to show that this kind of contextual information is useful and easy to plug into the proposed framework. For the sake of completeness we point out that the position of the horizon can also be estimated from a single image [18].

binaries[3] and set the parameters to the values proposed in [18]. The function p_L was evaluated only on the focus of attention p_C, which was extracted beforehand. To compare the detections to the ground truth, the criteria in [28] are implemented. They are *relative distance*, *cover* and *overlap*. A detection is considered correct, if the relative distance is less than 0.5 and cover and overlap are both above 50%. In our case, we only require 30% cover and overlap, due to the large variation in the pedestrians' width. Note that even though the detector by Dalal and Triggs is one of the best state-of-the-art pedestrian detectors, about 30% of its detections are incorrect (false positives), which was determined by using our Ljubljana urban image data set.

4.1 Extracting Contextual Features

For extracting the geometrical context probability maps, we used the publicly available executable by Hoiem [17][4]. The resulting context 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 [19] and is meant as a tradeoff between capturing the gross contextual information and being computationally efficient. The extraction of this geometrical context is rather time consuming as it is calculated out of the 1504×1000 pixel input images. The images could not be downsampled before the geometric context extraction since the resolution influences the results of Hoiem's algorithm.

The textural context features are calculated with the approach in [21], using the publicly available source code[5]. In this case we only calculate the textural context on a downsampled version of the images with a width of 80 pixels. The main idea is that at such low resolution the facades or vegetation are still highly textured whereas the ground plane looses their texture. In addition, calculating the features on these small downsampled images is computationally very 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 for geometrical and for textural context (see Fig. 4). 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 combination we use gives the best results. Note that one quarter of the input image contributes to the context feature extraction for each object.

4.2 Learning Contextual Feature Vectors

Using the pedestrians' ground truth and the evaluation criteria, half of the correctly detected pedestrians are used for training, the other half for testing. The initial set is split into these two sets via a random selection, to avoid a possible bias. To be able to learn this context, 5000 negative examples are drawn from

[3] http://pascal.inrialpes.fr/soft/olt/
[4] http://www.cs.cmu.edu/~dhoiem/projects/software.html
[5] http://elib.cs.berkeley.edu/src/blobworld/

Table 1. Thresholds used to calculate the focus of attention from the corresponding context probability maps

	geometry	texture	viewpoint
conservative	0.20	0.40	0.10
liberal	0.25	0.99	0.20

images containing no pedestrians, at random positions. Having positive and negative examples, a linear support vector machine (SVM) is trained. To verify the robustness of the SVM learning we use a cross-validation approach. Positive and negative examples are divided randomly into two sets, where the first is used for training and the second one for evaluation. In this test, the classification rate is very stable for geometrical and for textural context features (changes less than 1% over 100 iterations).

4.3 Using Learned Contextual Model

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. Since the SVM output is not probabilistic it is converted into a probability score. This is done by first zero-meaning the data, then setting the standard deviation to 3, clipping values below -1 and above $+1$ and finally scaling the data to $[0, 1]$. The basic reason for the clipping is to remove outliers. We also tested the approach from [29] to map SVM outputs to probabilities. However, this approach uses non robust fitting and is therefore not stable in our case. The two thresholds that are used to reject very unlikely areas in the context confidence maps are empirically set to the values listed in Table 1.

4.4 Viewpoint Prior

From a rough estimate of the horizon in the image, a viewpoint prior is determined. This prior is a probability density function and should hold for an average image. It is assumed that the standard deviation of the horizon estimate is known in degrees (e.g. by comparing the estimates to a ground truth). The angle of pedestrian occurrence with respect to the horizon is empirically extracted from the database (7° for Ljubljana urban image data set). In our setup, this 7° corresponds to 12 meters, which is the average distance from the camera's center to the pedestrians. To be able to convert the angles to image coordinates the field of view of the camera has be known ($88° \times 52°$ in our case). The focal length is used to convert the values into meters. We use information stored in the JPEG's EXIF header. The modeled pdf is a Gaussian with a standard deviation of the horizon estimate uncertainty. In addition, a plateau is put into the Gaussian's peak with the size of the uncertainty (compare with Fig. 6). It is assumed that the horizon is aligned to the horizontal axis of the image and therefore this 1D pdf is replicated over this axis.

The proposed viewpoint prior estimate is quite constrained and modeled for the detection of pedestrians only. This contextual information is not as general as the geometrical and textural context. Depending on the application modeling such a prior may or may not be feasible.

5 Experimental Results

Two sets of experiments were conducted to demonstrate and evaluate the approach. First, we show examples of context driven focus of attention in terms of images and in terms of a quantitative evaluation of the achieved speedup. Second, a performance comparison is given showing the increase of the detection rate.

5.1 Database

The novel concept described in the previous sections is tested on the Ljubljana urban image data set[6], comprised of 612 images of Ljubljana's center with a resolution of 3008×2000 pixels each. For our experiments, the images are downsampled to 1504×1000 pixels. The images are captured at three distinct tilt angles (therefore the horizon is in general not in the center of the images). This is a very challenging aspect, since methods like [19] will most likely fail, as they assume the horizon to be in the center of the image. To be able to compare the different pedestrian detection results to a ground truth, all 3803 pedestrians in the image database were manually labeled. Each pedestrian is defined by the corresponding bounding box, where the whole object is inside the box. Therefore, the deviation in width is quite significant since it varies with the pose of the pedestrian.

5.2 Speedup

Examples of the performance of the different contextual cues for context driven focus of attention estimation are given in Fig. 6. Shown are the input image I, the function p_C and for better visualization the image multiplied by this function $I \cdot p_C$ for context based on geometry, texture and viewpoint respectively. In general, the contextual cue based on geometry is able to reject unlikely position in the sky and on the ground plane. However, it is not able to reject regions of building facades. Using texture information as contextual cue, the focus of attention is set to positions where the objects of interest occur. Viewpoint also limits the search space but in a less constrained way than using texture.

Table 2 gives the average percentage of pixels of the test images where the function p_L based on local appearance has to be evaluated, after applying the *focus of attention* concept. Depending on the threshold (conservative versus liberal) the focus of attention is wider or smaller. In the case of setting the focus of attention in a liberal way and using textural context on average only 14% of the images has to be searched for pedestrians exhaustively. That yields a speedup of

[6] http://vicos.fri.uni-lj.si/LUIS34/

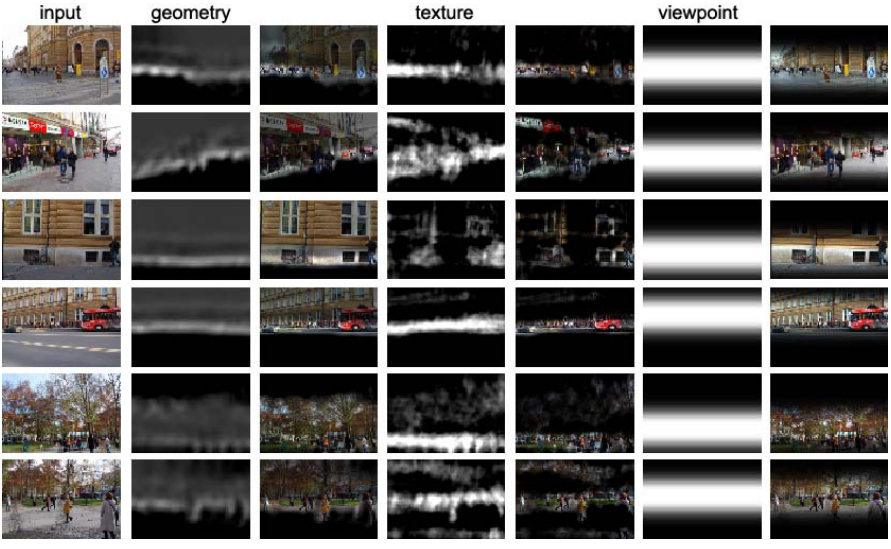


Fig. 6. Focus of attention for different contextual cues. Shown are the input image, the focus of attention and the image multiplied by the focus of attention, for context based on geometry, texture and viewpoint, respectively, for pedestrian detection.

Table 2. Average percentage of pixels in the images that are covered by the focus of attention. Note that, speedup using liberal filtering of the textural context estimation is 7.1.

	geometry		texture		viewpoint	
	conservative	liberal	conservative	liberal	conservative	liberal
Seemann	71%	53%	46%	14%	80%	68%
Dalal	68%	43%	45%	14%	80%	68%

7.1 in evaluating the function p_L. Using the other cues and setting the threshold in a conservative way, the speedup is less significant (maximal factor 2.0), since the geometric contextual cue and the viewpoint prior are less restrictive than context provided by texture.

5.3 Increase of the Detection Rate

When searching for objects in the focus of attention area only, it is obvious that some true detections may be missed, while we expect to discard many false detections. Therefore, we compare pedestrian detections based on local appearance alone (p_L) with the detections in the focus of attention ($p_C \cdot p_L$). In Table 3 the percentage of true positives (TP) and false positives (FP) are given, that were detected in the focus of attention area w.r.t. the detection on the whole image (not using the focus of attention). In general, the loss of TPs is small in comparison to the not detected FPs.

Table 3. Comparison of true positives and false positives in the focus of attention area w.r.t. the detections on the whole image (not using the focus of attention). The loss of TPs is small in comparison to the not detected FPs.

	TPs and FPs in % using conservative threshold	TPs and FPs in % using liberal threshold
Seemann: geometry	98.8 TP / 87.3 FP	98.6 TP / 85.4 FP
texture	96.5 TP / 47.9 FP	84.5 TP / 19.3 FP
viewpoint	99.2 TP / 78.6 FP	97.5 TP / 75.0 FP
all	94.4 TP / 39.2 FP	81.8 TP / 15.9 FP
Dalal: geometry	97.6 TP / 80.5 FP	96.5 TP / 77.8 FP
texture	96.8 TP / 51.8 FP	85.0 TP / 24.9 FP
viewpoint	98.7 TP / 79.3 FP	98.0 TP / 75.9 FP
all	93.8 TP / 41.5 FP	81.7 TP / 20.3 FP

To give more insights on how the individual contextual cues aid to the object detector's performance, the detection rate is plotted versus the false positives per image (FPPI) in Fig. 7. The local and contextual scores are fused as defined in Eq. (3). Also the contextual cues were fused by multiplying the individual functions, assuming that the features are statistically independent:

$$p_{C_{all}} = p_{C_{geometry}} \cdot p_{C_{texture}} \cdot p_{C_{viewpoint}} \tag{4}$$

These curves reveal that viewpoint is the weakest contextual cue, followed by geometry and texture, which is the strongest cue to aid local detection. The fusion of all three contextual cues with the local appearance-based method provides the best detection rate. In Table 4 the detection rates are shown for three fixed FPPI for simple comparison between the different pedestrian detectors. Finally, the same detection rate plot is given for Seemann's detector once applied on the whole image and once only applied to the focus of attention area obtained by

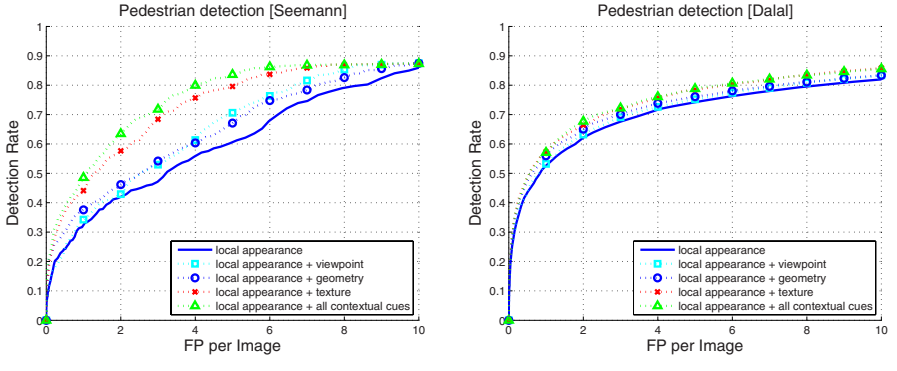


Fig. 7. Detection rate curves using local appearance in combination with contextual cues are shown for two different object detectors

Table 4. Detection rate at fixed false positives per image (FPPI). Contextual cues are sorted by the benefit they can provide to object detection.

	Seemann					Dalal				
	local	view-point	geo-metry	texture	all	local	view-point	geo-metry	texture	all
1 FPPI	32%	34%	38%	44%	49%	53%	53%	56%	57%	57%
2 FPPI	42%	43%	46%	58%	64%	62%	63%	65%	67%	68%
5 FPPI	61%	67%	71%	80%	84%	74%	75%	76%	79%	79%

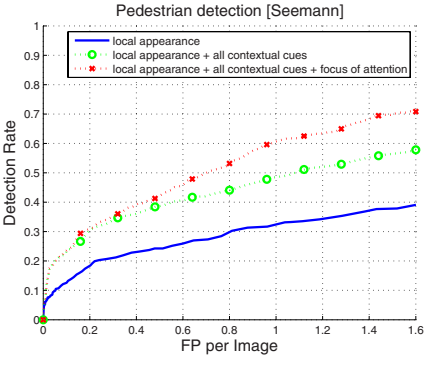


Fig. 8. Detection rate curves using local appearance in combination with all contextual cues using Seemann's pedestrian detector. Shown are the initial detection curve, the detection curve using all contextual cues, and the detection curve using all contextual cues with the focus of attention obtained by applying the liberal thresholds.

liberal thresholding in Fig. 8. Notice the different scaling of this plot in comparison to Fig. 7. When applying the detector to the focus of attention only, 15.9% of FPs remain (compare with Table 3). This shows that using the thresholded focus of attention estimation yields better results than the pure combination of p_L and p_C.

5.4 Discussion of Results

From the evaluation it is clear that context significantly aids object detection in general. First, the performance of both local appearance detectors was boosted by incorporating them into the proposed framework. The achieved benefit of using contextual information depends on the initial object detector's performance. Logically, context would not increase the accuracy of a detector if it would yield optimal detections anyway. However, limiting the search space will speedup any detector. Another interesting aspect is, that when boosting both detectors by context, the initially inferior detector by Seemann outperforms Dalal's detector at FFPI bigger than 3.

It also has turned out that the low-level visual cue based on very simple texture descriptors outperforms the high-level contextual cues based on semantic classification or on the viewpoint prior. At first glance this is quite surprising. Taking a closer look we see, that we implicitly encode the concept of figure/ground organization [30]. Namely, in the case of urban images, *figure* corresponds to regions behind the object (building, trees, etc.), which are highly textured objects. *Ground* corresponds to regions in front the objects, which are more or less homogenous (especially when analyzing these regions in a low resolution). In conclusion, we state that the textural cue is so strong that it outperforms the other ones and could therefore be used as the single contextual cue. The texture descriptors can also be calculated very efficiently. However, this evaluation was based on images containing urban scenes only. We cannot predict how the presented cues will perform when using a different set of images. But we are confident that our approach is general enough to handle other sets, as e.g. also park scenes provided good foci of attention (see Fig. 6, last two examples).

Finally, using the proposed concept with textural context and liberal threshold, pedestrian detection is speeded up by a factor of 7.1, while reducing the FP rate by 80% and only sacrificing 16% of TPs.

6 Conclusion

Context plays an important role in general scene perception and provides cues of information about an object's location within an image. However, object detectors typically ignore this information. We tackle this problem by presenting a concept of how to extract and learn contextual information from examples. This context was then used to calculate the focus of attention, that presents a prior for object detection. State-of-the-art local appearance-based object detection methods have been applied on selected parts of the image only. We demonstrated the performance on the task of pedestrian detection in urban scenes using a demanding image database. Results showed that context awareness provides complementary information over pure local appearance-based processing in object detection. In addition, the search complexity was decreased while the robustness of object detection was increased.

Acknowledgements. 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. The authors would like to thank Derek Hoiem and Navneet Dalal for providing the binaries of their algorithms. We would especially like to thank Edgar Seeman for providing us with the results of his algorithm run on our database. We would also like to acknowledge Dušan Omerčević, Matej Kristan, Barry Ridge, Jurij Šorli, Luka Fürst and Matjaž Jogan for useful comments and discussions on this manuscript.

References

1. Palmer, S.E.: The effects of contextual scenes on the identification of objects. Memory and Cognition 3, 519–526 (1975)
2. Biederman, I.: Perceptual Organization. In: On the semantics of a glance at a scene, ch. 8, pp. 213–263. Lawrence Erlbaum, Mahwah, NJ (1981)
3. Bar, M.: Visual objects in context. Nature Reviews, Neuroscience 5, 617–629 (2004)
4. Fei-Fei, L., Fergus, R., Torralba, A.: Recognizing and learning object categories. Tutorial presented at CVPR (2007), `http://people.csail.mit.edu/torralba/shortCourseRLOC`
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR (3), 993–1022 (2003)
6. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: CVPR, vol. 2, pp. 524–531 (2005)
7. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: ICCV, vol. 1, pp. 370–377 (2005)
8. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Transactions on Computers 22(1), 67–92 (1973)
9. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJRS 61(1), 55–79 (2005)
10. Friedman, J.H., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. Dept. of Statistics, Stanford University, Technical Report (1998)
11. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing features: Efficient boosting procedures for multiclass object detection. In: CVPR, vol. 2, pp. 762–769 (2004)
12. Perko, R., Leonardis, A.: Context awareness for object detection. In: OAGM/AAPR, pp. 65–72 (2007)
13. Oliva, A., Torralba, A., Guerin-Dugue, A., Herault, J.: Global semantic classification of scenes using power spectrum templates. In: CIR (1999)
14. Torralba, A., Sinha, P.: Statistical context priming for object detection. In: ICCV, vol. 1, pp. 763–770 (2001)
15. Torralba, A.: Contextual priming for object detection. IJCV 53(2), 153–167 (2003)
16. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. Technical Report AIM-2005-025, MIT AI Lab Memo (2005)
17. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: ICCV, vol. 1, pp. 654–661 (2005)
18. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. In: CVPR, vol. 2, pp. 2137–2144 (2006)
19. Bileschi, S.M.: StreetScenes: Towards Scene Understanding in Still Images. PhD thesis, Massachusetts Institute of Technology (2006)
20. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. PAMI 29(3), 411–426 (2007)
21. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. PAMI 24(8), 1026–1038 (2002)
22. Wolf, L., Bileschi, S.M.: A critical view of context. In: CVPR, pp. 251–261 (2006)
23. Schiele, B., Crowley, J.L.: Recognition without correspondence using multidimensional receptive field histograms. IJCV 36(1), 31–50 (2000)
24. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings 4th Alvey Visual Conference, pp. 189–192 (1988)

25. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 2, pp. 886–893 (2005)
26. Zhang, W., Košecká, J.: Image based localization in urban environments. In: 3DPVT, pp. 33–40 (2006)
27. Seemann, E., Leibe, B., Schiele, B.: Multi-aspect detection of articulated objects. In: CVPR, vol. 2, pp. 1582–1588 (2006)
28. Leibe, B.: Interleaved Object Categorization and Segmentation. PhD thesis, ETH Zurich, PhD Thesis No. 15752 (2004)
29. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers 10(3), 61–74 (1999)
30. Ren, X., Fowlkes, C., Malik, J.: Figure/ground assignment in natural images. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 614–627. Springer, Heidelberg (2006)