Computer Vision and Image Understanding 114 (2010) 700–711

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A framework for visual-context-aware object detection in still images

Roland Perko *, Aleš Leonardis

Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1001 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 9 September 2008
Accepted 22 March 2010
Available online 27 March 2010

Keywords:
Visual context
Object detection
Context integration

ABSTRACT

Visual context provides cues about an object's presence, position and size within the observed scene, which should be used to increase the performance of object detection techniques. However, in computer vision, object detectors typically ignore this information. We therefore present a framework for visual-context-aware object detection. Methods for extracting visual contextual information from still images are proposed, which are then used to calculate a prior for object detection. The concept is based on a sparse coding of contextual features, which are based on geometry and texture. In addition, bottom-up saliency and object co-occurrences are exploited, to define auxiliary visual context. To integrate the individual contextual cues with a local appearance-based object detector, a fully probabilistic framework is established. In contrast to other methods, our integration is based on modeling the underlying conditional probabilities between the different cues, which is done via kernel density estimation. This integration is a crucial part of the framework which is demonstrated within the detailed evaluation. Our method is evaluated using a novel demanding image data set and compared to a state-of-the-art method for context-aware object detection. An in-depth analysis is given discussing the contributions of the individual contextual cues and the limitations of visual context for object detection.

1. Introduction

Objects tend to co-vary with other objects and particular environments, providing a rich collection of contextual associations [29]. It is well known from the literature on visual cognition [31,10,4] and cognitive neuroscience [2,12,1], that the human and animal visual systems use these relationships to improve their ability of categorization. Consequently, context should be used in computer vision as well and can help object detection, as shown in [45,5,18,46,34,29].

In this paper we introduce a framework that uses two different types of visual contextual information to improve object detection. The first type is the spatial relation between an object and its surrounding. It is determined by exploring the visual content of a given scene. The second type are spatial relations between a specific object and other objects in the scene where the context is represented by spatial object co-occurrence (also called object-to-object priming in [16]). Both concepts are visualized in Fig. 1. In the first case semantic information could be extracted from images and exploited to find out if the object is in context, i.e. coherent w.r.t. the scene. In the second case, relative location priors could help to distinguish between correct and incorrect detections, e.g. in urban scenes pedestrians should be more or less at the same height or windows should occur above pedestrians. Overall, we employ the basic ideas as Biederman [3] when he stated that "*object identification is facilitated when an object is presented in a coherent scene*".

1.1. Our contributions

In this work we present a complete framework for visual-context-aware object detection. We answer two major questions which are essential for such a system, i.e. how to represent visual context and how to combine this information with object detection.

First, we propose methods of how to extract visual contextual information from single images and how this information can be learned from examples. For doing so, we utilize a method for sparse coding of contextual features using a spatial sampling technique. Appropriate image features based on geometry and texture are discussed as well. As a result a prior for object detection can be extracted. In addition, we define object co-occurrences and bottom-up saliency as further contextual cues.

Second, we introduce a concept to integrate the contextual information with a local appearance-based object detector. Our mathematical framework and the specific modeling of the conditional probability density functions used for integrating visual context with object detection is a crucial part of this work. As seen later, this modeling contributes significantly to the success of our method.

* Corresponding author. Fax: +386 1 4264 647.
E-mail address: roland.perko@fri.uni-lj.si (R. Perko).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.03.005

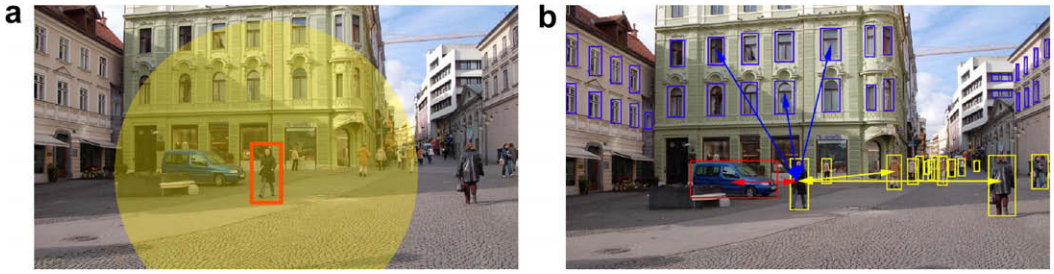

Fig. 1. Definitions of visual contextual information: spatial relations (a) between an object and its surrounding based on the visual content of the scene (e.g. the image content of the given yellow circular region) and (b) between a specific object and other objects in the scene (shown for the object categories pedestrians, cars and windows). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In contrast to other works the proposed integration technique avoids two common mistakes in literature: (i) Researchers assume statistically independence of local-appearance and visual context [47,45,5]. We model these dependency accordingly. (ii) Researches use the output of an object detector as a probability measure [47]. We do not rely on a detector's output, instead we model the probabilities correctly.

For modeling the underlying multi-dimensional probability density functions, we propose to use a *kernel density estimation (KDE)*.

To evaluate the system's performance, we chose the task of pedestrian detection in urban images using a state-of-the-art pedestrian detector on a demanding image database. The evaluation shows that our definition and integration of visual context increases the initial local appearance-based detection rate significantly and outperforms other frameworks for contextual processing. Using the gained insights we discuss the results of visual-context-aware object detection and show its limitations.

1.2. Organization of the paper

Related work will be discussed in detail in Section 2. After that our framework of context-aware object detection and implementation details are described in Section 3, followed by an in-depth analysis of results in Section 4. In Section 5 we discuss the limitations of contextual processing and conclude the paper with Section 6.

2. Related work

The influential work from Oliva and Torralba, e.g. [28,47,44,45], introduced a novel global image representation. An image is decomposed by a bank of multi-scale oriented filters, in particular four scales and eight orientation. The magnitude of each filter is averaged over 16 non-overlapping blocks in a 4×4 grid. The resulting image representation is a 512-dimensional feature vector, which is represented by the first 80 principal components. Despite the low dimensionality of this representation, it preserves most relevant information and is used for scene categorization, such as a landscape or an urban environment. Machine learning provides the relationship between the global scene representation and the typical locations of objects belonging to that category. To the best of our knowledge there exist no evaluation for the combination of this derived context priors with a state-of-the-art object detection algorithm. In a real scenario a coarse prior for the possible object location in the image does not automatically increase the performance of an object detector. As will be seen later, when combined just by multiplication, the results of the object detection may and often do degrade.

Hoiem et al. [17] provided a method to extract the spatial context of a single image. The image is first segmented into so called superpixels, i.e. a set of pixels that have similar properties. These regions are then described by low level image features, i.e. color, texture, shape and geometry, forming a feature vector. Each region is classified into a semantic class, namely *ground*, *vertical structures* and *sky*, using a classifier based on AdaBoost with weak decision tree classifiers. As a result each pixel in the input image is associated with the probabilities of belonging to these three classes. For the task of object detection this classification provides useful cues and they are exploited in [18,34]. Hoiem et al. [18] use the coarse scene geometry to calculate a viewpoint prior and therefore the location of the horizon in the image. The horizon, being the line where the ground plane and the sky intersect in infinity, provides information about the location and sizes of objects on the ground plane, e.g. pedestrians or cars. The scene geometry itself limits the location of objects on the ground plane, e.g. no cars behind the facade of a building. The innovative part of their work is the combination of the contextual information with the object hypotheses using inference. They construct a graphical model of conditional independence for viewpoint, object identities and 3D geometry of surfaces surrounding the objects. The inference is solved using Pearl's belief propagation algorithm [33]. Overall, the main idea is to find the object hypotheses that are consistent in terms of size and location, given the geometry and horizon of the scene. As a result, a cluster of object hypotheses is determined, that fits the data best. This contextual inference uses the global visual context and the relation between objects in that scene. The position of the horizon is an integral part of this system, limiting the approach to object categories that are placed on the ground plane and to objects of approximately the same size. E.g. the approach cannot be used to detect windows on facades or trees.

Bileschi [5] 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 [43]). Bileschi then defines the context using low-level visual features from the *Blobworld* system [6] (three color and three texture-based features). In addition ten absolute image positions are encoded followed by four binary sematic features, representing the four extracted classes (building, road, sky and tree). To extract a context vector for a given position in the image, the data is sampled relative to the object's 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 [53]. A more interesting outcome of the extensive studies by Bileschi is that using global position features (also applied 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, because the horizon is in the center of each image.

There are a couple of additional works on visual context, however, they are only sparsely related to our work. For example the

authors in [24] use local context to improve face detection. The main idea is that in the local neighborhood of a detected face also other body parts, e.g. shoulders, should exist. With this local context they are able to eliminate phantom faces. Or in [15] the spatial and topological relationships are exploited, which are learned via Non-Gibbsian Markov random field models. Manually segmented objects in images can then automatically be labeled using these models.

3. Our approach

An overview of the proposed framework is sketched in Fig. 2. The underlying concept is to derive context priors, from contextual feature maps or other cues, which are then combined with local-appearance based object detection within a fully probabilistic framework.

3.1. Mathematical formulation of object detection and context integration

In general, the problem of object detection in still images requires calculating a confidence score for an image patch given a set of features $\mathbf{v}$ (image measurement). In order to reduce the dimensionality of the vector $\mathbf{v}$, only a local neighborhood is used to calculate the object's presence (see e.g. [41]). This formalizes the main principle of classical object detection, stating that the only features relevant for detection of an object are the features that potentially belong to the object and not to the background [47]. In terms of mathematics an object detector is a function that maps a high-dimensional vector (holding image measurements) into a single value that represents the confidence score, also called the detection score s_L. Depending on the underlying object detection algorithm this score is e.g. determined by a *support vector machine* (SVM) [32,7], by the AdaBoost concept [48,40] or by a voting scheme [42,26]. When working within a probabilistic framework these detection scores have to be mapped to probabilities. This important aspect is neglected by other authors [47,53,34]. The standard way for performing this mapping is to evaluate the detector on a training set where the ground truth for object detection is known. Therefore, all detections can be classified into correct d_t and incorrect detections d_f, with all detections forming the set $d = \{d_t, d_f\}$. Given this ground truth information the characteristics of the object detector's scores s_L can be calculated and then mapped to probabilities. As we do not want to limit the notation to a single dimension we define the score $\mathbf{s}$ as the set $\mathbf{s} = \{s_L\}$. To get the mapping, we first model the conditional *probability density functions* (pdfs) $p(\mathbf{s}|d_t)$ and $p(\mathbf{s}|d_f)$. Second, we model the probability of a detection being correct given the detection score. According to the Bayes' theorem this probability, $p(d_t|\mathbf{s})$, can be written as

$$p(d_t|\mathbf{s}) = \frac{p(\mathbf{s}|d_t)p(d_t)}{p(\mathbf{s})} = \frac{p(\mathbf{s}|d_t)p(d_t)}{p(\mathbf{s}|d_t)p(d_t) + p(\mathbf{s}|d_f)p(d_f)} \tag{1}$$

where $p(d_t)$ and $p(d_f)$ are the prior probabilities of a detection being correct or incorrect respectively with $p(d_t) + p(d_f) = 1$. These prior probabilities are extracted from the training set. Note that, the challenge here is to model the conditional pdfs $p(\mathbf{s}|d_t)$ and $p(\mathbf{s}|d_f)$, to finally get $p(d_t|\mathbf{s})$. So far each score $\mathbf{s}$ is a scalar value so that $p(d_t|\mathbf{s})$ could be directly modeled via e.g. a polynomial fitting [35] or a sigmoid fitting [36].

To incorporate visual contextual information in this probabilistic framework, the contextual extraction is defined similarly as an object detector. Again, image measurements are taken and are used to calculate a confidence value for each location in the image. In contrast to a classical object detector, using a set of local image measurements on the object, the contextual information is gathered from the background or the surrounding of the object. As we will present different modalities of visual context, for each location in the image several contextual scores can be extracted $s_C = \{s_{C_1}, \ldots, s_{C_n}\}$. To fuse the local appearance score s_L with the contextual scores s_C, we construct a combined score vector $\mathbf{s} = \{s_L, s_C\}$. Now, the functions $p(\mathbf{s}|d_t)$ and $p(\mathbf{s}|d_f)$ are multi-dimensional probability density functions and therefore more difficult to model.

Other authors assume that the underlying features used for object detection and context extraction are statistically independent, as they are extracted from non-overlapping regions in the image [47,34]. In this case the conditional pdf of s_L and s_C given a correct detection (the same holds for incorrect detections), defined as

$$p(\mathbf{s}|d_t) = p(s_L, s_C|d_t) = p(s_L|s_C, d_t) \cdot p(s_C|d_t), \tag{2}$$

simplifies to

$$p(\mathbf{s}|d_t) = p(s_L|d_t) \cdot p(s_C|d_t), \tag{3}$$

since under the assumption that s_L is independent of s_C

$$p(s_L|s_C, d_t) \triangleq p(s_L|d_t). \tag{4}$$

Now instead of modeling the dependencies, the conditional pdf can be calculated by multiplication of the individual one-dimensional pdfs (cf. Eq. (3)). However, the assumption of statistical independence is rather vague and, as we will show later, even incorrect.

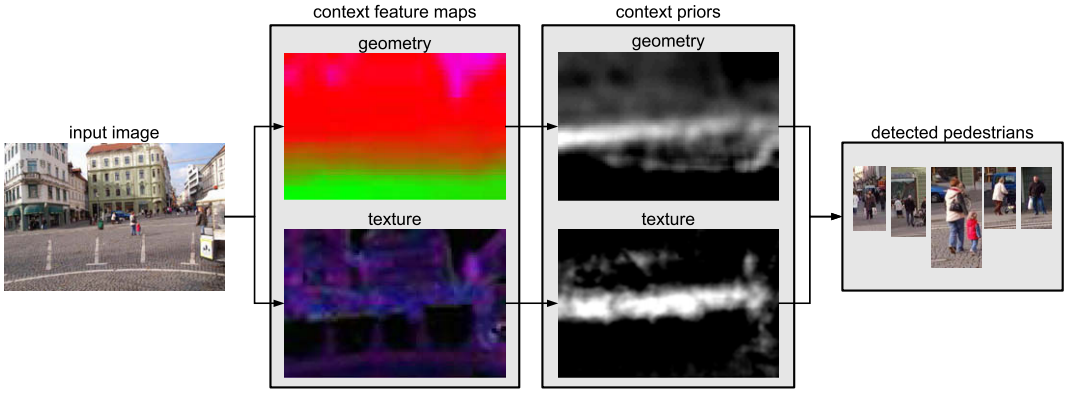

Fig. 2. Concept of context-aware object detection on the example of pedestrian detection. Context feature maps based on geometrical and textural features are extracted and are used to calculate a prior for object detection. Fusing the local appearance-based object detector scores with the contextual priors, ranks detections which are *in context* higher, yielding more accurate object detection.

To model the dependencies of object detection and context given in Eq. (1) adequately we introduce a novel approach based on *kernel density estimation* (KDE) [50,22], used to estimate $p(\mathbf{s}|d_t)$ and $p(\mathbf{s}|d_f)$ and therefore $p(d_t|\mathbf{s})$. The advantage of using KDE modeling in this specific setting is, that the conditional probabilities can also be learned for multi-dimensional input data without any manual modeling or setting of parameters, which is not the case for e.g. multi-dimensional histograms. For performing KDE the Matlab toolbox of Ihler and Mandel was utilized.[1] It should also be mentioned, that although KDE produces good probabilistic models for our kind of data, these models have a high number of components, which affects the computation time for evaluation. However the KDEs could be compressed using *reduced set density estimator* [11] or other techniques [23] which then allows efficient calculation. In particular, we apply a Parzen-window estimator with a Gaussian kernel, which is known to be a powerful tool in approximating distributions even when their form is far from Gaussian [50]. Starting from a set of data points, the Parzen estimator approximates the underlying distribution that generated those points by placing a Gaussian kernel on each data point. The covariances (also known as bandwidths) of the kernels are then adjusted by minimizing the asymptotic mean-integrated-squared error between the unknown distribution and its kernel-based approximation. In our approach we use diagonal covariance matrices and the well-known Silverman's rule-of-thumb rule [50, page 60] to estimate the optimal bandwidth automatically.

It should be noted here, that the main concept is to model the co-dependencies (cf. Eq. (1)). KDE is just used as a mathematical tool and could be replaced by any multi-dimensional modeling technique.

3.2. Visual context extraction

3.2.1. From image features to contextual information

3.2.1.1. Extracting contextual features. We assume that contextual information can be stored in feature maps, which are images themselves. These maps are stored in a lower resolution w.r.t. the input image since the contextual information is collected over a large part of the image. In our specific implementation these maps are stored so that the larger image dimension has 80 pixels. This specific width (respectively height) of 80 pixels was inspired by [6,5,49] and is meant as a tradeoff between capturing the gross contextual information and being computationally efficient. We also tried to reduce these maps to 60 pixels and lower, which results in strong degradation of the results. These maps can encode high-level semantic features or low-level image features. Examples for semantic feature maps could be the result of a semantic image classification (distinguishing classes like vegetation, sky or cars), where each feature map holds the classification score for the given class. Whereas low-level features could be information of e.g. image gradients, texture descriptors, shape descriptors or color descriptors. In this work, two complementary types of features are used to form contextual information. The first are geometrical features and encode geometrical properties of the given scene, i.e. vanishing points, parallel lines, spatial relation of semantic classes, etc. The second are texture features and represent intrinsic local image properties.

Geometrical features. The proposed context feature maps are the three semantic classes from Hoiem's classification approach [17] giving a confidence 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 confidences of the three semantic classes. An example is shown in Fig. 2 where these three layers are color-coded: ground (green)[2], vertical (red), sky (blue).
For extracting the geometrical context feature maps, we used the publicly available executable.[3] The resulting context feature maps are downsampled to a width of 80 pixels if the image is in landscape mode else the height is set to 80 pixels and smoothed with a 5×5 pixel average filter. The smoothing is performed to get rid of classification outliers, like done in [5]. Other filtering techniques, like a median filter, could be used as well.
The extraction of this geometrical context is rather time consuming as it is calculated from about 1.5 megapixel (972×1448 pixel) input images. The images can not be downsampled further before the geometric context extraction since the resolution influences the results of Hoiem's algorithm.

Texture features. For describing texture, three features proposed within the *Blobworld* system [6] are used, capturing information about the local structure and the gradient magnitude. They are polarity, anisotropy and texture contrast, measuring the likelihood of the local gradient to switch direction, the relative strength of the gradients in orthogonal direction and the roughness of the region. These features are extracted from the second moment matrix 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. [14]). An example is shown in Fig. 2 where these three layers are color-coded: anisotropy (red), polarity (green), texture contrast (blue).
The textural context features are calculated using the publicly available source code[4]. In this case we calculate the textural context on a downsampled version of the images with a width (respectively height) of 80 pixels, which is computationally very efficient.

To demonstrate that this kind of image features can assist object detection in urban environments, we visualize the average object of interest, the average gradient magnitude, the average geometrical context and the average textural context in Fig. 3 for different scales. Pedestrians were chosen as objects of interest given the data set in [30]. 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 can be learned. The geometrical context contributes to object detection as our objects of interest are strongly related to it. The texture cue contributes as it implicitly encodes the concept of *figure/ground organization* [38]. That is, in the case of urban scenes, *figure* corresponds to regions behind the object (buildings, trees, etc.), which are highly textured regions. Whereas *ground* corresponds to regions in front of the object, which are more or less homogeneous (grass, street, etc.).

3.2.1.2. 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 order to reduce the dimensionality of the contextual representation we perform a sparse sampling strategy instead of using the whole information encoded in the context feature maps similar to [5]. A feature vector is extracted by sampling the data of the context feature maps relative to the objects' centers for a certain number of radii and orientations. Since we deal with objects of an a-priori

[1] http://www.ics.uci.edu/~ihler/code/.

[2] For interpretation of color in Figs. 1–12, the reader is referred to the web version of this article.

[3] http://www.cs.cmu.edu/~dhoiem/projects/software.html.

[4] http://elib.cs.berkeley.edu/src/blobworld/.

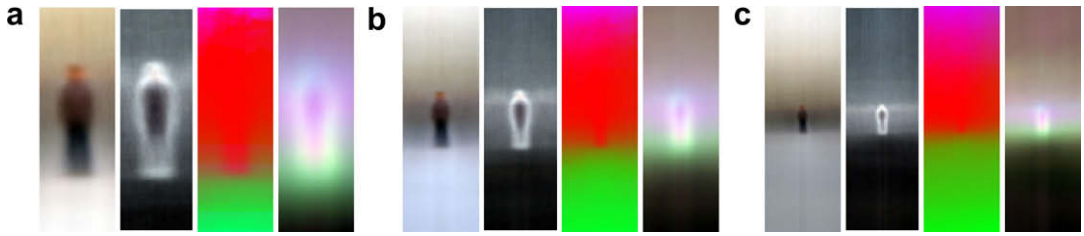

Fig. 3. The average pedestrian out of approximately 4000 manually labeled pedestrians from the data set in [30]. The average pedestrian, the average gradient magnitude image, the average geometrical context and the average textural context are shown for three different scales (a) 1.0, (b) 2.0 and (c) 4.0.

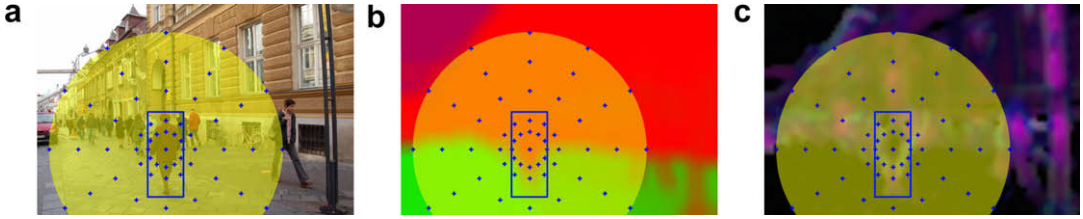

Fig. 4. Contextual feature sampling concept (the sampling locations are shown as crosses): one object is represented by a feature vector holding the contextual information, shown for (a) the input image, (b) the corresponding geometrical features and (c) the textural features.

known range of sizes in the real world, e.g. pedestrians or cars, these fixed sized regions are not representing the same semantic content for objects perceived at different scales. Therefore, these regions should be scaled with the object's size in the image, visualized in Fig. 5a, in contrast to [5,34], where a fixed sized region is used for sampling. In our previous work [35] we discuss the positive influence of the proposed scaling in detail. Actually, this scaling increases the similarity of the contextual feature vectors and yield a 1.3% increase of the detection rate. Fig. 4a illustrates the sampling concept, where the feature maps are sampled at 60 positions (5 radii and 12 orientations). The radii ($r \in [3\%, 5\%, 10\%, 15\%, 20\%]$ of the image diagonal) are scaled w.r.t. the object's height. We 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. Samples beyond the image borders are collected by border replication padding. The replication padding assumes that the semantics *continue* outside the image (e.g. a facade of a building will continue outside the given image). Then, these 60 values are stacked into a single feature vector for each layer of the confidence maps. These vectors are then concatenated and form the final feature vector, sparsely representing the contextual information surrounding the current object of interest. This yields a low-dimensional context representation, i.e. a 180-dimensional vector in the used implementation per object.

Such a contextual feature vector can be extracted for each object in a training set. These positive feature vectors together with negative feature vectors, randomly drawn from images not containing the specific object category, are passed to a supervised learning algorithm, i.e. a linear SVM in our case. The learned model should be capable of discriminating between a realistic or an unrealistic context for the object category of interest. Fig. 4 illustrates the workflow of how to extract a contextual feature vector from an image.

For testing the stability of the SVM modeling, the image data set is split into halves via random selection, where one half of the images is used for training and the other half for testing. Having positive and negative examples, a linear SVM is trained, where we use the SVMLight implementation [21].[5] A cross-validation approach shows, that the classification rate is very stable for geometrical and for textural context features (changes less than 1% over 100 iterations, where for each iteration the splitting into training and test set is performed randomly). This is a nice proof-of-concept for the generality of the extracted contextual information.

3.2.1.3. Using learned contextual model. To extract a context prior for a given image for each image position the corresponding contextual feature vector is extracted and supplied to the learned contextual model, i.e. a trained SVM in our case. As result this prior gives the likelihood of the presence of an object at this spatial location (see Fig. 2). Typically, the output of the machine learning algorithm is not probabilistic and spans over an arbitrary range of numbers. In order to make the fusion of different contextual scores simpler, we map the SVM outputs to the domain [0,1] using robust statistics. This is done by zero-mean normalization of the training 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. The specific parameters are tuned for the used SVMLight implementation [21] and have to be set accordingly for other machine learning techniques. We also tested the approach from [36] to map SVM outputs to probabilities. However, this approach uses non-robust fitting and therefore is not stable in our case.

Our method also enables the extraction of several context priors for predefined object sizes by scaling the radii in the sampling stage. An example of such priors is given in Fig. 5 for six scales $s \in [4, 2, 1.5, 1, 0.5, 0.25]$. The final context confidence score for a given detection can then be calculated by linear interpolation using the scores of the two adjacent scales. E.g. an object of size $s = 2.5$ gets the interpolated score $\mathrm{conf}_{2.5} = 0.25\,\mathrm{conf}_{4.00} + 0.75\,\mathrm{conf}_{2.00}$. As a result the prior based on context is not only able to provide regions where an object is likely to occur, it also gives the possible size of the object. In Fig. 5b the most likely locations for pedestrians for six different scales are visualized using geometrical context only. In general, objects of different sizes occur at different locations in the image, so that these priors can be directly used for a fixed sized template-based object detection technique.

3.2.2. Object Co-Occurrences

According to [29], a natural way of representing the context of a specific 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

[5] http://svmlight.joachims.org/.

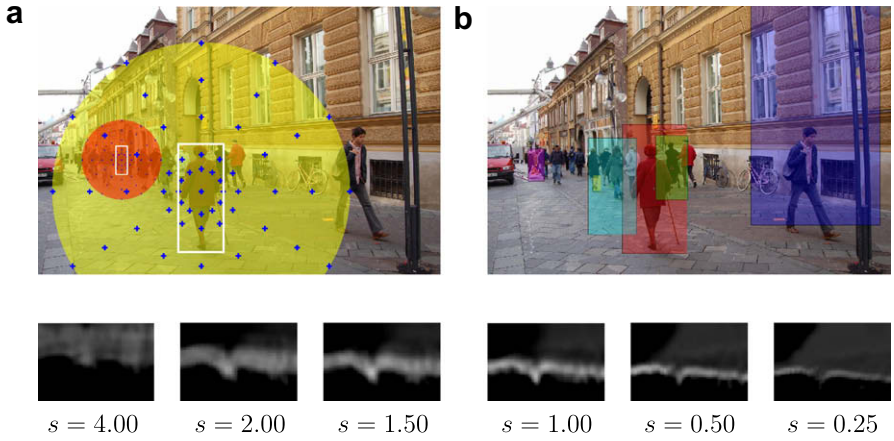

Fig. 5. Top row: The regions from which the contextual information is gathered in one feature vector is visualized with the red and yellow circle for the two marked objects. The blue crosses indicate the locations where the contextual information is sparsely sampled. (a) The regions are scaled according to the object's size so that they represent similar semantic information. (b) The most likely locations for pedestrians for six different scales are visualized using geometrical context only, showing that visual context provides a location and a scale estimate. Bottom row: Context priors based on geometry features for six scales s for the image in the top row. Bright regions indicate locations where pedestrians are likely to occur. It is visible that smaller objects are more likely to occur at different locations than bigger objects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sizes of other target objects are strongly restricted. For illustration purposes we conducted a proof-of-concept experiment. We fixed the location of one pedestrian in the image center and marked the relative occurrences of pedestrians and windows using the test image data set of [30]. As input approximately 4000 manually labeled objects were used for each class. Fig. 6 gives the probability distribution for the two object categories. In addition, some randomly drawn ground truth objects are visualized by their bounding boxes. It is clearly visible that the occurrence of different objects is not independent. Consequently, it makes sense to exploit this dependency for the task of object detection.

To model these spatial dependencies within our framework we extract the conditional probability function for object co-occurrences which is a 2D pdf like visualized in Fig. 6. This pdf is learned from ground truth information, i.e. manually labeled objects for the employed databases (see Section 4.1), using *kernel density estimation* (KDE) in the following way: All images in the data set are sequentially processed. For each image the ground truth objects are collected. Then, for each object the relative offsets to all other objects in the current image are calculated. These relative offsets w.r.t. the image center are then used to calculate the 2D pdf via KDE. The prior holds the likelihood of the presence of other objects given one known object in the center.

Once the co-occurrences are learned they are embedded in the test stage as follows: All object hypotheses from an object detector are used in a voting scheme. Each detection votes for all other detections in the image based on the spatial relation and according to the derived 2D pdf. For comparison we also perform a weighted voting, where the weight is defined as the object detection score. In this case weak detections, that are more likely to be incorrect, have a smaller influence in the voting process. Similar as the approach in [18] this object co-occurrence principle acts as a spatial clustering of the given object hypotheses.

In the special case of detecting pedestrians in urban scenes we introduce a second type of probability modeling. As clearly visible in Fig. 1b and in Fig. 6a pedestrians (more specifically their centers) are more or less aligned on a horizontal line. As a consequence we also model a 1D pdf by marginalizing the initial 2D pdf over the x-axis, yielding a 1D Gaussian-like shaped pdf. This kind of modeling reduces the computational costs and yield better results in this particular setting. Note that, the generality of the framework is lost when switching to 1D modeling. If our task, e.g., would be to detect cars in a circular traffic setup, then the original 2D pdf has to be used.

3.2.3. Bottom-up saliency

It is a common technique to focus on salient regions when searching for a particular object in an image. This kind of sequential processing is also known from the human visual system. In this work we are using pure bottom-up saliency as defined in [20] to calculate the prior for object detection. In contrast to the method described in Section 3.2.1 the extracted saliency map is task-independent and can therefore not adapt for specific object categories. Overall, it can be assumed that a purely task-independent bottom-

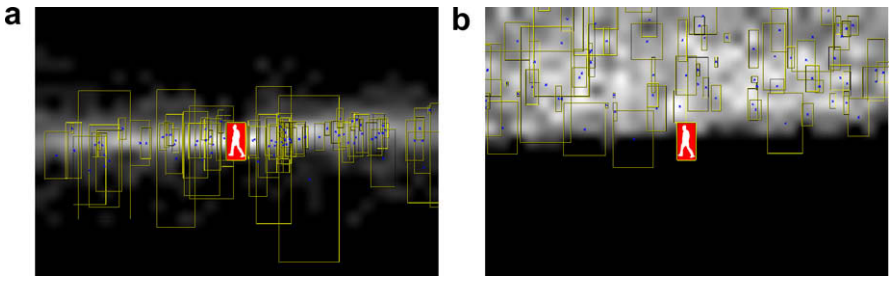

Fig. 6. Conditional dependencies of objects. The probability distribution of (a) pedestrians and (b) windows is visualized, under the condition of the presence of one pedestrian in the center of the image. A couple of object locations are shown by their corresponding bounding boxes.

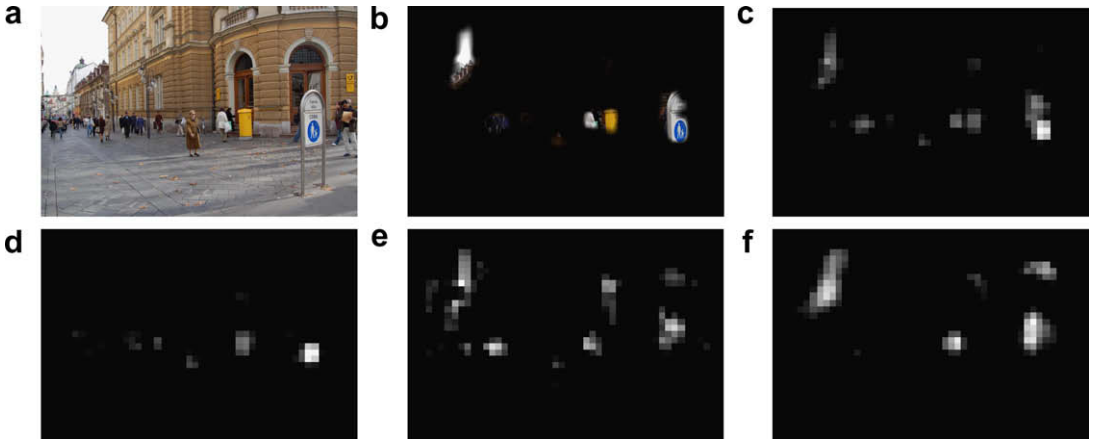

Fig. 7. Example of bottom-up saliency map extraction. (a) Input image, (b) input image with superimposed saliency map, (c) the saliency map, (d) conspicuity map based on color, (e) conspicuity map based on intensities and (f) conspicuity map based on orientations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

up saliency map will not contribute as much as the previously presented approach. However, it is included in our system to enable a comparison to other cues. The particular method to extract the saliency map of a still image is described in [49] which is available as a Matlab implementation.[6] As underlying conspicuity features color, intensity and orientations are used. The proposed default parameters in [49] are applied, i.e. using four orientations calculated via Gabor filtering and equally weighting between the different image features. Dyadic image pyramids are utilized with the lowest surround level of 3, highest surround level of 5, smallest center-surround delta of 3, largest center-surround delta of 4 and saliency map level of 4. For normalization the iterative scheme is used with 3 iterations (see [49] for details). The resulting saliency map is extracted at lower resolution setting the larger image dimension to 80 pixels and scaled to the domain [0,1]. An example is given in Fig. 7.

3.3. Object detector

Local appearance-based object detection is not the main focus of this paper. Nevertheless, the basic principle of the evaluated object detection algorithm is given. As the underlying object detector we apply the method of Dalal and Triggs [7], using Dalal's binaries.[7] The main arguments why this specific detector was chosen are: (i) It was developed for pedestrian detection which is exactly our use case. (ii) The resulting detection rates are very good and outperform several other methods [26,52]. (iii) It can be implemented on GPU and is therefore very efficient [51]. (iv) Threshold parameters can easily be modified so that also weak hypotheses can be used in the experiments.

This object detector is based on the *histograms of gradients* method, encoding the silhouette of an object. Like proposed in [18] we set the threshold for rejecting hypothesis very low, actually to the value -2.0. The detection scores are normalized to [0,1] using the given minimal and maximal values, where we used [1.0,1.7] (different scalings would not alter the final results due to KDE modeling, as long as minimal and maximal numbers are not clipped). Values above the thresholds are clipped. The detector's output are therefore normalized, however not probabilistic. To compare the detections to the ground truth, the criteria in [25] are implemented, being *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 30%. These classical criteria are also used in the PASCAL challenge [9] and in other works on pedestrian detection, e.g. [18,34,52].

4. Experimental results

After introducing the database used for evaluation, we analyze the influence of the different contextual cues on the task of pedestrian detection. Then our proposed method is compared to a state-of-the-art approach for pedestrian detection using contextual information. Finally, results are given for other object categories to show that our method is general and not limited to pedestrians only.

4.1. Database

We collected a demanding image databases for performing the evaluations. By *demanding* we mean images with a lot of background clutter and textured regions, where object hypotheses from local appearance-based object detection are often weak or even incorrect, and where objects occur at very different scales. Standard data sets like LabelMe [39] and especially its subset as used in [18,19] have a more or less homogeneous foreground. Therefore, classical object detectors perform better than on our data set, where the ground plane is often covered by cobblestones or grass.

4.1.1. Darmstadt urban image data set (DUIS131)

We collected the DUIS131 which is freely available[8] and first published in [35]. This image data set contains 1572 images of the city of Darmstadt in Germany with a resolution of 1944×2896 pixels each. For our evaluation the images were downsampled to 972×1448 pixels. For ground truth we manually labeled 4133 pedestrians. Each pedestrian is defined by its corresponding bounding box, encompassing the whole object. The object centered bounding boxes have a fixed aspect ratio of 1:2. On average, there are many small pedestrians with a height smaller than 128 pixels. However, also large objects are in the data set, the tallest having more than 900 pixels.

4.2. Evaluation of different contextual cues

We conducted tests to evaluate the performance of the proposed contextual cues, while also comparing the context integra-

[6] http://www.saliencytoolbox.net/.
[7] http://pascal.inrialpes.fr/soft/olt/.
[8] http://vicos.fri.uni-lj.si/duis131/.

tion techniques. For these tests we used local appearance alone (i.e. the object detector) as a reference and compared it to the fusion with geometry, texture, saliency, object co-occurrence, weighted object co-occurrence and to the fusion of all mentioned cues. The integration of context into the object detection framework is obviously only successful if it improves the detection rate. Therefore, we plot the detection rate versus the *false positives per image* (FPPI), i.e. the average number of incorrect detections (false positives) per image. For the case of object co-occurrences conditional dependencies of pedestrians are exploited and the evaluation is given only for the 1D prior, as the 2D prior yields slightly worse results and is computationally more complex.

In the first test, we assume that the local appearance scores and the context confidence scores are statistically independent and their conditional probability can therefore be calculated by multiplication (cf. Eq. (3)). In addition, the initial scores are treated as if they were probabilistic, as done in [47,5]. The detection rate plots are given in Fig. 8a showing that the contextual integration degrades the initial result significantly. The reason for this behavior is, that both assumptions used are incorrect. The initial confidence scores are not probabilistic and they are also not statistically independent. Nevertheless, we can interpret the influence of the different contextual cues. The worst cue is saliency, followed by non-weighted object co-occurrence, then weighted object co-occurrence, while the best cues are texture and geometry. The combined score "suffers" from the saliency score and is therefore not better than the best individual cue. The particular issue for saliency being more or less useless in this framework is, that it is task-independent. Pedestrians in urban scenes are often not very salient (e.g. dressed in black and walking on a gray pavement) and get a saliency score of exactly zero, which then also yields a combined score of zero.

In the second test, we again assume statistical independency but calculated the prior probabilities of the confidence scores. The difference to the previous test is that the confidence scores are mapped to probabilities with a prior function learned during training (performing a 1D KDE). Results are given in Fig. 8b. Obviously, the saliency maps are again degrading the results. However, all other cues contribute in a positive way to the object detection accuracy. The ranking of the cues is the same as in the previous test, with the exception that texture is aiding more than geometry. The combination of all cues is now better than any initial cue, which is an indication that the cues provide complementary information.

In the third test, we model the dependencies of the cues with the proposed KDE approach. To show the impact of the individual cues we modeled the conditional probabilities of local appearance and each of the contextual cues, which is done via a 2D KDE. Then all cues are modeled with a 6D KDE. Results are given in Fig. 8c, where two major changes to the curves in Fig. 8b are visible. First, the correctly modeled dependency of saliency and local appearance is not degrading the result anymore. However, the combined curve of these two modalities does not yield better accuracy then local appearance alone. The conditional pdf, generated via KDE, automatically models that saliency is not distinctive for this task. Second, the combination of all cues yields a significantly better outcome than the pairwise combinations. This is exactly what should be the case if the cues are complementary and the context integration is performed correctly.

To emphasize the importance of the context integration step we plotted the combined scores from all three tests in one graph (see Fig. 8d). At a fixed rate of two FPPI the naive approach (first test) decreases the detection rate by 3.8%, the second method modeling 1D pdfs increases the performance by 5.5% and the full KDE modeling boosted the results by 11.1%. To reduce the complexity in modeling the conditional pdfs we are not using saliency in the further test. This kind of contextual information could not assist object detection in our setting. From the object co-occurrence cues

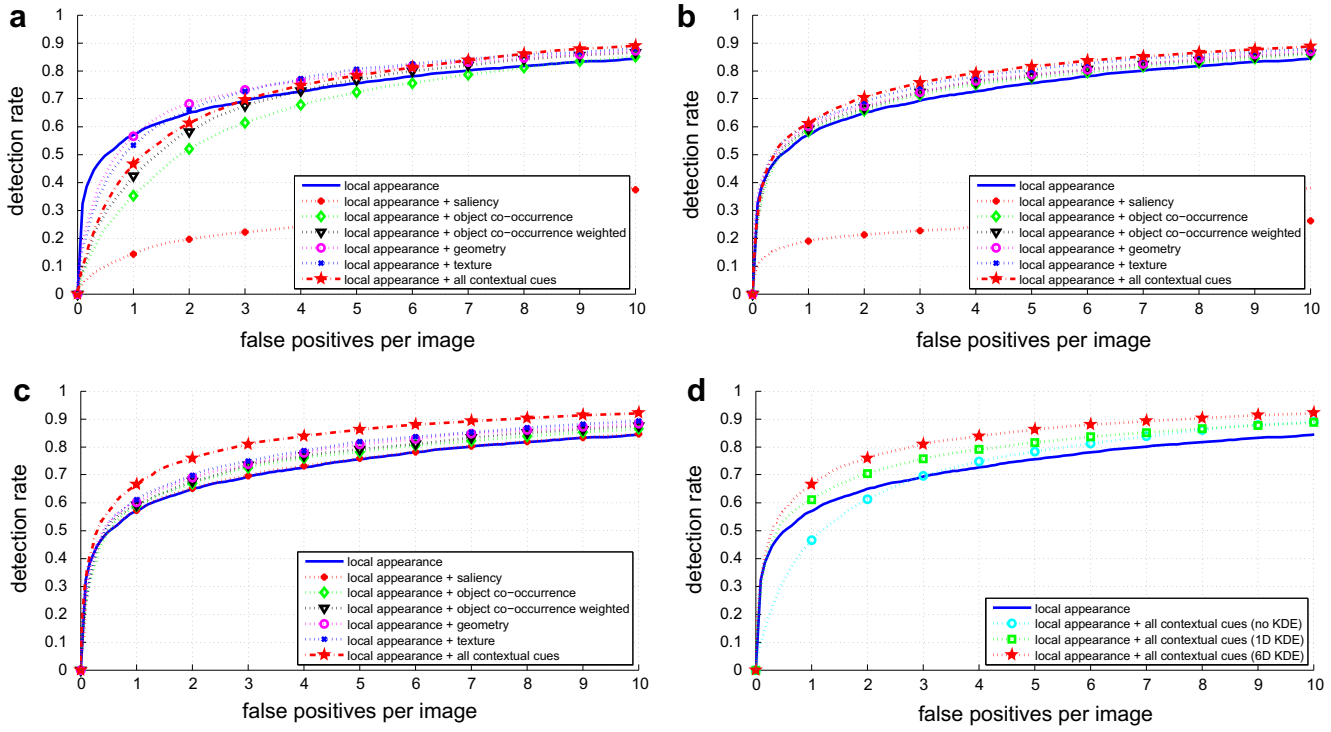

Fig. 8. Comparison of detection rates using different context integration techniques on the DUIS131 data set. Context integration based on (a) multiplication of the individual confidence scores, (b) multiplication of the individual probability scores and (c) modeling the total dependencies using the KDE approach. (d) Comparison of the fused detection rates from (a–c), where our KDE approach significantly outperforms the other context integration concepts.

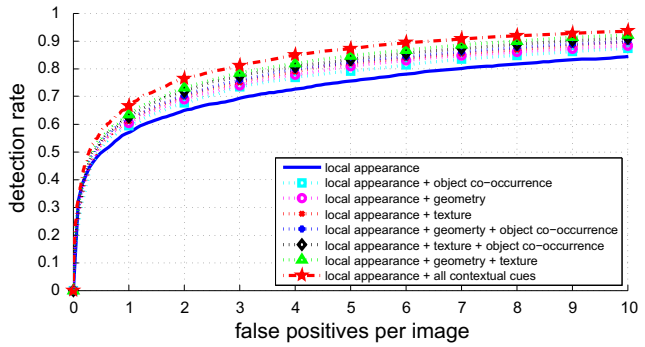

Fig. 9. Comparison of detection rates using different combinations of contextual cues on the DUIS131 data set. The cues are ordered by their impact to the object detection task, where the integration of all contextual cues yields the best result.

we use the weighted version only, as it always performed better. Therefore, in total three contextual features are used, namely texture, geometry and object co-occurrence.

Next, we evaluate if one specific combination of these features performs better than others. All combinations of the three contextual cues were combined with the local appearance information and results are shown in Fig. 9. The ordering of combination of cues is as following (from worst to best, average detection rate improvements given in brackets): object co-occurrence (3.4%), geometry (4.5%), texture (5.5%), geometry and object co-occurrence (6.2%), texture and object co-occurrence (7.4%), geometry and texture (8.5%), combination of all cues (11.3%).

These results are a proof-of-concept that local appearance and contextual information are not independent. The dependency can be modeled e.g. using the proposed KDE approach. The final results are significantly better than the results achieved by other authors.

4.3. Comparison to a state-of-the-art method

In this section we compare our framework for context-aware object detection to the well-known concept by Hoiem et al. [18]. For the tests their publicly available code is used.[9] We perform two different experiments to show that the contextual integration is of great importance. First, both methods are evaluated without modeling the probability functions. The results are given in Fig. 10a, showing that both methods degrade the initial detection rate (cf. Fig. 8a). For our approach we use contextual features based on texture, geometry and object co-occurrence. Since Hoiem's et al. approach is based on a probabilistic framework as well it cannot work correctly on this kind of data. Second, we convert the detection scores to probabilities and supply them to Hoiem's et al. algorithm. Then we compare the results to our framework using multi-dimensional KDE modeling. The detection rates are given in Fig. 10b. Both methods increase the accuracy of object detection if the contextual integration is performed accordingly. Our approach increase the detection rate by 11% on average and Hoiem's approach by 4%. At a fixed rate of two FPPI the increase is 11.5% and 3.8% respectively.

One reason for the rather poor performance of the approach in [18] can be found in their inference process, which sometimes produces an incorrect solution. A cluster of object hypotheses is determined that satisfies a viewpoint estimate, which however is incorrect. In such a case typically all resulting detections are incorrect and correct detections are mostly discarded. In the data set used for evaluation the viewpoint estimate is imprecise for 10.1% of all cases. We consider the horizon estimate as correct if its position w.r.t. the ground truth horizon deviates maximally by 10% of the image's height. In general, a cluster of incorrect object hypotheses from the object detection algorithm yields an improper solution in the contextual inference in [18]. However, in our approach only one out of three contextual cues, namely the object co-occurrence cue, is influenced by clusters of wrong hypotheses. The other cues based on geometry and texture are not intertwined with the spatial relations of the hypotheses, and therefore are not influenced negatively in this case.

[9] http://www.cs.uiuc.edu/homes/dhoiem/software/.

4.4. Generalization

In the experiments described thus far, the evaluation was performed using pedestrians as objects of interest. To prove that the concept is not limited to this kind of objects, we calculated context priors also for cars and windows using the LUIS34 data set presented in [30]. In addition to the 3803 pedestrians we manually labeled 803 cars and 3601 windows and run the context extraction with the same parameters. Results for visual interpretation are given in Fig. 11 using geometry as contextual cue. As the presence of an object also influences the visual contextual features, for 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. 11 it can be seen, that pedestrians should appear on pavements, cars occur on the road, which is next to the possible location of pedestrians. Windows are basically above the road on facades of buildings. These different image regions are correctly detected using our definition of visual context. The priors for cars and for windows will definitely support a local appearance-based detection algorithm, as it is the case for pedestrians.

When applying the framework for detecting a different object category using another local appearance-based detector, the following parameters have to be adjusted: The detector's output should be mapped to the domain [0,1] (cf. Section 3.3) and also weak hypotheses should be gathered (normally there is a threshold in the detection process). When switching to another learning scheme, instead of using the proposed SVM approach, the resulting values of this method have to mapped to [0,1] as well (cf. Section 3.2.1.2). All other internal parameters are calculated automatically from the given input data.

5. Discussion

We presented methods to extract visual contextual information from images and used it to increase the performance of local appearance-based object detection. The proposed framework is very general and evaluated on outdoor scenes. It is assumed that the benefit of using visual context for object detection would be lower for more complex scenes and objects, e.g. arbitrary indoor settings. However, the authors are confident that context also plays an important role in indoor recognition which was recently shown by e.g. [8,37,27]. A major contribution of this work was the context integration technique based on multi-dimensional KDE modeling. We compared our approach to the one state-of-the-art method by Hoiem et al. [18]. Both methods for performing visual-context-aware object detection improve the detection rate of the underlying detector. However, the increase of the detection rate is on average only about 4% for [18], while it is 11% for our method (see Fig. 10). Depending on the final application this boost is of interest or may be negligible. An interesting question is why these novel methods are not providing stronger cues to assist object detection. Part of the answer is illustrated in Fig. 12. In Fig. 12a two images of our data set are shown with all object hypotheses based on local appearance marked. In Fig. 12b all detections with a probability of being correct higher than 0.5 are given. In Fig. 12c the horizon estimate from [18] is visualized with the

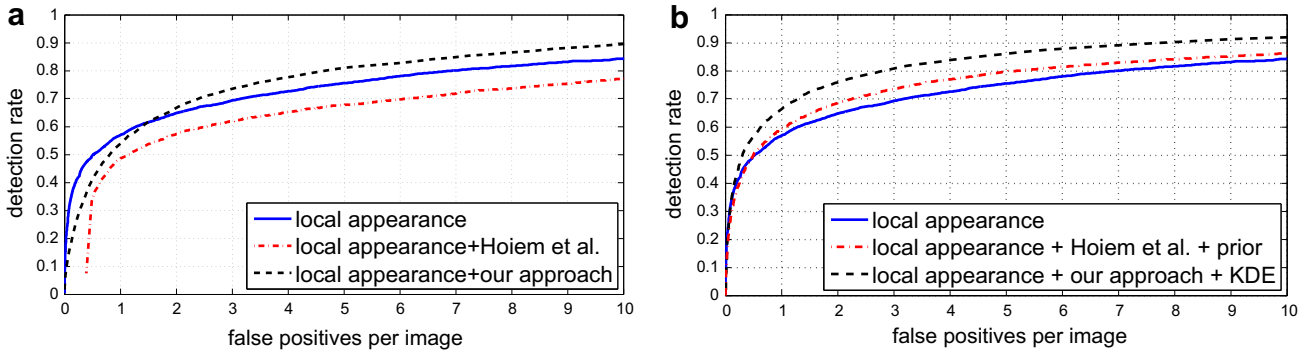

Fig. 10. Comparison to the state-of-the-art approach [18]. Detection rates (a) using the context integration based on pure multiplication and (b) using our KDE-based context integration technique. While the original methods decrease the accuracy, they yield good results when incorporating the visual context in a proper way.

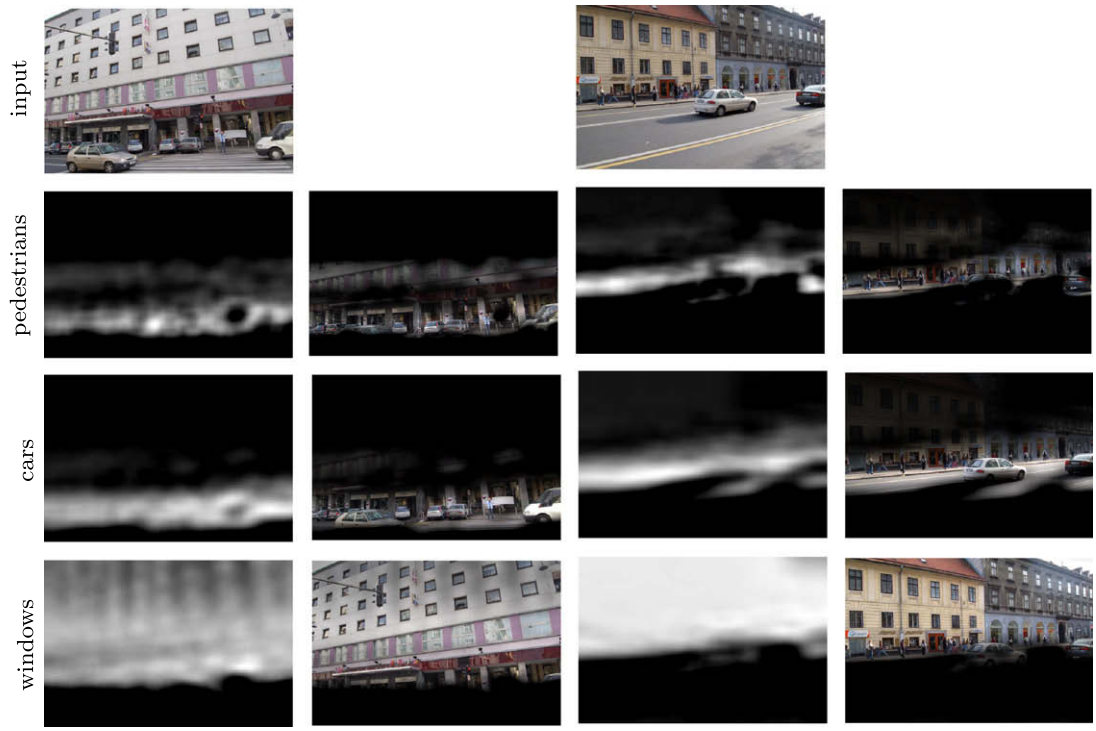

Fig. 11. Context priors for three different object categories shown for two images of the database in [30]. Shown are the context priors for pedestrians, cars and windows and the input image multiplied by these maps. Bright areas indicate locations where the probability for the presence of the given object category is high.

remaining object hypotheses after contextual inference. Even though the horizon is correctly estimated and all 11 (top row) respectively 9 (bottom row) detections satisfy the global scene geometry, only 1 of them is a correct detection in each row. In Fig. 12d the location priors from our method are shown for geometry features (shown for the scale $s = 1$, cf. Fig. 5). These priors are robust estimates, however they only down-rank a few detections with a high detection score, i.e. the hypothesis on the roof top in the second example. In general the problem is that there are many object hypotheses based on a local appearance measure that are incorrect but suit to the scene in terms of their position and size. Such hypotheses cannot be rejected or down-ranked by visual contextual information.

Additionally, it can happen, that a correctly detected object gets a low contextual score. This is the case when a specific situation is (i) not presented in the training set (e.g. a person standing on a balcony in our pedestrian scenario) and therefore not modeled or (ii) rare, i.e. an outlier in the contextual feature space. Such specific detections are basically "sacrificed" for an overall increased detection rate.

In total, visual context only provides priors for the position and size where an object of interest is likely to occur according to the given scene content. On the one hand, false object hypotheses fitting to the scene layout "survive" the contextual inference. On the other hand, hypotheses that are strongly out-of-context have a weak local appearance in many cases anyway. Due to this aspects, the boost of the detection rate is limited using visual context as an additional cue.

Next, we discuss the contributions of the proposed contextual cues for pedestrian detection.

First, we found that pure bottom-up saliency is not helping at all. To be fair, we have to point out that there is evidence that saliency in fact aids pedestrian detection, if motion is included in the saliency extraction. Obviously, moving objects in urban scene are often pedestrians, which is exploited in e.g. [13]. However, for this kind of processing video data is needed and non-moving pedestrians will not get detected.

Second, our concept of modeling object co-occurrences is straight forward and extremely efficient to implement. The detec-

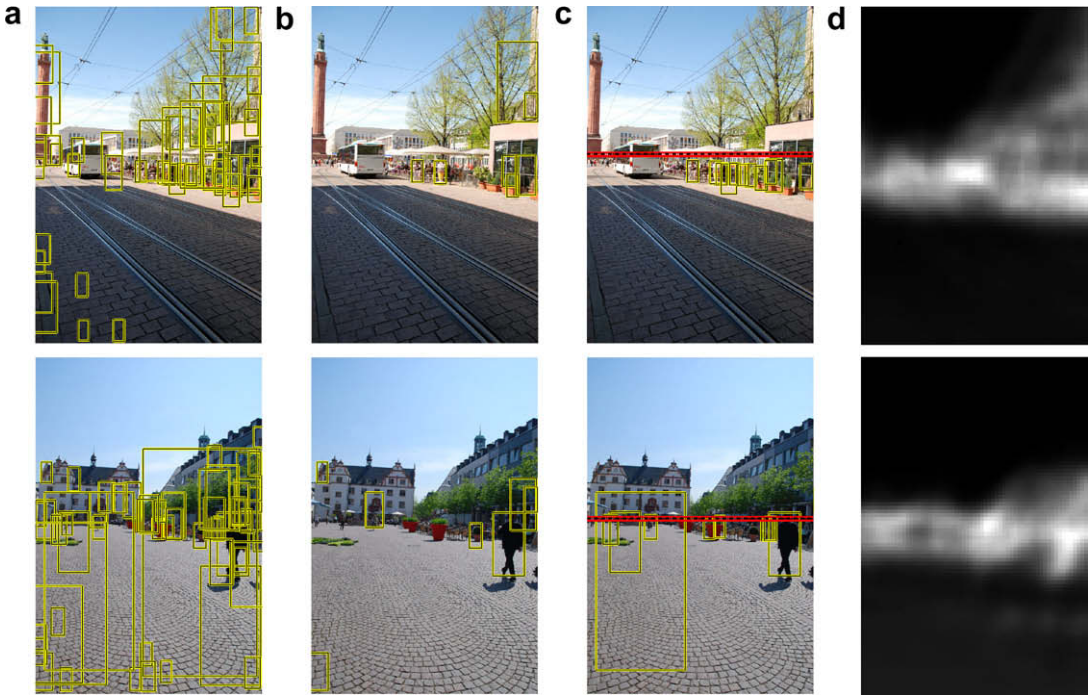

Fig. 12. Limits of visual-context-aware object detection. (a) Urban scene with hypotheses for pedestrians, (b) object hypotheses with a score larger than 0.5, (c) horizon estimate and detections supporting this estimate [18] and (d) context priors using geometry features.

tion rate increases by 3.4% in our evaluation using this contextual cue. The object co-occurrence method is limited due to two reasons: (i) The confidence score is not really telling if there is only a small number of objects detected in an image and (ii) clusters of incorrect hypotheses will get a high score even though all hypotheses are erroneous. The same two problems are also experienced with the algorithm in [18].

Third, the contextual cue based on the semantic image segmentation is contributing with 4.5% to the detection rate. The main limitations for the geometry are that many other objects share the same scene context which are not within the object category of interest. In the case pedestrian detection classical examples are trash cans or part of fences having coarsely similar shape as pedestrians and are often of similar height.

Fourth, the contextual cue based on a simple texture descriptor performs better than geometry, aiding 5.5% to the pedestrian detection. The beauty of these contextual features is that they can be estimated in a very low resolution (the input image can be downsampled to approximately 80×60 pixel) and are therefore computationally very cheap. However, it focuses on specific properties of urban scenes, i.e. that the ground is rather textureless in general. Therefore, this learned model is not general and will yield incorrect results in random outdoor scenes.

Overall, the individual contextual cues are not completely complementary, therefore the context integration via KDE yields on overall increase of 11.3% which is less than the sum of the single contributions. Nevertheless, it was shown that modeling the dependencies of local-appearance and the contextual cues via KDE yields very good results. This empirically proves that these cues are actually statistically depended and an independency assumption like e.g. in [47,45,5] is inadequate.

6. Conclusion

Visual context provides cues about an object's presence, position and size within the observed scene, which can be used to increase the performance of object detection techniques. Therefore, we presented a novel framework for visual-context-aware object detection in still images. Methods for extracting visual contextual information from still images were proposed, which were then used to calculate a prior for object detection. The concept is based on a sparse coding of contextual features, which are based on geometry and texture. In addition, bottom-up saliency and object co-occurrences were exploited, to define auxiliary visual context. To integrate the individual contextual cues with a local appearance-based object detector, a fully probabilistic framework was established. In contrast to other methods, our integration is based on modeling the conditional probabilities between the different cues using a kernel density estimation. We empirically proved that our dependency assumption is correct and that this integration is a crucial part of the framework which was demonstrated within the detailed evaluation. Our method was evaluated using a novel demanding image data set and compared to a state-of-the-art method for context-aware object detection. An in-depth analysis was given discussing the contributions of the individual contextual cues and the limitations of visual context for object detection. In our experiments the proposed framework increased the object detection rate by over 11%, which is a proof-of-concept that visual context, as defined in this work, aids object detection significantly.

Acknowledgments

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 acknowledge the CVIU reviewers, Matej Kristan and Martina Uray for useful comments and discussions on this manuscript.

References

[1] E. Aminoff, N. Gronau, M. Bar, The parahippocampal cortex mediates spatial and nonspatial associations, Cereb. Cortex 17 (7) (2007) 1493–1503.
[2] M. Bar, Visual objects in context, Nat. Rev. Neurosci. 5 (2004) 617–629.

[3] I. Biederman, Perceiving real-world scenes, Science 177 (4043) (1972) 77–80.
[4] I. Biederman, On the Semantics of a Glance at a Scene, Perceptual Organization, Lawrence Erlbaum, 1981 (Chapter 8, pp. 213–263).
[5] S.M. Bileschi, StreetScenes: Towards Scene Understanding in Still Images. PhD Thesis, Massachusetts Institute of Technology, May 2006.
[6] C. Carson, S. Belongie, H. Greenspan, J. Malik, Blobworld: image segmentation using expectation-maximization and its application to image querying, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1026–1038.
[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. Conf. Comp. Vis. Pattern Recog., vol. 2, 2005, pp. 886–893.
[8] S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros, M. Hebert, An empirical study of context in object detection, in: Proc. Conf. Comp. Vis. Pattern Recog., June 2009.
[9] M. Everingham, A. Zisserman, C.K.I. Williams, L. Van Gool, The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results, September 2006. <http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf>.
[10] A. Friedman, Framing pictures: The role of knowledge in automatized encoding and memory for gist, J. Exp. Psychol. Gen. 108 (8) (1979) 316–355.
[11] M. Girolami, C. He, Probability density estimation from optimally condensed data samples, IEEE Trans. Pattern Anal. Mach. Intell. 25 (October) (2003) 1253–1264.
[12] J.O.S. Goh, S.C. Siong, D. Park, A. Gutchess, A. Hebrank, M.W.L. Chee, Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation, J. Neurosci. 24 (45) (2004) 10223–10228.
[13] J. Harel, C. Koch, On the optimality of spatial attention for object detection, in: Int. Works. Attention in Cognit. Sys., vol. 5, May 2008, pp. 27–40.
[14] C. Harris, M. Stephens, A combined corner and edge detector, in: Proc. 4th Alvey Visual Conf., 1988, pp. 189–192.
[15] D. Heesch, M. Petrou, Non-Gibbsian Markov random field models for contextual labelling of structured scenes, in: Proc. British Mach. Vis. Conf., September 2007.
[16] J.M. Henderson, A. Pollatsek, K. Rayner, The effects of foveal priming and extrafoveal preview on object identification, J. Exp. Psychol. Hum. Percept. Perform. 13 (3) (1987) 449–463.
[17] D. Hoiem, A.A. Efros, M. Hebert, Geometric context from a single image, in: Proc. Int. Conf. Comp. Vis., vol. 1, October 2005, pp. 654–661.
[18] D. Hoiem, A.A. Efros, M. Hebert, Putting objects in perspective, in: Proc. Conf. Comp. Vis. Pattern Recog., vol. 2, June 2006, pp. 2137–2144.
[19] D. Hoiem, A.A. Efros, M. Hebert, Closing the loop on scene interpretation, in: Proc. Conf. Comp. Vis. Pattern Recog., June 2008.
[20] L. Itti, C. Koch, Computational modeling of visual attention, Nat. Rev. Neurosci. 2 (3) (2001) 194–203.
[21] T. Joachims, Making large-scale support vector machine learning practical, in: Advances in Kernel Methods: Support Vector Machines, MIT Press, Cambridge, MA, 1999, pp. 41–56 (Chapter 11).
[22] M.C. Jones, J.S. Marron, S.J. Sheather, A brief survey of bandwidth selection for density estimation, J. Am. Stat. Assoc. 91 (433) (1996) 401–407.
[23] M. Kristan, D. Skočaj, A. Leonardis, Incremental learning with Gaussian mixture models, in: Comp. Vis. Winter Works., February 2008, pp. 25–32.
[24] H. Kruppa, B. Schiele, Using local context to improve face detection, in: Proc. British Mach. Vis. Conf., September 2003.
[25] B. Leibe, Interleaved Object Categorization and Segmentation, PhD Thesis, ETH Zurich, PhD Thesis No. 15752, October 2004.
[26] B. Leibe, A. Leonardis, B. Schiele, Robust object detection with interleaved categorization and segmentation, Int. J. Comput. Vision 77 (1–3) (2008) 259–289.
[27] T. Malisiewicz, A.A. Efros, Beyond categories: the visual memex model for reasoning about object relationships, in: Neural Inf. Proc. Systems, December 2009.
[28] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vision 42 (3) (2001) 145–175.
[29] A. Oliva, A. Torralba, The role of context in object recognition, Trends Cogn. Sci. 11 (12) (2007) 520–527.
[30] D. Omerčević, O. Drbohlav, A. Leonardis, High-dimensional feature matching: employing the concept of meaningful nearest neighbors, in: Proc. Int. Conf. Comp. Vis., October 2007.
[31] S.E. Palmer, The effects of contextual scenes on the identification of objects, Mem. Cogn. 3 (1975) 519–526.
[32] C. Papageorgiou, T. Poggio, A trainable system for object detection, Int. J. Comput. Vision 38 (1) (2000) 15–33.
[33] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers, 1988.
[34] R. Perko, A. Leonardis, Context driven focus of attention for object detection, in: Int. Works. Attention in Cognit. Sys., vol. 4840, December 2007, pp. 216–233 (Chapter 14).
[35] R. Perko, C. Wojek, B. Schiele, A. Leonardis, Probabilistic combination of visual context based attention and object detection, in: Int. Works. Attention in Cognit. Sys., vol. 5, May 2008, pp. 166–179.
[36] J.C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Classifiers 10 (3) (1999) 61–74.
[37] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: Proc. Conf. Comp. Vis. Pattern Recog., June 2009.
[38] X. Ren, C.C. Fowlkes, J. Malik, Figure/ground assignment in natural images, in: Proc. European Conf. Comp. Vis., vol. 2, May 2006, pp. 614–627.
[39] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: a database and web-based tool for image annotation, Technical Report AIM-2005-025, MIT AI Lab Memo, September 2005.
[40] P. Sabzmeydani, G. Mori, Detecting pedestrians by learning shapelet features, in: Proc. Conf. Comp. Vis. Pattern Recog., June 2007.
[41] B. Schiele, J.L. Crowley, Recognition without correspondence using multidimensional receptive field histograms, Int. J. Comput. Vision 36 (1) (2000) 31–50.
[42] E. Seemann, B. Leibe, B. Schiele, Multi-aspect detection of articulated objects, in: Proc. Conf. Comp. Vis. Pattern Recog., vol. 2, June 2006, pp. 1582–1588.
[43] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 411–426.
[44] A. Torralba, Contextual modulation of target saliency, in: Neural Inf. Proc. Systems, vol. 14, 2002, pp. 1303–1310.
[45] A. Torralba, Contextual priming for object detection, Int. J. Comput. Vision 53 (2) (2003) 153–167.
[46] A. Torralba, A. Oliva, M.S. Castelhano, J.M. Henderson, Contextual guidance of attention in natural scenes: the role of global features on object search, Psychol. Rev. 113 (4) (2006) 766–786.
[47] A. Torralba, P. Sinha, Statistical context priming for object detection, in: Proc. Int. Conf. Comp. Vis., vol. 1, July 2001, pp. 763–770.
[48] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. Conf. Comp. Vis. Pattern Recog., December 2001.
[49] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (9) (2006) 1395–1407.
[50] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman & Hall/CRC, 1995.
[51] C. Wojek, G. Dorkó, A. Schulz, B. Schiele, Sliding-windows for rapid object class localization: a parallel technique, in: DAGM-Symposium, vol. 30, June 2008, pp. 71–81.
[52] C. Wojek, B. Schiele, A performance evaluation of single and multi-feature people detection, in: DAGM-Symposium, vol. 30, June 2008, pp. 82–91.
[53] L. Wolf, S.M. Bileschi, A critical view of context, Int. J. Comput. Vision 69 (2) (2006) 251–261.