Integrating Visual Context and Object Detection within a Probabilistic Framework

Roland Perko[1], Christian Wojek[2], Bernt Schiele[2], and Aleš Leonardis[1]

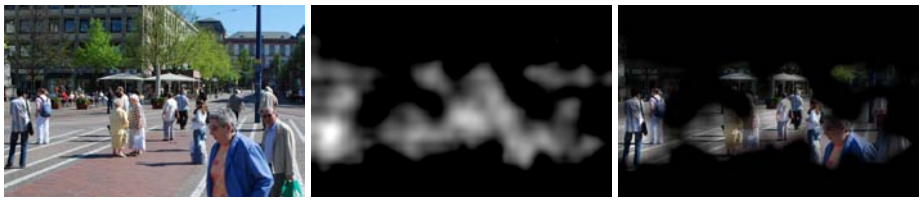
[1] University of Ljubljana, Slovenia
{roland.perko,ales.leonardis}@fri.uni-lj.si
[2] TU Darmstadt, Germany
{wojek,schiele}@cs.tu-darmstadt.de

Abstract. Visual context provides cues about an object's presence, position and size within an observed scene, which are used to increase the performance of object detection techniques. However, state-of-the-art methods for context aware object detection could decrease the initial performance. We discuss the reasons for failure and propose a concept that overcomes these limitations, by introducing a novel technique for integrating visual context and object detection. Therefore, we apply the prior probability function of an object detector, that maps the detector's output to probabilities. Together, with an appropriate contextual weighting, a probabilistic framework is established. In addition, we present an extension to state-of-the-art methods to learn scale-dependent visual context information and show how this increases the initial performance. The standard methods and our proposed extensions are compared on a novel, demanding image data set. Results show that visual context facilitates object detection methods.

1 Introduction

A standard approach for detecting an object of a known category in still images is to exhaustively analyze the content of image patches at all image positions and at multiple scales (see e.g. [1,2]). When a patch is extracted from an image, it is classified according to its local appearance and associated with a detection score. The score should correspond to the probability of the patch representing an instance of the particular object category and is usually mapped to a probability score. As it is known from the literature on visual cognition [3,4], cognitive neuroscience [5,6] and computer vision [7,8,9], the human and animal visual systems use relationships between the surrounding and the objects to improve their ability of categorization. In particular, visual context provides cues about an object's presence, position and scale within the observed scene or image. This additional information is typically ignored in the object detection task. Like in other promising papers on visual context for object detection [10,8,11,9], we define the context as the surrounding, or background, of the current object of interest. This context is used to focus the attention on regions in the image where the objects are likely to occur. Instead of searching the whole image at various

L. Paletta and J.K. Tsotsos (Eds.): WAPCV 2008, LNAI 5395, pp. 54–68, 2009.



(a) (b) (c)

Fig. 1. A concept of using visual context for object detection. (a) A standard image of an urban scene, (b) the focus of attention for the task of pedestrian detection using the method in [9] and (c) the image multiplied by the focus of attention.

scales for an object, visual context provides regions of interest, i.e. the focus of attention, where the search is restricted to. This results in speedup, but more importantly, it can increase the detection rate by not considering incorrect object hypotheses at unlikely positions in the first place. This concept is illustrated in Fig. 1. In this study, we conducted experiments with the approaches of Hoiem *et al.* [8] and Perko and Leonardis [9] on a specific task of detecting pedestrians in urban scenes. We found that the first approach may fail when too many incorrect object hypotheses are detected and the second may, in some cases, reduce the detection rate of the original detector. A detailed analysis revealed that these failure cases are linked to the assumption of these methods that visual contextual information *always* assists the detection step. Furthermore, the prior probability from the local appearance-based object detection method is ignored when combined with the contextual score. We state that the prior probability is an intrinsic property of the object detector used and is defined as the conditional probability of the detection being correct given the detection score. The function is used to map the detector's output to a probability space. In addition, the contextual information is weighted by a function depending on the probabilistic detection score. The basic idea is as follows: if an object is well defined by its local appearance, then context should not contribute much in the detection phase. It can even introduce additional errors by incorrectly re-ranking the detections. However, if the local appearance is weak, context can contribute significantly to improve detections. Therefore, we propose to learn this prior probability function, together with a contextual weighting, and embed it into the existing systems for object detection. An example of this concept is given in Fig. 2. The pedestrian in Fig. 2(b) is well defined by its local appearance and context is not important to get an unambiguous detection. However, the smaller pedestrians shown in Fig. 2(c) and (d) are not as easily detected based on local appearance alone, so that visual context provides more clues about these detections being correct.

Our contribution. We point out in which situations the current state-of-the-art methods for performing context aware object detection decrease the initial performance in practice and discuss the reasons for failure. In particular the approaches in [8] and [9] are evaluated and compared. We then propose a concept

(a) (b) (c) (d)

Fig. 2. A standard image of an urban scene. (a) Three pedestrians of different size are marked by the yellow arrows and boxes. (b-d) Close-ups of the marked objects using nearest-neighbor interpolation for resizing. Even though humans can recognize the objects at all three scales, it is obvious that the bigger pedestrian in (b) is easier to detect than the ones in (c) or (d).

that overcomes these limitations. More specifically, the contextual information is combined with the local appearance-based object detection score in a fully probabilistic framework. We also extend this framework to multiple object detectors trained for different object sizes. In addition, we present an extension to the method in [9] to learn scale-dependent visual context information and show how its performance increases. Then, the methods are compared on a novel demanding database.

Organization of the paper. Related work will be discussed in detail in Sec. 2 and the drawbacks of the state-of-the-art approaches are pointed out in Sec. 3. In Sec. 4 we describe the extension to standard methods and how the probabilistic framework is set up. Results are given in Sec. 5. In the discussion in Sec. 6 we analyze the limitations of contextual processing and conclude the paper in Sec. 7.

2 Related Work

In computer vision the combination of visual context with the task of object detection is a rather young field of research. The aim is to extract more global information from a single image and use it to improve the performance of classical object detection methods. Technically there are two issues to be solved. First, how to represent this kind of visual context within some data structure, i.e. a feature vector. There is a lack of simple representations of context and efficient algorithms for the extraction of such information from images [12]. And second, how to combine this information with an object detection technique. For the former the feature vectors holding contextual information are learned from a labeled database. The LabelMe image database [13] is often used for such purposes. Then for new images, the extracted feature vectors are classified using this learned model. For the latter it is assumed that the contextual information and the local information used for detecting objects are statistically independent.

Therefore, their conditional probability is equal to the product of the individual probabilities [10,9]. Namely, the combined probability p is calculated using the contextual probability p_C and the local appearance-based probability p_L as $p = p_C \cdot p_L$. Context could be also used in a cascade [11,9]. In this case, only pixels with a contextual confidence above a threshold are used in the object detection task. However, as the detection score is not re-ranked and therefore only out-of-context detections (e.g. pedestrians in the sky) are filtered, the increase of detection rate is negligible [14]. Before we present three methods in detail we want to point to the current review on visual context and its role in object recognition by Oliva and Torralba [12]. They also show how the focus of attention extracted using visual context can be combined with classical attention concepts, e.g. with the system of Itti and Koch [15]. Therefore, the results from visual context can be used as the top-down saliency in approaches like [16].

The influential work from Oliva and Torralba, e.g. [17,18,19,10], introduced a novel global image representation. The 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 the objects belonging to that category. To the best of our knowledge there exist no evaluation for the combination of this derived focus of attention 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 seen later, when combined just by multiplication the results of the detection may and often do degrade.

Hoiem *et al.* provided a method to extract the spatial context of a single image [20]. 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* 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 [8] and [9]. Hoiem *et al.* [8] 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. Now, the innovative part of their work is the combination of the contextual information with the object hypotheses using inference. Without going into detail, 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.

Perko and Leonardis [9] use the semantic classification of an image in [20] as one feature set, and low-level texture features, based on Carson's *et al. Blobworld* system [21] as a second set. Both types of features are probabilistic and extracted for each pixel in the image, which is downsampled for speedup. To define the visual context at a given position in the image, they sample those features radially for a given number of radii and orientations, like in [11]. The extracted feature vector is relatively low-dimensional, i.e. 180-dimensional as reported in [9]. A significant increase of the object detection rate is reported using this kind of contextual information, where the low-level texture-based scene representation is more important than the high-level geometry-based representation.

3 Drawbacks of the State-of-the-Art Approaches

The mentioned state-of-the-art methods for extracting and using visual context for an object detection task are reported to increase the performance of the initial object detection. However, we found that this is not always the case, especially when using a demanding image database. By *demanding* we mean images with a lot background clutter and textured regions where object hypotheses are often incorrect, and where objects occur at very different scales. We therefore collected an image data set in an urban environment and experimented with the methods in [8] and [9], using pedestrians as objects of interest. As done in the mentioned papers we plotted the detection rate versus the false positives per image (FPPI), and observed that the method by Hoiem *et al.* significantly decreased the initial detection rate. In the evaluation we used the publicly available Matlab source code[1]. Fig. 3(a) shows the initial detection rate using our own implementation of the Dalal and Triggs pedestrian detector [2] and the detection rate curves after applying the inference. Our analysis shows that there are two reasons for this behavior. First, the contextual inference process often produces an incorrect solution. A cluster of object hypotheses is determined that satisfies a viewpoint estimate which is however incorrect. In such a case typically *all* detections in that image are incorrect and correct detections are mostly discarded. An example is given in Fig. 3(b-c). In our database this happens for 10.1% of the images. We consider the horizon estimate as correct if its position w.r.t. the ground truth horizon position deviates maximal 10% of the image's height. Second, in this approach the object detection score is assumed to be probabilistic. Therefore, the support vector machine (SVM) outputs from object detection are mapped to probabilities using the approach in [22]. However, these mapped outputs are

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

only probabilistic in the sense, that they are in the range of $[0, 1]$. The relation from these scores to the probability that a detection is correct is still unknown. This additional mapping (called the prior probability) should be learned from training data and used to have real probabilities in this framework. As shown in Sec. 5.2 the method performs better when incorporating this function. The method of Perko and Leonardis [9] does not have problems when many incorrect hypotheses are given, as the detections are treated separately (no contextual inference). However, as the prior probability function is not modeled, locally well defined objects could be incorrectly re-ranked. In addition, the contextual influence is not specially related to the appearance-based score. Due to this two aspects the method performs poor at low FPPI rates, yielding even worse results than the initial detections not using contextual information at all. Fig. 3(a) gives the detection rate curves. In general we noticed that the current state-of-the-art methods for performing context aware object detection could fail in practice. We can also predict that the methods of Torralba [10] and Bileschi [11] will likewise lower the initial detection rate, as they are ignoring the prior object detection probability as well, and the underlying concept is similar to [9]. Using the prior probability from the object detector will fix these problems for all mentioned methods.

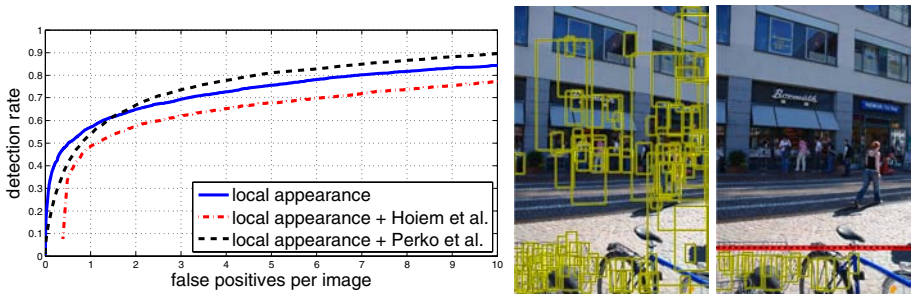


Fig. 3. Drawbacks of the state-of-the-art approaches that use visual context for object detection. (a) The initial detection rate based on local appearance only yield better results than the two evaluated methods that use contextual information. The approach in [8] decreases the performance in general, while the one in [9] has problems at low FPPI only. (b) An example image with object hypotheses from pedestrian detection. (c) Due to the cluster of incorrect hypotheses in the foreground the horizon estimate (horizontal line) is invalid, so that all detections after contextual inference are incorrect.

4 Our Extensions

Two extension to the existing methods are presented. First, we introduce a scale extension in Sec. 4.1 and second, we explain the probabilistic framework in Sec. 4.2.

4.1 Scale Extension

We implemented an extension to [11] and [9], i.e. to learn and use a scale-dependent visual context information. The mentioned methods use a fixed sized region, i.e. one quarter of the image area, to extract the feature vector holding the contextual information. However, when dealing with objects of an a priori known range of sizes in the real world, e.g. pedestrians or cars, these fixed sized regions are not representing the same semantic context for objects perceived at different scales. Therefore, these regions should be scaled with the object's size in the image. Smaller objects corresponds to objects in the distance, an effect of the projective geometry. Therefore, we simply scale the region of which the context is gathered for the given object of interest with its size, visualized in Fig. 4. For smaller pedestrians a smaller region is used to extract the features. These new features are learned using an SVM as in [9]. Then, instead of extracting only one context confidence map for one predefined size, multiple confidence maps are calculated for a given number of scales. An example of such confidence maps is given in Fig. 4 for six scales. The context confidence score for a given detection is then calculated by linear interpolation using the scores of the two adjacent scales. As a result the prior based on context is not only able to provide regions where an object is likely to occur, it also provides the possible size of the object. It

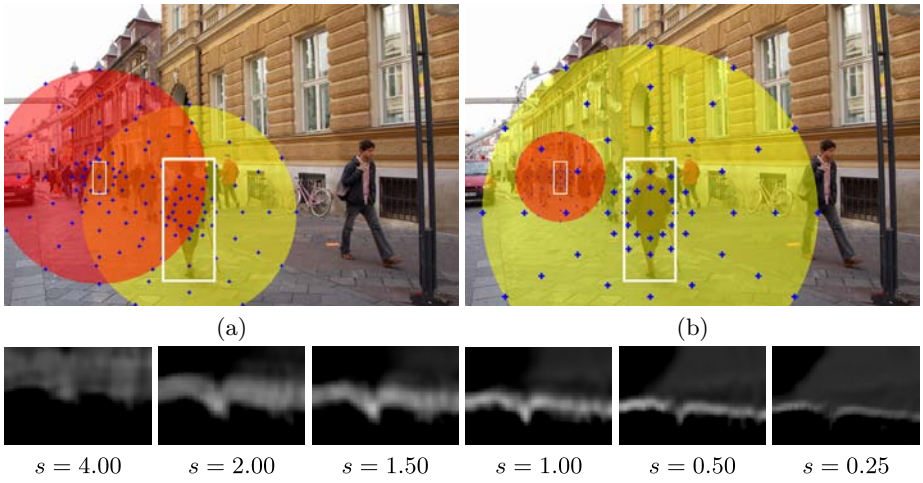


Fig. 4. Top row: The regions from which the contextual information is gathered in a feature vector is visualized with the red and yellow circles for the two marked objects. The blue crosses indicate the locations where the contextual information is sparsely sampled. (a) The regions are of constant size as proposed in [11,9]. (b) Regions are scaled according to the object's size so that they represent similar semantic information. Bottom row: Context confidence maps (foci of attention) 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.

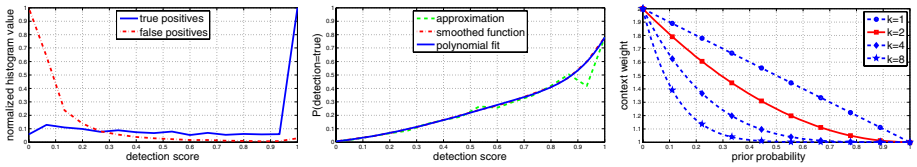
should be pointed out that extracting several foci of attention of course increase the computational expanses of the whole framework. However the confidence map extraction using the pre-learned SVM is faster than the feature extraction, so that the multiscale approach is also applicable in practise.

4.2 Probabilistic Framework

As stated before the prior probability of the object detector used should be modeled and applied to the whole framework. This probability function is an intrinsic property of the detector and can be learned in the training phase. It holds the conditional probability of the detection being correct given the detection score, which is in the probabilistic range. This function is only valid for one set of parameters, that means if, e.g., a detection threshold is changed the function has to be recalculated. In our approach we label the detections as true positives and false positives using the ground truth that exists in the learning phase of object detection. Two histograms with 16 bins are calculated holding the number of true and false detections. The prior probability is then extracted by dividing the number of true detections by the number of overall detections in each bin. To ensure a smooth function the values are filtered using average and median filtering, where the borders, i.e. values at 0 and 1 are preserved. Then an analytic function p_a is fitted describing the prior probability. In the current implementation a polynomial of order 6 is used. The concept is illustrated in Fig. 5(a-b). Instead of multiplying the local appearance-based detection score L with the contextual score p_C at the given position in the image as in [10,11,9], the final score is the product of the prior probability of the detection being correct $p_a(L)$ with the context confidence p_C weighted by the function w, defined as

$$p_{combined} = p_a(L) \cdot w \cdot p_C \quad \text{with} \quad w = (1 - p_a(L))^k + 1\,. \tag{1}$$

The parameter k defines the steepness of the weighting function w, where we use $k = 2$ in the experiments. We experimented with different values of k and



(a) (b) (c)

Fig. 5. Extraction of the prior probability function illustrated using Dalal and Triggs pedestrian detector [2] in (a-b) and the function used to weight the influence of the contextual information in (c). (a) Normalized histograms for true positives and false positives, (b) approximation of the conditional probability function, smoothed prior probability function and the polynomial fit p_a. (c) Function used to weight the influence of the contextual information, shown for different setting of the parameter k. Context gets a higher weight for detections with lower prior probabilities.

found out that the results improve when using the contextual weighting, i.e. $k \neq 0$. However, the specific value of k in rather unimportant as long as it is the range of $[1, 4]$. The function w is bounded in $[1, 2]$ and visualized in Fig. 5(c). This weighting models the concept that the contextual information gets higher weight for detections with a lower local appearance-based score, and lower weight for high-ranked detections.

5 Experimental Results

We conducted experiments using the proposed extensions described in Sec. 4 and show that the methods in [8] and [9] yield significantly better results with these extensions. First, we evaluate the scale extension in Sec. 5.1 and second, we analyze the probabilistic framework in Sec. 5.2. Additional results including videos can be found on our project page[2].

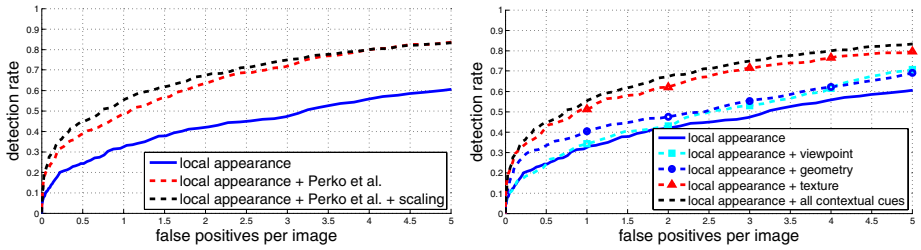
5.1 Scale Extension

To evaluate the results of the scale extension to [9] (see Sec. 4.1), we used the same data set as in the original paper, the Ljubljana urban image data set[3], and compared them to the initial results. The first result is, that the positive feature vectors are more similar (smaller standard deviation) compared to the original method. This indicates that the contextual information grasps a more similar semantic representation, when scaling the regions according to the object's size. The second result is an increase of the detection rate. Using the Seemann *et al.* detector [23] the increase is 4.1%, i.e. a relative increase of 19.2% over the original method. For the detections using Dalal and Triggs detector [2] the increase is 1.3%, i.e. relative increase of 22.3%. These numbers are calculated at a fixed rate of 2 FPPI. Fig. 6(a) shows the detection curves for the original approach and using the scale extension for the Seemann *et al.* detections. Fig. 6(b) visualizes the contributions of the three contextual cues to the final result, where the cue based on texture benefits most using the scale extension.

5.2 Probabilistic Framework

Darmstadt urban image data set. To test the new framework we collected a demanding image data set containing 1572 images of the city of Darmstadt[4] with a resolution of 1944×2896 pixels each. The images were downsampled to 972×1448 pixels for our evaluation. 4133 pedestrian were manually labeled and used as ground truth in the experiments. Each pedestrian is defined by the corresponding bounding box, where the whole object is inside. The bounding boxes have a fixed aspect ratio of $1 : 2$, centered at the object. For the small scale object detection task a subset of 121 images were taken (each 13^{th} image) and

[2] http://vicos.fri.uni-lj.si/research/visual-context/
[3] http://vicos.fri.uni-lj.si/luis34/
[4] http://vicos.fri.uni-lj.si/duis131/



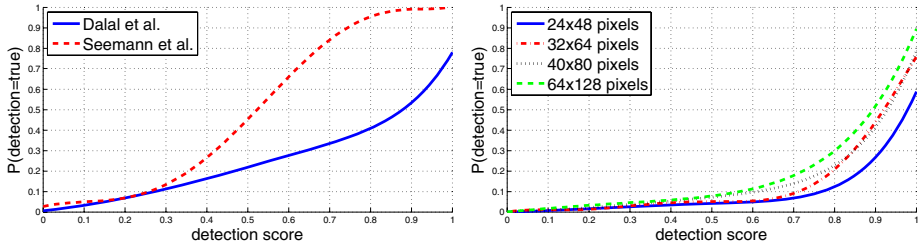
(a) (b)

Fig. 6. Detection rate curves for the scale extension. (a) Comparison of the original approach of [9] with our proposed scale extension. At 2 FPPI the original method boosted the performance by 21.3%, while the new approach increased it by 25.4%, an relative increase of 19.2%. At lower FPPI rates the boost is even more significant. (b) Contribution of the three contextual cues to the final result.

all pedestrians were labeled, down to 12×24 pixels, resulting in 661 pedestrians. This subset is called *sub13* in the rest of the paper.

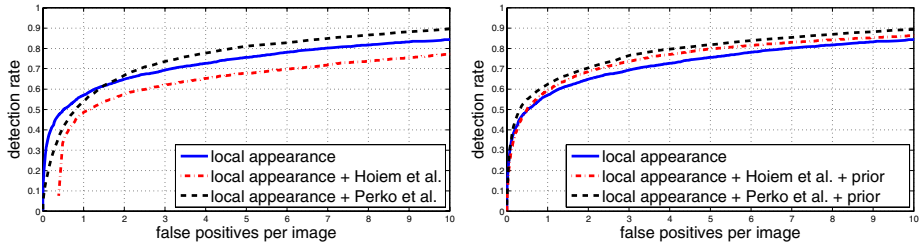
Object detectors. As seen in Sec. 5.1 and also addressed in [9] a weak object detector can easier be boosted using contextual information than a detector which gives very good results in the first place. As we aim for the more difficult task, we show that even the results of the best detectors can be significantly improved using visual context. Therefore, we use a detector based on Dalal and Triggs [2], which is one of the best pedestrian detectors currently available.

Prior probabilities of the object detectors. Fig. 7(a) shows the prior probabilities for the detectors in [2] and [23]. It is obvious, that the initial detection scores are rather different from the observed probabilities and that the prior probabilities vary for the two detectors. To experimentally prove our claim, that smaller objects are more difficult to detect than larger objects, we trained our own version of a *histogram of gradients (HoG)* based detector for different object sizes. Four detectors are trained for 24×48, 32×64, 40×80 and 64×128 pixels. Each detector then collects detections within its range, i.e. the first one collects detections with a height from 1 to 63 pixels, the next from 64 to 79 pixels and so forth. The four prior probability functions are shown in Fig. 7(b). As expected, the probability of a detection being correct is higher for larger objects. For objects larger than 128 pixels in height a detection score of 1 indicates that the detection is correct with 89%, while for smaller objects up to 63 pixels the same score indicates a correctness of only 59%. These prior probabilities are used to re-rank the detections. Like above, the initial detection score is quite different from the observed probabilities. For example scores up to 0.5 only indicate a true detection with less than 10% (see Fig. 7(b)). Therefore, all algorithms which take the initial detector's score within a probabilistic framework yield inaccurate results. The new detection score which corresponds to the observed probabilities is $p_a(L)$ and should be used in the approaches in [10,8,11,9].



(a) (b)

Fig. 7. Prior probability functions given for different detectors. (a) For Dalal and Triggs detector [2] and for Seemann *et al.* detector [23] and (b) for our HoG-based detector trained for four different object sizes.



(a) (b)

Fig. 8. Detection rate curves. (a) Plotted for the original approach of [8] and [9] and (b) for our proposed extensions using the prior probabilities and contextual weighting. While the original methods decrease the accuracy, they yield good results when incorporating the proposed extensions.

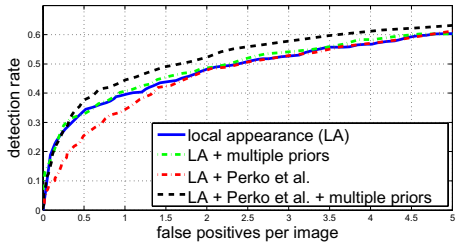


Fig. 9. Detection rate curves plotted for the *sub13* data set. Using multiple prior functions increase the performance of the local appearance-based object detection and of contextual inference. The detection rate is scaled to 0.7.

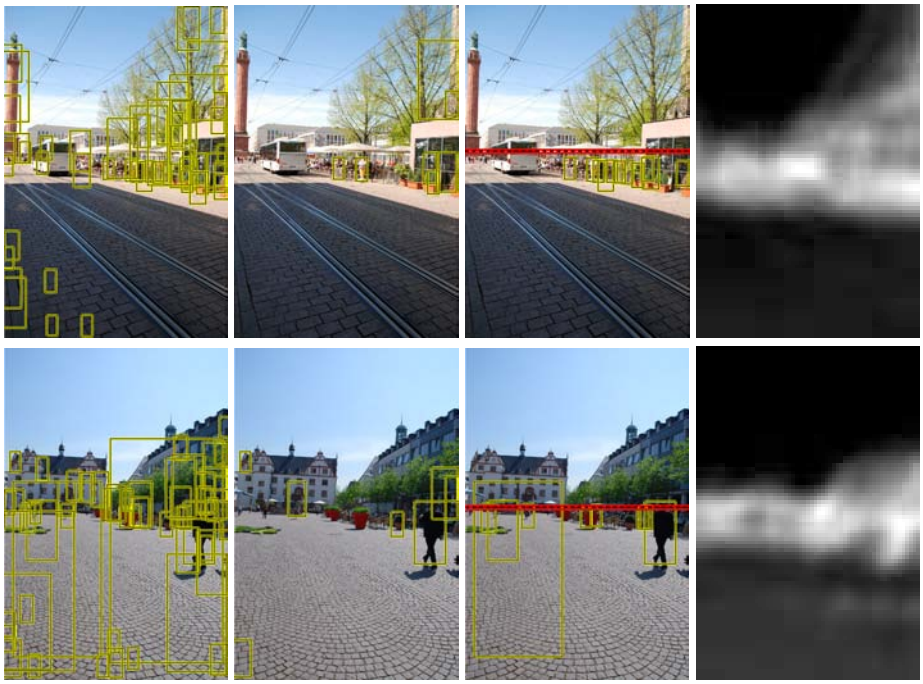
Results using visual context. The original results together with the results using our proposed extensions are given in Fig. 8 for [8] and [9]. With our extensions included both methods are able to increase the initial object detection performance, with an average boost of about 4% for [8] and 7% for [9]. Similar

results are achieved using the *sub13* data set using multiple prior functions, one for each trained detector. Results are given in Fig. 9, where we compare the initial detector's performance with our concept of multiple prior functions, once with and without using contextual inference. As expected the original method [9] performs poorly, while the performance increases when incorporating these probabilities.

6 Discussion

With our extension presented in Sec. 4 the two methods [8,9] for performing visual context aware object detection are improved. However, the increase of the detection rate is on average only about 4% for [8] and 7% for [9] (see Fig. 8). 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 within Fig. 10. In (a) two images of our data set are shown with all object hypotheses marked, and in (b) all detections with a score $p_a(L) > 0.5$. In (c) the horizon estimate from [8] is visualized with the 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 (d) the location priors from [9] are shown for geometry features (shown for the scale $s = 1$, cf. Fig. 4). These priors are robust estimates, however they will only down-rank a few detections with a high 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 and suit to the scene in terms of their position and size. Such hypotheses cannot be rejected or down-ranked by visual contextual information. Another aspect is the way how the contextual information is integrated with local appearance-based object detection. In Eq. (1) the prior probability of the object detector and a contextual weighting is introduced. However, the dependencies of the individual contextual scores and the object detection score are not modeled. Therefore, the next step would be to estimate the conditional probability density function of all cues, which could then be used to increase the overall performance.

To put it in simple words: 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 weak local appearance in many cases. Due to this aspects, the boost of the detection rate is limited using visual context as an additional cue. However, we assume that computer vision researcher will come up with more robust object detection methods based on local appearance. Visual context could then be used to prune the few out-of-context hypotheses with high detection score and to limit the search space for the detection.



(a) (b) (c) (d)

Fig. 10. 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 [8] and (d) focus of attention using geometry features [9]. Best viewed in color.

7 Conclusion

Visual context provides cues about an object's presence, position and size within the observed scene, which are used to increase the performance of object detection techniques. However, state-of-the-art methods [8,9] for context aware object detection could decrease the initial performance in practice, where we discussed the reasons for failure. We proposed a concept that overcomes the limitations, using the prior probability of the object detector and an appropriate contextual weighting. In addition, we presented an extension to state-of-the-art methods [11,9] to learn scale-dependent visual context information and showed how this increases the initial performance. The methods and our proposed extensions were compared on a novel demanding database, where the object detection rate was increased by 4% to 7% depending on the method used.

Acknowledgements. We thank all the authors who made their source code or binaries publicly available, so that we avoided painful re-implementation. In particular we thank Derek Hoiem, Navneet Dalal and Edgar Seeman. This

research has been supported in part by the following funds: Research program Computer Vision P2-0214 (RS), EU FP6-511051-2 project MOBVIS and EU project CoSy (IST-2002-004250).

References

1. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. Conf. Comp. Vis. Pattern Recog. (December 2001)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. Conf. Comp. Vis. Pattern Recog., vol. 2, pp. 886–893 (June 2005)
3. Palmer, S.E.: The effects of contextual scenes on the identification of objects. Mem. Cognit. 3, 519–526 (1975)
4. Biederman, I.: Perceptual Organization. In: On the semantics of a glance at a scene, pp. 213–263. Lawrence Erlbaum, Mahwah (1981)
5. Bar, M.: Visual objects in context. Nat. Rev. Neurosci. 5, 617–629 (2004)
6. Aminoff, E., Gronau, N., Bar, M.: The parahippocampal cortex mediates spatial and nonspatial associations. Cereb. Cortex 17(7), 1493–1503 (2007)
7. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of attention in natural scenes: The role of global features on object search. Psychol. Rev. 113(4), 766–786 (2006)
8. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. In: Proc. Conf. Comp. Vis. Pattern Recog., vol. 2, pp. 2137–2144 (June 2006)
9. Perko, R., Leonardis, A.: Context driven focus of attention for object detection. In: Paletta, L., Rome, E. (eds.) WAPCV 2007. LNCS, vol. 4840, pp. 216–233. Springer, Heidelberg (2007)
10. Torralba, A.: Contextual priming for object detection. Int. J. Comput. Vision 53(2), 153–167 (2003)
11. Bileschi, S.M.: StreetScenes: Towards Scene Understanding in Still Images. PhD thesis, Massachusetts Institute of Technology (May 2006)
12. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in Cognit. Sci. 11(12), 520–527 (2007)
13. 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 (September 2005)
14. Wolf, L., Bileschi, S.M.: A critical view of context. Int. J. Comput. Vision 69(2), 251–261 (2006)
15. Itti, L., Koch, C.: Computational modeling of visual attention. Nat. Rev. Neurosci. 2(3), 194–203 (2001)
16. Rasolzadeh, B., Targhi, A.T., Eklundh, J.O.: An Attentional System Combining Top-Down and Bottom-Up Influences. In: Paletta, L., Rome, E. (eds.) WAPCV 2007. LNCS, vol. 4840, pp. 123–140. Springer, Heidelberg (2007)
17. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision 42(3), 145–175 (2001)
18. Torralba, A., Sinha, P.: Statistical context priming for object detection. In: Proc. Int. Conf. Computer Vision, vol. 1, pp. 763–770 (July 2001)
19. Torralba, A.: Contextual modulation of target saliency. In: Neural Inf. Proc. Systems, vol. 14, pp. 1303–1310 (2002)

20. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: Proc. Int. Conf. Computer Vision, vol. 1, pp. 654–661 (October 2005)
21. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Patter. Anal. Mach. Intell. 24(8), 1026–1038 (2002)
22. 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)
23. Seemann, E., Leibe, B., Schiele, B.: Multi-aspect detection of articulated objects. In: Proc. Conf. Comp. Vis. Pattern Recog., vol. 2, pp. 1582–1588 (June 2006)