Toward Object Recognition with Proto-Objects and Proto-Scenes

Fabian Nasse, Rene Grzeszick and Gernot A. Fink
Department of Computer Science, TU Dortmund
{*fabian.nasse, rene.grzeszick, gernot.fink*}*@udo.edu*

Keywords: Object-recognition, visual attention, bottom-up detection, proto-objects, proto-scenes

Abstract: In this paper a bottom-up approach for detecting and recognizing objects in complex scenes is presented. In contrast to top-down methods, no prior knowledge about the objects is required beforehand. Instead, two different views on the data are computed: First, a GIST descriptor is used for clustering scenes with a similar global appearance which produces a set of Proto-Scenes. Second, a visual attention model that is based on hiearchical multi-scale segmentation and feature integration is proposed. Regions of Interest that are likely to contain an arbitrary object, a Proto-Object, are determined. These Proto-Object regions are then represented by a Bag-of-Features using Spatial Visual Words. The bottom-up approach makes the detection and recognition tasks more challenging but also more efficient and easier to apply to an arbitrary set of objects. This is an important step toward analyzing complex scenes in an unsupervised manner. The bottom-up knowledge is combined with an informed system that associates Proto-Scenes with objects that may occur in them and an object classifier is trained for recognizing the Proto-Objects. In the experiments on the VOC2011 database the proposed multi-scale visual attention model is compared with current state-of-the-art models for Proto-Object detection. Additionally, the the Proto-Objects are classified with respect to the VOC object set.

1 Introduction

Classifying objects in images is useful in many ways: Systems can learn about their environment and interact with it or provide detailed information to a user, e.g., in augmented reality applications. A precondition for classifying objects in a complex, realistic scene is the detection of objects that may be of further interest. This task usually requires detailed knowledge about the objects, e.g., by creating a model for every object category; cf. (Felzenszwalb et al., 2010). Typically, these models are moved over the scene in a sliding window approach. Such approaches have two disadvantages: First, object detection is computationally intensive and also a very specialized task, if a large set of possible objects is considered. Second, creating various object models typically requires tremendous amounts of labeled data.

In this paper we propose a more general approach using bottom-up techniques that do not require prior knowledge about the object classes at the detection stage. Basic information about a scene is gained by computing its *GIST* (Oliva et al., 2006). GIST refers to scene descriptors that model the coarse human perception of complex scenes. Within milliseconds a human observer is able to perform a brief categorization of a scene, for example, decide between indoor and outdoor scenes. In a first step scenes with similar GIST descriptions are clustered and described by a set of representatives. We refer to these representatives as *Proto-Scenes* since no further knowledge about them can be inferred. At object level a visual attention is applied for detecting Regions of Interest that are likely to contain an object, a so-called *Proto-Object*. For computing the visual attention a saliency detector that is based on the priciples of feature integration, object-based saliency and hierarchical segmentation is introduced.

In the resulting scene description it is not known what a scene shows, but whether it is similar to other scenes and where objects of interest may occur in this scene. Since no prior knowledge is used, this scene description is less accurate than the results of a specialized object detector but, therefore, it can be used as a layer of abstraction that can efficiently be computed and later be combined with an informed object recognizer. This is an important step toward the unsupervised analysis of complex scenes. Proto-objects can be found regardless of the actual instance and recognizers could be trained automatically, e.g.,

[0]The authors would like to thank Axel Plinge for his helpful suggestions.

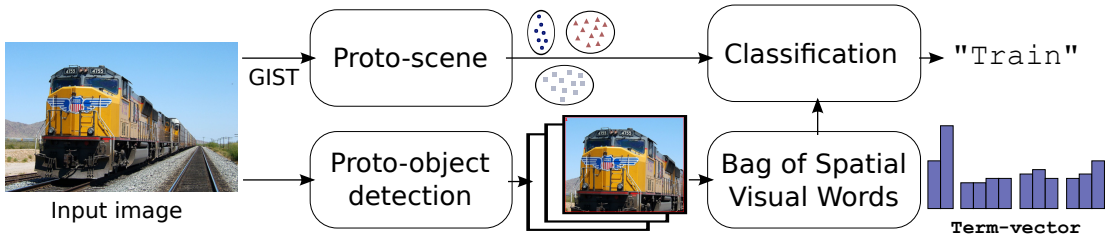

Figure 1: Overview of the proposed method: Proto-scenes are computed using a GIST representation. Proto-object regions are detected within the image. These regions are then described by a Bag-of-Features using Spatial Visual Words. The Proto-Scenes and regions are then evaluated in a classification step in order to obtain a class label. The "train" image is taken from the VOC2011 Database (Everingham et al., 2011).

from web sources, and used for evaluating the scene-contents with respect to different object classes. In this paper the quality of such a scene representation will be evaluated and, therefore, the VOC database that contains objects in complex scenes is used for evaluation.

After defining a set of object categories, a prior-probability for objects to occur in a given Proto-Scene is computed. Then, the Proto-Object regions can be used as input images for an object classifier. For the classification of Proto-Objects a Bag-of-Features representation using *Spatial Visual Words* (Grzeszick et al., 2013) is combined with a random forest. A Spatial Visual Word includes coarse spatial information about the position of a Visual Word within the Proto-Object region at feature level. The region itself holds the spatial information about the position within the scene. In (Grzeszick et al., 2013) it has been shown that this representation is more compact than the well known Spatial Pyramids (Lazebnik et al., 2006). It is therefore more suitable for an unsupervised approach that aims at recognizing arbitrary objects with low computational costs.

Summarizing, the contribution of this paper is two fold: 1. A novel Proto-Object detector that is based on a visual attention model is presented. It applies the priciples of feature integration at multiple scales to segmented Proto-Object regions. 2. The possibility of combining the computed detections and scene level information that were obtained completely unsupervised with a Bag-of-Features based object classifier is evaluated.

2 Related Work

A very elementary representation of a scene is its GIST (Oliva, 2005). The idea is to model the human ability to gather basic information about a scene in a very short time and to obtain a low dimensional representation for complex scenes. A common GIST descriptor, the Spatial Envelope, has been introduced by Olivia and Torralba in (Oliva et al., 2006). It models the dominant spatial structure of a scene based on perceptual dimensions like naturalness, openness, roughness or expansion. These are estimated by a spectral analysis with coarsely localized information. The advantages of using scene context for object detection has been shown in (Divvala et al., 2009). The recognition results of a part based object detector could be significantly improved by combining it with scene information.

Visual attention models steadily gained popularity in computer vision in recent years (Borji and Itti, 2013). Generally, they can be divided into two categories, top-down and bottom-up models. While top-down models are expectation- or task-driven, bottom-up models are based on characteristics of the visual scene. For bottom-up models several measures have been used in order to find salient image content, for example, center-surround histograms (Cheng et al., 2011; Liu et al., 2011), luminance contrast (Zhai and Shah, 2006) and frequency based measures (Achanta et al., 2009; Hou and Zhang, 2007). Locating objects by means of saliency detection is based on the assumption that there is a coherence between salient image content and interesting objects (Elazary and Itti, 2008). The presented approach is a bottom-up approach that explicitly follows the assumption that visual interest is stimulated by objects rather than single features as shown in (Cheng et al., 2011). Since there is no need for prior knowledge, such models can be used to re-evaluate top-down methods (Alexe et al., 2012) or integrated into methods for generical object detection as shown in (Nasse and Fink, 2012) using a region-based saliency model. Other approaches propose feature integration based attention models with subsequent use of image segmentation (Walther et al., 2002; Rutishauser et al., 2004).

For object classification Bag-of-Features representations are known for producing state-of-the-art results; cf. (Chatfield et al., 2011). Local appearance features, e.g. SIFT (Lowe, 2004), are extracted from a set of training images, clustered and quan-

tized. A fixed set of representatives, the so-called *Visual Words*, are used for describing the features. An image is then represented by a vector containing the frequencies of the occuring Visual Words, the *term-vector*. The Bag-of-Features discards all spatial information which originally was a major shortcoming when applying this approach to object recognition. Hence, context models that re-introduce coarse spatial information are able to significantly improve the classification results. The most common example are Spatial Pyramids (Lazebnik et al., 2006) which subdivide the image and create a term-vector for each region. Lately, the Spatial Pyramid representations became increasingly high dimensional using up to eight regions with 25.000 Visual Words each yielding 200.000 dimensional term-vectors (Chatfield et al., 2011). While these high dimensional models yield superior classification rates they also make it more difficult to handle large amounts of data. In order to reduce the dimensionality the presented approach computes Spatial Visual Words that directly encode spatial information at feature level (Grzeszick et al., 2013).

3 Bottom-up Recognition

The proposed method for bottom-up object recognition consists of three major steps that are also illustrated in Figure 1: First, given an input image, its Proto-Scene category is determined and Proto-Object regions are detected in the image. Then, each Proto-Object region is represented by a Bag-of-Features representation using Spatial Visual Words. Finally, the feature representations are used for computing a probability for an object to be present in the scene by using a random forest. The region based probabilities are weighted by a prior based on the Proto-Scene category.

3.1 Proto-Scenes

The most basic information that can be obtained about different scenes is a global similarity. Hence, the scenes are clustered, using Lloyd's algorithm (Lloyd, 1982), and represented by a set of M Proto-Scenes S_m.

For the global description the GIST of a scene is computed. Namely, the color GIST implementation described in (Douze et al., 2009) which resizes the image to 32×32 pixels and subdivides it into a 4×4 grid. This grid is used for computing the Spatial Envelope representation as introduced by Olivia and Torralba (Oliva et al., 2006). In an informed system, it is then possible to estimate the probability for an object of class Ω_c to occur in an image I_k based on the Proto-Scenes S_m:

$$P(\Omega_c|I_k) = \sum_{m=1}^{M} P(\Omega_c|S_m)P(S_m|I_k) \tag{1}$$

Here, the probability for an object class to occur in a Proto-Scene is estimated from a set of training images using Laplacian smoothing

$$P(\Omega_c|S_m) = \frac{1 + \sum_{I_k \in S_m} \sum_{O \in I_k} q(O,\Omega_c)}{C + \sum_{I_k \in S_m} \sum_{O \in I_k} \sum_{i=1}^{C} q(O,\Omega_i)} \tag{2}$$

for all classes C with

$$q(O,\Omega_c) = \begin{cases} 1 & O \in \Omega_c \\ 0 & else \end{cases}. \tag{3}$$

A Bayes classifier is trained on the Proto-Scenes that were uncovered by the clustering on the training images. It can then used for computing $P(S_m|I_k)$ for a given image I_k.

3.2 Detection

Regions of Interest are detected in an image using a bottom-up process that is based on visual attention models. The saliency of a region compared to the rest of the image is evaluated. Those regions are refered to as Proto-Object regions. The term is based on the fact that regions with a high visual interest are likely to contain an object but the content of the region is not identified yet. It does not become an actual object before the recognition process. The proposed saliency detector is based on three principles: feature integration, object-based saliency and hierarchical segmentation. Applying the theory of feature integration for a computational attention model was first proposed in (Itti et al., 1998) and is widely recognized. The theory suggests that in the pre-attentive stage the human brain builds maps of different kinds of features which compete for saliency and are subsequently integrated before reaching the attention of the spectator. Object-based saliency assumes that attention is stimulated by objects rather then by single features.

The presented approach combines the concept of object-based saliency with the region-contrast method proposed in (Cheng et al., 2011). First, a set of disjoint regions is determined using segmentation. Saliency values are then determined for each region by comparing a region with all other regions of the image. Hence, in contrast to other saliency approaches the method computes saliency values for regions instead of pixels.

The main disadvantage of this approach is that it is hardly possible to detect objects on a large scale

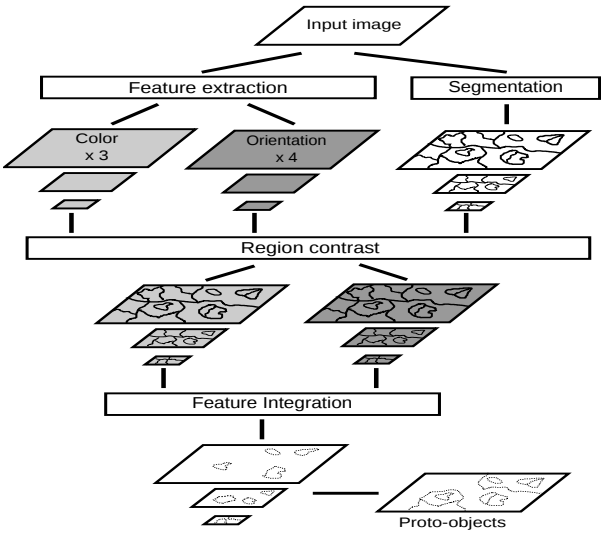

Figure 2: Overview of the saliency based detection approach. The principles of feature integration, object-based saliency and hierachical segmentation are combined for detecting Proto-Objects in a scene.

of different sizes. Therefore, the saliency algorithm is improved by using hierarchical segmentation in a scale space as proposed in (Haxhimusa et al., 2006). For computing the next image of the scale space I_{l+1} the image I_l is convoluted with a Gaussian function G of variance $\sigma = 0.5$:

$$I_{l+1}(x,y,\sigma) = G(x,y,\sigma) * I_l(x,y) \tag{4}$$

The overall concept is illustrated in Figure 2. In order to integrate a set of different features, three feature maps for color and four maps for orientation (4 bins with $\pm 45^\circ$) are computed, which is comparable to the approach described in (Itti et al., 1998). All feature maps are also computed in a scale space with three scale levels yielding 21 feature maps. The region-contrast method is applied on each of them independently, producing a set of regions with different saliency values for each feature map. The feature maps are integrated based on regions rather than pixel-wise. Since the saliency values in the different feature maps vary they need to be normalized. This is achieved by weighting each region with

$$w = (\mathcal{M} - \overline{m})^2, \tag{5}$$

where $\mathcal{M}$ is the highest saliency value in the map and $\overline{m}$ is the average saliency over all regions.

From the overall result a predefined number of the most salient regions is extracted and considered as Proto-Object regions. These are then processed by the object recognizer. Note, that the regions of the Proto-Objects can overlap if they are extracted from different layers.

3.3 Feature representation

Besides the visual interest, no additional knowledge about the Proto-Object regions is available. In order to allow for a classification of the Proto-Objects it is necessary to extract features from these regions. A Bag-of-Features representation using Spatial Visual Words is computed for each detected region. A Spatial Visual Word includes spatial information directly at feature level so that it is incorporated into the Bag-of-Features and does not need to be re-introduced. Also, redundancies that occur in the high dimensional representation of Spatial Pyramids are removed while keeping the benefits of incorporating spatial information (Grzeszick et al., 2013).

First, local appearance features are extracted from a set of training samples. In the following, densly sampled SIFT features (Lowe, 2004) are used. Then, unlike recent Bag-of-Features approaches, these appearance features are enriched by a spatial component at feature level as introduced in (Grzeszick et al., 2013). Spatial features s_i are appended to the descriptor so that similar appearance features in the same spatial region are clustered. In this paper quantized xy-coordinates based on 2×2 regions are considered for the spatial feature, which is similar to the Spatial Pyramid (Lazebnik et al., 2006). In this case the four regions can, for example, be represented by the coordinates $[(0;0);(0;1);(1;0);(1;1)]$. In order to increase the influence of the spatial component, the 128 dimensional SIFT descriptor a is divided by the average descriptor length, so that the sum of all dimensions becomes approximately one. Thus, a new feature vector v consisting of the appearance feature a

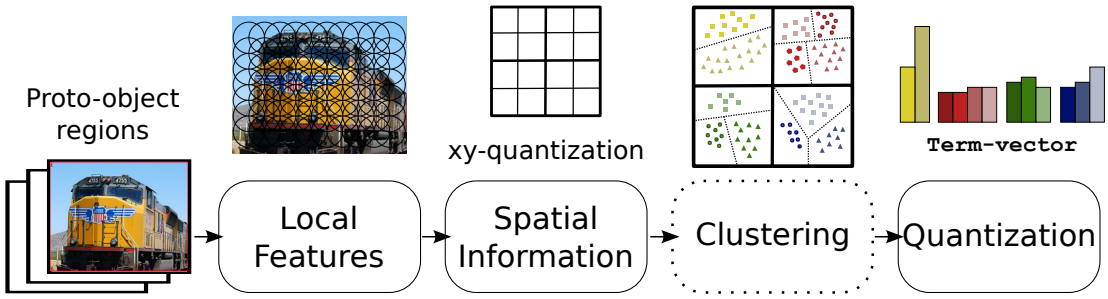

Figure 3: Overview of the feature representation: Given a Proto-Object region, local appearance features (e.g., SIFT) are extracted from the image based on a densly sampled grid. A spatial measure is used for combining the appearance features with spatial features, in this case, quantized *xy*-coordinates. During the training the modified descriptors from all training images are clustered in order to form a Spatial Visual Vocabulary that holds the important information of each spatial region. The features of each Proto-Object are quantized with respect to that vocabulary and represented by a set of Spatial Visual Words. The "train" image is taken from the VOC2011 Database (Everingham et al., 2011).

and the spatial feature vector s is constructed by:

$$v = (a_0, ..., a_{128}, s_0, s_1)^T \tag{6}$$

All features are then clustered to form a set of representatives. A single representative is refered to as a *Spatial Visual Word* and to the complete set as the *Spatial Visual Vocabulary*. For clustering the generalized Lloyd algorithm is applied (Lloyd, 1982). The features of each object from a training set are quantized with respect to the vocabulary and the sample is represented by a term-vector of Spatial Visual Words.

3.4 Classification

The classification is performed as a two step process. The Proto-Scenes are incorporated for computing $P(\Omega_c|I_k)$ as described in section 3.1. A random forest is trained on the Bag-of-Features representations from annotated object samples in order to estimate the probabilites $P(\Omega_c|R_j)$ of a region R_j to contain an object of class Ω_c. The probability of an object of class Ω_c to be present in a region R_j of image I_k is then defined by

$$P(\Omega_c|R_j, I_k) = P(\Omega_c|R_j)^{\alpha} P(\Omega_c|I_k) \; . \tag{7}$$

A weighting term α was introduced in order to account for different confidences of the Proto-Object and Proto-Scene classification. Experiments showed that there is a local optimum for $\alpha = 4$. In order to predict whether an object is present in a scene, the maximum probability of all regions R_j is computed.

The advantage of this approach is that two different views on the data, the very coarse scene representation and the more detailed object level information, are combined with each other.

4 Experiments

The method was evaluated on the VOC2011 database. The database contains 11,540 images of complex scenes with one or more objects that need to be recognized. In the following the quality of the detections is compared to state-of-the-art visual attention methods. In addition, the detections and the Proto-Scene information is used in order to classify the objects.

4.1 Detection experiments

The bottom-up object detector computes a set of Proto-Object regions that are completely independent of any task. Hence, the goal of the detection experiments is to evaluate the quality of the bottom-up detections for a given task. Some regions may contain objects that are of further interest while other regions may contain visually interesting objects that are not related to the task. The detector is applied on the VOC2011 dataset using the same overlap criterion that is used for the VOC-challenge:

$$a_0 = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} \tag{8}$$

where B_p is the detected region and B_{gt} is the ground-truth bounding-box of an object. Typically, an object is counted as detected if $a_0 \geq 0.5$.

In the experiments the maximum number of Proto-Object regions per image that are computed by the detector is limited by a parameter. Then, the ratio of detected regions compared to all objects of interest that are annotated in the dataset is determined. Figure 4 shows the results of the proposed approach compared with the single-scale region-contrast method (Cheng et al., 2011) and a feature integration approach presented in (Walther et al., 2002). The latter

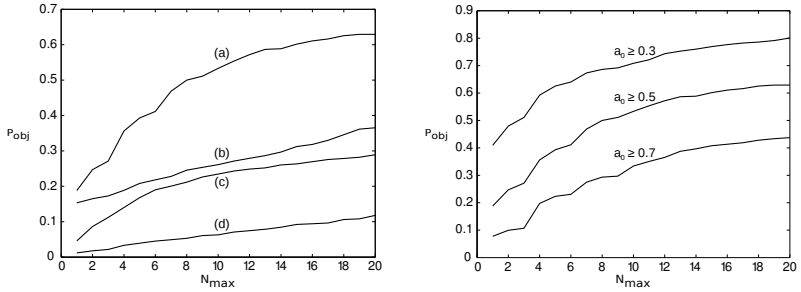

Figure 4: Results of the detection experiments. The graphs show the ratio of the detected objets, P_{obj}, over the maximum number of Proto-Object regions per image, N_{max}. Left: Comparison of different methods using the locating precision criterion $a_0 \geq 0.5$. (a) Proposed method. (b) region-contrast (Cheng et al., 2011) (c) feature integration (Walther et al., 2002) (d) randomly selected regions. Right: Results of the proposed method for different precision criteria.

is based on (Itti et al., 1998) and is using segmentation in the post processing in order to extend the most salient locations to regions. The results show that the proposed method clearly outperforms state-of-the-art visual attention approaches. Using only the most salient region computed by all approaches already shows that the proposed multi-scale approach detects the objects in the VOC2011 database more accurately than both other methods. Furthermore, increasing the number Proto-Object regions that are considered increases the performance improvement. With 20 Proto-Object regions about 20% more objects than with the single-scale region-contrast are detected.

The results also show that half of the objects of interest are among the ten most salient regions per image and are located with decent precision ($a_0 \geq 0.5$). Hence, in comparison with sliding-window detection approaches the proposed method is also computationally very efficient. Additionally, different overlap criteria are evaluated showing that more than 40% of the objects are detected with an overlap of 70% or more. When loosening the overlap criterion to 30% up to 80% of the objects are detected.

Examples for the detection approach are shown in Figure 5. They illustrate the difficulties of bottom-up detection. In the first two examples (car & cat) the object is split up in different parts at the finer scales. This also shows the advantages of using a multi-scale approach and explains the strong performance improvements compared to single-scale methods. In the third example (bird) the correct detection is at the finest scale. However, there is a more salient region that is created by noise at the bottom of the branch.

4.2 Recognition experiments

Using the bottom-up information that was obtained about the images in the VOC Database the next experiments combine the Proto-Scenes and Proto-Objects with an informed system that allows for object recognition as described in section 3.4. For the evaluation a confidence value for an object to be present in a scene is computed. The confidence is used in order to compute the average precision over a precision-recall curve for each object category; cf. (Everingham et al., 2011).

The annotated ground truth from the VOC2011 training dataset is used for modelling the 20 VOC object categories. Both, the ground truth objects and the detections are computed with a margin of 15% around the bounding box in order to catch a glimpse of back-

Category	10 Regions	10 Regions & Proto-scenes
Aeroplane	55.4%	**56.0%**
Bicycle	**33.4%**	33.3%
Bird	21.7%	**26.3%**
Boat	**26.1%**	23.4%
Bottle	**12.7%**	10.7%
Bus	51.3%	**53.0%**
Car	26.6%	**27.3%**
Cat	**42.0%**	41.1%
Chair	18.3%	**23.2%**
Cow	11.9%	**12.6%**
Diningtable	18.9%	**21.2%**
Dog	32.7%	**35.4%**
Horse	**25.5%**	24.2%
Motorbike	38.5%	**40.6%**
Person	60.0%	**60.9%**
Potted plant	5.9%	**7.6%**
Sheep	**21.8%**	19.8%
Sofa	22.3%	**22.6%**
Train	**36.7%**	36.3%
Tv-Monitor	35.5%	**37.3%**
mAP	29.9%	**30.6%**

Table 1: Average precision on the VOC2011 using a vocabulary size of 1.000 and 2×2 *xy*-quantization. Left column: 10 Proto-Object regions. Right column: Proto-Scene information obtained by 30 clusters is also incorporated.

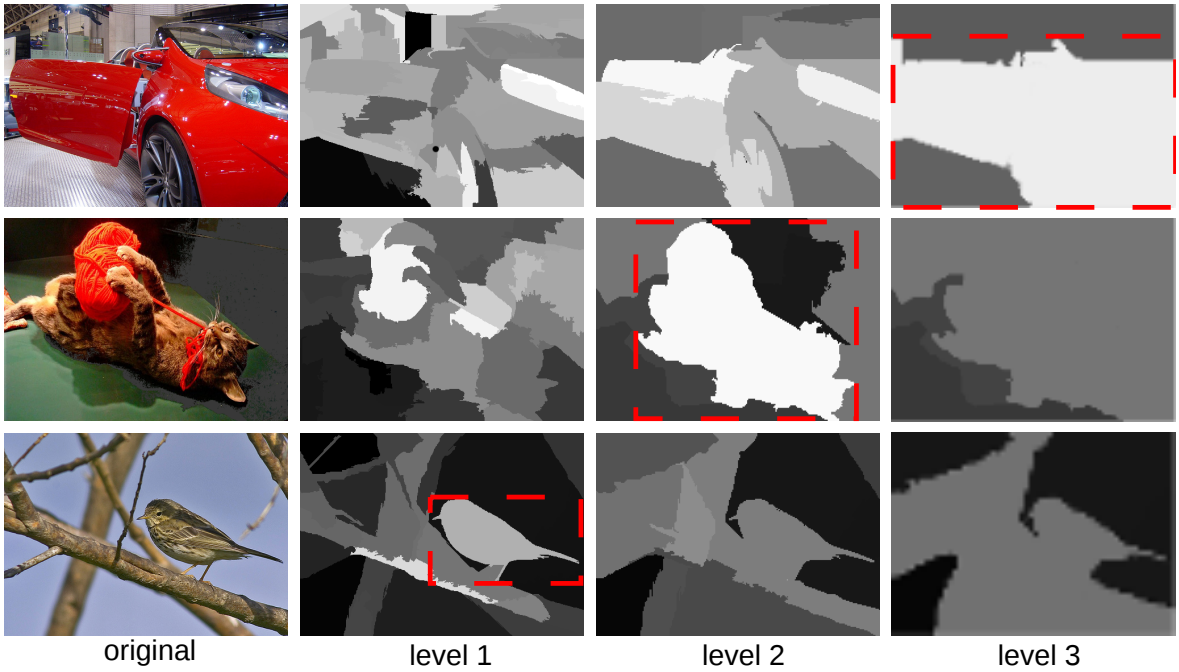

Figure 5: Detection of salient objects at different scale levels. The examples demonstrate the advantage of introducing hierarchical Proto-Object detection. Small objects (bird) will be detected on a higher reolution, i.e. level 1, while middle-sized (cat) and large objects (car) are detected on coarser scales, i.e. level 2 and 3, respectively. The images in the left column are taken from the VOC2011 database. This graphic is best viewed in color.

ground information that might be useful for the classification. The Bag-of-Features representations are computed using densly sampled SIFT features with a step width of 3px and bin sizes of $4, 6, 8$ and 10px, which is similar to the approach described in (Chatfield et al., 2011).

The results of the recognition experiments are shown in Table 1. In these experiments the ten most salient regions were considered in order to detect a high number of Proto-Objects while keeping a low false positive rate. For the Bag-of-Features a Spatial Visual Vocabulary with 1,000 Visual Words is computed and combined with spatial information from a 2×2 xy-quantization. As expected, the results are below state-of-the-art top-down approaches. There are mainly three reasons for this: First, the detected Proto-Object regions are not always completely accurate. The spatial information is distorted by translations and cropping. This is also why more detailed spatial information does not yield an improvement. The second reason is that some objects get rarely detected, since they do not show any visual interest. Note that about 55% of the object in the dataset are detected with an overlap $a_0 \geq 0.5$. While this is a good results for bottom-up detection it makes the classification very challenging. Third, some objects such as bottles are comparably small so that it is not always possible to compute a meaningful statistical representation like the Bag-of-Features on these regions. Large and visually more interesting objects like Planes, Busses or Persons show higher recognition rates. The results of these classes are, for example, comparable with the scene level pyramid models using 4.000 Visual Words described in (Chatfield et al., 2011), e.g. 60.6% for airplanes and 50.4% for busses.

The additional knowledge obtained from the Proto-Scenes improves the classification results. Especially categories that occur mostly in the same environment, such as airplanes or birds benefit from the Proto-Scene information. For the results presented in Table 1 30 Proto-Scenes were used. Note that this number of Proto-Scenes does not necessarily represent the number of natural scenes that occur in the dataset.

5 Conclusion

In this paper a bottom-up approach for object recognition based on proto-scenes and proto-object detection was presented. No prior knowledge about the object categories is required for creating an abstract representation of scenes and objects. This representation can later be used by an informed system for object classification.

The detection of real world objects with saliency based techniques has been evaluated showing that the presented multi-scale approach outperforms state-of-the-art visual attention models. The experiments confirmed that bottom-up recognition is more difficult but it is also easier to apply to arbitrary objects and more efficient than specialized detectors that need to be trained and applied separatly in a sliding window approach. These properties and the independence of annotations for most parts is an important step toward automated object recognizer training. It has also been shown that promising recognition rates can be obtained for some object categories on the VOC2011 database.

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1597–1604.

Alexe, B., Deselaers, T., and Ferrari, V. (2012). Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2189–2202.

Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207.

Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*.

Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., and Hu, S.-M. (2011). Global contrast based salient region detection. In *IEEE CVPR*, pages 409–416.

Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., and Hebert, M. (2009). An empirical study of context in object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1271–1278. IEEE.

Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., and Schmid, C. (2009). Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 19. ACM.

Elazary, L. and Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, 8(3):1–15.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2011). The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Grzeszick, R., Rothacker, L., and Fink, G. A. (2013). Bag-of-features representations using spatial visual vocabularies for object classification. In *IEEE Intl. Conf. on Image Processing*, Melbourne, Australia.

Haxhimusa, Y., Ion, A., and Kropatsch, W. G. (2006). Irregular pyramid segmentations with stochastic graph decimation strategies. In *CIARP*, pages 277–286.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H.-Y. (2011). Learning to detect a salient object. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(2):353–367.

Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110.

Nasse, F. and Fink, G. A. (2012). A bottom-up approach for learning visual object detection models from unreliable sources. In *Pattern Recognition: 34th DAGM-Symposium Graz*.

Oliva, A. (2005). Gist of the scene. *Neurobiology of attention*, 696:64.

Oliva, A., Torralba, A., et al. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23.

Rutishauser, U., Walther, D., Koch, C., and Perona, P. (2004). Is bottom-up attention useful for object recognition? In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–37–II–44 Vol.2.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., and Koch, C. (2002). Attentional selection for object recognition: A gentle way. In *Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, BMCV '02, pages 472–479, London, UK, UK. Springer-Verlag.

Zhai, Y. and Shah, M. (2006). Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, pages 815–824, New York, NY, USA. ACM.