Determining Patch Saliency Using Low-Level Context

Devi Parikh[1], C. Lawrence Zitnick[2], and Tsuhan Chen[1]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
[2] Microsoft Research, Redmond, WA, USA

Abstract. The increased use of context for high level reasoning has been popular in recent works to increase recognition accuracy. In this paper, we consider an orthogonal application of context. We explore the use of context to determine which low-level appearance cues in an image are salient or representative of an image's contents. Existing classes of low-level saliency measures for image patches include those based on interest points, as well as supervised discriminative measures. We propose a new class of unsupervised contextual saliency measures based on co-occurrence and spatial information between image patches. For recognition, image patches are sampled using a weighted random sampling based on saliency, or using a sequential approach based on maximizing the likelihoods of the image patches. We compare the different classes of saliency measures, along with a baseline uniform measure, for the task of scene and object recognition using the bag-of-features paradigm. In our results, the contextual saliency measures achieve improved accuracies over the previous methods. Moreover, our highest accuracy is achieved using a sparse sampling of the image, unlike previous approaches who's performance increases with the sampling density.

1 Introduction

Determining image patches of high saliency has recently received significant attention. The goal of saliency detection is to identify the image patches that are most informative of the image contents. A standard method for finding these patches is the use of interest point detectors based on local low-level image statistics [1–6]. Another class of saliency measures are discriminative in nature [7–12], where a patch is considered salient if it is informative from a classification perspective. The usefulness of a patch may be based on the mutual information between the presence of the patch and the scene categories [7] or the probability of misclassification of a patch [8]. Using these techniques, a relatively small number of patches can be sampled while still achieving high recognition accuracy.

In this paper, we explore the use of contextual information that is typically used for higher level reasoning for the low-level task of selecting informative or salient patches in an image. We consider a patch to be salient if it is predictive or representative of the other patches in the image. The relationships of image patches are modeled using co-occurrence and spatial information. Unlike

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 446–459, 2008.

previous saliency measures that rely only on local information, our approach incorporates contextual information using the patch statistics across the entire image. For recognition, the sampling of patches is performed using weighted random sampling based on patch saliency. In addition, we propose a sequential sampling approach based on increasing the maximum likelihood of patches given the set of previously selected patches.

Our saliency measure is evaluated within a bag-of-features framework [13–17]. This simple approach has shown good performance in a variety of recognition tasks and allows us to focus on the specific contributions of our paper. The bag-of-features approach consists of three components: a method for sampling patches from an image, a method for assigning the patches to a discrete patch vocabulary, and a method for classifying the resulting global descriptor. In this paper, we only address the first task of patch sampling, and use standard approaches for the other two components. A vocabulary of patch appearances, to which each patch is assigned, is constructed using K-means clustering [18, 19]. Classification is accomplished using an SVM classifier over the histogram of patch assignments [7, 15].

We compare our proposed contextual saliency measures to a variety of existing measures including interest points, discriminative approaches and random sampling. These measures are evaluated on both scene and object recognition tasks. In contrast to previous results that show recognition accuracy increases with the density of the sampling [7], the contextual measures achieve maximal accuracy using a sparse sampling. Moreover, in our experiments the accuracy of using contextual measures with sparse sampling is better than dense sampling using other methods.

The rest of the paper is organized as follows: Previous works are discussed in the following section. We describe our proposed contextual saliency measures in Section 3 and sampling methods in Section 4. In Section 5 we briefly describe the existing saliency measures used for comparison, followed by a description of the experimental setup in Section 6. Results and some discussion are provided in Section 7 and conclusions in Section 8.

2 Previous Work

Several works have explored the role of saliency measures for classification tasks. Nowak *et al.* [7] compare the interest operators to random dense sampling and find that random sampling performs comparable or superior to interest operators. Jurie *et al.* [20], apart from proposing a novel clustering approach to form codebooks, evaluate a discriminative saliency measure used for the feature selection problem. They find that when using smaller codebooks discriminative feature selection can be used to improve accuracies. However, using the full codebook for classification typically resulted in better performance. A related class of works [8, 21, 22] is visual search, where similar notions of saliency are important. The essence of visual search lies in the notion of active exploration,

in which saliency maps are dynamically updated or areas are marked for further exploration.

Most existing approaches [1, 4, 15, 17, 18, 23, 24, 19] for selecting a sparse set of image patches are based on interest point detectors [1–6]. These include those based on edge cornerness [2], difference of Gaussian convolutions [1], stable extremal regions [5] or local entropy [3]. While these measures are useful for obtaining reliable correspondences or matching, they do not relate directly to image understanding via classification or recognition. The strategies of dense sampling [13, 16, 25] or random sampling [12] have shown to provide comparable or even better performance than interest points [7]. Biologically inspired saliency measures following the "feature integration theory" [26] extract regions of the image that stand out from their surrounding as being salient [27, 28]. [3, 29] are based on a similar notion and [30] consider features to be salient if they are rare. While this is a plausible explanation to predict task-independent attention, they do not take into account task-dependencies. Walther *et al.* [31] incorporate such task dependencies and combine the biologically plausible saliency map of [28] with interest point operators [1] to show improved performances.

Using high-level contextual information for better image understanding has received significant attention in recent works [32–40]. Most of these approaches use context as a post-processing step to prune out false positives [32, 33], aid in detection by eliminating unlikely locations of objects [32, 34–37], or ensure semantically consistent labels to regions of an image [32, 33, 38–40].

3 Proposed Contextual Saliency Measures

Our goal is to select a sparse set of image patches that are most informative for classification. We propose that the patches, which are representative or predictive of other patches in the image, are also the patches most useful for classification. We measure the predictiveness of a patch using a contextual saliency measure based on co-occurrence and spatial information. As stated earlier, we examine our measure of saliency within the bag-of-features framework. In this framework, classification is achieved by selecting a set of image patches and assigning them to codewords. A histogram of codewords is then constructed and used for classification. In this paper, we address the first task of selecting image patches. We describe the standard method of K-means for codebook creation and Support Vector Machines for classification in Section 7.

For each image patch x_i, we compute a patch descriptor y_i. This descriptor can vary based on the application and properties of particular datasets. In this paper we examine two descriptors. The first is a 4×4 vector of average color values over a patch. This descriptor is useful for scenarios in which color information is important, such as in scene recognition. For object recognition in which edge information is more useful than color, we use the standard SIFT descriptor [1].

The codebook W consists of m descriptor templates. Each patch x_i in an image is assigned to a codeword w_a in the codebook. These assignments may

be soft or hard with α_{ia} being the probability of patch x_i being assigned to codeword w_a:

$$\alpha_{ia} = p(y_i|w_a) = \frac{1}{Z}\mathcal{N}(y_i; w_a, \sigma_w)$$

$\mathcal{N}$ is the standard normal distribution with mean w_a and variance σ_w. The value of Z is set so that the values of α_{ia} sum to one for all a, i.e. $\sum_{a=1}^{m} \alpha_{ia} = 1$. For hard assignments $\alpha_{ia} = 1$ for the codeword w_a that lies closest to y_i and $\alpha_{ia} = 0$ otherwise.

For each patch x_i in an image, we want to assign a saliency measure $\mathcal{S}_i$. In the following two sections we propose two saliency measures based on contextual information.

3.1 Occurrence-Based Contextual Saliency

Our measures of contextual saliency are based on how well each individual patch can predict the occurrence of other patches in the image. Our first saliency measure $\mathcal{S}^o$ uses co-occurrence information between codewords in images. Given a set of n patches in an image, we define the saliency of a patch x_i equal to the average likelihoods of the image patches conditioned upon y_i.

$$\mathcal{S}_i^o = \frac{1}{n}\sum_{j=1}^{n}\sum_{a=1}^{m} \alpha_{ia} p(x_j|w_a) \tag{1}$$

The value of $p(x_j|w_a)$ is computed by marginalizing over all possible codeword assignments for x_j

$$p(x_j|w_a) = \sum_{b=1}^{m} \alpha_{jb} p(w_b|w_a) \tag{2}$$

The value of $p(w_b|w_a)$ is the empirical conditional probability of observing codeword w_b given the codeword w_a has been observed somewhere in the image. These are learnt through MLE counts from the training images. Given hard assignments of patches to codewords, the two summations over m can be removed from equations (1) and (2).

Computing the above measure can be computational expensive, especially if the codebook size and number of patches is large. One method for reducing the computational complexity is to rearrange equations (1) and (2) as:

$$\mathcal{S}_i^o = \sum_{a=1}^{m} \alpha_{ia} \frac{1}{n}\sum_{j=1}^{n}\sum_{b=1}^{m} \alpha_{jb} p(w_b|w_a) \tag{3}$$

The value $\Phi_a = \frac{1}{n}\sum_{j=1}^{n}\sum_{b=1}^{m} \alpha_{jb} p(w_b|w_a)$ can then be pre-computed for each a, resulting in:

$$\mathcal{S}_i^o = \sum_{a=1}^{m} \alpha_{ia} \Phi_a \tag{4}$$

3.2 Location-Based Contextual Saliency

The previous contextual saliency measure was based solely on co-occurrence information without knowledge of the patch's location. In this section we propose a saliency measure that includes spatial information. The location of a patch in an image is modeled using a Gaussian Mixture Model with $c = 9$ components. For our experiments the Gaussian means are centered in a 3×3 grid evenly spaced across an image with standard deviations in each dimension equal to half the distance between the means. We define the value β_{iu} as the likelihood of x_i belonging to component l_u of the GMM, $u \in \{1, \ldots, c\}$, and $\sum_{u=1}^{c} \beta_{iu} = 1, \forall i$.

Similar to equation (1), we define our location-based contextual saliency measure $\mathcal{S}^l$ as

$$\mathcal{S}_i^l = \frac{1}{n} \sum_{j=1}^{n} \sum_{a=1}^{m} \sum_{u=1}^{c} \alpha_{ia} \beta_{iu} p(x_j | w_a, l_u) \tag{5}$$

The value of $p(x_j|w_a, b_u)$ is computed as

$$p(x_j | w_a, b_u) = \sum_{b=1}^{m} \sum_{v=1}^{c} \alpha_{jb} \beta_{jv} p(w_b, l_v | w_a, l_u) \tag{6}$$

The value of $p(w_b, l_v|w_a, l_u)$ is the empirical conditional probability of observing word w_b at location l_v given word w_a occurred at location l_u. These are learnt through MLE counts from the training images.

Similar to equation (4), we may pre-compute the values

$$\Psi_{au} = \frac{1}{n} \sum_{j=1}^{n} \sum_{b=1}^{m} \sum_{v=1}^{c} \alpha_{jb} \beta_{jv} p(w_b, l_v | w_a, l_u)$$

and find $\mathcal{S}_i^l$ as

$$\mathcal{S}_i^l = \sum_{a=1}^{m} \sum_{u=1}^{c} \alpha_{ia} \beta_{iu} \Psi_{au} \tag{7}$$

Since our proposed saliency measures are dependent on the codeword assignments of other image patches, a significant number of patches need to be sampled from the image for the measures to be reliable. However, we will only select a subset of these image patches for use in classification. While it may seem advantageous to use all the patches for classification, as we show later in our results, using a subset of the patches can actually lead to improved recognition rates.

As can be seen, there is no dependence of the saliency measures $\mathcal{S}^o$ or $\mathcal{S}^l$ on the class labels of the images, making the proposed contextual saliency measures unsupervised.

4 Sampling Strategies

Using the equations above we can compute a saliency measure for each patch in an image. In this section we discuss three methods for selecting a subset of these patches for use in classification. Let us assume s patches are desired for classification out of a possible n.

4.1 Sampling by Sorting

A naive approach to sampling is to pick the s patches with highest saliency. However, due to strong correlations in natural images, neighboring patches often have similar appearances, and would hence share similar saliency values. The result of using this technique is many neighboring patches being selected that convey similar information. As a consequence, classification rates may suffer.

4.2 Random Sampling

One method to reduce the odds of sampling neighboring or redundant patches is to use a weighted random sampling. The saliency map may be normalized to form a distribution over the patches from which samples can be drawn. This allows for patches with higher saliency to be sampled with a higher probability, without any one region dominating. This allows for a good balance between exploiting the highly salient regions, and exploring the rest of the image for other salient regions.

4.3 Sequential Sampling

The last strategy sequentially selects patches by considering the patches previously selected. Specifically, we pick the patch that is most predictive of the patches that were not highly likely given at least one of previously picked patches. Let us consider the saliency measures of equations (1) and (5), which compute the probability of $p(x_j|x_i)$ equal to $\sum_a \alpha_{ia} p(x_j|w_a)$ and $\sum_a \sum_u \alpha_{ia}\beta_{iu} p(x_j|w_a, l_u)$ respectively. Then given a set of previously picked patches $\{\acute{x}_1, \ldots, \acute{x}_t\}$ we compute our saliency measure as

$$\mathcal{S}_i(\acute{x}_1, \ldots, \acute{x}_t) = \frac{1}{n} \sum_{j=1}^{n} \max\left(p(x_j|x_i), p(x_j|\acute{x}_1), \ldots, p(x_j|\acute{x}_t)\right) \tag{8}$$

A each iteration, the patch with highest saliency is selected. This sequential approach selects patches that give the highest average increase in maximum predicted probability for the patches in the image. As a result, patches that convey similar information as those already chosen are unlikely to be selected.

5 Existing Saliency Measures

In our experiments, we compare our proposed contextual saliency measures with three classes of existing saliency measures. The first baseline measure is the uniform saliency measure across the entire image, where patches are sampled randomly from the image. Equivalently, this can be thought of as computing a distribution over the codewords using a normalized histogram. The distribution over codewords is then randomly sampled. The second measure is an interest-point based saliency measure. More specifically, we apply the Harris corner detector [2] to the image, and used its response at every location in the image as

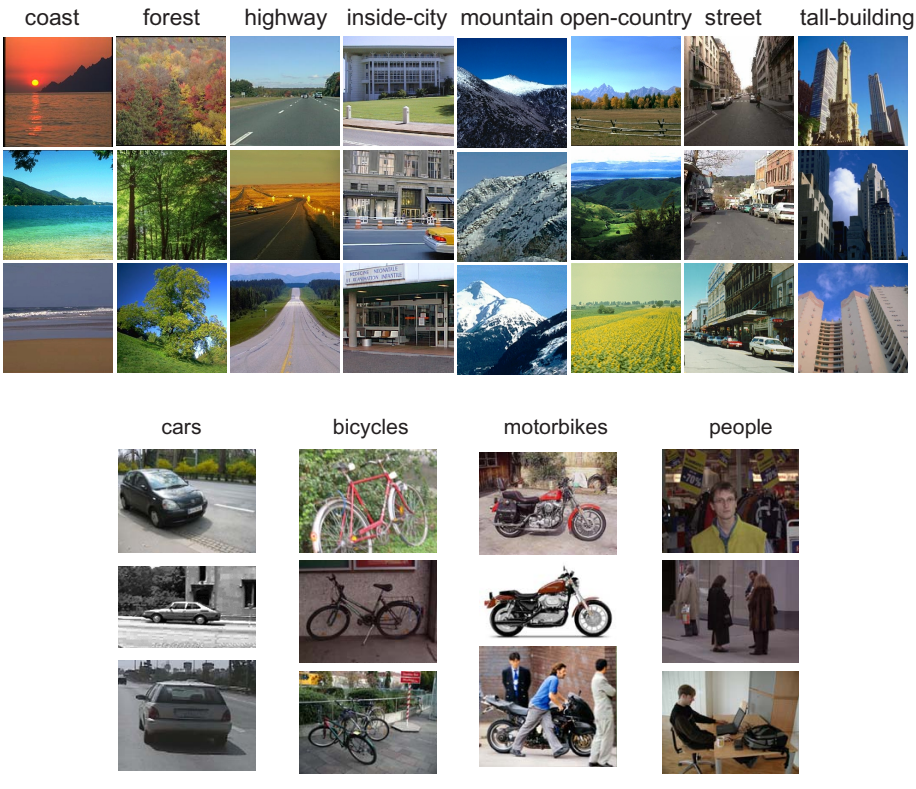

Fig. 1. Example images from the (top) outdoor scene category dataset [41] and (bottom) Pascal-01 object recognition dataset [43]

the saliency map. We also provide experiments using the patches found from the SIFT detector [1]. Finally, we compare against a discriminative saliency measure. The discriminative measure considers a patch to be salient if the mutual information of the patch and the class labels is high. More specifically, if $\mathcal{M}(w_a)$ is the mutual information of the a^{th} word with the class labels, the measure is defined as $\mathcal{S}_i^D = \sum_{a=1}^{m} \alpha_{ia} \mathcal{M}(w_a)$.

6 Experimental Setup

We evaluate our proposed contextual saliency measure for the tasks of scene and object recognition using the bag-of-features approach. In both scenarios we construct a codebook of feature descriptors using the standard K-means clustering technique with $K = 1000$. Classification is accomplished by first assigning each sampled patch to a codeword. A histogram of codewords is created as input into a Support Vector Machine (SVM) classifier. We use a Gaussian kernel SVM in all our experiments. We also experimented with adaptively thresholded histograms

by picking thresholds that maximize mutual information with the class labels as suggested by Nowak *et al.* [7]. However, the results were comparable, so we only report experiments using the normalized histograms in our experiments. Experiments using the nearest-neighbor classifier were also tested. The results were consistently inferior, and hence are not included here.

6.1 Scene Recognition

We evaluate our approaches on the outdoor scene category dataset from Torralba *et al.* [41]. Example images from this dataset are shown in Fig. 1 (top). It contains images from 8 categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways. There are a total of 2866 256×256 color-images. For scene recognition our 48 dimensional descriptor consists of the average color values sampled in a 4×4 grid. The patches were sampled evenly across the image on a 64×64 grid. The patch scale was set so that neighboring patches overlap by 75%. Each sampled patch is given a soft assignment to the codewords using equation (3) with $\sigma_w = 30$. Similar to Torralba *et al.* [42], we use 100 images per scene category for training, and the rest as testing.

6.2 Object Recognition

Our experiments on object recognition use the Pascal-01 [43] dataset which contains 4 object categories: cars, bicycles, motorbikes and people. Example images from the dataset are shown in Fig. 1 (bottom). A training set of 684 images and a test set of 689 images is defined. Since object recognition is more dependent on image gradients than color, we use SIFT [1] as our descriptor. The descriptor was sampled on a 64×64 uniformly spaced grid. The scale of the sampled patches was set so that horizontally neighboring patches overlap by 75%. In this scenario we used hard assignments of patches to codewords.

7 Results

7.1 Comparing Saliency Measures

Our first experiments test our contextual saliency measures and those described in Section 5 on the scene and object recognition datasets. Recognition accuracies are plotted relative to the density of samples used for classification in Fig. 2. The weighted random sampling strategy is used in all cases. For comparison, representative reported accuracies on these datasets are 84% on the outdoor scene recognition dataset by Torralba *et al.* [42] and $\sim$ 88% on the Pascal-01 object recognition dataset by Nowak *et al.* [7]. The highest accuracies achieved by the contextual saliency measures are 84% and 86% on the scene recognition task for $\mathcal{S}^l$ and $\mathcal{S}^o$ respectively, and 85% and 90% for the object recognition task. We also tested our algorithm using higher resolution grids for scene recognition. The highest accuracies for $\mathcal{S}^o$ were 55%, 68% and 81% for 8×8, 16×16 and 32×32 sampled grids respectively.

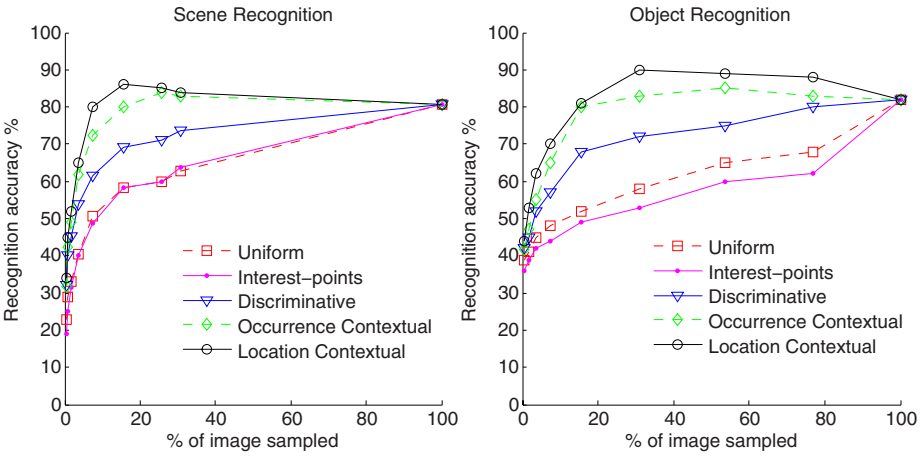

Fig. 2. Scene (left) and object (right) recognition accuracies for different saliency measures. The weighted random sampling strategy is used in all cases.

The contextual saliency measures have the best performance on both datasets. The discriminative measure using mutual information is next followed by similar results for both interest points and random saliency measures. We also ran experiments using the complete set of interest points found using the SIFT detector. On average the SIFT detector found 926 interest points and a recognition accuracy of 71% was achieved on the object recognition dataset. The performance of the saliency measures not using context increase monotonically with the density of the sampling. This is consistent with observations made by Nowak *et al.* [7]. However, with the contextual saliency measures a sparser sampling results in higher accuracy. This indicates that a sparser sampling is desirable not only for computational efficiency, but also higher recognition performance. With respect to the saliency measures, the usefulness of spatial information varies based on the dataset. The spatial information provides a larger performance boost for object recognition. We speculate that this is due to the increased spatial ambiguity of SIFT descriptors as compared to color descriptors. In scene recognition, color descriptors such as blue patches that correspond to sky or green patches that correspond to grass are strongly correlated with certain image locations. As a result, the spatial information may be redundant.

The various saliency maps for a set of sample images are shown in Fig. 3. In the scene recognition examples, objects that are unlikely given the scene category typically have lower saliency measures. As can be seen in Fig. 3(b), the saliency maps for the object recognition datasets have high saliency values even for the backgrounds. We believe this is due to several reasons: a strong correlation of objects and background, the higher entropy of the SIFT descriptor and the use of hard assignments.

The higher performance of contextual saliency measures over discriminative saliency measures may seem un-intuitive at first, since the discriminative saliency

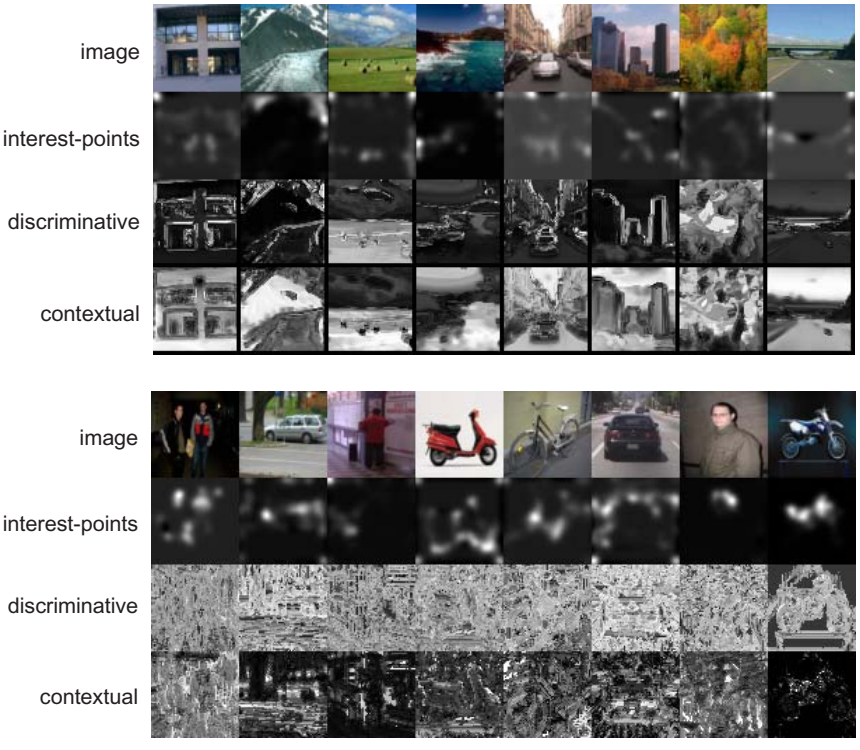

Fig. 3. Example saliency maps for images for the (top) scene recognition and (bottom) object recognition tasks using different classes of saliency measures. Maps are normalized to lie between 0 (least salient patch) and 1 (most salient patch).

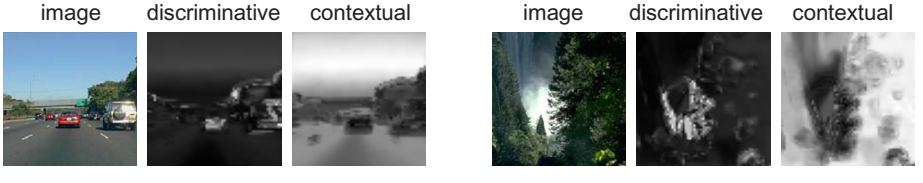

Fig. 4. Red patches on the car in the highway image (left) and the white patches from the light behind the trees in the forest image (right) are considered to be salient by the discriminative measure because they occur pre-dominantly in sunset coast and snow-covered mountain images respectively. However, the contextual saliency measure incorporates the context of the rest of the scene and thus considers the road, sky and tress to be salient instead.

measure is supervised and is optimized specifically for recognition accuracies. However, it should be noted that the discriminative saliency measure ignores the rest of the scene, or the context in which the patch is present. This can lead

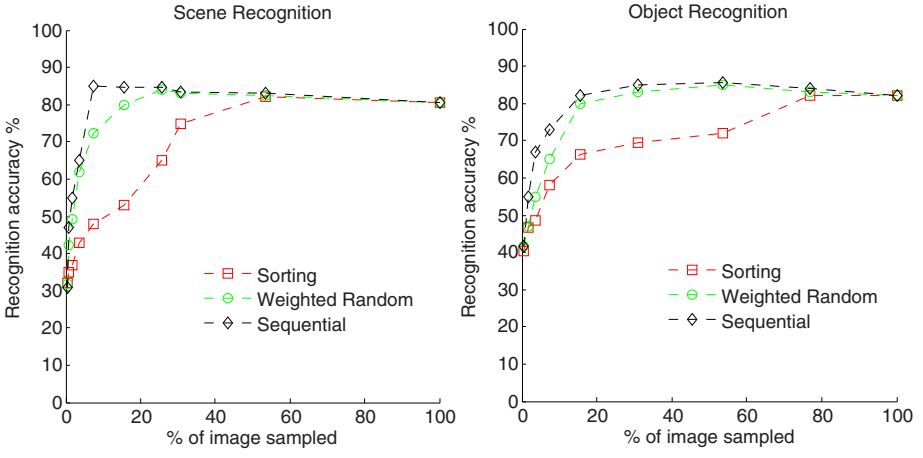

Fig. 5. Scene (left) and object (right) recognition accuracies for different sampling strategies. The occurrence-based contextual saliency measure $\mathcal{S}^o$ is used in all cases.

to undesirable artifacts. For instance, in a scene recognition task, red/orange patches may be considered to be salient by the discriminative measure since they occur pre-dominantly only in sunset (coast) images. Consider a highway test image that has a red car present in it, as seen in Fig. 4 (left). All the patches on the car will be considered highly salient by the discriminative saliency measure even though they are not representative of the scene. The saliency measure using context would identify that the red patches on the car are not representative of the image, and would not use them for classification. As a result, the contextual saliency measure is more likely to ignore clutter in the scene, resulting in higher accuracies. A similar result can be seen in Fig. 4 (right).

7.2 Comparing Sampling Strategies

To compare the different sampling strategies described in Section 4, we work with the occurrence-based contextual saliency measure. The scene and object recognition accuracies using the different sampling strategies are shown in Fig. 5. We can see that for scene recognition, the sorting strategy is much worse than the weighted random sampling. The features used for the scene recognition task are raw color patches, and hence neighboring patches in an image have very similar features and hence very similar saliency measures. While sequential sampling does not give higher accuracies, it reaches the peak accuracy using fewer patches than the weighted random sampling. Similar trends are seen for object recognition with the sequential and weighted random sampling results being more comparable. This may be due to the lower correlation of neighboring SIFT features as compared to the color descriptors used for scene recognition. Examples of how the saliency maps are updated after each iteration of the sequential sampling are shown in Fig. 6.

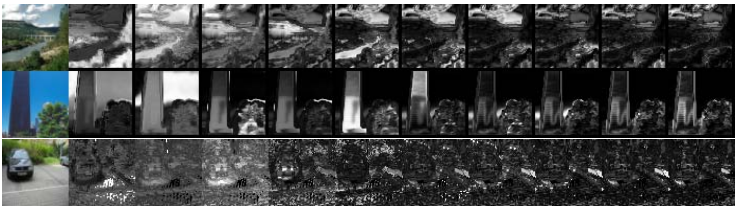

Fig. 6. Illustration of sequential sampling. Left: original image; Subsequent columns: saliency map being updated at each iteration; Top two rows: scene recognition; Bottom row: object recognition.

7.3 Discussion

Typically, contextual information is used for high-level reasoning about interactions between objects. In this paper, we demonstrate how contextual information may also be useful for low-level applications such as measuring patch saliency. While low-level contextual reasoning lacks semantic object information, even color or texture patches can supply useful contextual information as also shown in [44].

Discriminative saliency measures capture classification specific statistics of the patches. Our proposed contextual saliency measures capture contextual information of the entire image to determine saliency of patches. Both these aspects are complementary, and are both important to select representative patches that give good recognition accuracies. For the data sets we experimented with so far, the contextual information was more critical than the discriminative information, such as the example shown in Figure 4. However, one can imagine scenarios where the reverse is true. The balance between the two is task and domain dependent. A natural future direction, hence, is to combine discriminative and contextual information to design the optimum saliency measure for a given task. This is related to subjectiveness in the notion of saliency itself. Salient regions may be considered to be those that are representative of the image (as we do in this work), or those that are rare or unusual and hence draw attention. If we consider a more generic definition of saliency as being informative, it leads us back to the notion of task and domain dependency.

While our contextual saliency measure is unsupervised it is still dataset specific. That is, training images are needed to learn the co-occurrence statistics of the codewords. Other methods such as the use of interest points or random sampling may be better suited for applications in which the statistics of the images may not be known beforehand.

8 Conclusions

In this paper we propose two measures of saliency using contextual information. The first measure relies on co-occurrence information between codewords, while

the second measure includes spatial information. We test our saliency measures against several others using the bag-of-features paradigm. Our experiments show improved results over other saliency measures on both scene and object recognition datasets. In contrast to previous works that produce results with accuracies that monotonically increase with sampling density, the contextual saliency measures produce optimal results with a sparse sampling.

References

1. Lowe, D.: Distinctive image features from scale-invariant keypoints. In: IJCV (2004)
2. Harris, C., Stephens, M.: A combined corner and edge detector. In: AVC (1988)
3. Kadir, T., Brady, M.: Saliency, scale and image description. In: IJCV (2001)
4. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. In: IJCV (2004)
5. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: BMVC (2002)
6. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. In: IJCV (2000)
7. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: ECCV (2006)
8. Moosmann, F., Larlus, D., Jurie, F.: Learning saliency maps for object categorization. In: ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision (2006)
9. Walker, K., Cootes, T., Taylor, C.: Locating salient object features. In: BMVC (1998)
10. Fritz, G., Seifert, C., Paletta, L., Bischof, H.: Entropy based saliency maps for object recognition. In: ECOVISION (2004)
11. Serre, T., Riesenhuber, M., Louie, J., Poggio, T.: On the role of object-specific features for real world object recognition in biological vision. In: BMVC (2002)
12. Vidal-Naquet, M., Ullman, S.: Object recognition with informative features and linear classification. In: ICCV (2003)
13. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. In: IJCV (2001)
14. Lazebnik, S., Schmid, C., Ponce, J.: Affine-invariant local descriptors and neighborhood statistics for texture recognition. In: ICCV (2003)
15. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV workshop on Statistical Learning in Computer Vision (2004)
16. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: ICCV (2005)
17. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from googles image search. In: ICCV (2005)
18. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV (2003)
19. Sivic, J., Russell, B., Efros, A.A., Zisserman, A., Freeman, B.: Discovering objects and their location in images. In: ICCV (2005)
20. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: ICCV (2005)

21. Ye, Y., Tsotsos, J.K.: Where to look next in 3d object search. In: ISCV (1995)
22. Viola, P., Jones, M.: Robust real-time object detection. In: IJCV (2001)
23. Grauman, K., Darrell, T.: Efficient image matching with distributions of local invariant features. In: CVPR (2005)
24. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR (2005)
25. Agarwal, A., Triggs, B.: Hyperfeatures – multilevel local coding for visual recognition. In: ECCV (2006)
26. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology (1980)
27. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. In: PAMI (1998)
28. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology (1985)
29. Sebe, N., Lew, M.: Comparing salient point detectors. Pattern Recognition Letters (2003)
30. Hall, D., Leibe, B., Schiele, B.: Saliency of interest points under scale changes. In: BNVC (2002)
31. Walther, D., Rutishauser, U., Koch, C., Perona, P.: On the usefulness of attention for object recognition. In: ECCV (2004)
32. Hoiem, D., Efros, A., Hebert, M.: Putting objects in perspective. In: CVPR (2006)
33. Torralba, A., Murphy, K., Freeman, W.: Contextual models for object detection using boosted random fields. In: NIPS (2005)
34. Torralba, A., Sinha, P.: Statistical context priming for object detection. In: ICCV (2001)
35. Murphy, K., Torralba, A., Freeman, W.: Using the forest to see the trees: a graphical model relating features, objects, and scenes. In: NIPS (2003)
36. Bose, B., Grimson, E.: Improving object classification in far-field video. In: ECCV (2004)
37. Torralba, A., Murphy, K., Freeman, W., Rubin, M.: Context-based vision system for place and object recognition. AI Memo (2003)
38. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV (2007)
39. Parikh, D., Zitnick, C.L., Chen, T.: From appearance to context-based recognition: Dense labeling in small images. In: CVPR (2008)
40. Singhal, A., Luo, J., Zhu, W.: Probabilistic spatial context models for scene content understanding. In: CVPR (2003)
41. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. In: IJCV (2001)
42. Torralba, A.: Outdoor scene category dataset, `http://people.csail.mit.edu/torralba/code/spatialenvelope/`
43. Pascal01, `http://pascallin.ecs.soton.ac.uk/challenges/voc/`
44. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951. Springer, Heidelberg (2006)