Biologically Plausible Detection of Amorphous Objects in the Wild

Sunhyoung Han Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{s1han,nuno}@ucsd.edu

Abstract

The problem of amorphous object detection is investigated. A dataset of amorphous objects, Panda bears, with no defined shape or distinctive edge configurations is introduced. A biologically plausible amorphous object detector, based on discriminant saliency templates, is then proposed. The detector is based on the principles of discriminant saliency, and implemented with a hierarchical architecture of two layers. The first computes a feature-based top-down saliency measure tuned for object detection. The second relies on a similar saliency measure, but based on saliency templates, selected from the responses of the first layer. This architecture is shown to have a number of interesting properties for amorphous object detection, including the ability to detect objects characterized by the absence of features, and an interpretation as discriminant blob detection. Extensive experimental evaluation shows that it substantially outperforms state-of-the-art approaches for non-amorphous object detection, such as deformable parts models, sparse coded pyramid matching, detection based on the bag-of-features architecture, and the Viola and Jones approach. This brings into question some currently popular beliefs about object detection, which are discussed.

1. Introduction

Many object detection approaches have been proposed over the last decade. These range from the ubiquitous bag-of-features model [31, 27], to algorithms based on representations of shape [2, 10, 11], and models of configurations of parts [9, 5, 12], among others. The development of these algorithms is, in non-trivial part, guided by the portfolio of datasets available to compare different approaches. This portfolio has been considerably enriched and diversified since the early days of the UIUC carside and Caltech4 datasets. Recent benchmarks are much more extensive, covering a much larger number of object categories, viewpoints, and intra-class variation.

Nevertheless, it could be argued that current datasets only cover the extremes of the spectrum of scenarios faced by a practical object detector. One of these extremes corresponds to the detection of broad object categories, such as "airplane", "dog", "cat", etc., and datasets such as Caltech256, or PASCAL VOC. Each of the categories in these datasets is comprised of many distinguishable object subgroups of widely different appearance, e.g. the subclasses "siberian husky", "bulldog", and "yorkshire terrier" of the "dog" class. This constrains the number of training examples per sub-class, which is usually small, and bias the benchmarks towards specific classifier architectures. For example, because kernelized SVMs can store large portions of the training set as support vectors, it is not surprising that the combination of these classifiers with sophisticated kernels, and the loose bag-of-words representation, achieves good performance on these datasets.

Figure 1. Example images from the PandaCam dataset. Note the high variability of view points, illumination, and object pose.

On the other end of the benchmarking spectrum are tasks such as face or pedestrian detection. These refer to narrowly defined object classes, e.g. frontal faces or upright pedestrians, and datasets with a large ratio of training images to distinguishable object sub-classes. Such datasets sample the appearance of a more or less constant object in widely varying imaging contexts. It is thus not surprising that their most successful solutions are based on templates of object appearance, using shapelets [24], configurations of HOG [5, 9, 20], or Haar features [28].

In this work, we investigate object detection in the context of a real application which is not covered by these two scenarios. This application involves objects that lack many of the features commonly used for object detection, and which we denote *amorphous*. Strictly speaking, amorphous objects have no distinctive edge configurations, texture, or a well defined shape. They can be found in science fiction movies, in the form of jelly-like creatures that can take any desirable shape. While, in the real world, truly amorphous objects are rare, many real objects are close to amorphous (e.g. a jellyfish, a bean bag, etc.), and an even larger set *quasi-amorphous*. By this, we refer to objects that can have very characteristic appearance under some canonical poses, but appear amorphous under others. Many such examples exist in the animal world. This is illustrated by Figure 1, which shows a Panda bear under multiple poses. While the Panda face is very iconic, faceless poses tend to be quite amorphous. We will not worry further about these fine distinctions, simply referring to such objects as *amorphous*.

This work makes two contributions to the detection of amorphous objects. The first is an object detector, based on the idea of *discriminant saliency templates.* The intuition is that (at least in the natural world) the most distinctive property of amorphous objects is their lack of low level features, when compared to the surrounding scenes. This suggests modeling these objects as *blobs of feature absence*, i.e. regions where features that are usually active for natural images have a much weaker object response. One possibility for amorphous object detection is thus to rely on *discriminant blob detection,* by identifying blob-like regions in the responses of a set of *features that are discriminant* for object detection. A potential problem is, however, that discrimination may be due to the *absence* of feature responses. We overcome this problem by formulating blob detection as a form of top-down discriminant saliency [13].

Under this approach, detection is based on a two level classification architecture which implements a combination of *feature-based* and *template-based* discriminant saliency. The first level consists of a feature-based top-down saliency model, tuned for the detection of the target object. It is a robust classifier, that can detect the absence of a set of features, if this absence is informative of object presence. However, it is not highly selective, frequently generating false positives in background image regions. The second level learns *discriminant templates* of saliency response, which are then used to detect blobs of saliency compatible with the target object. This is again implemented with a top-down discriminant saliency model, tuned for object detection, which operates on saliency templates rather than image features. Altogether, this classifier is selective, yet robust enough to detect highly deformable objects of reduced visual structure. The use of saliency, rather than appearance, templates also makes it robust to pose variation.

The second contribution of this work is a *dataset* for the evaluation of amorphous object detectors. This dataset was assembled from video of a real animal habitat, the Panda bear exhibit of the San Diego Zoo, over the period of one year [1]. It resembles current pedestrian datasets, in that it requires the detection of a few objects under various imaging contexts. On the other hand, it is similar to object category datasets, in the sense that Pandas have wide variability of appearance. As can be seen from Figure 1, this is in part because they are highly deformable objects, and in part because the video feed is collected from multiple cameras, with multiple fields of view (varying backgrounds), at different distances from the Panda exhibit (varying scales), from different angles (varying poses), at different locations (indoors vs outdoors), at multiple times of the day, week, and year (different atmospheric conditions, variable shading, lighting, etc.), and with different potential occluders.

One of the attractives of the PandaCam dataset, is that it challenges currently prevalent beliefs about object recognition. For example, results on current datasets suggest that normalized representations of local image orientation are critically important for object detection. In fact, these representations are the only unifying link between the success of bag-of-features (almost invariably based on SIFT) on the PASCAL end of the spectrum, and template-based approaches (usually based on HOG) on the pedestrian end of the spectrum. On PandaCam, a comparison of the proposed detector with an equivalent approach built on templates of SIFT response shows that saliency templates achieve substantially higher localization performance. As an object detector, the proposed approach is also shown to achieve better performance than state-of-the art methods for template-based detection, namely the discriminant parts-based model of [9], detection based on the bag-of-features model [18], the sparse coded spatial pyramid matching method of [30], and the Viola Jones detector [28].

2. The PandaCam dataset

We start by introducing the PandaCam dataset, so as to motivate the challenges of amorphous object detection.

2.1. The dataset

The video feed provided by San Diego Zoo depicts the real time movement of a Panda family in a natural habitat,

Table 1. Edgeness statistics for object and background in the PASCAL VOC and PandaCam datasets.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	**PandaCam**
obj	.0111	.0176	.0135	.0153	.0181	.0164	.0202	.0092	.0112	.0167	**.011**
back	.0045	.0107	.0101	.0104	.0085	.0103	.0111	.0079	.0088	.0126	**.0146**
	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	–
obj	.012	.0106	.0134	.0181	.0113	.018	.0174	.0076	.0151	.0113	–
back	.0077	.0092	.0135	.0097	.0078	.0112	.0132	.0064	.0112	.0076	–

which includes bamboo trees, ponds, a small cave formed of rocks, and several other small structures. The dataset is divided into $5,018$ positive images containing Pandas and $2,987$ negative images without the animals. A bounding box is provided as detection ground truth for the positive images. The relative size of the objects varies from 2% to 90% of the image size. The variations in appearance are very large, due to the highly deformable shape of the Pandas, and the collection of the video from multiple cameras and multiple viewpoints. Illumination changes are also dramatic, since the dataset reports to a live cam that operates continuously, 24/7. Finally, the dataset is unique in the sense that the background clutter is much more structured than the objects to detect. Background trees, tree branches, rocks and leaves all have a rich combination of structure, shape, and texture. This is unlike the Pandas, which are mostly textureless and lack shape-defining edges.

2.2. Amorphous object statistics

To demonstrate this point, statistics of the PandaCam dataset were compared to those of PASCAL VOC. In particular, we considered a measure of "edgeness", and compared the relative amounts of this property in object and background, for the two datasets. To quantify edgeness, we filtered the image with a set of band-pass filters (Gabor functions of four orientations). It is well known that the responses X of such filters to natural images follows a generalized Gaussian distribution (GGD) [6]

$$P_X(x;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-\left(\frac{|x|}{\alpha}\right)^{\beta}}, \qquad (1)$$

where $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\mathrm{d}t,\ t > 0$ is the Gamma function, α a *scale* parameter, and β a parameter that controls the *shape* of the distribution. It is also known that β tends to be fairly stable, assuming values in the range $\beta \in [0.5, 0.8]$ [26]. We have confirmed this observation, and use $\beta = 0.5$ throughout this work.

Given a training sample $\mathcal{D} = \{x_1, \ldots, x_n\}$ of filter responses, the MAP estimate of the scale parameter α based on a conjugate prior is

$$\hat{\alpha}^{\beta}_{MAP} = \frac{1}{\kappa}\left(\sum_{j=1}^{n} |x_j|^{\beta} + \nu\right), \quad \text{with } \kappa = \frac{n+\eta}{\beta}, \qquad (2)$$

where η and ν are prior hyper-parameters [14]. The scale α is proportional to the variance of the responses, and a good measure of their activity. For the features considered here, it measures the image edgeness along the feature orientation.

Table 1 presents the α estimates obtained for object and background using the bounding boxes provided by PASCAL VOC and PandaCam. Larger values of α imply more edge structure. Note that the edgeness of the Panda object is much smaller than those of most object classes in PASCAL. The Panda detection problem is unique in the sense that the background has much richer structure than the object itself.

3. Discriminant saliency

The amorphous object detector proposed in this work is based on discriminant saliency. We next briefly review how this saliency principle can be used to implement a top-down measure of saliency, tuned for object detection. Saliency is formulated as optimal (in the minimum probability of error sense) classification of the visual stimulus into one of two hypotheses: a *target* ($Y = 1$) hypothesis of stimuli that are salient, and a *null* ($Y = 0$) hypotheses containing *background* stimuli [14]. Salient locations are those where target presence can be declared with largest confidence. Confidence is measured by the strength with which visual features in a region $\mathcal{A}(l)$, surrounding a location l, can be declared observations from the target class, by the optimal decision rule for target/background classification.

This is measured by the expected ratio of the likelihood of the observations under the target and null hypotheses, or Kullback-Leibler divergence,

$$S(l) = \int_{x\in\mathcal{A}(l)} P_{X|Y}(x|1)\log\frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)}dx.$$

Using the standard approximation of risk by empirical risk,

$$S(l) \approx \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}(l)} \delta_i \log\frac{P_{X|Y}(x_i|1)}{P_{X|Y}(x_i|0)} \qquad (3)$$

where $\delta_i = 1$ if x_i is a sample from the target class and $\delta_i = 0$ otherwise. The class-assignment variables are inferred with recourse to the Bayes decision rule, i.e. by replacing δ_i with

$$\hat{\delta}_i = \begin{cases} 1 & \text{if } P_{Y|X}(1|x_i) \geq 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

With these estimates, (3) can be written as [15]

$$S(l) = \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}(l)} \xi(\mathcal{P}(l)) \tag{4}$$

$$\xi(x) = \begin{cases} \frac{1}{2} \log \frac{x}{1-x}, & x \geq .5 \\ 0, & x < .5, \end{cases}$$

where

$$\mathcal{P}(l) = P_{Y|X}(1|x(l))$$

is the posterior probability of the target class given the feature responses observed at location l. This posterior probability is itself a function of the log-likelihood ratio, since

$$\mathcal{P}(l) = \sigma[g(x(l))] \tag{5}$$

with

$$g(x(l)) = \log \frac{P_{X|Y}(x(l)|1)}{P_{X|Y}(x(l)|0)} \tag{6}$$

$$\sigma(x) = 1/(1+e^{-x}). \tag{7}$$

When feature responses have GGD distribution of scales α_1 and α_0 under the target and null hypotheses, respectively, the log-likelihood ratio simplifies to

$$g(x(l)) = \frac{|x(l)|^\beta}{\alpha_0^\beta} - \frac{|x(l)|^\beta}{\alpha_1^\beta} + T, \tag{8}$$

where $T = \log\left(\frac{\alpha_0}{\alpha_1}\right)$. The parameters α_i^β are learned from training samples $\mathcal{D}_i$ of feature responses from the target and background classes, using (2).

4. The importance of feature absence

Feature absence has an important role in the detection of amorphous objects. We next discuss how discriminant saliency naturally accounts for this.

4.1. Amorphous objects and feature absence

Amorphous objects lack many of the features that are commonly used as cues for object detection. They do not have many distinctive edges, may not have a very distinguishable texture, and are characterized by a large shape variability. In fact, as shown in Figure 1, they can be thought of as blobs of low image complexity. However simple blob detection [4, 29] is unlikely to successfully find Pandas, as there are many blob-like regions in the backgrounds of Figure 1: smooth rocks, tree trunks, light reflections on interior walls, areas of the exhibit floor, etc. One possibility is to rely on *discriminant blob detection*, by identifying blob-like regions in the responses of features that are *discriminant for Panda detection*. These are regions of feature *absence,* i.e. where features that are usually active for natural images, have a much weaker response to the Panda stimulus. Feature absence is naturally detected by discriminant saliency.

4.2. Saliency and feature absence

We start by noting that this is not true for all formulations of saliency. While many saliency detectors have been proposed in the literature, most emphasize the detection of the *presence* of certain features in the visual field. This is, for example, the case of interest point detection, which is explicitly formulated as the detection of corners [16], points of image curvature [21], activity [17], or texture complexity [6]. A popular generalization of this idea [23, 32, 3], is to detect features of low probability within the visual field. Its decision theoretic implementation, for GGD feature statistics, is to find the locations l where feature responses have maximum entropy [17]

$$\begin{aligned} H(l) &= -\int_{\mathcal{A}(l)} P_X(x) \log P_X(x) dx \\ &= K + \frac{1}{2|\mathcal{A}(l)|} \sum_{i \in \mathcal{A}(l)} \left(\frac{|x_i|}{\alpha}\right)^\beta \end{aligned} \tag{9}$$

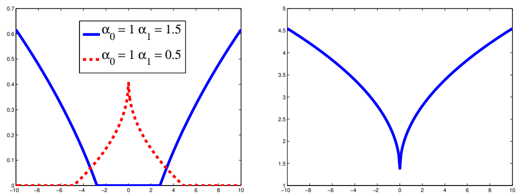
where $K = \log 2\alpha\Gamma(1/\beta)/\beta$, and α is the GGD scale for the feature responses accross the visual field. Note that this is, in many aspects, similar to the measure of (3). It is, in fact, equivalent in the limit of $\alpha_1 \to \infty$, if $\alpha = \alpha_0$ and the non-linearities $\sigma(.)$ and $\xi(.)$ are replaced by the identity map. While this may appear a small difference, it has a major impact on the ability of the saliency detector to switch between the detection of *feature presence* and *absence*.

Discriminant saliency can switch between these two detection strategies because it has access to *two scale parameters,* α_1 and α_0. When $\alpha_1 = \alpha_0$, the target and null distributions are identical, i.e. there is nothing to discriminate, and saliency is null for all x. When $\alpha_1 > \alpha_0$, the target distribution has a heavier tail than that of the null hypothesis, and saliency is high for *large feature responses*. Conversely, the null hypothesis has heavier tail when $\alpha_1 < \alpha_0$, and only *small feature responses* are salient in this case.

The two behaviors are illustrated in Figure 2 a), which presents the characteristic function of the discriminant saliency detector - curve of $\xi(\mathcal{P}(l))$ as a function of $x(l)$ - for different values of α_1, when $\alpha_0 = 1$. Note how 1) the *presence* of the feature X (large feature response $|x|$) elicits a strong saliency response when $\alpha_1 > 1$, but 2) strong saliency responses are reserved for feature *absence* (small $|x|$) when $\alpha_1 < 1$. For comparison, the characteristic curve of the entropy-based detector is shown Figure 2 b), for $\alpha = 1$. In this case, all degrees of freedom are exhausted once the background distribution is fixed, and the saliency detector is *always* a detector of feature presence.

5. Amorphous object detection

We next consider the design of an amorphous object detector based on saliency templates, derived from the dis-



(a) discriminant saliency (b) entropy

Figure 2. Characteristic function of (a) discriminant saliency and (b) entropy measure.

criminant saliency principle.

5.1. Discriminant saliency networks

The saliency computations of Section 3 can be implemented with a network that mimics the standard neurophysiological model of the visual cortex [15]. The network has two layers, one of simple and another of complex cells. The simple cell layer computes the *target posterior map* $\mathcal{P}(l)$, which is then transformed into the *saliency map* $\mathcal{S}(l)$ by the complex cell layer. A simple cell is associated with each location l of the visual field, and implements the computations of (5)-(8) and (2). This is the combination of filtering, divisive normalization, and a saturating non-linearity with which simple cells are modeled under the standard neurophysiological model. A complex cell then pools the simple cell responses within the region $\mathcal{A}(l)$, after application of the non-linearity $\xi(x)$, to implement (3). These are the operations of complex cells under the standard neurophysiological model. The resulting saliency value, $\mathcal{S}(l)$, is a decision-theoretic measure of the confidence with which the feature responses at l can be assigned to the target class. The two layer network of (simple and complex) units is denoted a *discriminant saliency network* [15].

5.2. Discriminant saliency templates

The proposed amorphous object detector is based on discriminant saliency templates. These are discriminant templates of saliency response, derived from features that are themselves discriminant for the detection of the target object. As illustrated in Figure 3, the detector is implemented as a two-stage discriminant saliency network. The two stages are identical up to the features used to evaluate saliency, i.e. the linear filtering implemented by their simple cells. The first layer relies on low level features, such as measures of image orientation [25, 22], color opponency, image intensity [8], projections into standard signal basis (wavelets or Gabor expansions), or even random projections [19]. These features can be deemed discriminant due to either their presence or absence in the target object. In this work, we use the first four filters in a discrete cosine transform (DCT) basis of size 8×8, other than the average (DC) filter. These filters are illustrated in Figure 3. For amorphous objects, saliency responses usually reflect a mix

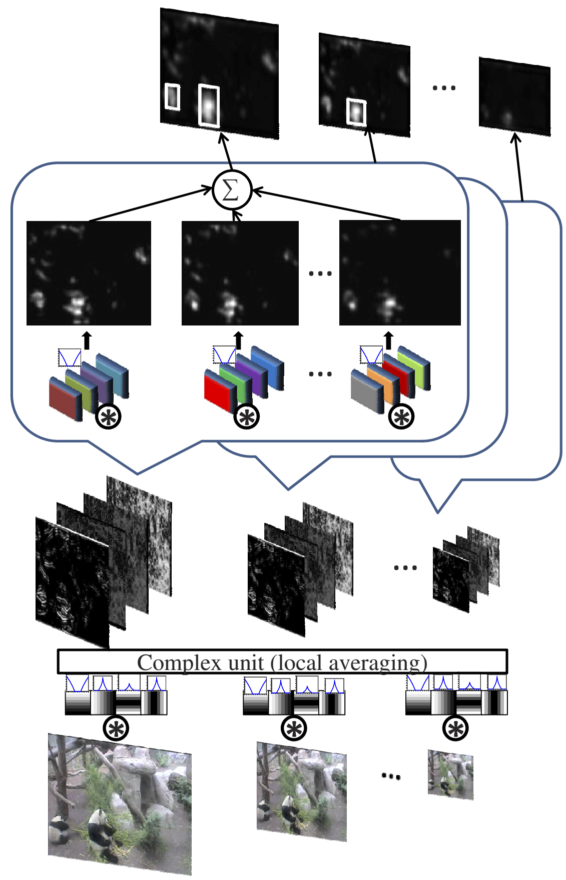


Figure 3. Architecture of the proposed template-based saliency detector. DCT features are used in the first network layer, and saliency templates in the second. The saliency computation is repeated at multiple scales.

of feature presence and absence, allowing the saliency map to be active in blob-like regions of low image complexity.

The second stage aims to detect configurations of saliency, produced by the first, which are distinctive for the target object. These salient configurations of salient feature responses capture information about object shape. The implementation of the second network stage has some resemblance to the second layer of the HMAX network of [25, 22]. A number of saliency templates are first randomly sampled from the outputs of the first network stage, during training. This is done by extracting patches, centered at random locations and scales, of response to random target images. Each patch has dimension $n \times n \times 4$, for $n \in \{4, 8, 12, 16\}$, and is normalized to zero mean and unit norm (over the 4 channels). During detection, the normalized patches (denoted patch filters) are correlated with first stage responses, to extract a second level of feature responses. These are then processed by the second stage of the network, to determine the saliency of these responses. Note that, unlike the HMAX network, which measures distances between patch filter and first stage response, this operation measures the discriminant power of the patch filter (saliency template) to classify first stage responses into object and background. This reinforces the responses of the templates that are discriminant for object detection and suppresses those of templates that are not.

6. Experiments

In this section, we describe a number of experiments designed to evaluate the performance of the proposed amorphous object detector.

6.1. Experimental Setup

In the PandaCam dataset, images are ordered by time stamp. In all experiments, the first $2,518$ positive images were used for training and the remainder $2,500$ as a test set. All positive training examples were cropped and normalized to a height of 80 pixels, while maintaining the original aspect ratio. The assembly of negative examples followed the iterative procedure of [5, 9]. Object detection was based on the saliency maps produced by the network of Figure 3. $5,000$ templates were randomly selected from first layer responses to positive training examples. The KL divergence between the GGD responses to object and background was then computed per template. The 500 templates of largest discriminant power were finally selected. Saliency maps produced by these templates were added into an overall saliency map, which was used for object detection.

Object detection was performed at 7 scales of a pyramid decomposition of each test image. More precisely, an image of size $H \times W$ was expanded into 7 pyramid layers of size $2^{0.5i}H \times W$, for $i \in \{-1, 0, 1, \ldots, 5\}$. This produced 7 saliency maps per image. The location of largest saliency was then found, at each scale, with a combination of box filtering and non-maximum suppression. A box filter of size $N \times N$ and amplitude $1/(N \times N)^{\gamma}$, was first convolved with all scale saliency maps, using $N = 80 \times 2^{0.5i}$ for scale i. The parameter γ was determined by cross validation. Non maximum suppression (size $N \times N$) was applied to the filtered saliency maps, to detect the location and the scale of largest saliency.

6.2. Saliency as focus of attention

We start by analyzing the performance of template-based saliency as a focus of attention mechanism. Its localization performance is compared to those of a SIFT-based saliency method, and a localization method based on discriminant visual words [7]. The SIFT-based saliency method is identical to that now proposed, but uses templates of SIFT response instead of saliency templates. Each image is represented by a collection of SIFT descriptors, extracted on a dense sampling grid of 16×16 patches, with 6 pixels of grid spacing. As for template saliency, $5,000$ templates of SIFT response were randomly chosen from the set of responses to positive examples, and normalized to zero mean

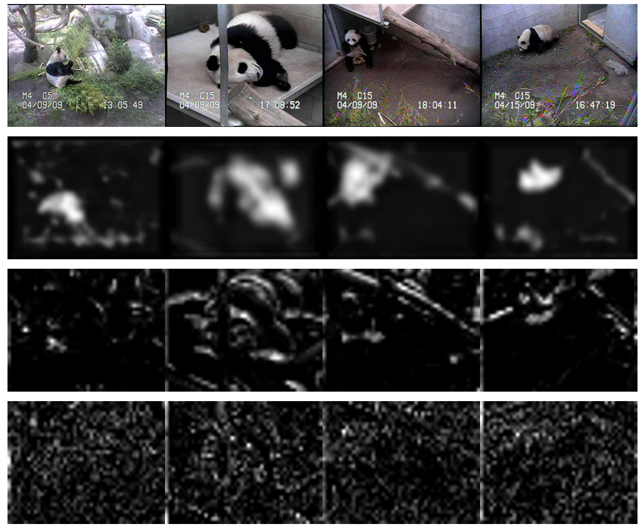


Figure 4. Images from the PandaCam dataset (top row), and saliency maps produced with saliency templates (second row), SIFT templates (third), and discriminant visual words (fourth).

and unit variance. Theses SIFT templates were then correlated with the SIFT responses to each training example, and the KL divergence between responses to object and background computed per template. The most discriminant 500 templates were finally used to produce SIFT-based saliency maps.

To compute saliency from visual words, images are represented as bags of SIFT descriptors, and quantized with a codebook of $1,000$ words, learned with k-means. The discriminant power of visual word, w, is then measured by the discriminability function proposed in [7],

$$D(w) = \frac{\text{\# target images containing } w}{\text{\# images containing } w}. \qquad (10)$$

The SIFT descriptor extracted from each image location l is then quantized into the closest visual word w^*. The saliency at l is the discriminability $D(w^*)$.

Figure 4 shows examples of saliency maps produced by the three methods. Test images are shown on the top row, template-based saliency maps on the second, SIFT-based saliency maps on the third, and saliency maps based on visual words on the fourth. Note that the latter are very noisy, with many false positives on the background, and few strong responses at target locations. SIFT-based saliency maps have much better localization, suppressing most responses from the background. However, while capturing the edge or contour structure of the target, they fail to respond to the object interior. This is not surprising, since SIFT is based on image gradients and the object interior is mostly smooth. Nevertheless, it is quite difficult to locate the object from these saliency maps. This task is much easier from the saliency maps produced by saliency templates, which 1) are

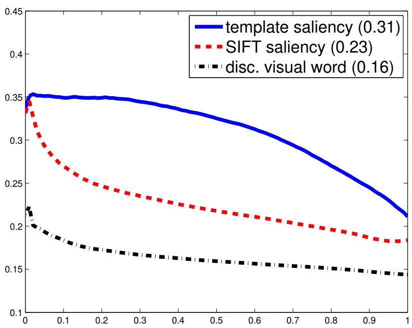


Figure 5. Precision recall curves for object localization.

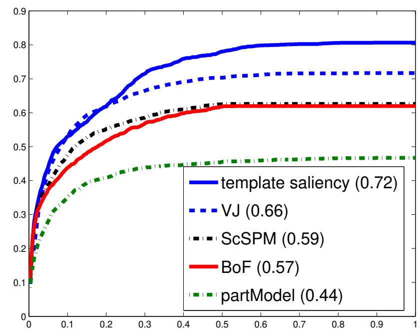


Figure 7. Curves of detection rate vs false positives per image.

active in the object interior, and 2) have even less false positives on the background. An objective comparison of localization performance in given in Figure 5, in the form of a precision recall curve. This curve is produced by thresholding the saliency map at various amplitude levels, measuring the overlap between the above-threshold region and the bounding box ground truth, and averaging over test images. The average precision is 0.31 for template saliency, 0.23 for SIFT saliency, and 0.16 for discriminant visual words.

6.3. Detection performance

The detection performance of template-based saliency was compared to those of the discriminatively trained part based model (partModel) of [9], the sparse coded spatial pyramid matching (ScSPM) method of [30], the bag-of-features (BoF) method of [18], and the Viola-Jones (VJ) detector, which combines boosting and Haar features [28]. The partModel was learned with 6 components, and the results reported were obtained with the non maximum suppression method of [9]. Detection with ScSPM, BoF, and VJ was based on a sliding window, using windows of seven scales, and a step size of 10 pixels. The non-maximum suppression scheme used for template saliency was also applied to these methods. For BoF and ScSPM, we used a spatial pyramid of 2 levels and a codebook of $1,000$ visual words.

Object detection performance was evaluated with the PASCAL measure, which requires an overlap greater than 50% between the bounding boxes of the detection area and ground truth. Figure 7 shows the curve of detection rate vs false positives per image (fppi) for all methods. The partModel was unable to model the Panda with the finite set of poses those were available, and achieved the worst performance of all methods. Both ScSPM and BoF produced a significant improvement, with ScSPM achieving slightly better performance. Another performance boost was achieved with the VJ detector. Finally, the template-based saliency detector produced the overall best performance. The detection rate at 0.3 fppi was 71.5% for template saliency, 66% for VJ, 58.6% for ScSPM, 56.8% for BoF and 43.8% for the partModel. Figure 6 shows detection examples by the template-based saliency detector. White windows indicate the ground truth and blue and red windows indicate detections. Blue is used for correct detections under the PASCAL measure, and red for false positives.

Besides the superiority of template-based saliency, an interesting conclusion from these experiments is the good performance of VJ. This is due to the fact that, unlike the partModel, BoF, and ScSPM, this method does not depend on image gradients. Instead, it relies of Haar features that can capture edgeless blobs. These results suggest that recognition approaches based on low level features other than the now predominant gradient based ones need to be studied rigorously.

References

[1] http://www.sandiegozoo.org/pandacam/.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on PAMI*, 2002.

[3] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Proc. NIPS*, 2006.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and it application to image quering. *IEEE Trans. on PAMI*, 2002.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[6] M. Do and M. Vetterli. Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE Trans. Image Processing*, 2002.

[7] G. Dorko and C. Schmid. Object class recognition using discriminative local features. *IEEE Trans. on PAMI*, 2004.

[8] L. Elazary and L. Itti. A bayesian model for efficient visual search and recognition. *Vision Research*, 2010.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on PAMI*, 2009.

[10] P. Felzenszwalb and J. Schwartz. Hierarchical matching of deformable shapes. In *Proc. CVPR*, 2007.

[11] V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *IJCV*, 2009.

[12] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computer*, 1973.

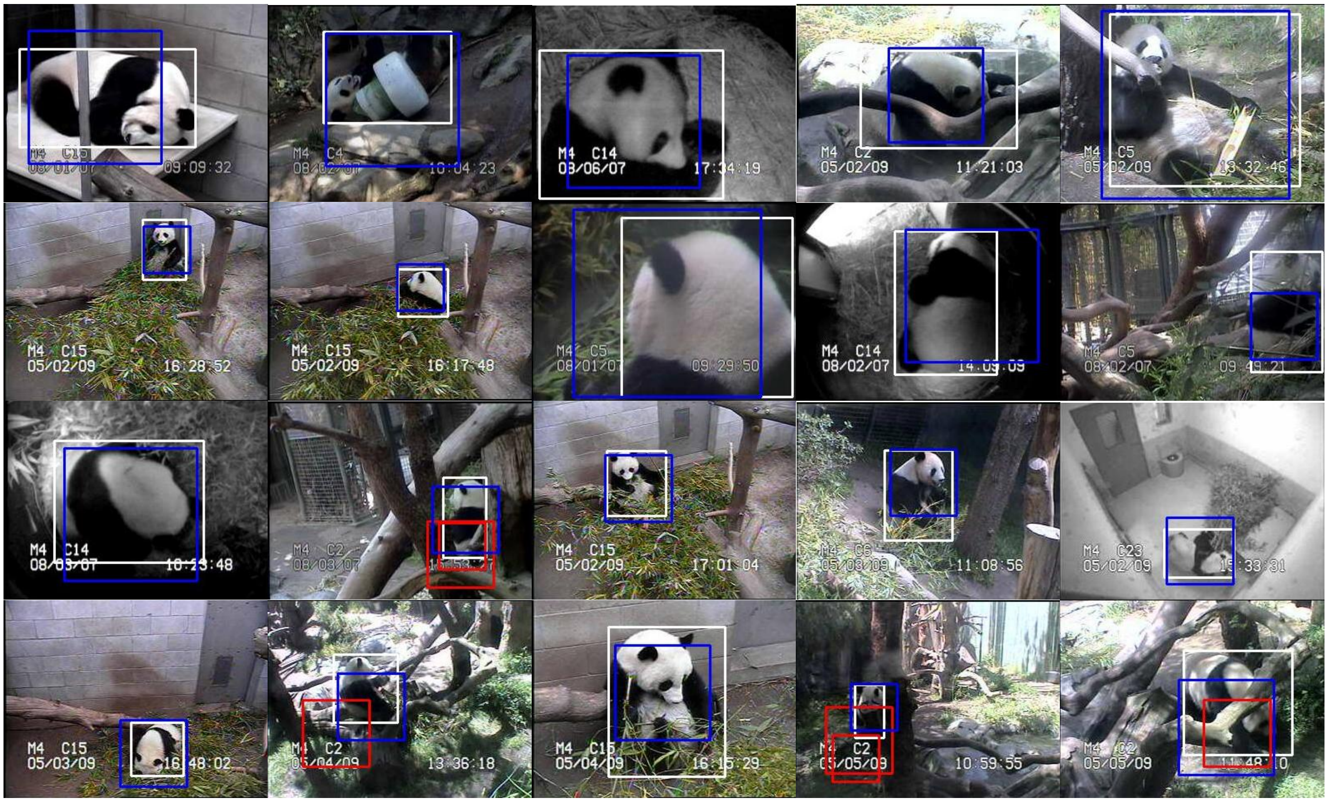


Figure 6. Examples of detections by the template-based saliency detector. White windows indicate ground truth, blue detections and red false-positives.

[13] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Trans. on PAMI*, 2009.

[14] D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, 2009.

[15] S. Han and N. Vasconcelos. Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research*, 2010.

[16] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.

[17] T. Kadir and M. Brady. Scale, saliency and image description. *IJCV*, 2001.

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffiner. Gradient-based learning applied to document recognition. In *Proc. IEEE*, 1998.

[20] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel svms is efficient. *Proc. CVPR*, 2008.

[21] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 2004.

[22] J. Mutch and D. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *IJCV*, 2008.

[23] R. Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 1999.

[24] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *Proc. CVPR*, 2007.

[25] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. on PAMI*, 2007.

[26] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu. On advances in statistical modeling of natural images. *Mathematical Imaging and Vision*, 2003.

[27] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proc. ICCV*, 2009.

[28] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 2004.

[29] C. Wren, A. Azarbayejani, and T. D. A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on PAMI*, 1997.

[30] J. Yang, K. Yu, and Y. Gong. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In *Proc. CVPR*, 2009.

[31] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture andobject categories: A comprehensive study. *IJCV*, 2007.

[32] L. Zhang, M. Tong, H. Marks, K. Tim, H. Shan, and G. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 2008.