Learning an Interest Operator from Human Eye Movements

Wolf Kienzle, Felix A. Wichmann, Bernhard Schölkopf, and Matthias O. Franz
Max-Planck Institute for Biological Cybernetics, Spemannstr. 38
72076 Tübingen, Germany
kienzle@tuebingen.mpg.de
http://www.kyb.mpg.de/~kienzle

Abstract

We present an approach for designing interest operators that are based on human eye movement statistics. In contrast to existing methods which use hand-crafted saliency measures, we use machine learning methods to infer an interest operator directly from eye movement data. That way, the operator provides a measure of biologically plausible interestingness. We describe the data collection, training, and evaluation process, and show that our learned saliency measure significantly accounts for human eye movements. Furthermore, we illustrate connections to existing interest operators, and present a multi-scale interest point detector based on the learned function.

1. Introduction

The past decade has seen an increasing interest in local feature based methods [31, 16, 7]. In this paradigm, visual objects are represented by sets of local image features, evaluated at particularly ”interesting” locations. This has several benefits. In contrast to methods that use global information (for example, a single image patch which covers the entire object), the parts-based approach is generally believed to be more effective in representing the complex variations that occur within visual object categories. In particular, articulated objects and partial occlusions can be handled naturally in this framework. Also, entire images can be encoded and processed efficiently, which is of great importance today as the size of databases keeps growing.

Before local features can be computed, interesting locations, or interest points, must be determined. To this end, one usually applies an interest operator, which takes the image as input and returns a set of coordinates where the image contains potentially interesting information. However, an appropriate notion of interestingness for, say, a categorization task, is rather complex and can only be vaguely defined. Most interest operators work under the simplifying (while physiologically plausible) assumption that interesting locations are points which are visually salient, i.e. ”pop out” in some way.

A variety of interest operators are currently used, each one implementing a different notion of saliency. For instance, the detectors by *Förstner* [9] and *Harris* [10] measure visual saliency by means of the local autocorrelation matrix. They were used in [31, e.g.] and [35] for image retrieval and object recognition, respectively. A scale invariant generalization of the *Harris* detector was applied in [17]. This idea was further extended to affine invariance in [18]. In the field of object detection, *Lowe* [16] proposed an interest operator based on maxima of the Laplacian (as in proposed in [15]) and local curvature. This approach has become popular also because in was introduced in connection with the widely used *SIFT* features which have been shown to be extremely robust under many image transformations [19]. Another widely used method was developed by *Kadir* and *Brady* [13] who derive their saliency measure from local entropy (usually of the intensity distribution) and changes in the local statistics over scale. An application of their method can be found in [7]. For more information on interest operators and a comprehensive comparison, see [30][20].

One of the original motivations for using interest operators is the fact that the human visual system samples images at discrete locations, called *fixation* points. It is only during a fixation — which typically lasts for a several tens to hundred milliseconds — that the incoming image is analyzed by the visual system. Between two fixations the gaze position moves extremely fast (within a few milliseconds) from the previous to the next fixation location. During these *saccades* the human visual system is known to be essentially ”blind” [5, 2, 29]. Additionally, the resolution at which the retinal image is processed decays with eccentricity [34, 28, 1]. This further underlines the special role of fixated locations, since the image at maximum resolution is only perceived around the fixation point.

While eye movements are known to depend on numerous *top-down* mechanisms, such as the observer's intentions, thoughts, etc. [36, 11], it is believed that *bottom-up* effects, i.e. the local image statistics at fixation points, play a role as well. Evidence for the importance of low-level vision was found in [27], where a significantly increased contrast around fixation points was reported. In [24], the authors study a large number of local features (such as edges, entropy, Laplacian, etc.), as to how they predict eye movements in different settings. They also present an actual application to image compression. In [21, 22], an explicit bottom-up saliency model based on color, intensity, and orientation filters (similar to the one proposed in [12]) is shown to account for saccade targeting. In [33], contrast and edges were found to be of particular importance. Recent studies with similar results are, [23, 33, 26, e.g.].

Note that the studies described above *construct* models and then test whether they account for eye movements. Similarly, existing interest point detectors are hand-crafted filters, even if they are physiologically motivated. In the work presented herein, we seek to *learn* the bottom-up saliency driving human eye movements, in order to build an interest point detector. The method is meant to complement existing algorithms for finding interest points which are rather ad hoc in terms of implementing actual interestingness. To this end, we deliberately *avoid* to construct features based on biological or physiological findings. Instead, we use a rather general, black-box type approach and simply train a classifier on fixated vs. randomly selected image patches.

2. Data Collection

2.1. Setup

Eye movement data was collected from 14 human subjects using 200 gray scale images of natural scenes. The 200 images were presented to each subject in four sessions (50 trials each). To prevent learning artifacts later on we proceeded as follows: First, the 200 images were randomly chosen from our database, which consists of 1626 calibrated natural images of size 4064×2704. From each image, we cut out a 1024×768 window at a random position, which resulted in "non-artistic" pictures, i.e. pictures that do not have the most salient object at the center and in focus. Effects of this subtlety are discussed in [27, e.g.]. The 200 images were presented to the subjects at full screen size on a linearized, 19 inch CRT in a medium darkened room at 60cm distance (which corresponds to approximately 37×27 degrees of visual angle, and to a resolution of 28 pixels per visual degree), with the instruction to perform *free viewing* (i.e. simply "look" at the pictures with no particular task in mind). Before each image, a pre-fixation target was shown on a uniform background, which subjects had been instructed to fixate. The pre-fixation target appeared at random locations before the presentation of each image to remove bias in initial gaze positions. For each subject, we randomized the order of images, the duration of pre-fixation targets ($\max(N(2, 0.5), 1)$ seconds, where $N(\mu, \sigma)$ is a normally distributed random variable with mean μ and standard deviation σ) and the image presentation time ($\max(N(3, 1), 1)$ seconds). Every ten images, a 4×3 grid of re-calibration targets was shown which the subjects had been told to fixate. These data were used off-line to account for drift and to estimate measurement errors, as described in the next section.

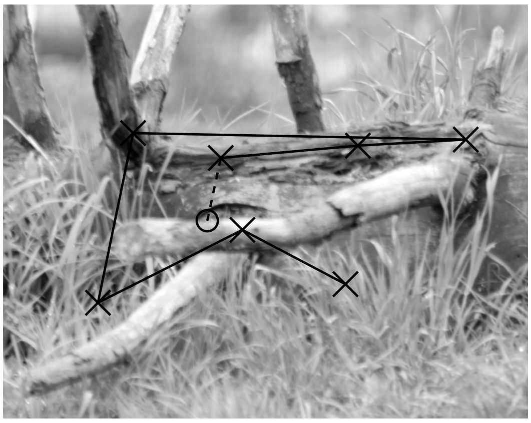

Figure 1. Sample eye tracking record. The black circle denotes the fixation position before the image appeared. The dashed line shows the first saccade, followed by further saccades (solid line). Fixations are marked by crosses.

We used a head-mounted video-based Eyelink II eye-tracker to record eye movements. Head tracking was activated, but we found it not sufficiently reliable, so we additionally used a chin rest and subjects were told keep their heads as still as possible. A careful calibration (using Eyelink software) was conducted at the beginning of each session, such that the accuracy was around 0.3 degrees or better. Subjects who did not achieve 0.5 degrees accuracy after several calibration attempts were excluded from the experiment and the subsequent analysis. The tracker was set to record gaze positions of both eyes at 250Hz. A sample eye tracking record is shown in Figure 1.

2.2. Drift Correction

A major concern in learning from eye movement data is the accuracy of the gaze position measurements. In practice, state-of-the-art eye tracking systems yield errors around 0.5 degrees of visual angle. For instance, [33] report a mean error of 0.40 ± 0.10 degrees during calibration. We obtained a similar result, 0.40 ± 0.14. Unfortunately, at a resolution of 28 pixels per degree, a deviation of 0.4 degrees corresponds to 11.2 pixels. This makes learning difficult, since even identical patches can seem uncorrelated when misaligned by this amount. Even worse, the system drifts over time, making the calibration error an optimistic estimate of the ”real” error. Finally, the features we are trying to learn may have a similar scale, if not smaller.

We therefore took great care to minimize and control measurement errors. For each set of re-calibration points (shown after every ten images), and for each eye separately, we fit an affine function to account for drift during the experiment by minimizing

$$\sum_i ||\mathbf{A}\mathbf{x}^{(i)}_{measured} + \mathbf{b} - \mathbf{x}^{(i)}_{target}||^2 \qquad (1)$$

w.r.t. $\mathbf{A} \in \mathbb{R}^{2\times2}$ and $\mathbf{b} \in \mathbb{R}^2$, where $\mathbf{x}_{target}$ and $\mathbf{x}_{measured}$ are the true positions of the re-calibration targets and their corresponding gaze measurements, respectively. Furthermore, we computed the leave-one-out error [32, e.g.] for each drift correction function $(\mathbf{A}, \mathbf{b})$. Based on this quantity, the worse performing eye was discarded (for each re-calibration stage separately). Then, the drift correction functions, as well as the error estimates were linearly interpolated (w.r.t. time) over the whole session of 50 images. Every fixation with an interpolated error above 1.0 degrees was discarded, the remaining fixations were corrected using the interpolated drift correction functions. Finally, we also discarded trials that were closer than 0.5 degrees to the screen boundary, or where the pre-fixation target had not been fixated within 2.0 degrees. This yielded 18065 fixations with an overall measurement error of 0.54 ± 0.19 degrees. Note that this estimate is more conservative than the calibration error, since it has been averaged over the entire experiment. This is similar to [23], where errors were estimated at the end of each session (0.54 ± 0.44 degrees).

2.3. Consistent Locations

As mentioned earlier, eye movements are not only driven by bottom-up saliency, but also by top-down mechanisms. For learning bottom-up saliency, this is clearly an undesirable effect, since it introduces trends to the data that cannot be captured by our simple model. Moreover, we have to expect that top-down effects vary among subjects (e.g., the *free viewing* task will be interpreted differently by different subjects) and change over time [21, 33]. To remedy this, we only considered image locations that were consistent among subjects, i.e. which had been fixated sufficiently many times. This was based on the assumption that top-down strategies tend to be too complex to yield overlapping fixations (the average number of fixations per subject and image was less than 6.5). Thus, we conjecture that consistent locations are generated by a type of saliency which is independent of subject, time, or top-down mechanisms.

To compute consistent locations, we proceed as follows (Figure 2). Consider one of our 200 training images. If

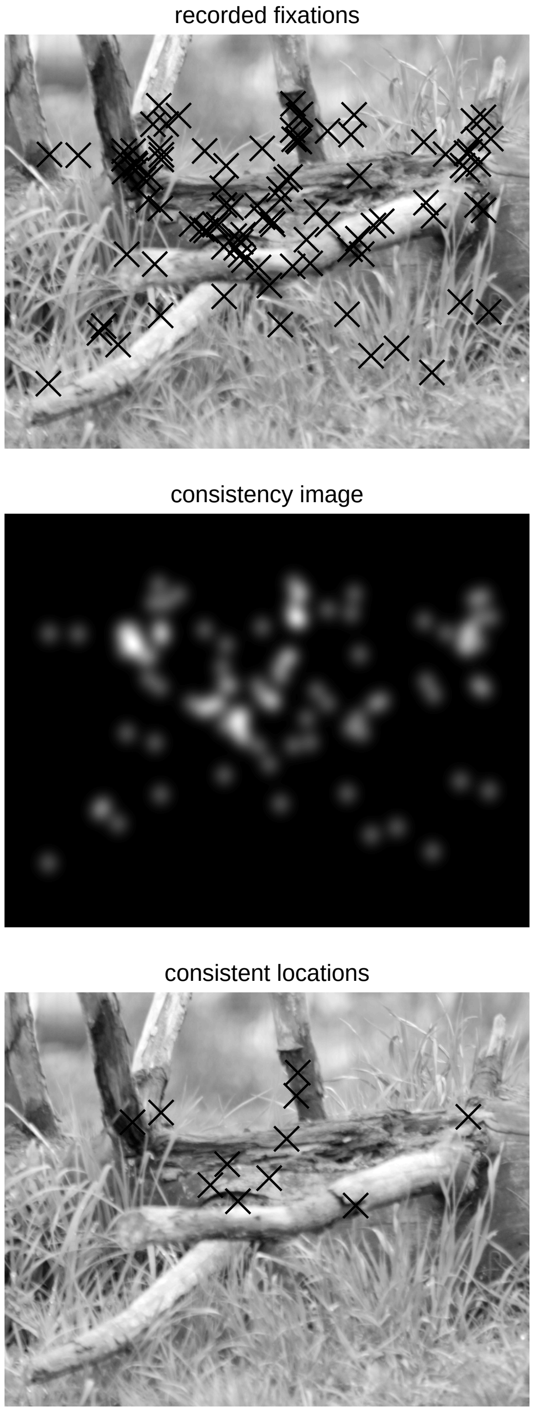

Figure 2. Finding consistent fixation points. The top panel shows the sample image from Figure 1 together with all recorded fixations from all subjects (the black ”x”s). The middle image shows the corresponding consistency image p(x,y|F). The thresholded maxima are plotted in the bottom picture.

we assume that the gaze measurement error is normally distributed with mean zero and standard deviation equal to $\sigma_m = 0.54$ (the overall measurement error), we can compute the likelihood that a position (x, y) was a fixation target, given the measured fixation locations (of all subjects) $F = \{(x, y)_i\}$ for that image. The likelihood — or consistency image (Figure 2, middle) — is given by

$$p(x, y|F) = \frac{1}{|F|} \sum_F \exp(-\|(x, y)_i - (x, y)\|/2\sigma_m^2) \quad (2)$$

All local maxima of $p(x, y|F)$ that were above a threshold $\delta = 1.8/|F|$ were considered as consistent locations. This choice for the threshold was chosen by hand. It corresponds to a location being consistent if there are, for example, at least two fixations within a 0.25 degree radius. This procedure yielded 1670 consistent locations, approximately eight per image.

Note that a consistent location can be seen as a weighted sum of nearby fixation measurements. This averaging effect brings further benefits in terms of measurement errors by reducing noise in the gaze measurements.

2.4. Patch Extraction

Regarding the representation of the local image structure, it is not clear on which scale to look for salient features. Ideally, one would use a multi-scale representation. This would, however, introduce additional parameters for each location, which would render the learning problem significantly harder. For the sake of simplicity, we chose a fixed patch size, which was determined by cross-validation. To this end, we cut out 10 centered square image patches at each consistent location, varying in size between 0.6 and 25 degrees of visual angle (if required, we mirrored the image at its boundaries). This range of sizes allows for an extensive search over scales, since 0.6 degrees is on the order of the measurement error, and 25 degrees covers most of the screen. All patches were low-pass filtered such that they could be down-sampled to 13×13 pixels without aliasing effects, and the pixel values were the stacked into 169-dimensional feature vectors $\mathbf{x}_i$. This yielded 10 data sets (one for each patch size) of size 1670. Finally, all data were associated with the label $y_i = 1$, denoting the positive class in our classification setting.

2.5. Negative Examples

The randomized image order, duration, pre-fixation position, etc. all aim at minimizing factors other than local image structure that may account for gaze positioning. To further reduce the risk of learning such artifacts, we collect the negative examples from the *same* locations as the positives, but with the image data taken from *different* images (as in [27]). That way, the negative examples have exactly the same position statistics as the positive examples. A possible bias due to the statistics of gaze positions (e.g., boundary effects) thus becomes invisible to a discriminative method.

We collected one negative example for each consistent location. As with the positive examples, this yielded 10 data sets of size 1670 (this time, with labels $y_i = -1$).

3. Training

3.1. Algorithm

We used a support vector classifier [4, 32] to learn the difference between positive and negative examples. Given the labelled training data $(\mathbf{x}_i, y_i) \in \mathbb{R}^{169} \times \{-1; 1\}$, $i = 1 \ldots m$, we solved the standard SVM problem

$$\begin{array}{ll} \min & \frac{1}{2}\sum_{i,j=1}^m \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^m \alpha_i \\ \text{s.t.} & 0 \leq \alpha_i \leq C \\ & \sum_{i=1}^m \alpha_i y_i = 0 \end{array} \quad (3)$$

for α_i, $i = 1 \ldots m$ using a *Matlab* wrapper for *LIBSVM* [6]. The regularization strength was set to $C = 1$ (based on preliminary experiments). The kernel function was a Gaussian $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, γ was found by cross-validation (see below). The solution of (3) is takes the form

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i y_i \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}\|^2). \quad (4)$$

Note that we use f as a real valued saliency measure, not a decision function (which would have an additional threshold parameter and a sign function), for reasons explained in the next section.

3.2. Performance Measure

Prediction performance was measured in ROC (receiver operator characteristic) score [25], also referred to as AUC (area under the ROC curve) or *Wilcoxon-Mann-Whitney* statistic. We preferred this measure over classification accuracy, since it does not depend on the specific choice of the decision threshold. The ROC score estimates the probability that two examples, drawn at random, are ranked in correct order. Random guessing therefore corresponds to a ROC score of 0.5, whereas a perfect predictor would have a ROC score of 1. Recent work by [33] showed that eye movements can be accounted for by a bottom-up saliency model with a ROC score of 0.63. Note that the performance of their predictor is rather low compared to what is typical to many computer vision and machine learning problems (e.g., face detection or character recognition). It is however not surprising, since bottom-up saliency can explain only part of a very complex mechanism.

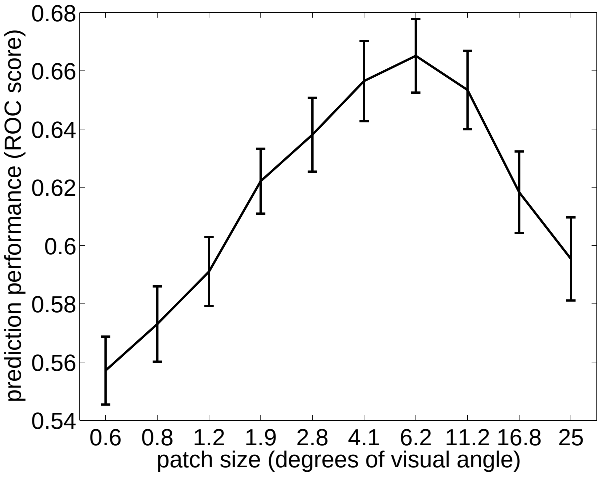

Figure 3. The performance of our model as a function of the patch size. Error bars denote one standard deviation of the mean performance.

3.3. Model Selection

We used 13×13 patches to encode the local image content. In a preliminary study, we found this a reasonable choice that trades off insufficient structure (too few pixels) for superfluous parameters (too many pixels). Note however, that salient structure may not only occur at pixel level but also on larger scales. Therefore, the actual visual *size* of the patch (measured in visual degrees) was determined by cross validation.

A problem with large patch sizes is that the extreme low-pass filtering makes patches of nearby fixations indistinguishable. Therefore, care must be taken that such clusters have all their members in either the training or the test set, never in both. Otherwise, the generalization performance may be overestimated. For our cross-validation procedure, we ensured this by splitting the data *image-wise* into training and validation folds, such that validation points always came from unseen images.

Prior to training and testing the classifier, all patches were preprocessed to have zero mean. We found that the performance did not degrade but improve slightly compared to raw pixel values. This is not surprising, since the saliency of the local mean intensity of a patch most likely depends on the intensities outside that patch, which is not encoded in our representation. Therefore, retaining the mean does not add relevant information and can only mislead the learning machine.

For the 10 patch sizes (0.6 up to 25 degrees, equally spaced on a log scale), we trained and tested our saliency model using a 100 fold cross-validation, i.e. we trained on 198 and tested on 2 images. The bandwidth of the Gaussian kernel σ was found by trying 16 values between $10^{-1} \ldots 10^{2}$ on a log scale. The mean ROC score of the best performing σ and its standard deviation as a function of patch size are shown in Figure 2. Please note that the prediction performance peaks around 6 degrees patch size. We performed a paired sign test, creating two ROC scores for each cross-validation fold, comprising 1) the performance of our learned model and 2) the performance of the same model, but learned with scrambled labels y_i. The p-values between 1.2 and 16.8 degrees were consistently below 10^{-6}, indicating that the learning machine was able to extract significant structure. Also, note that at 0.67, the performance of our black-box method is comparable to the results from [33] (ROC 0.63), although their approach uses hand-crafted, physiologically plausible features.

4. Experiments

4.1. A Comparison to Existing Methods

The purpose of the following experiment is to outline connections between the learned saliency function and existing interest operators. First, we tested how well frequently used saliency measures predict the eye movements recorded in our data set, compared to our model (ROC score 0.67). However, we did not use full implementations of the interest point detectors, since these usually do not yield saliency values at all image position, but only at isolated points (e.g., scale space maxima of the Laplacian [15, 16]). Instead, six saliency measures were computed:

- The local r.m.s. contrast (or signal energy). For zero-mean patches, this corresponds to the standard deviation of the 13×13 pixel values.
- The entropy of the local intensity histogram, computed from a 13×13 neighborhood. This feature is used by *Kadir* and *Brady*'s method [13])
- The determinant and trace of the local Hessian $\mathbf{H}$, as used in *Lowe*'s [16] detector and in *Lindeberg*'s [15] scale selection principle (note that the trace is equivalent to the Laplacian). Before computing the Hessian, the image was down-sampled (after appropriate low-pass filtering) by a factor of $3/13$ in order to make the scale of the derivative filter comparable to that of the learned operator. For both of these quantities the absolute value was used.
- The determinant and trace of the local autocorrelation matrix $\mathbf{M}$, as used in the *Harris* [10] or Förstner [9] detector. The size of the neighborhood was adapted as in the Hessian case.

Note that the choice of corresponding scales for the different filters is by no means optimal, it is rather meant to allow for a qualitative comparison. Still, we found that all

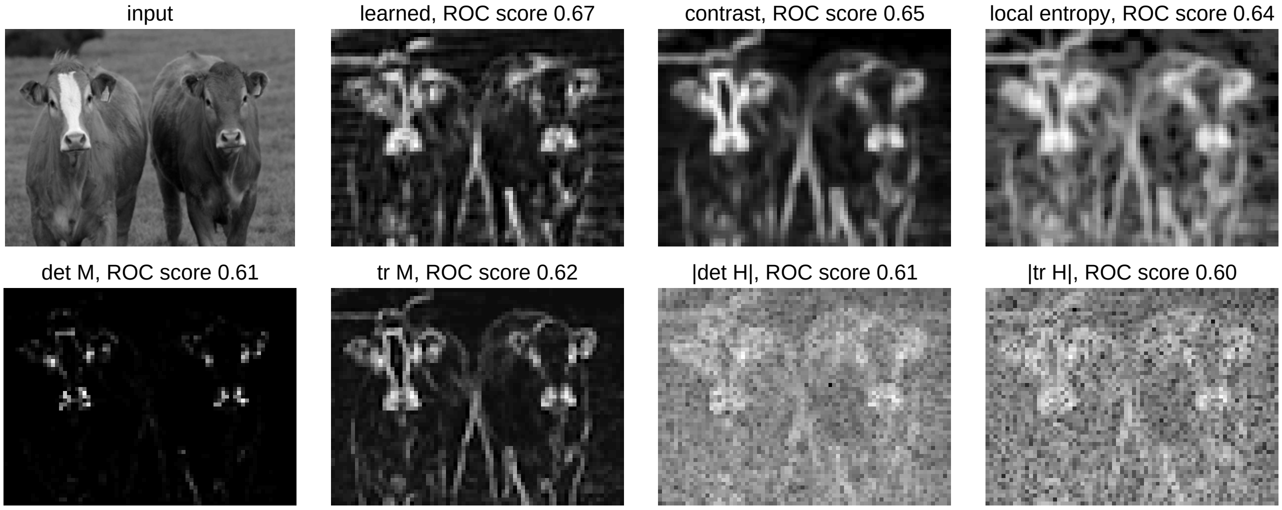

Figure 4. A sample image (top left) and various notions of local saliency.

features yielded ROC scores above 0.60, the best ones being entropy (0.64) and r.m.s. contrast (0.65). A possible explanation for this is that all tested measures (including our learned function) are correlated with contrast. This would be in agreement with the fact that contrast is arguably one of the most salient low-level features [27, 21, 33]. Interestingly, entropy performs about as well as contrast. But *Kadir* and *Brady* [13] argue that entropy alone cannot be good saliency measure, since it uses no structural information, i.e. pixels within a patch can be permutated without changing the entropy. This does not contradict our results, however. Since our patches were taken from natural images, the probability of obtaining a patch which is completely unstructured *and* has high entropy is low [8]. We might even conjecture that in the domain of natural images, entropy is mainly correlated with contrast, which would explain the good prediction performance of this otherwise poor saliency measure.

To illustrate the connections between the tested features, Figure 4 shows saliency maps evaluated on a test image. Note that all saliency measures are somewhat correlated, but there are subtle differences. These will become clearer in the following experiment.

4.2. The Learned Detector

Most of the current interest point detectors [16, 13, 18] search for salient locations not only on one scale, but on a scale space over the image [15]. The scale space is often *Gaussian*, i.e. the third (scale) dimension corresponds to the standard deviation of a Gaussian filter with which is image is convolved. It is common to associate interest points with a *characteristic scale*, at which some saliency measure is maximal over scale. For example, a necessary condition for an interest point in *Lowe*'s method is that it is a local maximum of the Laplacian in a Gaussian scale space. It will be associated with this scale, if it passes further saliency tests and becomes an actual interest point (this *scale selection mechanism* was first proposed in [15]).

To turn our learned saliency model into a multi-scale interest point detector, we follow a similar approach. We construct a Gaussian scale space (with three intermediate scales per octave, as in [16]) and simply output all scale space maxima of our saliency measure. The top 50 (w.r.t. saliency) interest points on our sample image are shown in Figure 5. For a comparison, outputs of the *Lowe* and *Kadir/Brady* method are also given. In the former, the r parameter (the maximal ratio of the principal curvatures) was set to 10, and the top 50 (w.r.t. the magnitude of the Laplacian at the characteristic scale) outputs are shown. For the latter, the saliency and inter-saliency thresholds were set to 0.25 and 1.2, respectively. This yielded exactly 50 interest points. Note that despite the similarity of the saliency maps in Figure 4, the selected interest points in Figure 5 are considerably different.

5. Discussion

In this paper, we described our ongoing work on learning interest operators from eye movements. In order to make a learning approach feasible at all, we had to take great care in thoroughly setting up the experimental and data collection procedure. We found that the prediction performance of our approach (ROC score 0.67) is comparable to state-of-the-art methods for modelling eye movements despite that we did not use any biological prior knowledge. When comparing the learned function to saliency measures derived from existing interest operators, it turned out that all of them could

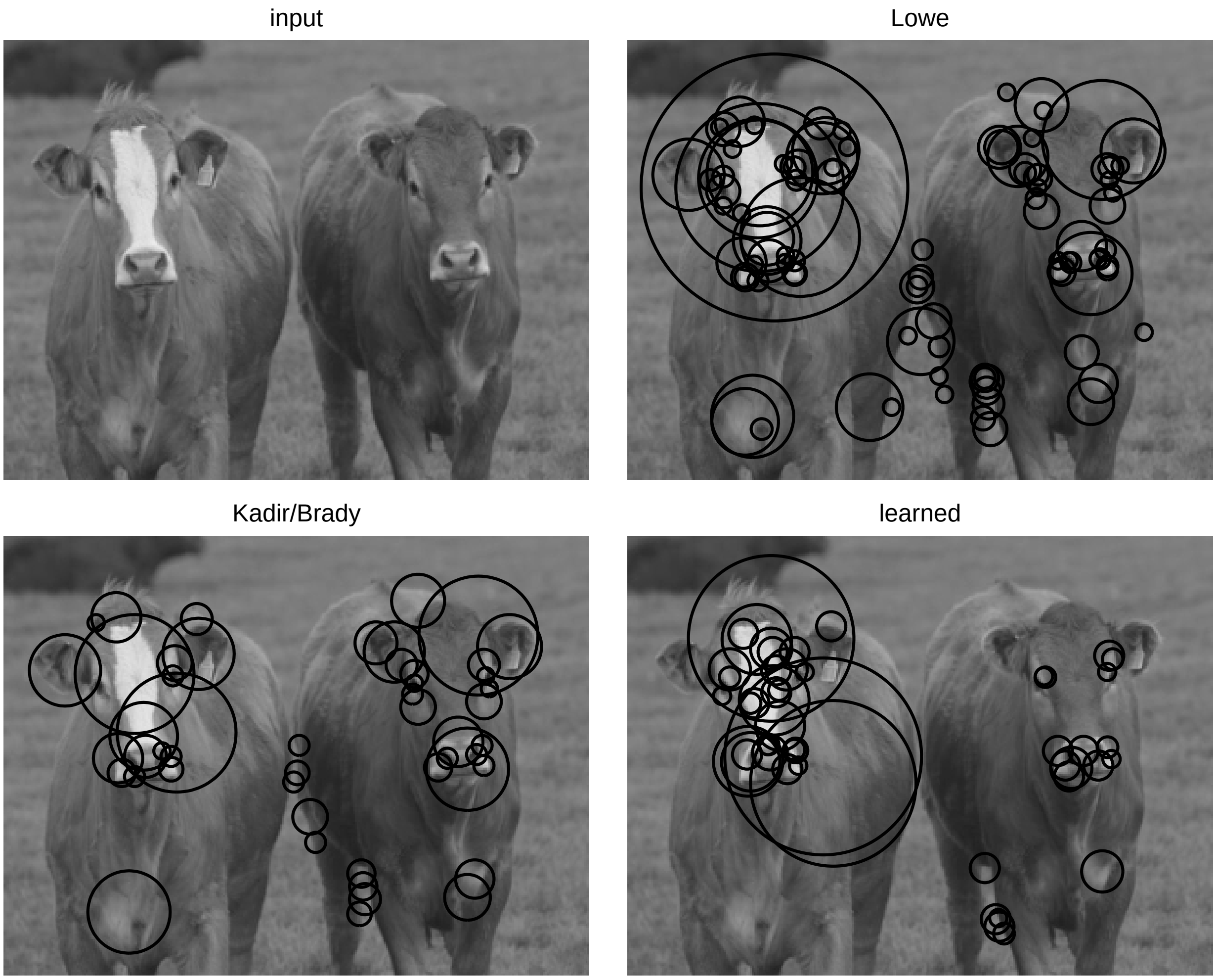

Figure 5. A sample image (top left) and the top 50 interest points for Lowe's method (top right), the Kadir/Brady detector (bottom left) and our learned interest point detector (bottom right).

predict our eye movement data reasonably well. This is probably due to the fact that the investigated saliency measures strongly correlate with contrast which is known to be one of the most salient features in human low-level vision.

We also presented a preliminary multi-scale detector derived from our learned saliency function. Before it can be applied in computer vision applications, some issues have to be resolved: first, the support vector classifier is computationally inefficient. We are currently incorporating ideas from [3] and [14] to remedy this; second, the performance has to be tested in connection with parts-based object detection and categorization methods.

References

[1] M. S. Banks, A. B. Sekuler, and S. J. Anderson. Peripheral spatial vision: limits imposed by optics, photoreceptors, and receptor pooling. *Journal of the Optical Society of America*, 8(11):1775–1787, 1991. 1

[2] B. Bridgeman, D. Hendry, and L. Stark. Failure to detect displacement of the visual world during saccadic eye movements. *Vision Research*, 15:719–722, 1975. 1

[3] C. J. C. Burges. Simplified support vector decision rules. In *International Conference on Machine Learning*, pages 71–77. Morgan Kaufmann, 1996. 7

[4] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998. 4

[5] D. Burr, M. Morrone, and J. Ross. Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature*, 371:511–513, 1994. 1

[6] C.-C. Chang and C.-J. Lin. LIBSVM – a library for support vector machines, version 2.71, `www.csie.ntu.edu.tw/∼cjlin/libsvm/`. 4

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, page 264, 2003. 1

[8] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, 4(12):2379–2394, 1987. 6

[9] W. Förstner. A framework for low level feature extraction. In *European Conference on Computer Vision*, 1994. 1, 5

[10] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988. 1, 5

[11] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005. 2

[12] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000. 2

[13] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001. 1, 5, 6

[14] W. Kienzle, G. H. Bakir, M. O. Franz, and B. Schölkopf. Face detection — efficient and rank deficient. In Y. W. Saul, L.K. and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 673–680. MIT Press, 2005. 7

[15] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998. 1, 5, 6

[16] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, 1999. 1, 5, 6

[17] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, volume 1, pages 525–531, 2001. 1

[18] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, pages 128–142, 2002. 1, 6

[19] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005. 1

[20] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005. 1

[21] D. J. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002. 2, 3, 6

[22] D. J. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16(2):125–154, 2003. 2

[23] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(8):2397–2416, 2005. 2, 3

[24] C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000. 2

[25] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Third International Conference on Knowledge Discovery and Data Mining*, 1997. 4

[26] R. Raj, W. S. Geisler, R. A. Frazor, and A. C. Bovik. Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *Journal of the Optical Society of America A.*, 22(10):2039–2049, 2005. 2

[27] P. Reinagel and A. M. Zador. Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, 10(4):341–350, 1999. 2, 4, 6

[28] J. G. Robson and N. Graham. Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, 21:409–418, 1981. 1

[29] J. Ross, M. C. Morrone, M. E. Goldberg, and D. C. Burr. Changes in visual perception at the time of saccades. *Trends in Neurosciences*, 24(2):113–121, 2001. 1

[30] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000. 1

[31] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), 1997. 1

[32] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002. 3, 4

[33] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5p):643–659, 2005. 2, 3, 4, 5, 6

[34] V. Virsu and J. Rovamo. Visual resolution, contrast sensitivity, and the cortical magnification factor. 37(3):475–494, 1979. 1

[35] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *European Conference on Computer Vision*, 2000. 1

[36] A. Yarbus. Eye movements and vision. *Plenum Press*, 1967. 2