How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements

Wolf Kienzle[1], Bernhard Schölkopf[1], Felix A. Wichmann[2,3], and Matthias O. Franz[1]

[1] Max-Planck Institut für biologische Kybernetik, Abteilung Empirische Inferenz, Spemannstr. 38, 72076 Tübingen
[2] Technische Universität Berlin, Fakultät IV, FB Modellierung Kognitiver Prozesse, Sekr. FR 6-4, Franklinstr. 28/29, 10587 Berlin
[3] Bernstein Center for Computational Neuroscience, Philippstr. 13 Haus 6, 10115 Berlin

Abstract. Interest point detection in still images is a well-studied topic in computer vision. In the spatiotemporal domain, however, it is still unclear which features indicate useful interest points. In this paper we approach the problem by *learning* a detector from examples: we record eye movements of human subjects watching video sequences and train a neural network to predict which locations are likely to become eye movement targets. We show that our detector outperforms current spatiotemporal interest point architectures on a standard classification dataset.

1 Introduction

Interest point detection is a well-studied subject in the case of still images [14], but the field of spatiotemporal detectors for video is fairly new. Currently, there exist essentially two methods. The earlier one is a spatiotemporal version of the *Harris* corner detector [4] proposed by *Laptev* [9]. This detector has been shown to work well in action classification [15]. However, spatiotemporal corners are a relatively rare event, resulting in overly sparse features and poor performance for many real-world applications [1, 11]. To remedy this, the *periodic* detector was introduced by *Dollár* [1]. It responds to simpler spatiotemporal patterns, namely intensity changes in a certain frequency range. The authors show that a simple recognition framework based on this detector outperforms the Harris-based approach of [15].

As both of these approaches are relatively new, they are still far from being as well-understood and empirically justified as their spatial counterparts. Clearly, spatiotemporal corners and temporal flicker of a single frequency are only a subset of all potentially interesting events in a video. Here, we present a new approach to spatiotemproal interest point detection. Instead of *designing* new interesting spatiotemporal features, we *learn* them from an already working, and very effective interest point detector: the human visual system. Our basic idea

is to record eye movement data from people watching video clips and train a small neural network model to predict where people look. Used as an interest point detector, the neural network is shown to outperform existing methods the same dataset which has been used as a benchmark for both the Harris and the Periodic detector.

The connection between eye movements and interest operators has been made before by several authors. In [13], a biologically inspired attention model was used for object recognition in images. The idea of designing an interest point detector directly from eye movement data was recently proposed in [7, 8]. They found that humans attend to center-surround patterns, similar to what is already being used in some engineered detectors, e.g., [10]. However, their approach only considers still images, and they do not report how their system performs on typical computer vision tasks.

2 Eye Movements

The human visual system has its highest resolution at the center of gaze, or *fovea*, which covers about one degree of visual angle. In fact, a disproportionately large amount of cortical processing power is devoted to this small area. Towards the periphery, both resolution and processing power decay quickly [17]. As a consequence, a visual scene does not enter the visual system as a whole, but is *sampled* by the eyes moving from one location to another. During eye movements, the center of gaze is either held fixed over a constant image area during *fixations*, follows moving objects during *smooth pursuit*, or changes rapidly during *saccades* in which visual input is mostly turned off (*saccadic suppression*) [2]. The choice of which image regions become saccade targets is not random, however. In part, it can be explained by typical patterns in the local image structure occuring at fixated image locations [12, 8]. Thus, the human eye movement mechanism bears a resemblance to interest point detectors in that it uses local image statistics to decide where to sample visual input for subsequent processing.

The aim of this work is to build an interest point detector that imitates this effect. To this end, we recorded eye movement data from 22 human subjects. Each subject viewed 100 short clips from the movie *Manhattan (1979)*, presented on a 19" monitor at 60cm distance at 24 frames per second with a resolution of 640×480 pixels. Each clip was 167 frames long (about seven seconds), and the clips were sampled uniformly from the entire film (96 min) such that no cuts occurred during a clip. Each subject viewed all 100 clips in random order and with blanks of random duration in between. No color transform was applied, since the movie is black and white. Eye movements were recorded using an *Eyelink II* tracker, which, after careful calibration, yielded measurements of typically 0.3 degrees accuracy. Figure 1 shows three frames from an example clip together with the recorded fixations from all 22 subjects.

In a post-processing step we discarded all fixations that occurred before frame 38 or after frame 148 to ensure a sufficient number of video frames both before and after each fixation. Also, a set of background (negative) examples was gen-

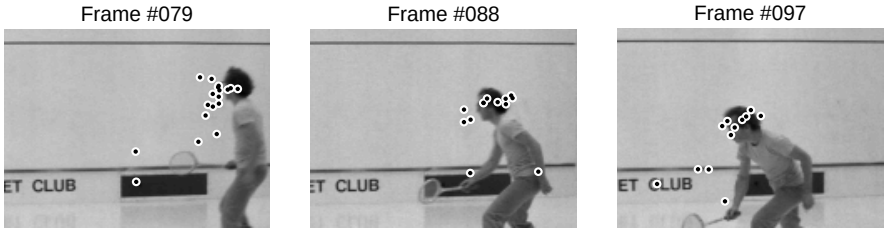


Fig. 1. Recorded eye movements on a sample video from our dataset (Section 2). Fixations from all 22 users are shown as circles (there are no markers for subjects which did not fixate, but moved their eyes during the respective frame).

erated by using the same fixation positions, but with the video data taken from *wrong*, i.e., randomly chosen clips. This way of generating background examples is common practice in eye movement analysis and prevents artifacts caused by the non-uniform sampling prior due to the limitations of the viewing field and head motion in the eye tracking setup [12]. Finally, we split the set of all fixations and background points into a training set (18691 samples), and a test set (9345 samples). The training set was used for designing the *learned* detector (Section 3.3), the test set was used to compare the three interest point algorithms in terms of how well they predict human fixations (Section 4.1).

3 Spatiotemporal Interest Point Detectors

3.1 The Spatiotemporal Harris Detector

The spatiotemporal *Harris* detector is due to *Laptev* [9], and extends the widely-used Harris corner detector [4] to the time axis. Analogously to the spatial case, the spatiotemporal Harris detector is based on the 3×3 second-moment matrix M, which describes the local gradient distribution, spatially at scale σ and temporally at scale τ. Interest points are computed as the local maxima of the quantity

$$S_H = \det M - k(\text{trace } M)^3, \tag{1}$$

where $k = 0.005$ is an empirical constant [9], corresponding to the well-known magic number 0.04 in the original spatial detector [4]. Here, we refer to S_H as the *saliency function* of the detector, according to the biological term *saliency* [5] which is used to describe the *interestingness* of locations in an image. Note that the output of the detector is a discrete set of locations, while S_H is defined on the entire video clip. In practice a second set of scales σ_i, τ_i is used for integration of the moment matrix over a spatiotemporal window [9], usually taken to be a multiple of σ, τ. Throughout this paper we used the implementation from [1], with the default setting of $\sigma_i = 2\sigma$, $\tau_i = 2\tau$. Thus, the detector has two free parameters, the spatial scale σ and the temporal scale τ.

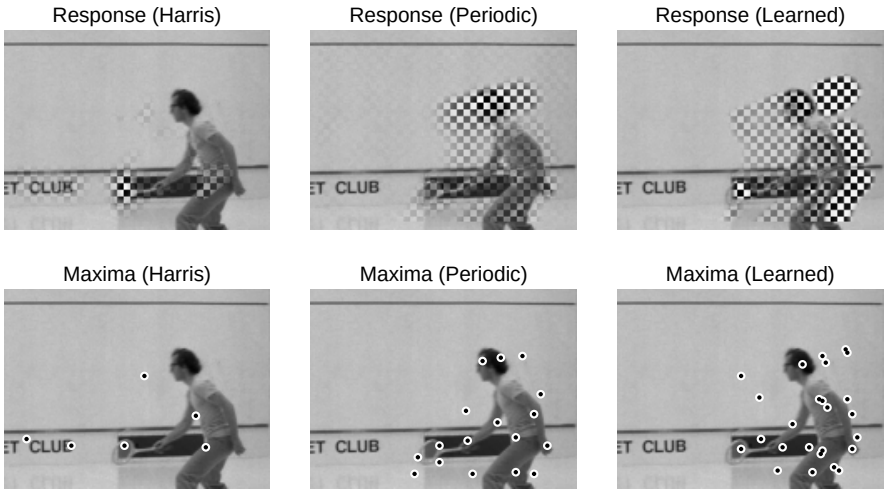


Fig. 2. Qualitative comparison of detector responses S_H (eq. 1), S_P (eq. 2), and S_L (eq. 3). The blended checkerboard texture in the top row illustrates detector responses on frame 88 from Figure 1. The bottom row shows the corresponding regional (2D) maxima. Parameters were set to $\sigma = 2$, $\tau = 3$ for all detectors.

The response of the spatiotemporal Harris detector can be characterized similarly to the 2D case: the saliency function S_H, or *cornerness*, is large if the spatiotemporal gradient varies significantly in all three dimensions. *Laptev* intuitively describes the detected events as split or unification of image structure and as spatial corners changing direction. The applicability of this concept to action classification was shown in [15]. S_H computed on the center frame of our sample sequence in Figure 1 is shown in Figure 2 (left column). The highest values are achieved where the racket passes the black bar in the background.

It should be mentioned that in the conceptual simplicity of the spatiotemporal Harris detector lies also a possible drawback. Clearly, the time axis is not just a third image dimension, such as in volume data [3], but it describes a very different entity. Perhaps not surprisingly, it was found that the 3D-Harris detector can lead to unsatisfactory results, in that it tends to produce too few interest points [1, 11]. This has given rise to the development of the *Periodic* detector, which we describe in the following section.

3.2 The Periodic Detector

The so-called *Periodic* detector was proposed by *Dollár* [1] as an alternative to *Laptev*'s method. In *Dollár*'s approach, the image is smoothed spatially and then filtered temporally with a quadrature pair of one-dimensional Gabor filters. The

squared outputs of the two Gabor filters are added to get the saliency function

$$S_P = \sum_{i=1}^{2} (I * G(\sigma) * F_i(\tau, \omega))^2 \tag{2}$$

where I denotes the 3D image, $G(\sigma)$ is a 2D spatial Gaussian filter with standard deviation σ, and $F_1(\tau, \omega)$, $F_2(\tau, \omega)$ are 1D temporal Gabor filters with frequency ω and scale τ (with odd and even phase, respectively, as illustrated in Figure 3, right plot). Interest points are again the local maxima of the saliency function over space and time. In the implementation we use [1] the frequency is fixed to $\omega = 0.5/\tau$. In effect, this detector has the same parameters as the Harris detector, namely σ for the spatial scale and τ for the temporal scale.

Intuitively, the saliency S_P is large where image intensity changes temporally at a rate $\omega = 0.5/\tau$. Accordingly, the authors [1] refer to this detector as the *Periodic* detector. Figure 2 shows its output on the frame 88 of our example sequence (Figure 1). This suggests that, as intended [1], S_P takes significant values in larger regions than the Harris measure S_H.

3.3 The Learned Detector

The Harris and Periodic detector are based on analytic descriptors of local image structure assumed to be useful for computer vision applications. The interest point detector we propose here is instead based on image features selected by the human visual system.

The architecture of our detector is movitated by that of the Periodic detector (2). It consists of a simple feed-forward neural network model with sigmoid basis functions

$$S_L = b_0 + \sum_{i=1}^{k} \alpha_i \tanh(I * G(\sigma) * W_i + b_i), \tag{3}$$

i.e., the input video I is first convolved with a spatial Gaussian low pass G of width σ, then by k temporal filters W_i. The k filter outputs are fed into tanh nonlinearites (with bias b_i) and then added together using weights α_i and a global bias term b_0. Note that this generalizes the Periodic detector to an arbitrary number of arbitrarily shaped input filters: instead of two quadratic basis functions we now have k sigmoids, and the temporal filters will be fitted to the eye tracking data instead of being fixed Gabor filters. Additionally, each basis function contributes to the output S_L with a different weight and bias.

In the learning step, we fit the saliency function (3) to our recorded eye movement data: we optimize the filters W_i, the weights α_i, and the biases b_i using regularized logistic regression, i.e., by minimizing

$$E = \sum_{i=1}^{m} (y_i s_i - \log(1 + \exp s_i)) + \lambda \sum_{i=1}^{k} \alpha_i^2 \tag{4}$$

Here, the s_i are the values of S_L at the training samples (see Section 2). The corresponding labels y_i are set to 1 if i is a fixation, and 0 if it is a background

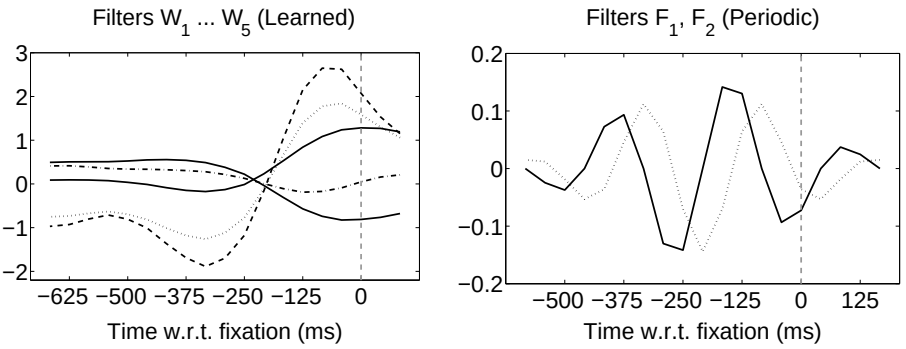


Fig. 3. The 19-tap temporal filters from the Learned (*left*) and the Periodic (*right*) detector. Shown on the horizontal axis is the time relative to the beginning of a predicted fixation (horizontal gray line). Note that both detectors have different offsets in time, corresponding to the values which are optimal in terms of predictivity (cf. Table 1): -7 and -5 frames (w.r.t. the central tap) for the Learned and Periodic detector, respectively .

example. Note that this corresponds to a maximum a posteriori estimate in a logit model, i.e., the learned saliency function S_L has a probabilistic interpretation: it equates to the logarithmic odds ratio of a fixation by a human observer, $S_L = P(Y = 1|I)/P(Y = 0|I)$. To carry out the optimization of (4) we used a scaled conjugate gradient method [16]. Prior to training, the training data were denoised to 95% variance by projecting out the least significant PCA components. The network weights were initialized to random values.

During learning, several design parameters have to be set: the regularization parameter λ, the number of filters k, and the spatial scale σ of the Gaussian. The size of the temporal filter W_i was set to 19 frames which corresponds to three times the value of $\tau = 3$ in the Harris and the periodic detector, the standard setting used in [1, 11] and also throughout this paper. Additionally, we introduce a temporal offset Δt, which denotes the position of the center of the temporal filters W_i relative to the beginning of a fixation. The rationale behind this is that the time at which a fixation is made does not necessarily coincide with the time at which the video contains the most useful information to predict this. As an example, the typical *saccade latency*, i.e., the time between seeing something interesting and making a saccade is 150−200ms (6−8 frames at 24 fps) [2]. The design parameters were found via 8-fold cross-validation, where the performance was measured in area under the ROC curve (ROC score), the standard measure for predicting eye movements [7]. The search space was a 4D grid with $\log_2 \sigma \in [-1 \dots 8]$ in steps of 2/3 ranging from single pixels to the full screen, $\Delta t = -29 \dots 9$ in steps of 2, $k = 1, 2, 5, 10, 20$, and $\log_{10} \lambda = -4, -2, 0, 2$. We found a clear performance peak at $\sigma = 1$, $\Delta t = -7$, $k = 5$ and $\lambda = 0.01$. We will refer to the detector trained with these parameters in the following as the *learned* detector.

The right plot in Figure 2 shows the output S_L on our example sequence from Figure 1. Note that, similarly to the periodic detector, our detector has a large response over extended areas. Interestingly, the largest Harris response (at the racket) leads to a high response, too. The five learned filter kernels W_i are shown in Figure 3 (left plot). As found during learning, the optimal tenmporal aligment of the filter kernels is at $\Delta t = -7$, which centers them at about 300ms before the fixation event. Examining the shape of the learned kernels, we find that all kernels have a steep slope 200ms before the fixation event, which means that the detector is tuned to temporal intensity changes occuring at that time. Interestingly, this matches very well with the typical saccade latency of 150-200 ms, i.e., the time between deciding to make and making a saccade (the saccades themselves are typically very short (20-50ms)). Note that we did not put any such assumption into the design of our detector. Therefore, this property must stem from the data, meaning that our detector has in fact *learned* a biologically plausible feature of bottom-up saliency.

4 Experiments

4.1 Eye Movement Prediction

For still images it has been shown that simple local image statistics such as increased RMS contrast attract the human eye [12]. As most spatial interest point detectors strongly respond to local contrast, they do in fact explain some of the variance in human eye movments. For time-varying images, it is known that flicker and motion patterns attract our attention [2]. Since the Harris and Periodic detector respond to such features, we expect a significant correlation with human eye movements in this case as well. To quantify this, we computed ROC scores of the saliency functions R_H (Harris), R_P (Periodic), and R_L (Learned) on our testset (Section 2). ROC scores are the standard measure for eye movement prediction [8]. In still images, the state-of-the-art for purely *bottom-up* (based on image content only) models is around .65 [8]. Note that this seemingly low score makes perfect sense, since eye movements are also controlled by more high-level, *top-down* factors, such as the observers thoughts or intentions [18], which are not considered by bottom-up models by construction.

Here, we compare the three detectors in terms of how well they predict human fixation locations. To reduce the inherent advantage of the *Learned* detector—which was built for this task—we also trained the free parameters of the Harris and the *Periodic* detector: analogously to Section 3.3, we fixed $\tau = 3$ and optimized σ and Δt on the training set via cross-validation. Test ROC scores (averaged over eight random subsets of the test set, $\pm$ standard error) are shown in Table 1, together with the optimal values for σ and Δt found in cross-validation. This shows that our detector outperforms the two others by a large margin, reaching state-of-the-art performance. This is not surprising since we specifically designed the Learned detector for this, while the others were not. Another observation is that the optimal temporal offset Δt is very similar in all three cases, and in agreement with the typical saccadic latency of $6 - 8$ frames (cf. Section

Detector	ROC score	$\log_2\sigma$	Δt
Learned	.634 $\pm$.007	0.0	-7
Periodic	.554 $\pm$.015	-1.0	-5
Harris	.522 $\pm$.005	3.3	-8

Table 1. How human eye movements are prediced by spatio-temporal interest point detectors (Section 4.1).

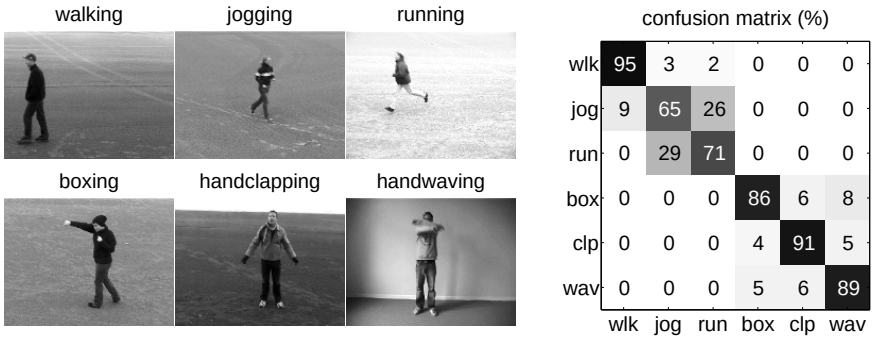
3.3). Also, all detectors have scores significnatly above chance level, which means that they are indeed related to the spatiotemporal features that the human eye is attracted to.

4.2 Action Classification

We have seen that the Learned detector outperforms existing methods in terms of predicting eye movments. This, however, should be regarded only as a proof of concept, since our main interest is to solve actual computer vision problems, not to predict eye movements. To make a fair comparison, we tested our detector within the domain for which the Harris and Periodic detectors were designed. We used the KTH action classification dataset [15], which was also used by the inventors of the Harris and Periodic detector to test their approaches. The dataset contains 598 videos (160×120 pixels, several seconds long) of 25 people performing 6 different actions (walking, jogging, running, boxing, handwaving, handclapping) under varying conditions (indoor, outdoor, different scales). Figure 4 shows one example frame from each class.

In this experiment, we adapt *Dollár*'s method for video classification, as used in [1, 11]). The original method is based on the periodic detector. At each interest point, a block of video data (a *cuboid*) is extracted. Then, a codebook is built by applying PCA and K-means clustering. That way, a video is described by the histogram of its cuboids, quantized to the codebook entries. As multiclass classifier on top of this feature map, [1] train RBF (Radial Basis Function) SVMs and [11] use pLSA (probabilistic Latent Semantic Analysis). To test our approach we use *Dollár*'s *Matlab* code with all settings to standard (in particular $\sigma = 2$, $\tau = 3$), but with the *Periodic* detector replaced with our *Learned* detector. The periodic detector uses a threshold of 0.0002 on S_P below which all local maxima are rejected. For our detector, a natural choice for this threshold is zero, since S_L can be interpreted as the log odds of a fixation where $S_L = 0$ corresponds to a fixation probability above .5.

As in [11], we compute a leave-one-out estimate of the test error by training on the data of 24 persons, and testing on the remaining one. This is repeated 25 times. Codebooks are generated using 60,000 random samples of the training cuboids, 100 PCA components and 500 centers in K-means. Classification is done with a hard margin linear SVM. The confusion matrix and the average accuracy (the mean of the diagonal elements of the confusion matrix) are shown in Figure 4. This shows that our method outperforms previous approaches. Note that we



Algorithm	**Accuracy**
our method	82.8
Niebles et al. [11]	81.5
Dollár et al. [1]	81.2
Schüldt et al. [15]	71.7
Ke et al. [6]	63.0

Fig. 4. Action classification results. *Top left*: The KTH action classification dataset [15]. *Top right*: The confusion matrix of our classification system, which uses the Learned interest point detector. *Bottom left*: A comparison against existing algorithms.

intentionally kept most of the settings in *Dollár*'s original method in order to isolate the effect that the new interest point detector has on the performance. We therefore expect that our results improve further if we tune the entire system to suit our detector best.

4.3 Real-Time Demo and Matlab Implementation

For many applications it is vital that interest points can be computed very efficiently. Being conceptually similar to the periodic detector, the learned detector also very efficient. With five (eq. 3) instead of two (eq. 2) temporal filters, we expect the number of operations to be about 2.5 times higher. A demo application which shows the learned saliency function S_L superimposed onto a webcam feed in real-time (as in Figure 2, top right) can be downloaded at `http://www.kyb.mpg.de/~kienzle`. The *Matlab* code for detecting interest points, which plugs into *Dollár*'s feature extraction framework [1], is provided at the same location.

5 Discussion

We have presented a new spatiotemporal interest point detector based on a very simple neural network which predicts where a human observer would look in a given video. The detector was trained on real eye movement data and we showed that it predicts the location of human eye movements on independent test clips

with state-of-the-art accuracy. We also tested our approach in a computer vision environment. We found that the learned detector, plugged into a simple classification framework, outperforms previous action classification methods on a large real-world dataset. This indicates that a biologically inspired measure of interestingness can be indeed beneficial for computer vision applications. This is a nontrivial result, since existing detectors were specifically designed for computer vision problems, whereas our detector was designed to mimic human eye movements. A possible drawback of our present approach is that the detector is spatiotemporally separable, which makes it blind to time-varying spatial patterns, such as the direction of motion. We are currently working on an improved version which takes this into account.

References

1. P. Dollar, V. Rabaud, G. Cottrell, and S. J. Belongie. Behavior recognition via sparse spatio-temporal features. In *International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.
2. J. M. Findlay and I. D. Gilchrist. *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press, 2003.
3. S. Frantz, K. Rohr, and H. S. Stiehl. On the Localization of 3D Anatomical Point Landmarks in Medical Imagery Using Multi-Step Differential Approachesb In *Proc. DAGM*, pages 340–347, 1997
4. C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.
5. L. Itti, Koch C., and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20(11):1254–1259, 1998.
6. Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. ICCV*, pages 166–173, 2005.
7. W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. O. Franz. Learning an interest operator from eye human movements. In *IEEE CVPR Workshop*, page 24, 2006.
8. W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. In *Proc. NIPS 19*, 2007 (in press).
9. I. Laptev. On space-time interest points. *IJCV*, 64:107–123, 2005.
10. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
11. J. C. Niebles, H. Wang, H. Wang, and L. Fei Fei. Unsupervised learning of human action categories using spatial-temporal words. In *Proc. BMVC*, 2006.
12. P. Reinagel and A. M. Zador. Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, 10(4):341–350, 1999.
13. U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *IEEE Proc. CVPR*, pages 37–44, 2004.
14. C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *IJCV*, 37(2):151–172, 2000.
15. Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proc. ICPR*, pages 32–36, 2004.
16. The Netlab Toolbox. available at http://www.ncrg.aston.ac.uk/netlab/.
17. B. A. Wandell. *Foundations of Vision*. Sinauer Associates, Inc., 1995.
18. A. Yarbus. Eye movements and vision. *Plenum Press*, 1967.