A Learned Saliency Predictor for Dynamic Natural Scenes

Eleonora Vig[1], Michael Dorr[1,2], Thomas Martinetz[1], and Erhardt Barth[1]

[1] Institute for Neuro- and Bioinformatics, University of Lübeck,
Ratzeburger Allee 160, 23538 Lübeck, Germany
{vig,dorr,martinetz,barth}@inb.uni-luebeck.de
[2] Schepens Eye Research Institute, Dept. of Ophthalmology, Harvard Medical School,
20 Staniford Street, Boston, MA 02114, USA
michael.dorr@schepens.harvard.edu

Abstract. We investigate the extent to which eye movements in natural dynamic scenes can be predicted with a simple model of bottom-up saliency, which *learns* on different visual representations to discriminate between salient and less salient movie regions. Our image representations, the geometrical invariants of the structure tensor, are computed on multiple scales of an anisotropic spatio-temporal multiresolution pyramid. Eye movement data is used to label video locations as salient. For each location, low-dimensional features are extracted on the multiscale representations and used to train a classifier. The quality of the predictor is tested on a large test set of eye movement data and compared with the performance of two state-of-the-art saliency models on this data set. The proposed model demonstrates significant improvement – mean ROC score of 0.665 – over the selected baseline models with ROC scores of 0.625 and 0.635.

Keywords: eye movements, visual saliency, low-level image properties, natural dynamic scenes, structure tensor.

1 Introduction

The active nature of seeing has been of great interest to both the biological and computer vision community. In human vision, the environment is sampled with 2-3 gaze shifts per second, by foveating relevant, so called "salient" items in the scene. Insights into the processes that guide eye movements towards relevant targets are of importance also for their utility in various image processing applications such as image and video compression [5,16] and active vision [1]. Such visual processes are believed to be influenced by two types of factors: on the one hand, eye movements are driven involuntarily, by stimulus "salience"; on the other hand, they are also influenced by the task at hand and contextual knowledge [17]. Here, we are dealing with the former, stimulus-driven, type of visual attention.

K. Diamantaras, W. Duch, L.S. Iliadis (Eds.): ICANN 2010, Part III, LNCS 6354, pp. 52–61, 2010.

Over the recent years, a number of studies have investigated the relationship between eye movements and low-level image features at fixations[1], e.g., [12,13]. These so-called bottom-up models of attention center on the concept of a "saliency map", which topographically encodes stimulus conspicuity [11]. Inspired by the Feature Integration Theory of Treisman and Gelade [14], in the first stage of visual processing separate low-level features such as edges, contrast, or color are extracted on multiple scales. Next, normalized center-surround difference maps are computed for individual features and combined together to obtain the global saliency map. From this map, the next location to be fixated is chosen by a selection process, which combines winner-take-all and inhibition-of-return strategies.

Most existing bottom-up saliency models are biologically inspired and they differ in their underlying computational principles – mainly from information theory – they use to formally define the concept of saliency: information maximization [18,3], Bayesian surprise [6], self-information of visual features [19], efficient coding [15], and optimal decision making under constraints of computational parsimony [4]. Since these biological models tend to be complex, a large number of parameters need to be tuned manually.

Learning techniques are often employed as a practical solution to the parameter tuning problem. Still, only very recently, approaches to derive saliency-based interest operators from human fixation data have been pursued. In [8], the authors learned optimal parameters for an attention model based on low-, mid- and high-level features calculated by existing saliency methods. Kienzle et al., on the other hand, employed machine learning algorithms to learn directly on the pixel intensities of static scenes [10] and Hollywood videos [9], with the goal to find the structural differences between fixated and non-fixated locations. Although the model structure was inferred from the data and not set manually, predictability was constrained by the reduced ability of learning algorithms to operate in high-dimensional spaces given a limited number of training samples. To account for the noise in both the eye tracking and the biological system, studies usually need to consider a spatio-temporal neighborhood around the fixation. Therefore, feature space dimensionality, in which the learning is conducted, is determined by the number of pixels of the neighborhood around the fixation, which for a reasonably sized neighborhood (e.g., 64 by 64 pixels, 2.5 deg) grows rapidly (more than 4000 dimensions). Therefore, the algorithms in [10,9] were limited to a single spatial scale.

Most work on gaze allocation has dealt with static stimuli, and only recently has the number of studies dealing with videos increased. Incorporating temporal information is not always straightforward, and in a learning context the task of eye movement prediction is further complicated by the increased number of dimensions.

[1] In the process of seeing, our eyes alternate between fixations, when they are aimed at a fixed point, and rapid reorienting movements called saccades.

1.1 Overview of the Proposed Approach

In this paper, we combine machine learning techniques with structure tensor-based multi-scale image features[2] to predict eye movements on natural dynamic scenes. Figure 1 gives a graphical overview of the approach. We use a large set of eye movements to label video patches as either attended or non-attended. Our goal is to learn the structural differences between these two classes of space-time patches, and apply the learned model to predict the saliency (or class membership) of novel, unlabeled video regions. To do so, we compute low-level feature maps of the videos, derived from the tensor invariants, on several spatio-temporal scales of an anisotropic image pyramid. Next, instead of learning on the high-dimensional feature patches (e.g., 64 by 64 pixels on a single scale), we use a simple method to reduce dimensionality. In the neighborhood of each fixation, we extract the local feature energy: the root-mean-square of the pixels in the spatio-temporal feature patch. We compute such an average energy (a single scalar) on each scale of the feature pyramids. Thus, for each fixation, we obtain a low-dimensional representation consisting of the energy values on the different scales. Together with the class labels (attended or non-attended), the feature energy vectors form the training data of a classifier, which learns a mapping between feature energy vectors and their class membership (i.e., salient or not). Finally, we use ROC analysis to test the model's predicting power on unlabeled test data.

In [15], an earlier version of this model was used to predict gaze behavior of new observers on videos that have already been "seen" (i.e., learned on) by the classifier. In other words, the complete movie set was used both for training and testing, but eye movement data of any particular viewer were only present in either the training or the test set. In this paper, the model is put to the test in terms of how well it can predict eye movements on new, unseen videos. Furthermore, we extend this model by extracting the image features (geometrical invariants) on an anisotropic instead of an isotropic pyramid, which incorporates more information at the cost of moderate increase in the number of dimensions. Also, improvements are suggested on how the negative examples, i.e., locations that were not fixated, are labeled. Finally, the approach is placed in the context of existing methods: the model's performance is compared to the predictive power of two state-of-the-art saliency models of dynamic scenes [7,19].

2 Methods

2.1 Geometrical Invariants

The original video representation is considered to be highly redundant. Therefore, learning on some low-level properties of a movie that are known to be predictive for saliency is beneficial, as it removes the redundancy which appears

[2] Features here denote some low-level quantifiable properties of the image/video, such as color, edginess, contrast, etc.

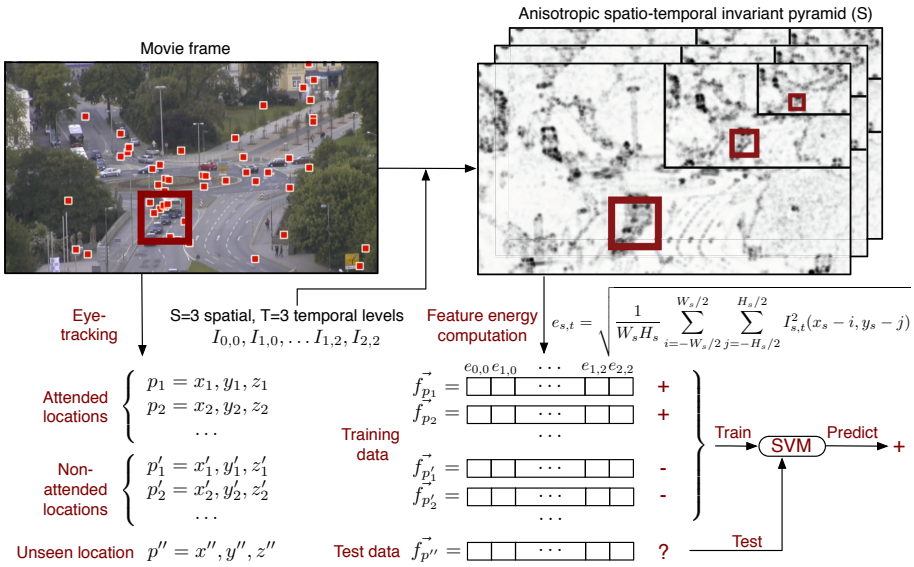

Fig. 1. Flow diagram summarizing our approach. Using eye tracking data (fixations are denoted by small filled squares in the movie frame from the left), we label movie regions as attended or non-attended. Image features – the geometrical invariants – are extracted on multiple scales of an anisotropic spatio-temporal pyramid. For a neighborhood (large unfilled square shown schematically) around each location, the average feature energy is computed on each scale of the spatio-temporal pyramid. An SVM is trained on the obtained energy vectors and is then used to predict whether an unseen location will be attended or not.

as noise to the classifier. Thus, instead of learning on the raw image intensities, we choose to use structure-tensor based image representations for our analysis. The structure tensor has been extensively used in image processing for solving a number of problems from motion and orientation estimation to occlusion detection, etc. Its geometrical invariants have been proven to be good bottom-up predictors of eye movements [15].

Considering the video as a function of space and time $f(x, y, t)$ ($f : \mathbb{R}^3 \to \mathbb{R}$), the structure tensor is defined as

$$\mathbf{J} = \int_{\Omega} \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} \mathrm{d}\Omega \,, \tag{1}$$

where the integral over Ω is a Gaussian smoothing function and f_x, f_y, and f_t stand for the first order partial derivatives.

Typical image or movie structures can be characterized based on the geometrical invariants of the structure tensor, H, S, and K, which correspond to the minimum "intrinsic dimension" (iD – local signal variation) of a region:

$$\begin{aligned} H &= 1/3 \text{ trace}(\mathbf{J}) \\ S &= |M_{11}| + |M_{22}| + |M_{33}| \\ K &= |\mathbf{J}| \end{aligned} \tag{2}$$

where M_{ij} are minors of $\mathbf{J}$. If $K \neq 0$, the intrinsic dimension is 3, i.e., K encodes space-time corners and non-constant motion in the movie. If $S \neq 0$ the iD is at least 2, i.e., S additionally encodes stationary corners, and straight edges that change in time. Finally, if $H \neq 0$ the iD is at least 1, and H also responds to stationary edges, and uniform regions that change in time.

Previously, we have found that the degree to which eye movements can be predicted increases with the intrinsic dimension: the higher the intrinsic dimension, the higher the predictive power.

2.2 Multiscale Features

The above described invariants were computed on each scale of an anisotropic spatio-temporal multiresolution pyramid. As opposed to an isotropic pyramid, where space and time are subsampled simultaneously, in case of an ansiotropic spatio-temporal pyramid each level of a spatial pyramid is decomposed further into its temporal bands. The resulting finer partition of the spectrum allows for the consideration of more information, but also comes at a computational cost. Partial derivatives of J in Eq. 1 were calculated by first smoothing the video with spatio-temporal 5-tap binomial kernels, and then applying $[-1, 0, 1]$ kernels. The smoothing kernel Ω was another 5-tap binomial. The details of the computation of invariants are the same as in [15] but for the fact that here an anisotropic pyramid is used.

2.3 Dimensionality Reduction

In the introduction, we argued that, on the one hand, models of saliency have to consider a spatio-temporal neighborhood (or "patch") around the fixation to account for location uncertainty. On the other hand, in a machine learning context, even small neighborhood sizes quickly become intractable; therefore, in practice, learning can only be performed on a single scale. To overcome this problem and allow incorporating information from multiple scales, we "compress" raw pixel information by averaging image feature energy in the neighborhood around the fixation, so that we obtain a single scalar value for the whole neighborhood. This allows us now to compute such feature energy on every scale of the above spatio-temporal pyramids, as the dimensionality is still kept low. Here, we use a spatial neighborhood only, as the uncertainty induced by measurement errors and saccade imprecision is higher in the spatial domain than in the temporal one.

For an attended (or not attended) movie location $p = (x, y, z)$ (with spatial coordinates x and y, and frame number z), we compute a feature vector $\boldsymbol{f_p} = (e_{0,0}, e_{0,1}, \cdots, e_{S-1,T-1})$ consisting of the feature energies extracted on each scale of an anisotropic pyramid with S spatial and T temporal levels. The feature

energy of a neighborhood (centered around the location p) computed on the s-th spatial and t-th temporal pyramid level is thus

$$e_{s,t} = \sqrt{\frac{1}{W_s H_s} \sum_{i=-W_s/2}^{W_s/2} \sum_{j=-H_s/2}^{H_s/2} I_{s,t}^2(x_s - i, y_s - j)} \,, \tag{3}$$

where W_s and H_s stand for the (subsampled) spatial width and height of the neighborhood on the s-th spatial scale (independent of the temporal scale). $I_{s,t}$ represents the s-th spatial and t-th temporal level of one of the invariant pyramids, H, S, and K. The spatial coordinates of the location are also subsampled on the spatial scale s: $(x_s, y_s) = (x/2^s, y/2^s)$.

2.4 Learning

Given a collection of videos together with a set of attended and not attended locations on these videos, we can now formulate the task of predicting salient locations as a binary decision problem, allowing to apply efficient classification algorithms from machine learning.

Thus, the problem of learning salient locations consists in finding, based on the locations' feature energy vectors and associated class labels (attended or not) $(\boldsymbol{f}_{\boldsymbol{p_i}}, l_i) \in \mathbb{R}^{S \times T} \times \{-1, 1\}$, a function $g : \mathbb{R}^{S \times T} \to \mathbb{R}$, that is returning for a previously unseen movie patch $\boldsymbol{p}$, based on its energy vector $\boldsymbol{f}_{\boldsymbol{p}}$, a confidence value quantifying the patch's level of saliency.

The data is partitioned "movie-wise" into a training and a test set: gaze data of all viewers on one movie are retained for testing, while the fixations on the remaining movies are used for the training. Thus, we are predicting gaze behavior on new movies that the classifier has not yet "seen".

For the classification we use a standard soft margin Support Vector Machine with Gaussian kernels. Prior to training, we linearly scale each attribute to $[-1, 1]$. Optimal model parameters are found with cross-validation on the training sequence. To measure the quality of prediction, we perform an ROC analysis using the collected human gaze data as ground truth.

3 Experimental Evaluation

The following section evaluates the predictive power of the proposed approach as compared with two standard models of bottom-up saliency for videos: that of Itti and Koch [7], and SUNDAy [19]. The saliency model of Itti & Koch is perhaps the most well-known model, against which all other approaches are compared. It is an implementation of the classical saliency map described in the introduction. SUNDAy uses a Bayesian framework to analyze fixations: novelty is defined as self-information of the visual features, but the natural statistics used to detect outliers are learned from previous examples, and are not based only on the current movie.

3.1 Videos and Eye Tracking Data

For our analysis, eye movements were acquired from 54 human subjects free-viewing 18 high-resolution (HDTV standard, 1280×720 pixels, 29.97 Hz) videos of natural outdoor scenes of about 20 s duration each. The video clips depicted populated streets and roundabouts, people in a pedestrian area, on the beach, playing in a park, and animals. They were presented on a screen covering 48 by 27 degrees of visual angle, so that the maximum displayed frequency was 13.3 cycles per degree. The recordings were conducted in Karl Gegenfurtner's lab at the Dept. of Psychology of Giessen University using an SR Research EyeLink II eye tracker running at 250 Hz. About 40,000 saccades were extracted from the raw gaze data using a dual-threshold velocity-based procedure [2].

3.2 Data Set Labeling

Whereas the class of salient locations is well defined by the set of fixations (more precisely by the locations where saccades land), the selection of non-fixated locations is not trivial. A seemingly intuitive choice would be to generate random locations from a uniform distribution either from the entire sequence or from regions that were not fixated. However, this method ignores a common phenomenon inherent in such data sets: the central fixation bias [13]. A tendency of human subjects is observed to fixate more in the central part of the display rather than in the periphery. Therefore, it has been proposed that negative examples should also be drawn from this specific distribution of human fixation locations. To assure identical distribution, a standard method used in the vision literature [13] is to consider (temporally aligned) fixations of another, randomly selected video, so that the negative training examples of movie A are chosen using the scanpaths of a randomly selected movie B. Previously, we followed this approach to generate an equal number of (about 40,000) non-attended locations. However, a drawback of this approach is that the negative set of locations, i.e., the fixations on a single random movie are clustered within the spatio-temporal volume of the movie because of the spatio-temporal coherence of eye movements on natural scenes. Thus, comparing two clustered set of movie locations leads to a slight overestimation of the prediction performance. Here, we corrected for this effect, by considering randomly selected scanpaths from randomly chosen movies, rather than taking all scanpaths of only one random video. This procedure ensures a more scattered distribution of the non-attended locations.

3.3 Experimental Results

To compute the geometrical invariants on different spatio-temporal scales, we constructed an anisotropic pyramid with $S = 3$ spatial and $T = 3$ temporal levels, as described in Section 2.2. Thus, for each attended and non-attended location (about $40,000$ per class) we obtained a 9-dimensional feature energy vector. Feature energy was computed in a neighborhood of 128 by 128 pixels on the highest scale and accordingly smaller sizes on lower scales, so that the

Table 1. ROC scores of the different models

Model	ROC score	std
Chance	0.5	–
Itti & Koch	0.625	0.003
SUNDAy	0.635	0.002
SVM with H	0.647	0.003
SVM with S	0.650	0.002
SVM with K	0.665	0.002

effective window size was about 4.8 deg on all scales. These two design parameters were found by systematically evaluating, in terms of ROC analysis, a range of different numbers of pyramid levels and window sizes.

A classifier was trained on all but one video from the movie set and testing was performed on the withheld movie. The optimal parameters of the kernel SVM (i.e., the width of the Gaussian γ and the penalty term C) were found by 8-fold cross-validation on the training sequence. Testing was repeated 18 times so that each movie served as test data once. Next, we created a single ROC curve for the complete set of movies, by varying a threshold on the decision values returned for the 18 test sets. This analysis was performed for all three invariants of the structure tensor.

For comparison, the saliency maps computed by the two state-of-the-art algorithms were treated as maximum likelihood binary classifiers for discriminating between fixated and non-fixated movie locations. To create such maps, the default model parameters were used, detailed on the web pages of the toolboxes[3]. For the Itti & Koch model, different fusion schemes of the individual feature maps into a master saliency map were tested. Best results were obtained with the Maxnorm normalization scheme, in which the fusion of the feature maps is based on normalized summation. By thresholding these maps, movie regions above the threshold were classified as salient. A systematic variation of the threshold, again on the complete movie set, resulted in a single ROC curve per model.

Based on the ROC curve, a single measure, called the ROC score (or the Area Under the Curve – AUC), provides an estimate of the prediction quality. ROC scores for the different models are shown in Table 1. Performance is shown as mean $\pm$ standard deviation over 5 realizations of the data set of negative locations (i.e., random pairing of scanpaths and movies). Note that the set of positive locations is well defined by the saccade end-points, i.e., has no random component.

A random classifier has an ROC score of 0.5, whereas an optimal predictor reaches an AUC of 1.0. As seen in Table 1, all models have an average ROC score higher than chance. Predictability using the invariants reaches an ROC score of up to 0.665 (invariant K), which is significantly higher than the AUC values obtained for the Itti & Koch (0.625) and SUNDAy (0.635) models. Also, invariant

[3] `http://ilab.usc.edu/toolkit/`
`http://mplab.ucsd.edu/~nick/NMPT/`

K outperforms the invariants H and S by a fair margin. This is, however, to be expected, as we have previously shown that prediction performance increases with the intrinsic dimension: movie regions that change in more spatio-temporal directions are more predictive than more uniform regions, because they are more informative, and therefore, draw attention.

4 Discussion and Conclusion

In this paper, we have presented a simple model of bottom-up visual saliency, which combines tensor-based multiscale image representations with machine learning to predict where people look in natural dynamic scenes. To enable the computation of visual features on multiple scales of an anisotropic multiresolution pyramid, we compressed the raw pixel information in the neighborhood around the fixation to a single value: the neighborhood's average feature energy. Certainly, the use of all available (i.e., uncompressed) information would favor the more accurate prediction, but the increased dimensionality poses a problem to existing machine learning algorithms: their performance drops when the data dimensionality is high relative to the number of training data. Thus, even though we are aware of the amount of information loss introduced by such an averaging, we argue that this significant loss is counterbalanced by the increased ability of classifiers to operate in low-dimensional spaces.

We obtained very favorable prediction results as compared with two standard methods. We argue that the efficiency of the proposed approach is due to the use of relevant image features computed on high-resolution videos, combined with state-of-the-art machine learning techniques which operate on a very large data set of eye movements. Note that the model of Itti & Koch, and SUNDAy do not profit from the advantages of a large data set. Furthermore, our method is robust with respect to varying video content.

We have, however, restricted ourselves to a single type of image feature, namely the geometrical invariants of the structure tensor. We expect that an extension of the model to incorporate a number of different features, such as color and orientation, would positively influence the predictability even further.

Finally, note that eye movements are more predictable than indicated by our ROC scores since here we have removed a number of biases, used only a very low-dimensional representation, and did not consider temporal correlations of the scanpaths.

Acknowledgments. We would like to thank Karl Gegenfurtner: data were collected in his lab at the Dept. of Psychology of Giessen University. Our research has received funding from the European Commission within the project GazeCom (contract no. IST-C-033816) of the 6th Framework Programme. All views expressed herein are those of the authors alone; the European Community is not liable for any use made of the information.

References

1. Aloimonos, Y., Weiss, I., Bandyopadhyay, A.: Active Vision. International Journal of Computer Vision 1(4), 333–356 (1988)
2. Böhme, M., Dorr, M., Krause, C., Martinetz, T., Barth, E.: Eye Movement Predictions on Natural Videos. Neurocomputing 69(16-18), 1996–2004 (2006)
3. Bruce, N., Tsotsos, J.K.: Saliency, Attention, and Visual Search: An Information Theoretic Approach. Journal of Vision 9(3), 1–24 (2009)
4. Gao, D., Vasconcelos, N.: Decision-Theoretic Saliency: Computational Principles, Biological Plausibility, and Implications for Neurophysiology and Psychophysics. Neural Computation 21(1), 239–271 (2009)
5. Geisler, W.S., Perry, J.S.: A Real-Time Foveated Multiresolution System for Low-bandwidth Video Communication. In: Human Vision and Electronic Imaging: SPIE Proceedings, pp. 294–305 (1998)
6. Itti, L., Baldi, P.: Bayesian Surprise Attracts Human Attention. Vision Research 49(10), 1295–1306 (2009)
7. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
8. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to Predict Where Humans Look. In: Proceedings of IEEE International Conference on Computer Vision (2009)
9. Kienzle, W., Schölkopf, B., Wichmann, F.A., Franz, M.O.: How to Find Interesting Locations in Video: a Spatiotemporal Interest Point Detector Learned from Human Eye Movements. In: Proceedings of the 29th Annual Symposium of the German Association for Pattern Recognition, pp. 405–414. Springer, Berlin (2007)
10. Kienzle, W., Wichmann, F.A., Schölkopf, B., Franz, M.O.: A Nonparametric Approach to Bottom-Up Visual Saliency. In: Advances in Neural Information Processing Systems, pp. 689–696. MIT Press, Cambridge (2006)
11. Koch, C., Ullman, S.: Shifts in Selective Visual Attention: towards the Underlying Neural Circuitry. Human Neurobiology 4(4), 219–227 (1985)
12. Reinagel, P., Zador, A.M.: Natural Scene Statistics at the Centre of Gaze. Network 10, 341–350 (1999)
13. Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.: Visual Correlates of Fixation Selection: Effects of Scale and Time. Vision Research 45, 643–659 (2005)
14. Treisman, A.M., Gelade, G.: A Feature-Integration Theory of Attention. Cognitive Psychology 12(1), 97–136 (1980)
15. Vig, E., Dorr, M., Barth, E.: Efficient Visual Coding and the Predictability of Eye Movements on Natural Movies. Spatial Vision 22(5), 397–408 (2009)
16. Wang, Z., Lu, L., Bovik, A.C.: Foveation Scalable Video Coding with Automatic Fixation Selection. IEEE Transactions on Image Processing 12(2), 243–254 (2003)
17. Yarbus, A.L.: Eye Movements and Vision. Plenum Press, New York (1967)
18. Zetzsche, C., Schill, K., Deubel, H., Krieger, G., Umkehrer, E., Beinlich, S.: Investigation of a Sensorimotor System for Saccadic Scene Analysis: an Integrated Approach. In: Pfeifer, R., Blumenberg, B., Meyer, J., Wilson, S. (eds.) Proc. 5th Intl. Conf. Soc. Adaptive Behavior, vol. 5, pp. 120–126. MIT Press, Cambridge (1998)
19. Zhang, L., Tong, M.H., Cottrell, G.W.: SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes. In: Proceedings of the 31st Annual Cognitive Science Conference, Amsterdam, Netherlands (2009)