Hooligan Detection: the Effects of Saliency and Expert Knowledge

D. Endres*, H. Neumann+, M. Kolesnik×, M.A. Giese*

*Theoretical Sensomotorics, Cognitive Neurology, University Clinic Tübingen,
University of Tübingen, CIN Tübingen and HIH Tübingen,
Frondsbergstr. 23, 72070 Tübingen, Germany
dominik.endres@klinikum.uni-tuebingen.de,
martin.giese@uni-tuebingen.de@uni-tuebingen.de
+Institute for Neuroinformatics, University of Ulm, 89069 Ulm, Germany,
heiko.neumann@uni-ulm.de
×Fraunhofer FhG-FIT, 53754 St. Augustin, Germany
marina.kolesnik@fit.fraunhofer.de

Keywords: crowd monitoring, optic flow, motion patterns, saliency, expert knowledge

Abstract

We investigated differences in visual search of dangerous events between security experts and naïve observers during the observation of large scenes, typically encountered on the grandstand of stadiums during soccer matches. Our main technical objective was the reduction of computational effort required for the detection and recognition of such events. To overcome the scarcity and legal issues associated with real footage, we designed a new algorithm for the synthesis of crowd scenes with well-controlled statistical properties. We characterize the relative importance of saliency and expert knowledge for the generation of correct detections and the visual search strategies for both security experts and naïve observers. We found that during the first few seconds of this search task, experts and naive observers *look* at the scenes in a similar fashion, but experts *see* more. We compare the results with theoretical models for saliency and event classification. We show that the recognition model can deliver reasonable classification/detection performance even when operating under real-time constraints. When real-time operation is not a concern, performance can be improved further by allowing the model to grow.

1 Introduction

The detection of security-relevant events (SRE) in large crowds is a difficult vision problem, both for human observers and for computerized monitoring systems. In particular, the detection and recognition of such events in crowds comprised of hundreds or thousands of people may incur a computational cost which is prohibitive for real-time applications on current commodity hardware when event recognition is carried out on every part of a visual scene. Yet, human experts can perform this task quite adeptly. To draw inspiration for system design, we therefore investigated differences in visual search of dangerous events between security experts and naïve observers during the observation of large scenes, typically encountered on the grandstand of stadiums during soccer matches.

Ideally, we would have liked to use real footage for our experiments. However, since security relevant events are both rare and unavailable for legal reasons, we designed a new algorithm for the synthesis of crowd scenes, which we call the "Tübingen hooligan simulator" (THS). Like [1, 12], we built a statistical generator for dynamic crowd scenes containing > 1000 people with realistic variability and well-defined statistics of the occurrence of SREs. We did **not** attempt to simulate single people, because generating realistic interactions (e.g. brawl) is too difficult. Instead, we recorded small groups of interacting people as the 'building blocks' for our generator. This algorithm is described in section 2.

Subjects were eye-tracked during the observation of the synthesized scenes. In section 3, we characterize the relative importance of saliency and expert knowledge for the generation of correct detections and the visual search strategies for both types of observers. We found that during the first few seconds of this search task, experts and naive observers look at the scenes in a similar fashion, but experts see more. This suggests that the fixation behavior of both observers types is driven by (low-level) saliency, whereas event classification performance is strongly influenced by expert training.

Since THS allows for the generation of large amounts of SRE video data, it can also be used for training and benchmarking computer monitoring approaches. We compare the results with theoretical models for saliency and event classification in section 4: an approach for saliency computation based on low-level features (section 4.1), and a non-parametric graphical Bayesian recognition model that was trained with expert knowledge derived from scenes containing SREs (section 4.3), exploiting optic flow features extracted with a neurally plausible algorithm described in section 4.2 and the Bayesian optic flow from [15].

We show that the recognition model can deliver reasonable classification/detection performance even when operating un-

Normal	Security-relevant
waving arms	brawl (x3)
waving flags	fast dispersal
hopping	moving up/down over seats
swaying	pushing others (or or many)
angry gesturing	walking along filled seat rows
sitting	vandalism against chairs
standing	throwing objects
orderly exiting/entering	lighting/passing Bengal torches

Table 1. Events which typically occur during a soccer matches in the audience. *Left column*: normal events, i.e. not security relevant. *Right column*: security relevant. The 3 types of brawl are: 1.) all people fight, 2.) two groups converge to fight and 3.) two people fight within a peaceful group.

der real-time constraints. When real-time performance is not a concern, performance can be improved further by allowing the model to grow.

2 The Tübingen Hooligan Simulator

A data set for the investigation into the detection and recognition of behavioral patterns should be comprised of a collection of videos pertinent to the application domain. For a soccer stadium application, one would therefore like to use recordings from real soccer matches. However, this approach is problematic for several reasons: firstly, most soccer matches are (fortunately) relatively peaceful affairs. Thus, security-relevant scenes are hard to obtain. Secondly, even if they were obtainable, legal constraints prohibit their storage beyond a short time interval, making them unsuitable as benchmark data. Thirdly, as a consequence of the scarcity of security-relevant scenes, a data set compiled from stadium recordings would almost certainly not contain all events which are considered security-relevant in the stadium context by security professionals. To deal with these problems, we created a benchmark data set by re-enacting relevant events. We contacted officers of the Düsseldorf police (Polizeiinspektion Nord) in charge of stadium security at the Esprit arena to obtain the expert knowledge necessary to decide which events to include. Past experience had shown that the communication between police officers and scientists can be fraught with difficulties, stemming mostly from differences in experience with SREs. We decided to overcome these difficulties by implementing a prototyping approach. We began with a telephone interview, during which we compiled a list of normal (i.e. not security relevant) and SREs. The full list of SREs is shown in table 1.

Subsequently, we re-enacted these events in a lecture theater with a group of $\approx$ 10 lay actors. We repeated each event multiple times in different parts of the lecture theater, the resulting videos were overlaid to create the impression of a larger crowd. Two frames from the videos can be seen in fig. 1. We showed these videos to the police officers, asking them for feedback with regard to:

- realism,
- completeness of both normal events and SREs,
- and the correctness of the labels (normal vs. security-relevant).

Virtually all videos were deemed sufficiently realistic by the police officers. However, they pointed out some missing events, such as smoke bombs or the burning of flags, for which we have yet to devise a viable re-enactment strategy. Moreover, some events which we initially considered security-relevant are part of the 'normal' set and vice versa. This feedback, which would not have been obtainable without our reenaction of the events, highlights the importance of refining such data sets through prototyping.

To construct a realistic and sufficiently difficult detection task, we built visual scenes by embedding the SREs in neighborhoods of normal events. This was accomplished by randomly filling a 7x7 grid of event patches with normal events. Spatial contiguity between patches was promoted by populating the grid with events drawn from a Markov random field with nearest-neighbor interactions [5], see fig. 2. We performed Gibbs sampling with ≈ 50000 burn-in iterations to obtain samples from the equilibrium distribution. For the generation of stadium-like blocks of behavior, we used attractive potentials that were stronger vertically than horizontally. Isolated behaviors (e.g. flag-waving) were connected to their neighbors through repulsive potentials.

In half of the scenes thus generated, we placed a security relevant event somewhere on the grid. Figure 3 shows an example frame from the resulting scenes.

Figure 1. Two frames from the Tübingen hooligan simulator. *Left*: waving crowd, a normal behavior. *Right*: brawl, a security-relevant event.

3 Human Psychophysics

For a quantitative investigation into the search strategy employed by human observers, we conducted eye-tracking experiments with one of the officers and 13 naïve subjects. We used a Tobii X120 mobile eye-tracking system. The stimulus material consisted of 10 blocks of 22 synthesized movie clips, each of which had a duration of 4.7s. Half of the clips contained a SRE, we randomized the event order in each block. Subjects were acquainted with the stimuli during an initial training phase. Subsequently, they were asked to report whether they saw a relevant event, and if so, *which* and *where*. Furthermore, subjects were instructed to keep their fixation on the relevant event until the end of the clip.

We computed several gaze statistics from the eye-tracking data and evaluated whether there was a significant difference

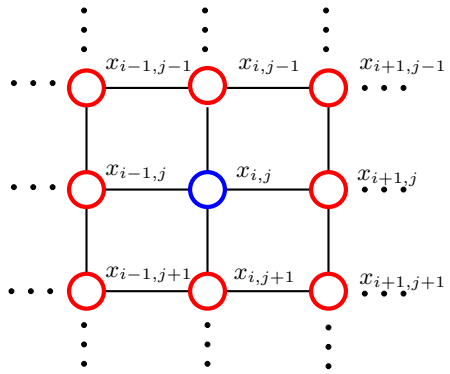


Figure 2. We synthesized large crowd scenes by drawing events (see table 1) from a Markov random field with regular grid topology [5]. Each node represents a discrete random variable $x_{i,j} \in \{1, \ldots, M\}$ specifying the patch type (scene) at position (i, j). Spatial contiguity was promoted through nearest-neighbor interaction potentials $\Psi(x_{i,j}, x_{k,l})$.

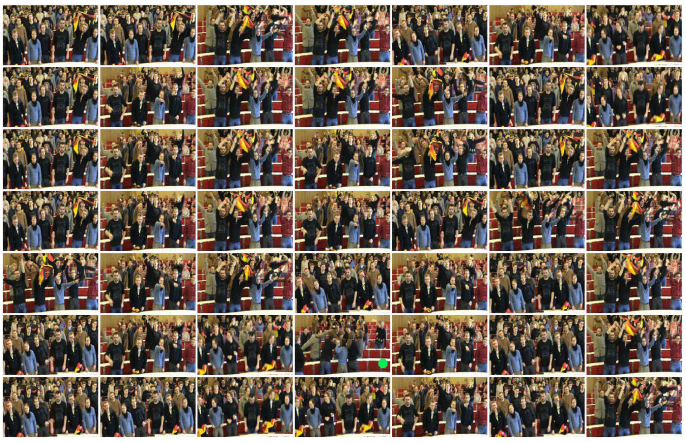


Figure 3. A typical frame generated by the simulator. The green circle (not visible in the experiment) indicates the position of the security-relevant event, here: brawl.

($p = 0.05$, two-tailed t-test) between experts and naïve observers. The results are summarized in table 2. For the majority of events, we did not find a significant difference between the two observer types, i.e. experts and naïve observers use very similar fixation strategy. We hypothesize that both are driven by **saliency** [8].

The results of the behavioral task (verbal event reporting) are shown in table 3. The expert's classification and detection accuracies are clearly higher than those of the naïve observers. In other words, experts and naïve observers **look** in a **similar** fashion, but experts **see more**. Furthermore, if the expert erred, it was always a miss and never a false positive, as indicated by his high precision. This type of bias was explained by the expert by the cost of intervention: deploying a task-force to deal with a SREs requires a lot of effort. On the other hand, if e.g. a brawl dies down after a few seconds, no intervention is necessary at all. Hence, misses are more tolerable than false positives.

gaze statistic	fraction of significant differences
number of fixated events	5/11
fixation time	2/11
time to fixation	1/11

Table 2. Fraction of significant differences between human experts and naïve observers. '5/11' means: we found a significant difference between expert and naïve observers for 5 out of 11 events (see table 1). Significance was determined with a 2-tailed t-test, $p = 0.05$. For the majority of events, we did not find a significant difference between the two observer types.

human recognition performance

	expert	naïve
classification	0.90	0.75
detection	0.90	0.79
recall	0.79	0.66
precision	1.0	0.89

Table 3. Human recognition performance measure. *Classification* rate: fraction of correctly named events. *Detection* rate: fraction of correct distinctions between 'security-relevant' and 'normal'. *Recall*: probability that an event is detected, given that it is relevant. *Precision*: probability that an event is relevant, given that it is detected.

4 Bio-inspired crowd monitoring

Our above observations about the roles of saliency and expert knowledge during the visual search for SREs show that it might not be necessary to perform (computationally expensive) recognition on the whole video stream. Rather, it appears to be sufficient to run recognition only on those parts of the stream which are highly salient. While this approach runs the risk of missing SREs if they are not salient enough, both the expert's behavior as well as his verbal feedback indicate that a certain amount of misses is tolerable. We therefore compared a machine vision saliency approach to human observers, to test whether it provides the required level of (fast) filtering of the video stream. The results are described in section 4.1. In section 4.3, we detail experiments with a simple, but real-time capable recognition system which processes the salient regions of the video stream. The input to this recognition system is (bio-inspired) optic flow, which we describe in section 4.2.

4.1 Comparison of human and machine vision saliency

Of the numerous machine vision saliency approaches available, we used *Attention based on Information Maximization* (AIM) [8]. AIM was reported to yield human-like results in visual search tasks on static natural images [8]. In a nutshell, AIM learns an ICA decomposition of the images, and collects local statistics of the ICA coefficients. It then computes a saliency measure for each point in an image by evaluating the self-information [9] of the ICA coefficients at this point under the distribution of the surround.

Since we do not have direct access to the perceived

'saliency' of the human observers, we compare humans and AIM via the *saliency rank* of the SREs. For humans, we defined the saliency rank of an SRE as the number of patches which were fixated for at least 100 ms until the target SRE was fixated for the first time. The AIM saliency rank is obtained by ordering the patches by their average saliency.

saliency ranks

	median	25%	75%
expert	6	3	10
naïve	4.8	2.9	6.2
AIM	6	2	10

Table 4. Comparison of human and AIM [8] saliency averaged across all SREs. For humans, we defined the saliency rank as the number of fixations $\geq$ 100 ms until the target event was fixated. For AIM, we computed the saliency value at every pixel, averaged across patches and ordered the events according to those averages. Saliency ranks of both AIM and human observers were comparable.

Table 4 shows the results. Median saliency ranks and 1st and 3rd quartiles have comparable values between humans and AIM. Moreover, recall that our video clips were comprised of 49 event patches. A median saliency rank of 6 implies that on average, saliency-based filtering of the scene provides a reduction of computation effort by a factor > 8.

We therefore find that an AIM spatio-temporal saliency approach is well suited as a pre-processing step for detecting SREs in crowd scenes.

4.2 Optic flow

We aim at comparing the behavioral detection results with automatic processing results derived from simulating a model of visual perception. We pursue a biologically inspired modeling approach to achieve human-like performance in visual processing. The model utilized here makes use of spatio-temporal flow patterns as primary features for event detection and differentiation. The architecture consists of several hierarchically organized stages of processing, each of which is an abstract representation of a cortical area with cells and representations of different selectivities. The mechanism is an algorithmic variant of a neurodynamical model previously described in detail in [3, 7], which has recently been extended by considering form-motion interaction [4]. The present algorithmic variants of the model architecture considers the initial stages of motion detection and subsequent integration. To achieve the necessary efficiency the individual stages of processing along the neural hierarchy were realized by utilizing discrete image processing mechanisms. This gained a significant increase in processing speed and improvement of robustness concerning the application to real-world sequences as well as the achievement of manageable storage requirements.

Initial candidate motions are detected by a discrete matching mechanism based on an extended class of rank-order approaches using the Census transform [2]). Here, numerical directional derivatives of the luminance function are calculated at each spatial location where each difference value is mapped into one of three classes depending on the sign of the slope function, i.e. whether it is positive, negative, or approximately zero (given a tolerance value). The binarized values for each directional derivative constitute a vector to represent the local structure of the input. Correspondences of image patches between two frames of an image sequence are established at locations using the Hamming metric as distance measure. Candidates with zero Hamming distance (same Census values) determine an initial motion correspondence, or hypothesis, which includes a weight which indicates the likelihood (confidence) of a particular velocity at a given position.

The initial motion detection is run for two successive frames in a backward reference fashion, namely for frames t_0 and $t_{-\Delta t}$. The matching is efficiently calculated by evaluating the Hamming distance for image shifts ranging from 1 pixel up to the diagonal image size D. Consequently, the detectable speeds for two-frame matches are given by $\|u\| = \Delta x/\Delta t$. In order to detect movements with sub-pixel speed, the matching is calculated between frames with reduced temporal sampling rate, i.e. $n \cdot \Delta t$, with $n = \{2, 3\}$. The utilization of additional frames with larger temporal distances re-scales the low-amplitude velocities to detectable speeds, leading to increased direction and speed resolution and, thus, smoother flow field representations.

The processing of image flow is organized in a modular fashion involving a three-stage processing cascade. In the first stage the motion likelihoods derived from a processing stage (e.g., initial motion detection in model area V1) are fed forward to the subsequent stage. In our model architecture this next level corresponds to area MT where initial estimates from model area V1 are integrated. The computation of a likelihood representation of motion estimation in model area MT operates on a coarser spatial scale (V1:MT ratio is 1:5) by integrating feed-forward activities from the previous stage. In other words, the bottom-up integration of likelihoods can be considered as a stage of input filtering over space and velocity. In accordance with the visual cortical architecture MT activities generate a modulation signal that is fed back to the previous stage in order to enhance the likelihood of predicted hypotheses. The modulation that defines the second stage of the cascade works in accordance with the linking hypothesis [11] to multiplicatively enhance the likelihoods of matching motion hypotheses. Top-down enhancement can only be effective where bottom-up input activation exists. If feedback activity is absent the bottom-up feed-forward likelihood is retained. Thus, the driving bottom-up activation and modulating top-down activities have an asymmetric role as feedback cannot generate new hypotheses on their own. In order to enhance as well as suppress hypothesis another stage is required that reduces and contrast enhances local distributions of activations. This is achieved by the third stage of the cascade that realizes a divisive inhibition of likelihoods for hypotheses between a pool of motion selective units in a given neighborhood around a target position. Such a process of mutual inhibition keeps the likelihoods within bounds. In other words, divisive inhibition tends to nor-

malize the overall activation of cells. In conjunction with the stage of modulatory enhancement the normalization allows the likelihoods to be increased or suppressed. If a likelihood has been increased by matching feedback it subsequently receives a competitive advantage during the divisive inhibition. In case it has not received substantial modulatory enhancement the unit contributes less to the pool of activations and is thus inhibited in the competition. This functionality implements basic elements of the biased competition theory as proposed for attention selection by [13].

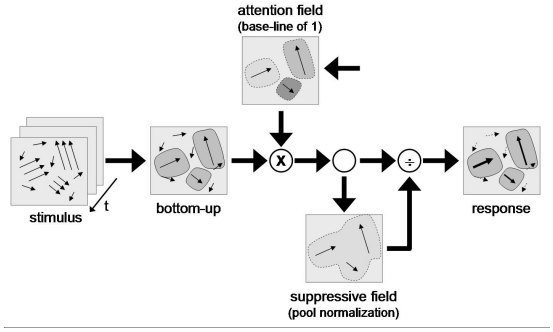


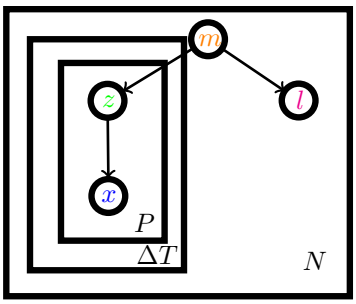
Figure 4. Generic stages of the motion computation architecture (for details, see text).

Overall, the computational stages of proposed scheme is directly related to the recent proposal of the normalization model of attention selection [14]. Our model mechanism demonstrates that mechanisms of biased competition generalize to the earlier stages of sensory processing as well (see Fig.4). Also the generic mechanism can be interpreted in a Bayesian framework. The filtered input generates the bottom-up driving input activation which is subsequently multiplied by the top-down attention field. This field is driven by a spatially homogeneous tonic input which is overlaid by top-down signals generated by higher levels of processing. The attention field can be considered as a spatio-temporal prior of the motion field that is continuously updated through the course of recurrent motion computation. After the modulation the likelihoods are spatially pooled and subsequently divided by the summed activity. This defines the suppressive field to calculate the pool normalization such that the activities are rescaled to the range between 0 and 1. Thus, the likelihoods after enhancement and normalization can be considered as probabilities for the confidence of the presence of a motion hypothesis.

4.3 Recognition results

To carry out the recognition within a salient region of the video stream, we experimented with a spatio-temporal bag-of-features model. Such models are popular in computer vision for offering a good compromise between reasonable recognition performance and computational simplicity, consider e.g. histograms of oriented gradients for pedestrian detection [10]. A graphical model representation of our approach is shown in figure 5. Since we are interested in motion patterns, the input x to our model are small patches of optic flow, computed either with the bio-inspired approach described in section 4.2 or the Bayesian optic flow from [15]. Both give comparable results.

A salient region of a scene is decomposed into P such patches whose optic flow x within a time interval ΔT is modeled as a space-time bag m of features z. Each bag is associated with a label l. A video stream is modeled as a sequence of N such labeled feature bags. We experimented with 1st order features, here: Gaussian clusters with diagonal covariance matrices, and 2nd order features, here: Gaussian clusters having full covariance matrices.



recognition perf.

classification	0.73
detection	0.91
recall	0.92
precision	0.95

Figure 5. *Left:* a graphical model representation of the crowd motion pattern recognition approach. A salient region of a scene is decomposed into P patches whose optic flow x within a time interval ΔT is modeled as a space-time bag m of features z. Each bag is associated with a label l. A video stream is modeled as a sequence of N such labeled feature bags. *Right*: the model's recognition performance measure for a model with 60 full-covariance Gaussian features. While the classification performance is significantly lower the human expert's (see table 3), the detection rate is comparable.

Since we did not know the number of features z nor the number of feature-bags m in advance, we equipped both with a (truncated) Dirichlet process prior, whose parameters were learned via variational Bayesian expectation maximization [6]. We used half of our data for training ($\approx$ 7500 frames), the other half for validation. Rates are conditional on event occurence statistics after saliency based pre-filtering. The results for a model with at most 60 full covariance Gaussian features per bag is shown in fig 5, right. While the model's classification performance can not match a human expert, its detection performance can. Its recall is higher than that of the expert, but at the expense of some false positives.

We also investigated the scaling of run-time and classification/detection performance as a function of the upper bound on the number of features per bag. Furthermore, we experimented with the effect of using 1st (diagonal covariance) and 2nd (full covariance) order features. As fig. 6 shows, 2nd order features are clearly superior. Real-time operation, i.e. processing more than 10 frames per second, can be achieved with ≤ 30 features per bag[1], when some recognition performance is sacrificed.

[1] C++ implementation running on 4 cores

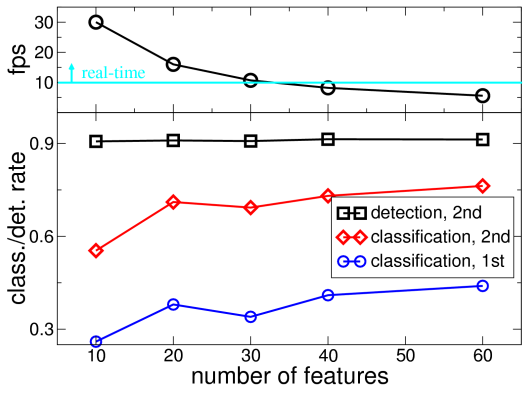


Figure 6. *Top:* scaling of runtime with number of features z (see fig. 5). Real-time operation can be achieved with ≤ 30 features. *Bottom*: scaling of classification and detection rates with the number and type of features. 1st-order features: Gaussian clusters with diagonal covariance matrices. 2nd-order features: Gaussian clusters with full covariance matrices, i.e. marginalized PCA decompositions of the optic flow patches.

5 Conclusions

We have demonstrated that the computational effort required for the detection and recognition of SREs can be effectively reduced by focusing only on those regions of the video stream which are highly salient. This approach is also taken by human (expert) observers. It will invariably generate some misses, but those are generally found tolerable. The costly mistakes are usually false positives, since they may trigger intervention measures. Future work should therefore focus on reducing the false positives to near zero, while keeping the miss rate in the acceptable range. In addition, we have begun to synthesize SRE scenes in a real soccer stadium to further increase the realism and relevance of our experiments.

Acknowledgments

This work was supported by the EU project FP7-ICT-215866 SEARISE, see http://www.searise.eu. We thank F. Vintila, S. Cavdaroglu, H. Alhumsi and L. Tuchscherer for help with the psychophysical data collection. We are particularly indebted to two officers of the Düsseldorf police, PHK G. Mainda and PHK H.J. Berg, for providing us with their expert knowledge about security-relevant events during soccer matches, and for their patience during the experiments. H.N. further acknowledges support from the Transregional Collaborative Research Center SFB/TRR62 *Companion Technology for Cognitive Technical Systems* funded by the German Research Foundation (DFG).

References

[1] E. Andrade, S. Blunsden, and R. Fisher. Simulation of crowd problems for computer vision. In *First International Workshop on Crowd Simulation (V-CROWDS '05), Lausanne*, pages 71–80, 2005.

[2] P. Bayerl and H. Neumann. A fast biologically inspired algorithm for recurrent motion estimation. *IEEE Transactions on Pattern Analyis and Machine Intelligence*, 29(2):246–260, 2007.

[3] Pierre Bayerl and Heiko Neumann. Disambiguating visual motion through contextual feedback modulation. *Neural Computation*, 16:2041–2066, 2004.

[4] Cornelia Beck and Heiko Neumann. Interactions of motion and form in visual cortex: A neural model. *Journal of Physiology - Paris*, 104:61–70, 2010.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[6] David M. Blei and Michael I. Jordan. Variational methods for the dirichlet process. In *In Proceedings of the 21st International Conference on Machine Learning*, 2004.

[7] Jan D. Bouecke, Emilien Tlapale, Pierre Kornprobst, and Heiko Neumann. Neural mechanisms of motion detection, integration, and segregation: From biology to artificial image processing systems. *EURASIP Journal on Advances in Signal Processing*, 2011, 2011. Article ID 781561.

[8] N. Bruce and J. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24, 2009.

[9] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[10] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005.

[11] R. Eckhorn, H.J. Reitboeck, M. Arndt, and P.W. Dicke. Feature-linking via synchronization among distributed assemblies: Simulation of results from cat cortex. *Neural Computation*, 2:293–307, 1990.

[12] K. Lee, M. Choi, Q. Hong, and J. Lee. Group behavior from video: a data-driven approach to crowd simulation. In *SCA '07: Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 109–118, Aire-la-Ville, Switzerland, Switzerland, 2007. Eurographics Association.

[13] J.H. Reynolds and L. Chelazzi. Attention modulation of visual processing. *Annual Reviews in Neuroscience*, 27:611–647, 2004.

[14] J.H. Reynolds and D.J. Heeger. The normalization model of attention. *Neuron*, 61:168–185, 2009.

[15] Eero P. Simoncelli. Bayesian multi-scale differential optical flow. In B. Jähne, H. Haussecker, and P. Geissler, editors, *Handbook of Computer Vision and Applications*, volume 2, chapter 14, pages 397–422. Academic Press, 1999.