BIOLOGICALLY INSPIRED ATTENTIVE MOTION ANALYSIS FOR VIDEO SURVEILLANCE

Florian Raudies and Heiko Neumann

Institute of Neural Information Processing, University of Ulm, 89069 Ulm, Germany

Abstract: Recently proposed algorithms in the field of vision-based video surveillance are build upon directionally consistent flow (Wixson and Hansen, 1999; Tian and Hampapur, 2005), or statistics of foreground and background (Ren et al., 2003; Zhang et al., 2007). Here, we present a novel approach which utilizes an attention mechanism to focus processing on (highly) suspicious image regions. The attention signal is generated through temporal integration of localized image features from monocular image sequences. This approach incorporates biologically inspired mechanisms, for feature extraction and spatio-temporal grouping. We compare our approach with an existing method for the task of video surveillance (Tian and Hampapur, 2005) with a receiver operator characteristic (ROC) analysis. In conclusion our model is shown to yield results which are comparable with existing approaches.

1 INTRODUCTION

Video surveillance is a recent field of research addressing the tasks of detection, localization, recognition, and tracking of specific objects. Existing approaches are based upon the following assumptions to detect attentional regions in spatio-temporal image sequences: (i) directionally consistent image flow were used to separate coherent object movement from spatio-temporal fluctuations of scene events (Wixson and Hansen, 1999; Tian and Hampapur, 2005), or (ii) temporally non-deformable image features were matched between subsequent frames (Zhou and Aggarwal, 2001). In contrast, our model utilizes moving features which could be further updated over time due to an increasing gain of evidence for the presence of the spatio-temporal structure or event.

Several mechanisms in our model are motivated by neurophysiological evidence. At first, simple features are extracted, according to the early visual processing in area V1 (Hubel and Wiesel, 1968). These features are temporally differentiated to extract on- and offset of temporal changes in the image structure (Marr and Ullman, 1981). The result of feature extraction and differentiation are then integrated and temporally smoothed which in turn leads to a lower temporal signal resolution. This result is referred to as *streak image*, representing traces from temporal

changing features. For the extraction of activity for a specified orientation of those traces, we employ Gabor filters (Daugman, 1988). Grouping of these activities is realized by long-range interaction filters utilizing a biologically inspired scheme (Neumann and Sepp, 1999). Motion information is processed along a pathway parallel to the form features. This division of segregated form and motion processing is reminiscent of the ventral and the dorsal stream in cortical visual processing in primates.

The main contributions of our model architecture are: (i) the formulation of a biologically inspired model for the task of video surveillance, (ii) the construction of a motion *streak image*, integrating only salient temporally changing features, (iii) a general grouping mechanism, and (iv) feature binding at various stages within the model. To the best of our knowledge, this is the first model which incorporates motion streaks as an attentional signal and furthermore constructs salient streak representations, which contain no temporally static features. Additionally, the combination of three attentive signals is new in the field of video surveillance. Processing in two parallel pathways is motivated by the representation of changes on two temporal scales, whereas motion streaks integrate over several time steps, and image flow contains mostly information of two temporally subsequent snapshots. These two temporal scales combine

long-time motions with actually motions in an appropriate way to identify attentional regions, which is one task of surveillance.
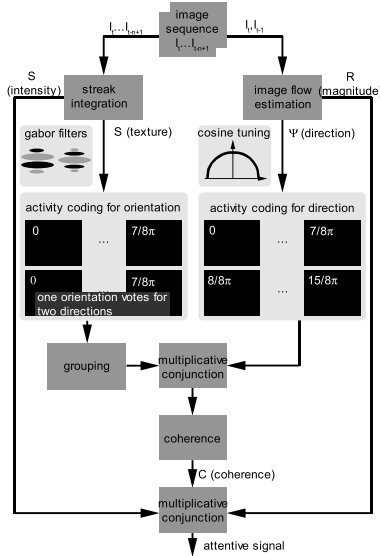

Figure 1: Schematic overview of the model. Details about the functionality of all parts are reported in Sec. 3.

2 OVERVIEW OF THE MODEL

This model computes an attention signal from an image sequence $I_t, ..., I_{t-n+1}$ of n frames. The *streak image* is the result of a temporal integration of all frames while the image flow is estimated on the basis of the two most recent frames. These two mechanisms form four signal channels: (i) the intensity of the motion streaks, (ii) the specific texture of the motion *streak image*, (iii) the direction of image flow, and (iv) the magnitude of image flow. At the very beginning an activity coding is constructed, for orientations in the texture channel and for directions of the image flow. Subsequently, a grouping mechanism is employed to further process the activity code of the texture channel. This mechanism enforces aligned structures getting the same orientation. An incorporation of both activity codes is realized by a multiplication. The result contains information about the direction of feature displacements, which is then transformed into a coherence signal. Finally, again a multiplication of this coherence signal together with the streak intensity signal and motion magnitude signal results in the final attention signal. Fig. 1 gives an schematic overview about the processing pathways of the signal channels.

3 SPECIFIC MECHANISMS

Motion Detection and Integration. For the estimation of image flow we used the algorithm of Lucas & Kanade (Lucas and Kanade, 1981), with a window size of 21. Spatial derivatives are calculated using the discrete kernel $[-1, 8, 0, -8, 1]/12$ according to (Barron et al., 1994). Further processing uses a polar representation of the image flow with magnitude R and direction Ψ. For the interaction with the streak texture signal a population code for the direction Ψ of motion is constructed by a rectified cosine-tuning

$$A_\Psi^{motion, dir} = \max(\cos(\Psi - \psi), 0), \quad \psi = 0, ..., 15/8\pi \quad (1)$$

for each sampling direction ψ. Here the sampling is equidistant and is defined for sixteen directions, an example is shown in Fig. 2.

Temporal Change Detection and Motion Streaks. The computation of salient motion streaks starts with the extraction of corner features, resulting in a representation of more distinctive and localized image properties. These features are detected with the Förstner corner detector (Förstner, 1986). In a preprocessing stage each image is smoothed with a Gaussian kernel (σ_{pre}=1.0). The structure tensor is computed by calculating the vector product of the intensity gradient that is averaged over a local neighborhood (smoothing with a Gaussian kernel σ_{tensor}=0.5). Let λ_1 and λ_2 denote the eigenvalues of this tensor, then a continuous valued corner is characterized by

$$F = \lambda_1 \lambda_2 / (\lambda_1 + \lambda_2), \quad \bar{F} = F * G_{\sigma_t}^{gauss}. \quad (2)$$

These corner features are temporally smoothed with a Gaussian ($\sigma_t = 2.0$), which suppresses temporal noise in the input sequence. A logarithmic transformation is applied to enhance features with low response amplitudes, resulting in $\bar{F}^{space}$. For an integration of moving features a temporal derivative

$$\bar{F}^{time}(x, y; t) \approx \bar{F}^{space}(x, y; t) * [1, -1](t) \quad (3)$$

of each image is additionally calculated. Together with the previous temporal smoothing this mechanism eliminates pure static image features. The conjunction of feature representations $\bar{F}^{space}$ and $\bar{F}^{time}$ given by $\bar{F}^{comb} = \bar{F}^{space} \cdot \bar{F}^{time}$ results in a high spatial resolution and a sufficient temporal resolution of features. Temporal integration of the combined signals is realized by the weighted sum

$$S = \frac{\alpha}{1 - (1 - \alpha)^n} \sum_{i=0}^{n-1} (1 - \alpha)^i \bar{F}^{comb}(x, y; t - i), \quad (4)$$

where S denotes the motion *streak image*. The parameter α avoids an infinite temporal integration

(α=0.05). In addition to the intensity image the *streak image* contains a specific texture generated by moving corner features. An analysis of this texture is realized by oriented derivatives of the *streak image S*

$$A_\phi^{streak,ori} = S * G_{\phi,\sigma,\lambda}^{gabor}, \quad \phi \in [0,\pi) \qquad (5)$$

computed with a Gabor filter approach (DC level free). The Gabor filters are parameterized by eight orientations ϕ, with $\Delta\phi = \pi/8$, one scale σ, and wavelength λ (σ=4, λ=6 px).

Grouping Mechanism. Grouping is performed for the motion streak activities $A_\phi^{streak,ori}$ serving as input activity A_ϕ^{in}, and the result of grouping A_ϕ^{out} is further referred to as $\hat{A}_\phi^{streak,ori}$. The grouping method employed is a simplified version of a mechanism that has been proposed to account for the neural mechanisms of surface boundary formation and the extraction of invariant surface features (Weidenbacher et al., 2006; Neumann and Sepp, 1999). The method consists of four steps forming an iterative loop (for notational simplification we omit parameters denoting 2-D spatial location):

$$x_\phi^{ori} = A_\phi^{in}(1 + cA_\phi^{out}) \qquad (6)$$

$$y_\phi^{ori} = x_\phi^{ori}/(\mu + \sum_\Upsilon x_\Upsilon^{ori}) \qquad (7)$$

$$x_\phi^{group} = y_\phi^{ori} * G_{\phi,\sigma_1,\sigma_2}^{bipol} \qquad (8)$$

$$A_\phi^{out} = x_\phi^{group}/(\mu + \sum_\Upsilon x_\Upsilon^{group}). \qquad (9)$$

In Eq. 6 the initial input signal A_ϕ^{in} is nonlinearly enhanced by the feedback term $(1 + cA_\phi^{out})$, with the feedback constant c (c=100). This modulatory coupling of feedback guarantees stability, therefore only existing feature activities at given location and orientation are enhanced by feedback, since the term $1 + cA_\phi^{out}$ is gated by the driving signal activation A_ϕ^{in}. However, feedback alone cannot generate any feature activities. Substages at the levels of orientation and grouping computation (Eq. 7 and Eq. 9) consist of a normalization step to keep activities within bounds. The parameter μ affects the influence of the total activation in the normalization process ($\mu = 10^{-2}$). Eq. 8 realizes the propagation (grouping) of orientational responses along their corresponding orientation axis ϕ. In this equation, $G_{\phi,\sigma_1,\sigma_2}^{bipol}$ denotes a bipolar filter consisting of two Gaussian kernels which are spatially offset along the orientation axis at the target location. Each kernel has an elongation σ_1 for the major axis, σ_2 for the minor axis, and orientation ϕ (σ_1=3.5, σ_2=1.5). These kernels are combined to form a grouping filter along the orientation ϕ. In Eq. 9 a normalization similar to Eq. 7 is applied. To quantify the

improvement caused by the iterative grouping mechanism we report the orientation error for a scene with existing ground-truth motion, shown in Fig. 4 (B).

Combination of Motion Responses with Streaks and Attentional Signal. A general motivation for the need of an attentional signal in many applications of computer vision is given in (Rothenstein and Tsotsos, 2007), suggesting that an attentional signal reduces the complexity for visual search tasks. Here, the activity of motion direction is multiplicatively combined with the grouped activity of motion streaks, each serving as an attentional signal. For an combination activities corresponding to orientations are replicated $\hat{A}_\psi^{streak,dir} = \{\hat{A}_\phi^{streak,ori}, \hat{A}_{\phi+\pi}^{streak,ori}\}$, by this means that one orientation votes for the two corresponding directions. The final multiplication $A_\psi^{comb,dir} = A_\psi^{motion,dir} \cdot \hat{A}_\psi^{streak,dir}$ are motivated for two reasons: (i) activities corresponding to a specific direction from streaks and motion are independent sources, (ii) streak activity acts as a bias for motion activity or vice versa.

Coherence and Attentional Signals. The combined activity is used for the computation of a coherence signal, based on large patches with similar movements. Therefore, a maximum likelihood estimate is calculated from the ensemble (population) of activities (Deneve et al., 1999), given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \sum_\psi A_\psi^{comb,dir} \begin{pmatrix} \cos\psi \\ \sin\psi \end{pmatrix} / \sum_\psi A_\psi^{comb,dir}, \quad (10)$$

with the direction $\Psi^{comb,dir} = \arctan 2(y,x) \in (-\pi,\pi]$, where *arctan*2 denotes the four quadrant inverse tangent. Coherent motion

$$C = \left\| \begin{pmatrix} \cos(\Psi^{comb,dir}) * G_{\sigma_{coh}} \\ \sin(\Psi^{comb,dir}) * G_{\sigma_{coh}} \end{pmatrix} \right\|_2 \qquad (11)$$

is detected by large integrating Gaussian filter kernels (σ_{coh}=11.25). As in Fig. 1 all three attentive signals, namely the flow magnitude R, the streak intensity S, and the coherence of directions C, are normalized and multiplied to provide the final attentional signal. This multiplication is a simple mechanism to avoid additional complexity. The mechanism could be extended by incorporating weighing factors for the three attentional signals.

4 PROCESSING RESULTS AND EVALUATION

Results are presented for motion detection and for motion streak integration. Subsequently, we present

an evaluation of our model for the task of video surveillance.

Motion Detection. Here, results for the Hamburg Taxi Sequence are shown (http://i21www.ira.uka.de, 03/2007). In this scene the background flickers, and the main attentive motion signals which should be detected are the movement of the cars and the pedestrian. The activity code of the flow field is shown in Fig. 2. Activity representing the motion of the vehicles are present in several maps, due to the broad tuning. For example, the car approaching from the left side induces activity in four neighboring maps, according to the directions $\psi = 0, \pi/8, 14\pi/8, 15\pi/8$ (compare with Fig. 2).
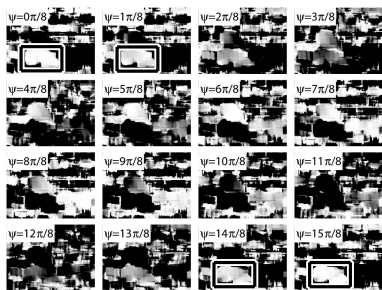

Figure 2: Motion activity for the Hamburg Taxi Sequence. Activities are normalized and linearly coded from black (zero activity) to white (unit activity).

Motion Streak Integration. Fig. 3 (A) shows the last frame of the Hamburg Taxi sequence and Fig. 3 (B) shows the *streak image*. Additionally, lines are superimposed on this *streak image*, where the length reflects the streak intensity and the orientation results from the texture analysis. Therefore, the activities $\hat{A}_\phi^{streak,ori}$ corresponding to orientations ϕ are interpreted as described in Eq. 10, where directions ψ are substituted by orientations ϕ. The resulting orientation Φ defines a vector $(\cos(\Phi), \sin(\Phi))$ which is weighted with the streak intensity S, forming the lines.
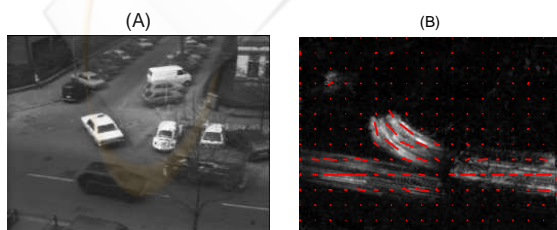

Figure 3: Motion *streak image*: (A) Last frame of the Hamburg Taxi sequence. (B) Temporally integration of features produces the *streak image*, where the extracted and grouped orientation of the streaks is superimposed by lines (41 frames, 10 iterations, sampled 15 times).
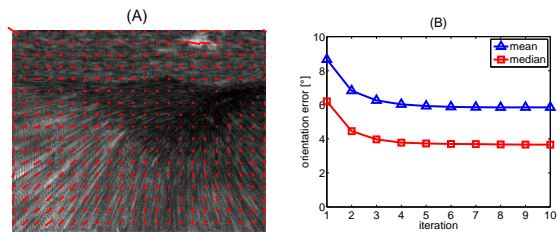

Figure 4: Motion streaks for Yosemite sequence with clouds: (A) Streak lines with superimposed extracted orientations from texture and grouping (15 frames used for integration, 10 iterations, sampled 15 times). (B) Decreasing error over iterations of grouping.

Figure 5: Masked last frame of sequences: (A) Durlacher Tor. (B) Indoor. Attentional regions are plotted with full intensity and the background with an intensity of at most thirty percent.

In general the construction of a motion *streak image* and the analysis of the streak texture for orientations is not restricted to object motion. Thus, we present results for an ego-motion sequence, a simulated flight through the valley of the Yosemite park. Temporal integration results in the motion *streak image* in Fig. 4 (A). From this characteristic streak texture the orientation is determined and former refined by grouping. The final result (lines) is superimposed to the streak texture. To show the effect of iterative grouping the mean and median orientation error are reported in Fig. 4 (B). This orientation error is defined as the deviation between the ground-truth orientation and the estimated orientation. After only few iterations (approx. five) of recurrent grouping this error saturates at the level of 3.9 deg median and 6.3 deg mean.

Evaluation for Video Surveillance. For video surveillance an evaluation based on four different sequences has been conducted (results are shown only for two of them). In this evaluation regions with attentive motion should be correctly detected which refer to suspicious activity. Accordingly, those regions are masked in the last frame of the processed sequence, shown in Fig. 5. The first scene is a characteristic traffic scenario, the second a typical indoor sequence (from (Brown et al., 2005)).

For the evaluation of robustness receiver operator characteristics (ROC) with decisions at pixel level are investigated. Within this analysis the false positive

ratio is defined as FPR $= \| N_{attentive}^{det} \setminus N_{attentive}^{gt} \| / \| N_{inattentive}^{gt} \|$ where $N_{attentive}^{det}$ is the set of pixels detected as salient, $N_{attentive}^{gt}$ is the set of salient pixels according to the masked regions, and $N_{inattentive}^{gt}$ for inattentive pixels. The numerator contains the set difference between $N_{attentive}^{det}$ and $N_{attentive}^{gt}$. A FPR of zero indicates that no inattentive pixel is detected as salient. The true positive ratio is defined as TPR $= \| N_{attentive}^{det} \| / \| N_{attentive}^{gt} \|$. In our model the attentional signal is finally thresholded, as the only decision parameter that needs to be adjusted in the algorithm. In the algorithm of (Tian and Hampapur, 2005) the parameter T_d that is applied to the accumulated temporal difference image is critical and is therefore selected as reference parameter in the ROC analysis. Additionally, the parameter W_{accum} for the weighted summation of the accumulated temporal difference seems critical. For this reason we applied the analysis for two representative values $W_{accum} = 0.5$ (as stated in (Tian and Hampapur, 2005)) and $W_{accum} = 0.125$. In Fig. 6 results for the analysis are shown. Beside this, results for the attentive signals from streak saliency $\tilde{S}$, motion saliency $\tilde{R}$, saliency from coherence $\tilde{C}$, and their multiplicative conjunction are visualized, where the symbol 'tilde' imposes that the signals are scaled to the interval $[0, 1]$. For the traffic scene (Fig. 6 (A)) our model performs better than that of (Tian and Hampapur, 2005) and in the second sequence our model performs better for specific ratios of FPR/TPR. Another measure is the area under the ROC curve, where the our model has the best performance value for the traffic sequence (0.982) and for the indoor sequence (0.970). For the method of (Tian and Hampapur, 2005) the calculation of the area under the ROC curve is not possible due to the two variable thresholds W_{accum} and T_d.

5 DISCUSSION

The model proposed in this contribution follows the main idea of (Tsotsos et al., 2005) for the decomposition of an image sequence into single features (here the channels intensity, texture, flow direction, and magnitude) and their re-combination into an attention signal. In contrast to the approach of (Tsotsos et al., 2005), which suggests a feed-forward processing pyramid, our model contains feedback signals in the grouping stage.

For the construction of motion streaks a related idea was reported in (Majchrzak et al., 2000). Unlike to our approach the authors simply add the images in a specific time window. Then they extract edges in the resulting motion *streak image*. On the basis of
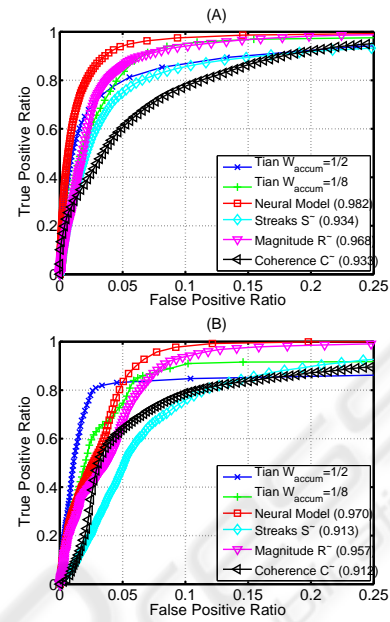

Figure 6: ROC analysis: (A) Durlacher Tor (frames 11-21). (B) Indoor, (frames 45-55, parameter σ_{pre}=2). Our model constructs the attentive signals streaks $\tilde{S}$, magnitude $\tilde{R}$, coherence $\tilde{C}$, and a combination denoted by the neural model. The x-axis ranges from 0 to 0.25. Values in brackets denote the area under the ROC curve.

the orientation of these edges they define a decision mechanism to differentiate between a rotational field, sidewards translation, or a translation near the line of sight (FOE/FOC). Compared to our method, the feature extraction and subsequent weighted integration results in more salient streaks. For this specific streak texture we then extracted and grouped orientations. A heading estimation on the basis of these orientations as proposed by (Majchrzak et al., 2000) is also possible.

Our model behaves robust compared to assumptions in video surveillance: First, the assumption of directionally-consistent movements or coherent temporal motion (Wixson and Hansen, 1999; Tian and Hampapur, 2005), which is here included within the temporal integration for the construction of the *streak image*. If this assumption is not fulfilled the streak intensity and flow magnitude supports an attention signal. Second, non-stationary backgrounds (like wiggling trees, waves, fountains, rain, snow) are assumed. Generally, those backgrounds are outlined through spatial and temporal statistics (Ren et al., 2003; Zhang et al., 2007). Within this predictive mechanism only the most important variations are captured through a sub-space analysis of the input sequence. Our method outlines distracting background movements by temporal smoothing and integration.

6 CONCLUSIONS

The proposed model combines information from motion streaks and image flow, forming the signals magnitude of image flow, intensity of motion streaks, and coherence of motion directions. These three signals are combined for the final attention signal. An ROC analysis for scenarios of video surveillance is conducted and shows that three attentive signals and the final signal are appropriate for the division between attentive motion and background noise. Compared with the method of (Tian and Hampapur, 2005) our model has only one critical threshold and shows better results for two analyzed scenes. Main challenges solved by our model are the robust processing and analysis of noisy background and locally incoherent motions of walking persons.

Within our model motion streaks are used, which provide information about orientation and speed for object and self-motion. In addition to the approach of (Majchrzak et al., 2000) our model provides a dense field of orientations and speeds. Therefore, motion streaks could serve as a robust prior or bias for the estimation of image flow and also for the estimation of ego-motion. Future work will pursue incorporations of motion streaks into those estimation tasks.

ACKNOWLEDGEMENTS

REFERENCES

Barron, J., Fleet, D., and Beauchemin, S. (1994). Performance of optical flow techniques. *IJCV*, pages 43–77.

Brown, L., Senior, A., Tian, Y.-L., Connell, J., Hampapur, A., Shu, C.-F., Merkl, H., and Lu, M. (2005). Performance evaluation of surveillance systems under varying conditions. *IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance*.

Daugman, J. (1988). Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *Trans. Acoustics, Speech, and Signal Proc.*, 26(7):1169–1179.

Deneve, S., Latham, P., and Pouget, A. (1999). Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2:740–745.

Förstner, W. (1986). A feature based correspondence algorithm for image matching. *ISP Comm. III, Rovaniemi 1986, International Archives of Photogrammetry*, pages 26–3/3.

Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, (195):215–243.

Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Proc. DARPA Image Understanding Workshop*, pages 121–130.

Majchrzak, D., Sarkar, S., Sheppard, B., and Murphy, R. (2000). Motion detection from temporally integrated images. In *Proc. IEEE 15th ICPR*, pages 836–839.

Marr, D. and Ullman, S. (1981). Direction selectivity and its use in early visual processing. *Proc. Royal Soc. of London*, B, 211:151–180.

Neumann, H. and Sepp, W. (1999). Recurrent V1-V2 interaction in early visual boundary processing. *Biol. Cybernetics*, 81:425–444.

Ren, Y., Chua, C.-S., and Ho, Y.-K. (2003). Motion detection with nonstationary background. *Machine Vision and Applications*, 13:332–343.

Rothenstein, A. and Tsotsos, J. (2007). Attention links sensing to recognition. *Image and Vision Computing*. (in press).

Tian, Y.-L. and Hampapur, A. (2005). Robust salient motion detection with complex background for real-time video surveillance. *Proc. IEEE Workshop on Motion and Video Computing*, pages 30–35.

Tsotsos, J., Liu, Y., Martinze-Trujiloo, J., Pomplun, M., Simine, E., and Zhou, K. (2005). Attending to visual motion. *Computer Vision and Image Understanding*, 100:3–40.

Weidenbacher, U., Bayerl, P., Neumann, H., and Flemming, R. (2006). Sketching shiny surfaces: 3D shape extraction and depicting of specular surfaces. *ACM Trans. on Applied Perception*, 3:262–285.

Wixson, L. and Hansen, M. (1999). Detecting salient motion by accumulating directionally-consistent flow. In *Proc. of the Seventh IEEE ICCV*, pages 797–804.

Zhang, W., Fang, X., Yang, X., and Wu, Q. (2007). Spatiotemporal gaussian mixture model to detect moving objects in dynamic scenes. *J. of Electronic Imaging*, 16.

Zhou, Q. and Aggarwal, J. (2001). Tracking and classifying moving objects from video. In *IEEE Int. Workshop on PETS*.