OOP: Object-Oriented-Priority for Motion Saliency Maps

Anna Belardinelli, Werner X. Schneider and Jochen J. Steil

Abstract The ability to attend to motion is paramount for living beings. The human visual system is able to detect coherent motion and select within multiple moving objects the most conspicuous or most relevant to the task at hand. Similarly, any artificial agent operating in dynamic environments needs to be endowed with a mechanism for rapid detection and prioritization of moving stimuli in its field of view. In this paper, we present a biologically and psychologically inspired model of this ability and tune it for the extraction of motion at different scales and velocities. Unlike many computational models that compute saliency pixelwise, we extract moving proto-objects through segmentation of motion energy features. These perceptual units, so called proto-objects, are identified as consistently moving blobs. A proto-object based priority map is hence obtained by assigning a single saliency value to the region confining a segmented object. Priority stems from a combination of bottom-up saliency, evaluated in a center-surround fashion, and from top-down biasing of motion features or motion saliency. Experimental simulations on synthetic displays and real sequences show the effectiveness of the proposed approach.

Anna Belardinelli
Cognitive Interaction Technology Center of Excellence (CITEC), Bielefeld University, Universitätsstrasse 25, Bielefeld, Germany e-mail: anna.belardinelli@cit-ec.uni-bielefeld.de

Werner X. Schneider
Neuro-cognitive Psychology and Cognitive Interaction Technology Center of Excellence (CITEC), Bielefeld University, e-mail: wxs@uni-bielefeld.de

Jochen J. Steil
Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, e-mail: jsteil@cor-lab.uni-bielefeld.de

1 Introduction

Both research on visual attention and modelling thereof in the past decades has concentrated mostly on selective processing of static stimuli, characterized through a wealth of features. Yet, we live in a highly dynamic world, populated with moving things, which call for a selective mechanism much more urgently than static objects do. Early detection and selection of the most salient kind of motion can sometimes make the difference in the struggle for survival. Even simple insects do have some form of motion perception [8], but usually quite limited color vision. In a very cluttered scene, moving objects are supposed to attract our gaze very effectively, as shown by [4], where motion contrast accounts for most of the fixations. On a neurophysiological level, motion information is indeed processed even along a different, more direct pathway, i.e. the dorsal pathway, as opposed to other features needed for object recognition in the ventral pathway [13]. It can be argued that attending to static objects is the prerequisite of perception for action (like searching for a cup and grasp for it), whereas attending to motion elicits perception for prompt reaction and interaction, being tied to events evolving in time and triggering our response (such as an approaching danger or person). Embedding motion in a visual attention model therefore targets gaze orienting in real-world environments instead of modeling picture viewing, as recently suggested by [28].

Our model builds upon a novel approach for extracting and prioritizing moving objects in a scene. In previous work [2], we introduced a basic framework for producing motion saliency maps from spatiotemporal filtering. That model proved to be effective in selecting relevant motion but was not broadly tuned in the frequency domain and produced a pixel-based saliency map. A purely pixel-based approach to computational modeling of attention has inherent limitations and differs fundamentally from the way the visual brain of humans and other animals processes information. Indeed, there is growing evidence in psychology for an object-based account of attention [26, 22]. Features and properties are namely not perceived per se, but as belonging to distinct *object files*. Object correspondence is then maintained through spatio-temporal continuity [17]. In his Theory of Visual Attention (TVA) [3], Bundesen defines mathematically how our visual system could assess top-down relevance of each object in the stimulus. The first implementation of object based attention according to TVA and to static features has been presented in [30]. Proto-objects (or perceptual files, which consist of selected regions) are the basic units of attention, upon which a priority value is computed. Objects are then fed into a Winner-Take-All (WTA) network, providing access to working memory for those winning the race. Such proto-objects can be defined in a flexible way. In some related work, object-based attention was modelled by extracting object candidates upon color [27] or edge features [20], or spreading of activation around a salient location [29]. On the other hand, as a matter of fact motion is a quite distinctive property which naturally induces segmentation of the scene within foreground and background (see [18] for an application to background subtraction), hence provides a more straightforward extraction of object units.

As usual when designing an attention architecture, in the case of prioritize motion the problem is to identify and to sort salient regions, that is, not just detecting moving objects but defining which one requires to be first attended. Priority is not intrinsic in the location nor in the object but it is defined relatively to its surround, in a contrast based way (bottom-up), and according to relevance to the task (top-down). We use here the term priority to comply with the proposal of [9] of composing bottom-up salience and top-down relevance in a single priority map. In this paper, we combine all these ideas to extend our previous model to account for multiscale motion, proto-object extraction and object priority evaluation. Saliency is given by means of center-surround computations both on a location-based and an object-based level. Relevance is given by tuning the model according to the given task. Proto-objects (in the following termed objects) are defined as blobs of consistent motion energy and coherent direction. Objects standing out from the surrounding in terms of amount of energy or direction are hence given larger saliency. In the next sections, we describe the components of our system and present some results. Section 2 explains our implementation of the energy model [1] for motion perception, Section 3 proposes the definition and characterization of moving proto-objects and how to compute their saliency. Finally, Section 4 shows some experimental simulations and results.

2 An energy framework for motion feature extraction and prioritization

In the human visual system, motion is perceived through a series of parallel local motion detectors. Two neighbour detectors are delayed one with respect to the other and their outputs are combined to obtain direction selection. The *Reichardt detector* [23] models this mechanism with two components implementing spatial asymmetry (two units apart) and temporal asymmetry (one unit output is delayed) and a cross-correlation stage. This is substantially equivalent to spatiotemporal filtering [1] . We extend the implementation of Adelson's and Bergen's energy model [1] for coherent motion perception that we introduced in [2]. The basic idea is that coherent motion can be selected inside an intensity frame buffer by filtering the oriented edges and bars, left by objects moving in the spatiotemporal volume. Instead of just one couple of Gabor filters in quadrature for extracting right/left-ward and up/down-ward motion from $x-t$ and $y-t$ planes respectively, we designed a Gabor filter bank to extract motion information at different spatio-temporal scales (frequency bands) and velocities (filter orientations), trying to sample most of the spatiotemporal frequency domain included in the window $u, v \in [0, 0.5]$, to comply with the sampling theorem. That is, we code each voxel in a Gabor space, according to its oriented energy response, analogously to the coding suggested for modeling our visual system [10]. Gabor filters have been long known to resemble orientation sensitive receptive fields present in our visual cortex and represent band-pass functions conveniently localized both in the space and in the frequency domain [6]. This is still valid in the

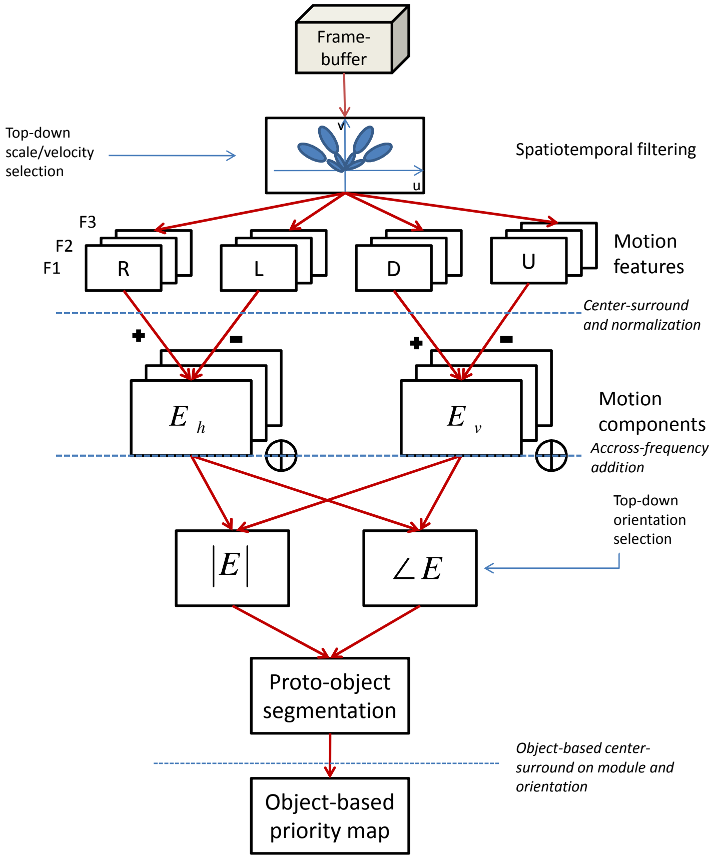

Frame-buffer
Top-down scale/velocity selection
v
u
Spatiotemporal filtering
F3
F2
F1
R
L
D
U
Motion features
Center-surround and normalization
+
−
+
−
E_h
E_v
Motion components
Accross-frequency addition
Top-down orientation selection
$|E|$
$\angle E$
Proto-object segmentation
Object-based center-surround on module and orientation
Object-based priority map

Fig. 1: The processing flow for motion extraction and saliency computation. In the beginning a frame buffer is filtered by a Gabor filter bank and direction based feature maps are obtained (R, L, U, D). Afterwards, horizontal and vertical components of motion energy are computed (E_h, E_v), and from those energy magnitude and phase ($|E|$,) are extracted. These allow the segmentation of proto-objects upon which priority is finally computed.

spatiotemporal domain, as measured by [7] in the receptive fields of simple cells in V1 and as obtained via ICA (Independent Component Analysis) computation on video sequences by [14]. In both studies, resulting receptive fields resemble 3D Gabor filters (whose central slices are 2D Gabor filters as well) at different orientations and frequencies.

The overall schema of our model is depicted in Fig. 1 and is explained in detail in the following.

Basically, given a frame buffer $\mathscr{B}$, we filter any vertical (column-temporal dimensions) or horizontal (row-temporal dimensions) plane $I(s,t)$ in $\mathscr{B}$ with every filter $G_{\theta,f}$ in the bank, in its odd (superscript o) and even (superscript e) component:

$$E_{\theta,f}(s,t) = \sqrt{(G^{o}_{\theta,f}(s,t) \star I(s,t))^2 + (G^{e}_{\theta,f} \star I(s,t))^2} \tag{1}$$

where $s = \{x,y\}$, $f = \{0.0938, 0.1875, 0.3750\}$ (the max spanned frequency is 0.5 cyc/pixel, the frequency bandwidth is 1 octave), $\theta = \{\pi/6, \pi/3, 2/3\pi, 5/6\pi\}$. That is, we designed a filter bank with 4 orientations ($\theta = 0, \pi/2$ were left out, as corresponding to static or flickering edges) and 3 frequency bands. The energy function $E_{\theta,f}$ describes the vector length of the combined odd and even filter responses at (s,t). Maxima of the energy function have been shown to occur where interesting feature (such as edges and corners) are present in static images too [19].

From combination of opponent filter pairs (i.e filters with same slope but opposite orientation, θ and $(\pi - \theta)$) we can extract a measure of direction-selective energy at a specific velocity. For instance, in our case right-sensitive filters have $\theta_r = \{\pi/6, \pi/3\}$, while left-sensitive filters have $\theta_l = \{(\pi - \pi/6), (\pi - \pi/3)\}$. A measure of the total rightward (leftward) energy at a specific frequency can hence be obtained by summing rightward (leftward) energy accross velocities:

$$R_f = \sum_i \left| \frac{E_{\theta_{r_i},f} - E_{\theta_{l_i},f}}{E_{\theta_{r_i},f} + E_{\theta_{l_i},f}} \right|_{\geq 0} \quad L_f = \sum_i \left| \frac{E_{\theta_{r_i},f} - E_{\theta_{l_i},f}}{E_{\theta_{r_i},f} + E_{\theta_{l_i},f}} \right|_{\leq 0} \tag{2}$$

where the $|\cdot|$ operator selects points greater/less than zero, corresponding to rightward/leftward motion. The numerator gives a measure of local motion contrast, while the denominator, which represents flicker energy, serves the purpose of divisive normalization, improving direction discrimination [12]. The same can be done for upwards (downwards) energy computation, by taking $s = y, \theta_u = \theta_r$ and $\theta_d = \theta_r$. In this way we obtain 4 feature volumes R, L, U, D at different frequencies, as displayed on the second level of the overall schema of our framework, in Fig.1.

Subsequently, we operate a first attentional processing by applying normalization and center-surround operators to the frames of each feature buffer. Due to receptive field center-surround interactions, indeed, ganglion cells are usually described as firing more strongly whenever a central location is more contrasted with respect to its surroundings. Again, this holds in the motion domain as well, as shown by [24]: locations displaying different motion in terms of energy module or direction pop out from the surroundings and are enhanced in the saliency map. Center-surround inhibition is usually obtained via accross-scale differences [16] or center-neighborhood differences at the same scale [11]. We chose the second way, as faster due to the use of integral images. At the same time, feature maps need to be normalized to the same range and weighted according to the number of occurring local maxima, so that a feature map with many activation peaks is given less weight than one with

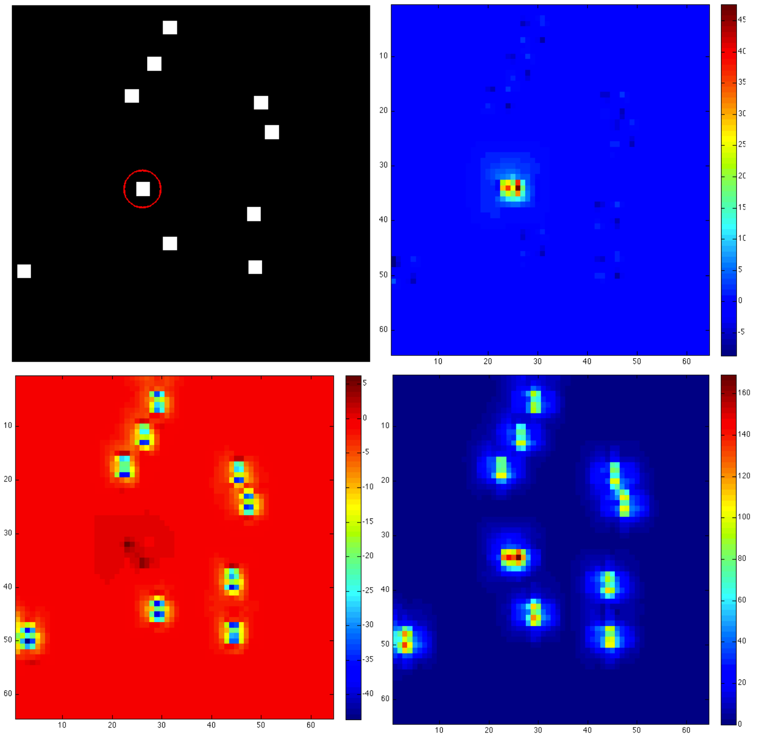


Fig. 2: Application of the salient energy extraction framework to a synthetic display (top left) containing a pop-out object, represented by the red-circled square, moving horizontally, while the other squares move vertically. Top right, the horizontal motion feature map and the vertical motion feature map (bottom left) are shown, both at f=0.3750. Bottom right, the temporal average of the module of salient energy, achieved by merging horizontal and vertical energy at different frequencies.

few peaks. This can be realized in a biological plausible manner by iteratively filtering the feature frames with a DoG (Difference of Gaussians) filter and taking each time just the non negative values [15]. We then compose horizontal and vertical features to obtain a measure of horizontal and vertical energy and sum accross frequencies:

$$E_h = \sum_f (\mathscr{N}(CS(R_f)) + \mathscr{N}(CS(L_f))), \;\; E_v = \sum_f (\mathscr{N}(CS(U_f)) + \mathscr{N}(CS(D_f))).$$

Here $\mathscr{N}(\cdot)$ and $CS(\cdot)$ denote the normalization and center-surround operator, respectively, which are applied to each $x-y$ frame of the feature buffers.

To illustrate the effectiveness of our procedure we use a purely bottom up synthetic stimulus, depicted in Fig.2. The sequence (256 x 256 x 5) displays nine squares at random positions moving downwards at 1 $pixel/frame$ velocity and just one square moving rightwards at the same velocity, representing the oddball (marked by a circle). The horizontal feature map (top, right) correctly highlights the oddball, while in the vertical feature map (bottom, left) the vertical moving squares are shown. Due to normalization these latter have less energy (see colorbar), albeit moving at the same velocity as the horizontally moving one.

E_h and E_v can be regarded as the projection on the x and y axes of the salient motion energy present in the frame buffer. Hence from these components we can achieve for every voxel magnitude and phase of the salient energy:

$$|E(x,y,t)| = \sqrt{E_h(x,y,t)^2 + E_v(x,y,t)^2}, \ \angle E(x,y,t) = \arctan(E_v(x,y,t)/E_h(x,y,t)).$$

A module-based saliency map can be seen in Fig.2, bottom right, obtained by averaging the $|E|$ frame buffer along time. Top-down modulation at this level can be implemented by selecting the filter parameters (number of orientations, number of frequency bands, orientation and frequency bandwidths) according to the current task. In this way, one can decide to attend just to a particular direction of movement, to a particular scale of objects or to a particular velocity range.

3 Proto-object construction and selection

In the previous section, we have shown how to obtain a saliency/priority map enhancing relevant motion zones. Such a map is still pixel-based but, as pointed out in the introduction, we aim at an object-based map to facilitate subsequent processing for object recognition and action selection. We need to evaluate priority of an object with respect to its entirety [3] and with respect to the surrounding background, not just by considering each single pixel it is composed of. Indeed, even if motion processing attains to the dorsal, or ”where”-, pathway, nevertheless attentional processes can modulate segregation and grouping of the visual input into ”object tokens” across both pathways [25].

To this end, we extracted proto-object patches defined as blobs of consistent motion in terms of module and direction. As the Gestalt law of common fate states, points moving with similar velocity and direction are perceptually grouped together in a single object (e.g.,[21]). A simple segmentation on the module map would not be sufficient, since adjacent objects moving in different directions would be merged. Hence, we take the last frame of the magnitude volume $|E(x,y,t)|$ (which is the only one with causal responses, depending just on the previous frames). We threshold it to discard points with too low energy and apply the mean shift algorithm to the corresponding phase of the remaining active points in the last frame of the phase volume $\angle E(x,y,t)$, weighted according to their magnitude. The mean shift algorithm is a kernel-based mode-seeking technique, broadly used for data clustering and segmentation [5]. Being non-parametric, it has the advantage that it does not need the number of clusters to be specified beforehand. We thereby cluster pixel

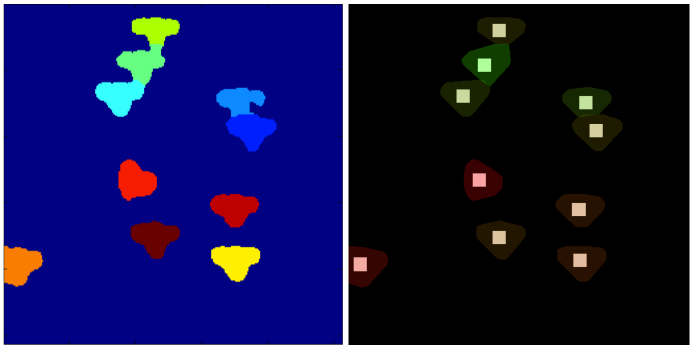


Fig. 3: Object segmentation and prioritization. Left, the result of the mean shift segmentation on directions relative to salient energy is displayed. Each cluster is denoted by a different color. Right, convex hull patches corresponding to segmented objects are superimposed to the original frame: color is determined by saliency, with the least salient object having RGB=(0, 1, 0) and the most salient being displayed in pure red with RGB=(1, 0, 0).

regions with a certain amount of energy according to their direction. Application of this procedure to the synthetic stimuli presented above gave the results presented in Fig.3 on the left. Each square is assigned to a cluster as it is an independently moving object.

Once we have labelled regions we can extract the object convex hulls by means of morphological operations and can compute their saliency. Again, we define object saliency as proportional to motion contrast in terms of module and orientation, in a center-surround fashion. Given an object $\mathbf{o}$, defined by its bounding box, and given its surround $N(\mathbf{o})$ of size proportional to the area of $\mathbf{o}$, similarly to [20], we have:

$$S_{mag}(\mathbf{o}) = \langle |E(x,y)| \rangle_{(x,y)\in(o)} - \langle |E(x,y)| \rangle_{(x,y)\in N((o))} \tag{3}$$

where the $\langle\cdot\rangle$ operator computes the mean of the points in the subscript set.

For orientation saliency, since some non rigid objects can display more than one direction but still a dominating general direction, we compute the histograms of the orientations of the object $\mathbf{o}$, weighted according to the energy module, as:

$$h(i) = \sum_{\angle E(x,y)\in i} |E(x,y)| \tag{4}$$

where i represents the i-th bin. In so doing, the more likely orientations are the ones relative to high energy points. Orientation saliency is hence given by the similarity between the orientation distributions of the object and of its surround. Similarity is evaluated through the Bhattacharyya distance:

$$S_{or}(\mathbf{o}) = 1 - \sum_i (\sqrt{h_o(i) * h_{N(o)}(i)}) \tag{5}$$

Hence, the more the orientation distribution of the object differs from that of the surrounding, the greater the orientation saliency.

Finally, the overall saliency of the object is calculated as linear combination of the two components:

$$S(\mathbf{o}) = \alpha S_{mag}(\mathbf{o}) + \beta S_{or}(\mathbf{o}) \tag{6}$$

Both S_{mag} and S_{or} are normalized to the interval $[0,1]$. α and β are taken equal to 0.5 in the case of pure bottum-up selection, but can be top-down biased for task-driven selection. In Fig.3, the segmented patches with color intensity proportional to the overall saliency are superimposed on the original frame. The oddball is correctly shown as the object with the highest saliency, the most reddish.

4 Experimental simulations and discussion

In the following, we perform experiments with real world sequences. We chose the dataset presented in [18][1], originally to the aim of background subtraction. The dataset contains 18 sequences displaying single or multiple objects moving, taken either by a fixed or moving camera. All sequences have some frames manually annotated by subjects asked to segment what was moving in the foreground. We produced a ROC curve to compare in 30 frames of each sequence our saliency

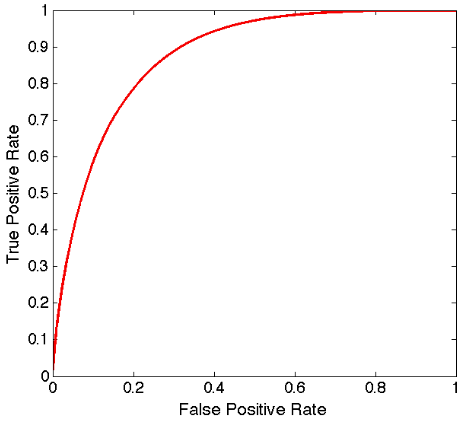


Fig. 4: ROC curve comparing our system against human performance in a foreground motion classification task. The AUC is 0.87.

[1] http://www.svcl.ucsd.edu/projects/background_subtraction/

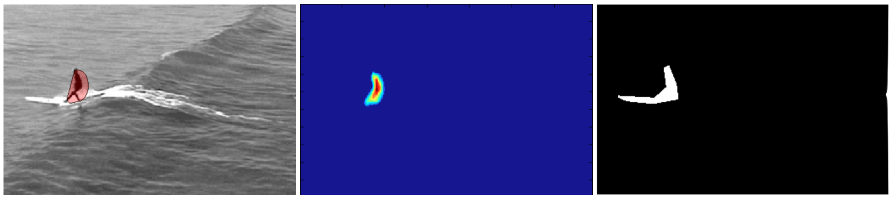


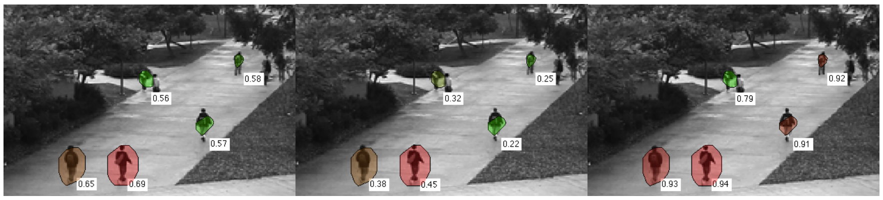
Fig. 5: One sequence of the tested dataset. On the left, original frame with segmented moving object superimposed. In the middle, the corresponding saliency map. On the right, the foreground motion mask annotated by a human subject.

maps (at different thresholds) against the binary masks segmented by human subjects. Albeit the task given to subjects was different from our system's (attending to motion) and albeit in some sequences the camera was movin, while our systems cannot discriminate between background and foreground motion, the overall classification performance was rather good, as shown in Fig.4. Our system performed best in the sequences with fixed camera, while performance droppedto some degree in sequences with moving camera (the saliency maps did not discriminate between foreground and background motion while the subjects were asked to do so). When tested singularly, the best sequence produced an Area-Under-the-Curve (AUC) of 0.98 while the worst sequence produced an AUC of 0.73. An example is presented in Fig.5, and shows a frame belonging to the ”surfers” sequence.

This kind of test tells more about the ability of the system to detect rather than to prioritize motion. In a crowded scene, however, such as a station or a road (see Fig. 6), there is a wealth of moving objects competing for attention capture and therefore a prioritization and selection mechanism is extremely useful. In the experiments depicted in Fig. 6, it can be noticed how differently moving objects can be discriminated according to their distinctiveness from other motion patterns in the surroundings. Since the final saliency is evaluated on the whole object region, it is not said that the object containing the most salient pixel is the most salient object too. In this case we show how the top-down biasing can affect priority and hence the selected object. Priority can for instance be given to energy, if a conspicuously moving object is sought, or to phase if an object standing out for its direction is sought. In Fig.6 on the left, magnitude and phase saliecy are given equal weights. In the middle, top-down biasing was achieved by assigning $\alpha = 0.9$ and $\beta = 0.1$, on the right, conversely, $\alpha = 0.1$ and $\beta = 0.9$, that is, higher priority was given to direction saliency.

5 Conclusions

The presented framework can be tuned and refined in a number of ways to make it more or less selective and task-oriented. A major limitation, at the moment, is the



(a) (b) (c)

Fig. 6: Motion prioritization on an other sequence of the chosen dataset. Proto-objects are labelled with their priority weights. (a), bottom-up priority. (b) top-down biasing, higher weight is given to magnitude. (c), top-down biasing, higher weight is given to phase.

constraint of stationary camera. This limits its current biological plausibility, since humans are able to discriminate scene motion from ego-motion when moving the head or the body. Similarly, this limit can be overcome by applying stabilization techniques to the buffer frame, or modelling the motion distribution of the background and applying background subtraction as in [18].

The main novelty of our system is the definition of moving proto-objects which is related to the their amount of motion and direction distinctiveness. We have shown how this approach can successfully select and prioritize relevant motion within a crowded scene. This is based on low-level processing and relies on the extraction of coherent motion in different directions. Further higher-level processing will have to be combined with specific task descriptions and a more elaborated description of motion patterns in terms of frequency and spatiotemporal signatures. Interesting issues still to be investigated are the temporal scale and resolution that are needed to recognize these patterns (we arbitrarily took a 5 frames temporal span for computational needs) and how far such a system can get without object continuity and indexing [22].

References

1. Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A **2**(2), 284–299 (1985)
2. Belardinelli, A., Pirri, F., Carbone, A.: Motion saliency maps from spatiotemporal filtering. Attention in Cognitive Systems pp. 112–123 (2009)
3. Bundesen, C.: A theory of visual attention. Psychological review **97**(4), 523–547 (1990)
4. Carmi, R., Itti, L.: Visual causes versus correlates of attentional selection in dynamic scenes. Vision Research **46**(26), 4333–4345 (2006)
5. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE PAMI **24**(5), 603–619 (2002)
6. Daugman, J.G.: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J. Opt. Soc. Am. A **2**(7), 1160–1169 (1985)

7. DeAngelis, G.C., Ohzawa, I., Freeman, R.D.: Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. i. general characteristics and postnatal development. Journal of Neurophysiology **69**(4), 1091–1117 (1993)
8. Egelhaaf, M., Borst, A., Reichardt, W.: Computational structure of a biological motion-detection system as revealed by local detector analysis in the fly's nervous system. J. Opt. Soc. Am. A **6**(7), 1070–1087 (1989)
9. Fecteau, J.H., Munoz, D.P.: Salience, relevance, and firing: a priority map for target selection. Trends in cognitive sciences **10**(8), 382–390 (2006)
10. Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Am A **4**(12), 2379–2394 (1987)
11. Frintrop, S., Klodt, M., Rome, E.: A real-time visual attention system using integral images. In: Proceedings of the 5th International Conference on Computer Vision Systems (2007)
12. Georgeson, M.A., Scott-Samuel, N.E.: Motion contrast: a new metric for direction discrimination. Vision Research **39**(26), 4393 – 4402 (1999)
13. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. Trends in neurosciences **15**(1), 20–25 (1992)
14. van Hateren, J.H., Ruderman, D.L.: Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. Proceedings: Biological Sciences **265**(1412), 2315–2320 (1998)
15. Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging **10**(1), 161–169 (2001)
16. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE PAMI **20**(11), 1254–1259 (1998)
17. Kahneman, D., Treisman, A., Gibbs, B.J.: The reviewing of object files: object-specific integration of information. Cognit Psychol **24**(2), 175–219 (1992)
18. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**, 171–177 (2009)
19. Morrone, M.C., Burr, D.C.: Feature detection in human vision: A phase-dependent energy model. Proceedings of the Royal Society of London. Series B. Biological Sciences **235**(1280), 221–245 (1988)
20. Orabona, F., Metta, G., Sandini, G.: A proto-object based visual attention model. In: Attention in Cognitive Systems, pp. 198–215 (2008)
21. Palmer, S.E.: Vision Science: Photons to Phenomenology, 1 edn. The MIT Press (1999)
22. Pylyshyn, Z.W.: Visual indexes, preconceptual objects, and situated vision. Cognition **80**(1-2), 127–158 (2001)
23. Reichardt, W.: Autocorrelation, a principle for evaluation of sensory information by the central nervous system. In: W. Rosenblith (ed.) Principles of Sensory Communications, pp. 303–317. John Wiley, New York (1961)
24. Rosenholtz, R.: A simple saliency model predicts a number of motion popout phenomena. Vision Research **39**(19), 3157 – 3163 (1999)
25. Schneider, W.X.: VAM: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action. Visual Cognition **2**(2-3), 331–376 (1995)
26. Scholl, B.J.: Objects and attention: the state of the art. Cognition **80**(1-2), 1–46 (2001)
27. Sun, Y., Fisher, R., Wang, F., Gomes, H.M.: A computer vision model for visual-object-based attention and eye movements. Computer Vision and Image Understanding **112**(2), 126–142 (2008)
28. Tatler, B.: Current understanding of eye guidance. Visual Cognition pp. 777–789 (2009)
29. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks **19**(9), 1395 – 1407 (2006)
30. Wischnewski, M., Steil, J.J., Kehrer, L., Schneider, W.X.: Integrating inhomogeneous processing and proto-object formation in a computational model of visual attention. In: Human Centered Robot Systems, pp. 93–102 (2009)