Attending to Motion: an object-based approach

Anna Belardinelli

Abstract Visual attention is the biological mechanism allowing to turn mere sensing into conscious perception. In this process, object-based modulation of attention provides a further layer between low-level space/feature-based region selection and full object recognition. In this context, motion is a very powerful feature, naturally attracting our gaze and yielding rapid and effective shape distinction.
Moving from a pixel-based account of attention to the definition of proto-objects as perceptual units labelled with a single saliency value, we present a framework for the selection of moving objects within cluttered scenes. Through segmentation of motion energy features, the system extracts coherently moving proto-objects defining them as consistently moving blobs and produces an object saliency map, by evaluating bottom-up distinctiveness of each object candidate with respect to its surroundings, in a center-surround fashion.

1 Introduction

Cognitive architecture for autonomous robotics often rely on the capability to deal with objects, act upon them, recognize scenes and partners and hence behave properly in a given situation. These higher level cognitive functions can take place only if the flow of the huge and multimodal information stream coming from the sensory system is firstly processed and filtered by an attention mechanism, which extracts relevant patterns and conveniently codes and prioritizes what has to be further processed.

Although most computational models of attention are location-based, there is growing evidence for an object-based account of attention [21]. In his Theory of Visual Attention [3], Bundesen defines matematically how our visual system could

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

Dagstuhl Seminar Proceedings 10081
Cognitive Robotics
http://drops.dagstuhl.de/opus/volltexte/2010/2628

assess top-down relevance of each object in the stimulus. That is, 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 WTA network, providing access to working memory for those winning the race. Such proto-objects can be defined in a flexible way and upon different features.

Research on visual attention, and modelling thereof, has concentrated in the past decades mostly on static stimuli, characterized through a wealth of features, accounting for bottom-up attentional capture and accordance with task-related requirements. Yet, we live in a highly dynamic world, populated with moving things, which call for a selective mechanism much more compellingly 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, the dorsal pathway, as opposed to other features needed for object recognition [12]. If attending to static objects is the prerequisite of perception for action (like searching for a cup and grasp for it), attending to motion fosters perception for 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 would then move into the direction suggested by [23] of considering gaze orienting in real-world environments instead of end up with a model of picture viewing.

Without disregarding the importance of the deployment of attention to static features, our model builds upon a novel approach for extracting and prioritizing moving objects in a scene. In a previous work [2], we introduced a basic framework for producing motion saliency maps from spatiotemporal filtering. That model was not broadly tuned in the frequency domain and produced a pixel-based saliency map. Motion is a quite distinctive property which naturally induces segmentation of the scene within foreground and background (see [16] for an application to background subtraction), hence provides a more straightforward extraction of object units than color [22] or edge features [17], or spreading of activation around a salient location [24].

As usual when designing an attention architecture, in the case of attending to motion the problem is to identify and prioritize salient regions, namely, not just detecting moving objects but defining which one requires to be first attended. Saliency is not intrinsic in the location nor in the object but it is defined relatively to its surround, in a contrast based way, and according to relevance to the task. In this paper, we try to bring all these ideas together and extend our model to account for multiscale motion, proto-object extraction and object saliency 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 surround-

ing with respect to 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 Motion feature extraction and prioritization

We extend the implementation of the energy model for coherent motion sensing by [1] introduced in [2]. The basic idea is that coherent motion can be selected inside an intensity frame buffer by filtering in 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 to represent band-pass functions conveniently localized both in the space and in the frequency domain [6]. This is still valid in the 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 [13]. In both studies, resulting receptive fields resemble 3D Gabor filters (whose central slices are 2D Gabor filters as well) at different orientations and frequencies.

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:

$$E_{\theta,f}(s,t) = (G^o_{\theta,f}(s,t) \star I(s,t))^2 + (G^e_{\theta,f} \star I(s,t))^2 \tag{1}$$

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.

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 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.

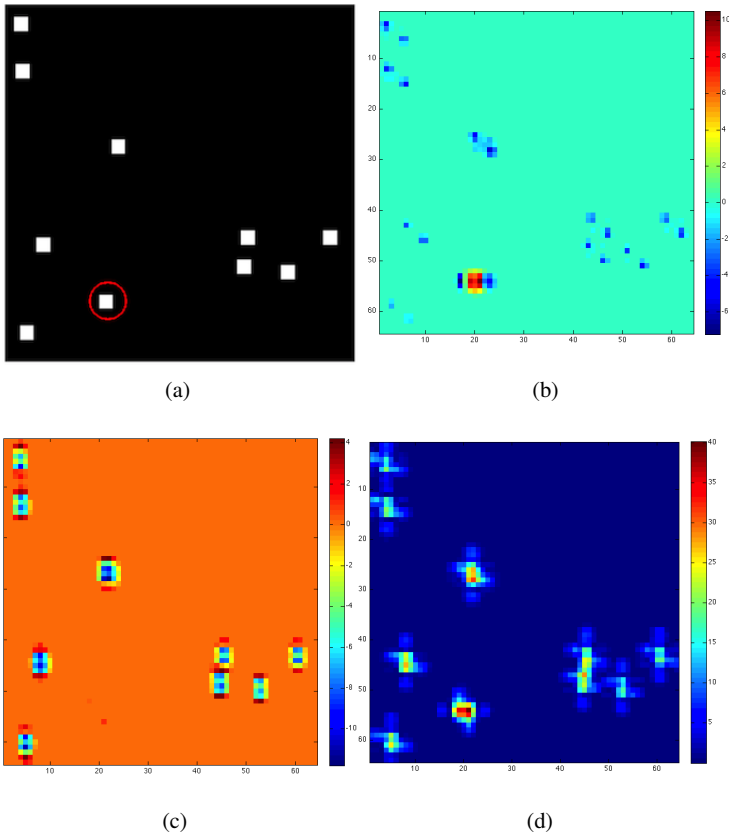
Subsequently, we operate a first attentional processing by applying normalization and center-surround operators to the frames of each feature buffer. Due to receptive field center-surround interactions, indeed, ganglion cells are usually described as firing more strongly whenever a central location is more contrasted with respect to its surroundings. Again, this holds in the motion domain as well, as shown by [19]: 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 [15] 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 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 [14]. We then compose horizontal and vertical features to obtain a measure of horizontal and vertical energy and sum accross frequencies:

$$E_h = \sum_f (\mathcal{N}(CS(R_f)) + \mathcal{N}(CS(L_f))) \quad E_v = \sum_f (\mathcal{N}(CS(U_f)) + \mathcal{N}(CS(D_f))) \tag{3}$$

Here $\mathcal{N}(\cdot)$ 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.1a. 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 red circle). In the horizontal feature map (Fig. 1b) correctly just the oddball is shown, while in the vertical feature map (Fig. 1c) just the vertical moving dots 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:



(a) (b)

(c) (d)

Fig. 1: Application of the salient energy extraction framework to a synthetic display (a) containing a pop-out object, represented by the red-circled square, moving horizontally, while the other squares move vertically. (b), the horizontal motion feature map and (c) the vertical motion feature map are shown, both at f=0.3750. (d), the temporal average of the module of salient energy, achieved by merging horizontal and vertical energy at different frequencies.

$$|E(x,y,t)| = \sqrt{E_h(x,y,t)^2 + E_v(x,y,t)^2} \tag{4}$$

$$\angle E(x,y,t) = \arctan(E_v(x,y,t)/E_h(x,y,t)) \tag{5}$$

A saliency map derived from magnitude map be seen in Fig.1d, 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 formation and saliency evaluation

In the previous section, we have shown how to obtain a saliency map enhancing relevant motion zones. Such a map is yet pixel-based and, as said in the introduction, an object-based map would best help subsequent processing for object recognition and action selection. We need to evaluate the saliency of an object with respect to its entirety 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 [20].

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. A simple segmentation on the module map would not be sufficient, since adjacent objects moving in different directions would be merged. Hence, we threshold the temporally averaged magnitude map $|E(x,y)|$ to discard points with too low energy and apply the mean shift algorithm to the phase of the remaining points in the average phase map $\angle E(x,y)$. 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 previously. We cluster in this way objects 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.2a. The vertically moving squares are assigned to a class while the horizontally moving square belongs to a different class.

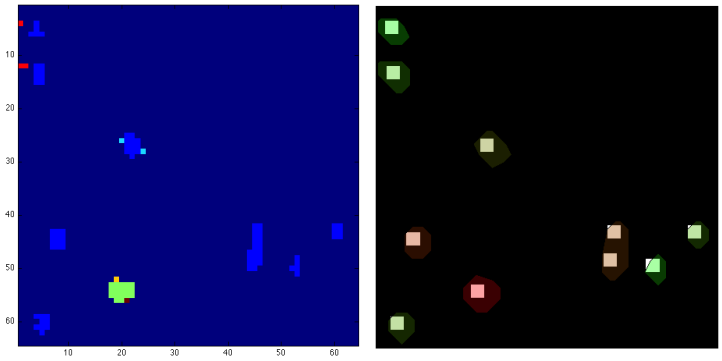
Once we have labelled regions we can extract the object convex hulls by means of morphological operations and can compute their saliency. Again, we define object saliency as proportional to motion contrast in terms of module and orientation, in a center-surround fashion. Given an object $\mathbf{o}$, defined by its bounding box, and given its surround $N(\mathbf{o})$ of size proportional to the area of $\mathbf{o}$, similarly to [17], we have:

$$S_{mag}(\mathbf{o}) = \langle |E(x,y)| \rangle_{(x,y)\in(o)} - \langle |E(x,y)| \rangle_{(x,y)\in N((o))} \tag{6}$$

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:

$$h(i) = \sum_{\angle E(x,y)\in i} |E(x,y)| \tag{7}$$



(a) (b)

Fig. 2: Object segmentation and prioritization. (a), the result of the mean shift segmentation on directions relative to salient energy is displayed. Each cluster is denoted by a different color. (b), 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).

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:

$$S_{or}(\mathbf{o}) = 1 - \sum_i (\sqrt{h_o(i) * h_{N(o)}(i)}) \tag{8}$$

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:

$$S(\mathbf{o}) = \alpha S_{mag}(\mathbf{o}) + \beta S_{or}(\mathbf{o}) \tag{9}$$

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.2b, 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 Experiments and discussion

Having tested the effectiveness of our framework on synthetic stimuli, where the pop-out target can be easily and univocally identified by every subject, we made some experiments with real world sequences. In particular, we took some sequences from the Getty Image footage [1], those taken with fixed camera and displaying multiple moving objects. In a crowded scene, indeed, such as a station or a crossroad (see Fig. 3 and 4), 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. 3 and 4, it can be noticed how differently moving objects even very close to each other 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.

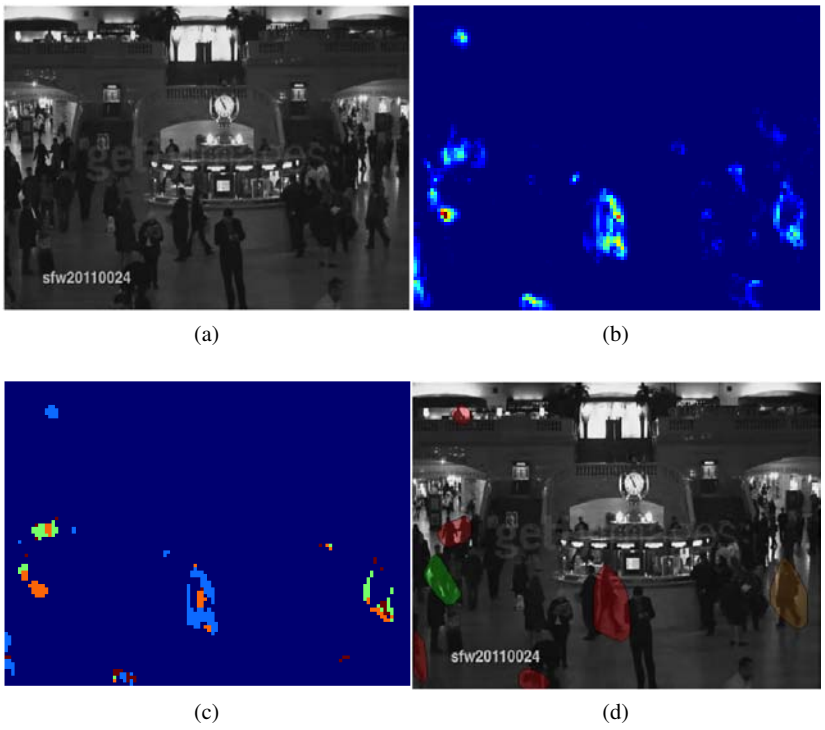
The presented framework can be tuned and refined in a number of ways to make it more or less selective and task-oriented. A major limitation, at the moment, is the constraint of stationary camera. This limits its current biological plausibility, since humans are able to discriminate scene motion from ego-motion when moving the head or the body, due to the Vestibular-Ocular Reflex (VOR) present in our visual system. 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 [16].

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 [18].

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)

[1] http://www.gettyimages.com/

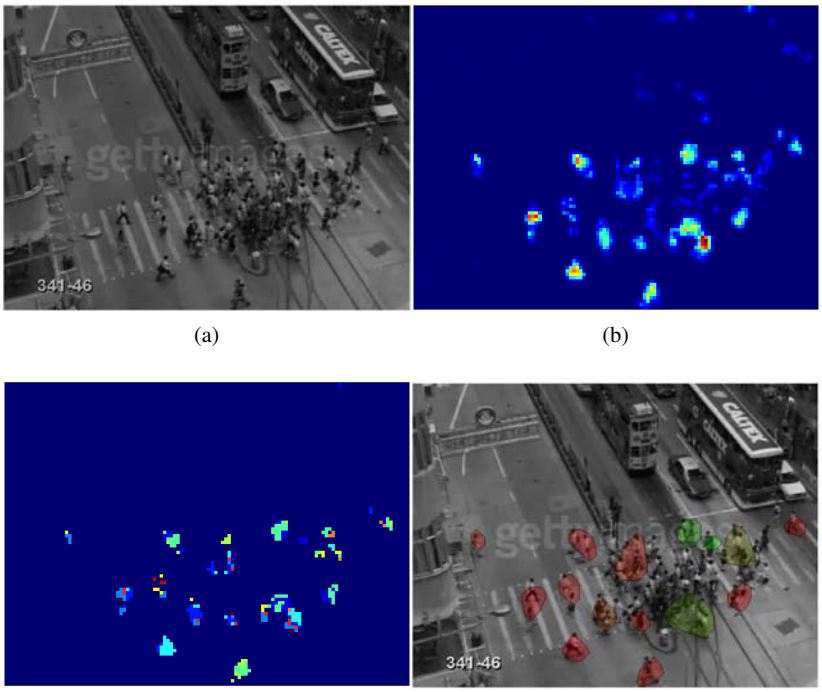


(a) (b)

(c) (d)

Fig. 3: Object-based saliency selection applied to a real world sequence. (a), an original frame. (b) the temporal average of the energy module. (c), the segmented phase map and (d) objects with their saliency are shown.

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. Pattern Analysis and Machine Intelligence, IEEE Transactions on **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. 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)
10. 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)
11. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. Trends in neurosciences **15**(1), 20–25 (1992). URL http://view.ncbi.nlm.nih.gov/pubmed/1374953



(a) (b)

(c) (d)

Fig. 4: Object-based saliency selection applied to a second real world sequence. (a), an original frame. (b) the temporal average of the energy module. (c), the segmented phase map and (d) objects with their saliency are shown.

12. 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)
13. Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging **10**(1), 161–169 (2001)
14. 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)
15. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**, 171–177 (2009)
16. Orabona, F., Metta, G., Sandini, G.: A proto-object based visual attention model. In: Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint, pp. 198–215 (2008)
17. Pylyshyn, Z.W.: Visual indexes, preconceptual objects, and situated vision. Cognition **80**(1-2), 127–158 (2001)
18. Rosenholtz, R.: A simple saliency model predicts a number of motion popout phenomena. Vision Research **39**(19), 3157 – 3163 (1999)
19. 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)

20. Scholl, B.J.: Objects and attention: the state of the art. Cognition **80**(1-2), 1–46 (2001)
21. 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)
22. Tatler, B.: Current understanding of eye guidance. Visual Cognition pp. 777–789 (2009)
23. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks **19**(9), 1395 – 1407 (2006)