Biologically Inspired Cognitive Architectures (2013) 5, 10–28

Available at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/bica

RESEARCH ARTICLE

Saliency prediction in the coherence theory of attention☆☆

Valsamis Ntouskos [a], Fiora Pirri [a,*], Matia Pizzoli [b,1], Arnab Sinha [a], Bruno Cafaro [a]

[a] ALCOR Lab., DIIAG, University of Rome, ''Sapienza'', Rome, Italy
[b] Artificial Intelligence Lab., University of Zurich, Zurich, Switzerland

KEYWORDS
Visual attention;
Saliency prediction;
Proto-objects;
Visual search;
Cognitive robotics;
Cognitive vision

Abstract
In the coherence theory of attention, introduced by Rensink, O'Regan, and Clark (2000), a coherence field is defined by a hierarchy of structures supporting the activities taking place across the different stages of visual attention. At the interface between low level and mid-level attention processing stages are the proto-objects; these are generated in parallel and collect features of the scene at specific location and time. These structures fade away if the region is no further attended by attention. We introduce a method to computationally model these structures. Our model is based experimentally on data collected in dynamic 3D environments via the Gaze Machine, a gaze measurement framework. This framework allows to record pupil motion at the required speed and projects the point of regard in the 3D space (Pirri, Pizzoli, & Rudi, 2011; Pizzoli, Rigato, Shabani, & Pirri, 2011). To generate proto-objects the model is extended to vibrating circular membranes whose initial displacement is generated by the features that have been selected by classification. The energy of the vibrating membranes is used to predict saliency in visual search tasks.

1. Introduction

Saliency prediction in visual search requires to understand which features of the scene are processed and how, and in which way this processing delivers a structure that is overtaken by attention, which then induces focusing on a selected region of the scene.

In artificial systems this is a crucial concept. There are two main reasons for that. On the one hand the complexity of searching the visual field is too high to be managed by

☆☆ The research has been supported by EU project NIFTi.
* Corresponding author.
E-mail addresses: ntouskos@dis.uniroma1.it (V. Ntouskos), fiora.pirri@dis.uniroma1.it (F. Pirri), pizzoli@ifi.uzh.ch (M. Pizzoli), sinha@dis.uniroma1.it (A. Sinha), cafaro@dis.uniroma1.it (B. Cafaro).
[1] The author has contributed to the paper while he was at Alcor Lab, in Rome.

2212-683X
http://dx.doi.org/10.1016/j.bica.2013.05.012

processing the whole visual input at the resolution of the fovea, as indicated by Tsotsos et al. (1995). On the other hand feature detectors and orientation filters handle pre-attentive processing by partially discarding the visual input, but they cannot handle the further integration processing required to lift up the low-level structures to focused attention.

We should note that artificial systems suffer of several limitations due to the mechanic, electronic and software components. Yet artificial systems need to learn to predict saliency to find targets in crowded scenes, without overloading teir resources. This is a necessary step in the design of efficient cognitive systems, to avoid memory or reasoning being clogged and paralyzed by the huge amount of visual information acquired at possibly high frame rate. A tacit assumption is that artificial computational models rely on psychophysical, neurophysiological and psychological studies (PNP) on pre-attentional and attentional processing, and then add further constraints to these models to cope with the above mentioned limitations.

This is the line of research mainly taken so far, though following two main directions, namely predicting saccade directions and predicting saliency from the features standpoint. Predicting saccades directions has been analyzed in Koch and Ullman (1985), Tsotsos et al. (1995), Itti, Koch, and Niebur (1998), Minato and Asada (2001), Belardinelli, Pirri, and Carbone (2007). Predictions of saccade targets with a number of features, via bottom-up models, has been tested in Carmi and Itti (2006).

In general, approaches have exploited the simulation of saccades either by active cameras, as in Butko, Zhang, Cottrell, and Movellan (2008), Mancas, Pirri, and Pizzoli (2011), or via biologically founded prior models of saliency as in Pichon and Itti (2002), Ackerman and Itti (2005), Hügli, Jost, and Ouerhani (2005), Cerf, Harel, Einhäuser, and Koch (2007), Sala, Sim, Shokoufandeh, and Dickinson (2006), Mahadevan and Vasconcelos (2010), to cite some of the works from the wide literature on saliency prediction.

In this paper we focus on the steps between features analysis and collection and their integration into a coherent structure that is then passed to attention, basing our approach purely on collected data and the concept of proto-object developped within the coherence theory of attention by Rensink (2000).

Indeed, since Treisman and Gelade (1980) foundational work on feature integration, it became clear that in the pre-attentive, early vision phase, primitive visual features can be rapidly accessed in searching tasks. For example colors, motion, and orientation can be processed in parallel and effortlessly, and the underlying operations occur in within hundreds of milliseconds. So the pre-attentive level of vision is based on a small set of primitive visual features organized in maps, that are extracted in parallel while the attentive phase serves to group these features into coherent descriptions of the surrounding scene. When attention takes on the control, processing passes from parallel to serial.

Since Treisman's feature integration theory, several models have been further provided in the literature, for feature integration. Among those that led to a concept of representation we consider Duncan and Humphreys (1989) who have observed that there is a large differentiation *in search difficulty, observed across different stimulus material*. On this basis Duncan introduces the theory of visual selection as distinguished into three stages: the *parallel one*, that produces an internal structured representation, a *selective one* matching the internal representation, and the *transduction one* providing the input of selected information to the visual short term memory. This theory relies on the evidence of low efficiency of basic features parallel processing, in the presence of heterogeneous distractors. On the basis of this observation Duncan introduces the concept of *structural unit* as an internal representation given to the visual input (close to 3-D model of Marr & Nishihara (1978)). Further, Wolfe (1992) has shouldered the concept of structural units, by noting that visual search might need grouping and categorization. Indeed, Wolfe, Friedman-Hill, Stewart, and O'Connell (1992) suggest that categorization is a strategy that is invoked when it is useful and that it could affect different features of the visual input. Wolfe (1994) makes clear that attentional deployment is guided by the output of earlier parallel processes, but its control can be exogenous, *based on the properties of the visual stimulus* or endogenous, based on the subject task, and he introduces the notion of feature maps (see also Treisman, 1985) as independent parallel representations for a set of basic limited visual features. Finally, activation maps, both bottom-up and top-down, serve in Wolfe (1994) model to guide attention toward distinctive items in the field of view. In summary Wolfe suggests that information extracted in parallel, with loss of details, serves to create a representation for the purpose of guiding attention.

The huge amount of literature that has studied how, from parallel processing, across large areas of the visual field, focused attention emerges (see also Neisser & Becklen, 1975 & Julesz, 1986) has led to the quest for a virtual representation that could explain the way input is discarded and selected features are integrated in a coherent representation.

According to these principles, in this paper we propose a methodology, suitable for computational artificial-attention, to study saliency for visual search, in dynamic complex scenes, motivated by the concept of virtual representation developed in the coherence theory of attention of Rensink (2000), Rensink et al. (2000), Rensink (2002). Rensink introduces the concept of *proto-object* as a volatile support for focused attention, which is actually needed to see changes, see Rensink, O'Regan, and Clark (1997). Rensink (2000) assumes that proto-objects are formed in parallel across the visual field and form a continuously renovating flux that is accessed by focused attention. Proto-objects are collected by focused attention to form a stable object temporally and spatially coherent, which provides a structure for perceiving changes.

In Fig. 1 Rensink's triadic architecture is illustrated. In this architecture the lower level corresponds to the retinotopic mapping and, going up, proto-objects are structures for more complex feature configurations formed in parallel across the visual field and lying at the interface between low-level vision and higher attentional operations. These structures are said to be volatile, and fading away as new stimuli occur, within ''few hundreds of milliseconds'', as detailed in Rensink et al. (2000). Focused attention, in Rensink's triadic architecture, accesses some of the generated proto-objects to stabilize them and form individual objects ''with both temporal and spatial coherence'',

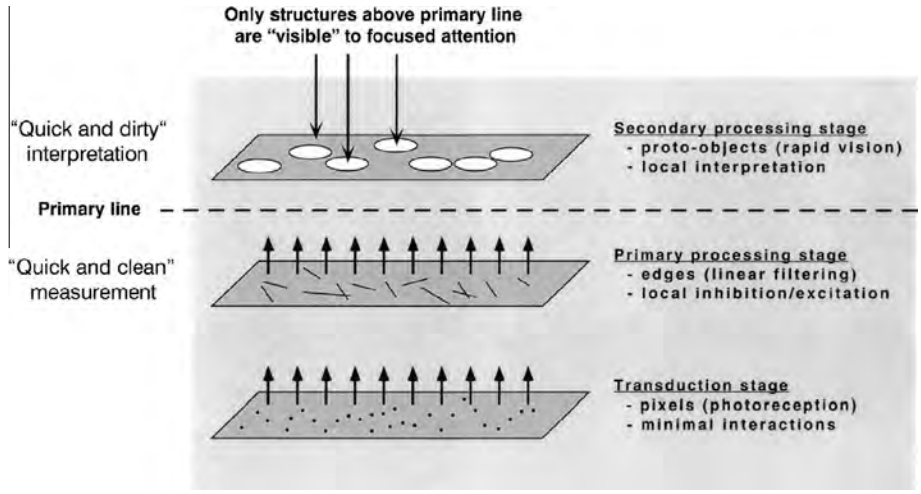

Fig. 1 The image above, taken from Rensink (2000), illustrates Rensink low-level vision architecture whose output are proto-objects *that become the operands for attentional objects* Rensink (2000).

Rensink (2000). Proto-objects are linked within a coherence field to the *nexus*, a structure coarsely summarizing the properties of the stabilized ones. Proto-objects have been explored in computational attention for modeling how object recognition can use their representation and generation, thus at the high-level interface, in Walther and Koch (2006), and in Orabona, Metta, and Sandini (2008). Here, instead, we are interested in the other side of the interface, namely we model their generation and study their spatial and temporal persistence across the visual fields in visual search tasks. Note that we take into account real dynamic environments. Furthermore we show that these structures can be used to learn the parameters of the underlying process and predict saliency distribution across the scene.

The paper is organized around the problem of modeling the data acquisition, for a freely moving subject, the recovery of the point of regard in the scene and the proto-object generation, as follows. In the next section we illustrate how to obtain the scanpath of a subject searching for some objects in the scene. Namely how to obtain the position of the head and the direction of the gaze in the scene, using a wearable device, the Gaze Machine (GM). In the section *Coherent features for point saliency*, we illustrate how features are learned from the data acquired by the GM, specifically for a set of search tasks. Then, in section *Generating Proto-Objects*, we introduce a model for the generation of proto-objects based on vibrating membranes to account for their volatility, according to the learned features. Finally we provide some experimental validation.

2. Acquisition model for search strategy estimation

To model saliency prediction, computational studies have quite limited resources available, as data acquisition is based on uncertain measurements and ground truth is available only if experiments are rather constrained. The realization of a wearable device that allows to register the Point of Regard of a subject in an unconstrained condition has made possible to collect a great amount of data, see Fig. 2.

We aim at exploiting these data for modeling the features that are selected during a search task, whether these specify general properties that are preserved across tasks or local properties closely related to the target. These properties characterize the spatial and temporal relations inducing the stimulus to be triggered. As highlighted in Serences and Yantis (2006) the V4 area displays neural activity with features similar to the target, and this is the area involved in the formation of a coherence field, according to the coherence theory of attention. Indeed, the interaction between stimuli-driven and voluntary factors becomes further and further relevant in the later stages of attentional processing, where more complex coherent fields of features configurations are formed. From the stand point of computational attention a *proto-object* can be described as a *configuration of features having relative time and spatial coherence, directly affected by attention, and generating a motion field pulling the gaze toward the target*.

Proto-objects in this sense are dynamic and relatively volatile feature structures related both to fast eye movements, namely saccades, and to saliency. These feature structures are precursors of attention and further used by attention to drive recognition — this is the double face of proto-objects between pre-attentive and selective attention, as highlighted in Duncan and Humphreys (1989) and Rensink (2000) — and can be localized in time and space: proto-objects may last few milliseconds up to hundreds of milliseconds.

We recall that the POR, namely the Point of Regard, is *the point on the retina at which the rays coming from an object regarded directly are focused*. In particular, we assume that PORs are the point on the fovea, sub-tending a visual angle of about 1.7°.

Saccades are fast eye movements that can reach peak velocities of 1000°/s. While a subject is moving, like in our framework, saccades do not exceed 30°, but the velocity follows an exponential function. According to Bahill and Stark (1979), the range in the duration of 30° saccades can be up to 100 ms. Saccades models rarely explain the role of saliency, being mainly motivated by the need to model the motion control (see Bahill, Bahill, Clark, & Stark, 1975;

Fig. 2 The Gaze Machine (GM) worn by the subject collecting PORs in an outdoor search task.

Bahill & Stark, 1979; Zhou, Chen, & Enderle, 2009, and for a review see Kowler, 2011 and the references therein). It follows that saccade models do not contribute to the interpretation of proto-objects, although saccades direction and speed are substantial to explain the motion field a proto-object generates and how it fades away.

Similarly, saliency models not grounded in the 3D visual scene lack to explain the coherence of proto-objects, their motion field, hence their dynamics. To measure the volatility of proto-objects we rely on two models: a model of the scan path, and a model of the surface response to the POR. To obtain meaningful data from which parameters can be estimated, we use an acquisition device, the Gaze Machine specified in Pirri et al. (2011), here denoted GM. In particular we present below a novel method to recover the scan path of the the head and eyes of a subject wearing the device.

2.1. Scan path estimation

The formal model for scene acquisition, PORs projections into the retinal plane (image plane) and their registration into the scene structure, while the subject explores the environment, is the Gaze Machine (GM) model, described in Pizzoli et al. (2011) and Pirri et al. (2011). Here we are mainly concerned with the scan path of the head; namely of the subject's head, while she/he is moving across the environment to perform a search task. The task implies possible return to previously focused regions, in so inducing relations among the PORs at different time periods. In other words the scan path model has to establish whether a set of PORs belongs to the same saliency region, according to the process deployed during search. Some results of scanpath estimation, namely of the projection of the gaze on the visual field, are illustrated in Fig. 5.

First note that the GM enables good controlled experiments, as the device can be well fitted on the head, the pupil rate acquisition can reach 180 Hz, ensuring to get good saccades approximation, while the visual field can be acquired at a rate up to 30 Hz, the association with the much faster acquisition of gaze is maintained by time-stamping. The GM calibrated stereo rig records the experimental stimuli, allowing for dense 3D reconstruction from multiple views. Moreover, the localization of the subject in the 3D experimental scenario is based on the visual data acquired by the GM scene cameras.

The above statement assesses that the model we propose is quite general and allows a calibration procedure that is efficient and easy to perform *on field*, with little intervention from the subject. After the calibration, the parameters for the model of eye positions are recovered and the gaze direction $\hat{\rho}(t)$ is computed, on the basis of the imaged pupil at time t, and the geometry of the multi-camera system. The estimated POR is relative to the acquisition device and a localization step is needed in order to measure gaze behaviors in the 3D world taking into account the changes in the pose of the subject's head.

To build a map of gazed 3D points requires the following steps:

1. estimating the 3D POR π^c in the reference frame of the GM left scene camera.
2. estimating the 3D pose (6 degrees of freedom) of the GM left scene camera in the reference frame of the experiment at hand, in terms of translation $\mathbf{t}$ and orientation $\mathbf{R}$;
3. computing the 3D POR in the world reference frame as $\pi^w = \mathbf{R}\pi^c + \mathbf{t}$.

Note that the 3D PORs are naturally attached to 3D points that are imaged in the retinal plane, and the 3D points generate the 3D global map. For an abstract structure of the hierarchical construction see Fig. 4.

2.2. Subject localization

Most of the issues affecting the localization of a camera system, see Hartley and Zisserman (2000), Faugeras, Luong, and Papadopoulou (2001), also apply to the GM, with some notable differences. Indeed, the main concern of the GM localization is high precision in the estimation of the whole trajectory, needed to correctly estimate the 3D POR, see Fig. 3, to see the head poses of a subject performing a search task.

We follow an efficient hierarchical approach subdividing the whole trajectory into sets of frames, that we specify as *coherent subsequences*. Indeed, subsequences are characterized by a high level of coherence in terms of what the subject is attending in the course of the experiment. More specifically, the pose estimation is performed sequentially,

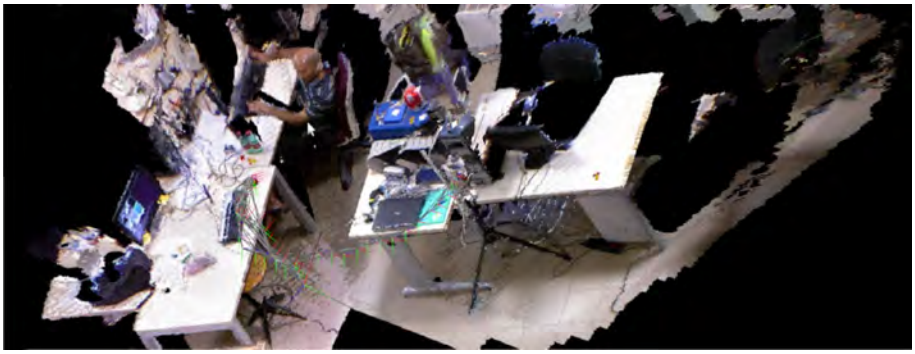

Fig. 3 The Figure illustrates the reconstruction of the scene where the subject is performing the experiment *searching for the J*, wearing the GM. The head poses are projected on the scene, head poses are computed with the described localization algorithm.

adding a new frame to the last acquired set, denoted *subpath*, as long as the estimation is sufficiently accurate, performing sparse bundle adjustment to enforce consistency and to avoid drifting, see Triggs, McLauchlan, Hartley, and Fitzgibbon (2000); Hartley and Zisserman (2000).

Subsequences are induced by the selection of a *keyframe* to delimit the coherence of head poses. Namely, the set of keyframes constitutes a subset of the whole frame sequence and a new keyframe, eliciting a new subsequence, is created upon the event of a change in the visual scene.

The sequence of images collected by the GM scene cameras is used to localize the subject in the experimental environment. The estimation of the subject's pose relies on matching descriptors from visual features corresponding to the current view with those recorded in the map built so far. The overall process is summarized as follows:

1. Take the first frame of the sequence as the first keyframe. A map of 3D feature points is initialized by triangulating matched image features in the first pair of stereo frames.
2. For each new pair of stereo frames, compute matched feature points and descriptors among left and right views; triangulate to get a new set of unoptimized 3D points. Match the computed descriptors with the current map. Estimate the pose w.r.t. the current map and compute the POR in 3D. Check if a new keyframe has to be selected, if not repeat 2.
3. Upon the selection of a new keyframe, add the current frame to the keyframe list. Optimize by a local bundle adjustment w.r.t. unoptimized 3D points and cameras from the subsequence. Add the optimized points to the map and empty the set of unoptimized points.

Let us call $(\tilde{\mathbf{x}}_i, \widetilde{\mathbf{X}}_i)$, $i = 1, \ldots N$, the N pairs of matched retinal plane and map points, $\tilde{\mathbf{x}}_i \in \mathbb{R}^2$ and $\widetilde{\mathbf{X}}_i \in \mathbb{R}^3$ respectively. The pairs $(\mathbf{x}_i, \mathbf{X}_i)$ represent the same points in homogeneous coordinates: $\mathbf{x}_i \in \mathbb{R}^3$ and $\mathbf{X}_i \in \mathbb{R}^4$. The goal is to compute the pose, expressed by the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$, of the camera that is projecting the 3D points $\mathbf{X}_i$ into the retinal points $\mathbf{x}_i$. We refer in general to cameras specified by a translation $\mathbf{t}$, a rotation $\mathbf{R}$ and a calibration matrix $\mathbf{K}$ as $\mathbf{P} = \mathbf{K}[\mathbf{R}|\,\mathbf{t}]$. The rotation, translation and calibration might be decorated by superscripts specifying whether they involve the left (l), the right (r), or the scene (s) cameras. According to Hartley and Zisserman (2000), let us define the matrix $\mathbf{K}$ expressing the intrinsic camera parameters, namely the focal lengths f_x and f_y and the position of the principal point in image coordinates (p_x, p_y), as

$$\mathbf{K} = \begin{pmatrix} f_x & 0 & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{1}$$

Fiore's linear algorithm for exterior orientation Fiore (2002) has been used to generate multiple hypotheses in a RANSAC-based, robust estimation process (Fischler & Bolles, 1981). The core routine estimates the camera pose by solving

$$Z_i \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = s\mathbf{K}\mathbf{R}(\widetilde{\mathbf{X}}_i + \mathbf{t}) \qquad i = 1, \ldots, N. \tag{2}$$

Here Z_i, $i = 1, \ldots, N$ are the depth parameters and s is the scale parameter. Note that these last parameters can be recovered up to an arbitrary common scale factor, and that the calibration matrices (likewise those of the eye cameras) are pre-estimated. The algorithm first estimates Z_i in order to subsequently solve the problem of absolute orientation with scale. The model selection process makes use of an error function that takes into account re-projection errors in both the left and right retinal planes of the stereo pair. Using the l and r superscripts to identify quantities related to the left and right scene cameras, respectively, and assuming the relative pose $\mathbf{R}^s$ and $\mathbf{t}^s$ of the scene cameras fixed to the GM stereo rig known from calibration, the error function is:

$$\epsilon_i = d\left(s\mathbf{K}^l\mathbf{R}(\tilde{\mathbf{X}}_i + \mathbf{t}), \mathbf{x}_i^l\right)^2 + d\left(s\mathbf{K}^r\mathbf{R}^s[\mathbf{R}(\tilde{\mathbf{X}}_i + \mathbf{t}) - \mathbf{t}^s], \mathbf{x}_i^r\right)^2 \tag{3}$$

where d is the Euclidean distance and $\mathbf{K}^l$, $\mathbf{K}^r$ are the calibration matrices of the left and right scene cameras, see Pirri et al. (2011). The two distance terms in Eq. (3) account for reprojection errors in the left and right scene camera planes. The largest consensus set is selected by RANSAC according to Eq. (3) and used to estimate a model. A final Levenberg-Marquardt optimization is carried out to refine the linearly estimated pose by iteratively minimizing ϵ_i with respect to $\mathbf{R}$ and $\mathbf{t}$:

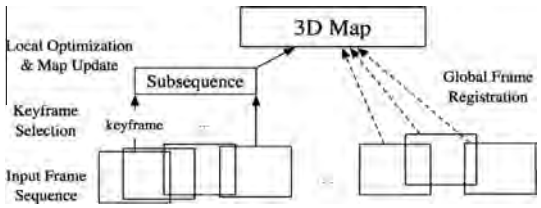

Fig. 4 Visual Localization of the subject. Local consistency is enforced by optimization on frame subsequences, limited by keyframes. Frame registration with the 3D map ensures global consistency.

$$\mathbf{R}, \mathbf{t} = \arg\min_{\mathbf{R},\mathbf{t}} \sum_i \epsilon_i. \tag{4}$$

Details of the suggested minimization can be found, for example, in Hartley and Zisserman (2000).

Keyframe selection Upon the acquisition of a new pair of scene frames, the pose of the subject is estimated from matched features among the current frames and the 3D map. This method guarantees a global consistency across the whole experiment and it is accurate as long as the global map is accurate.

At this point the goal is to detect the change in space of the focus of overt attention in order to identify sequences of PORs that exhibit a coherence in space and time.

The collected scene frames are clustered into subsequences according to the subject's POR and *keyframes* are used to delimit coherent subsequences. Roughly speaking, keyframes consist of scene frames corresponding to time steps in which the focus of overt attention changes and a new sequence of PORs starts. Therefore, a strategy is required to select keyframes when no knowledge of the pose and, thus, of the 3D point of regard of the subject is retained. We introduce a keyframe selection method that evaluates the *novelty* of a view in the experiment by measuring how different it is from the last selected keyframe. The quantities involved in the keyframe selection are the n matched pairs of visual features $\{(\mathbf{x}, \mathbf{x}')_i, i = 1 \dots n\}$, between the current scene frame and the last keyframe, and the pair (γ, γ') of gaze positions as projected into to the current frame and into the last keyframe. Note that in this phase the correspondences $(\mathbf{x}, \mathbf{x}')_i$ are drawn among frames collected by one of the scene cameras at different timesteps and the pair (γ, γ') refers to coordinates on the image plane.

A change in the subject's vantage point induces a motion of the camera acquiring the scene and a variation of the POR in space. Suppose that the subject, during a search task, is focusing on a particular object in the scene and that her pose, in the experiment frame, can be described by a certain motion model. This will induce a sequence of PORs that is consistent with the given motion model. Therefore, we evaluate the opportunity to instantiate a new keyframe by checking the consistency of the current POR with a motion model estimated on the basis of frame to keyframe correspondences. We characterize the subject's change in head pose by means of two types of motion models that can be estimated from the scene frames: a planar homography, represented by the **H** matrix, and the *fundamental matrix* **F** (see Hartley & Zisserman, 2000 for a comprehensive

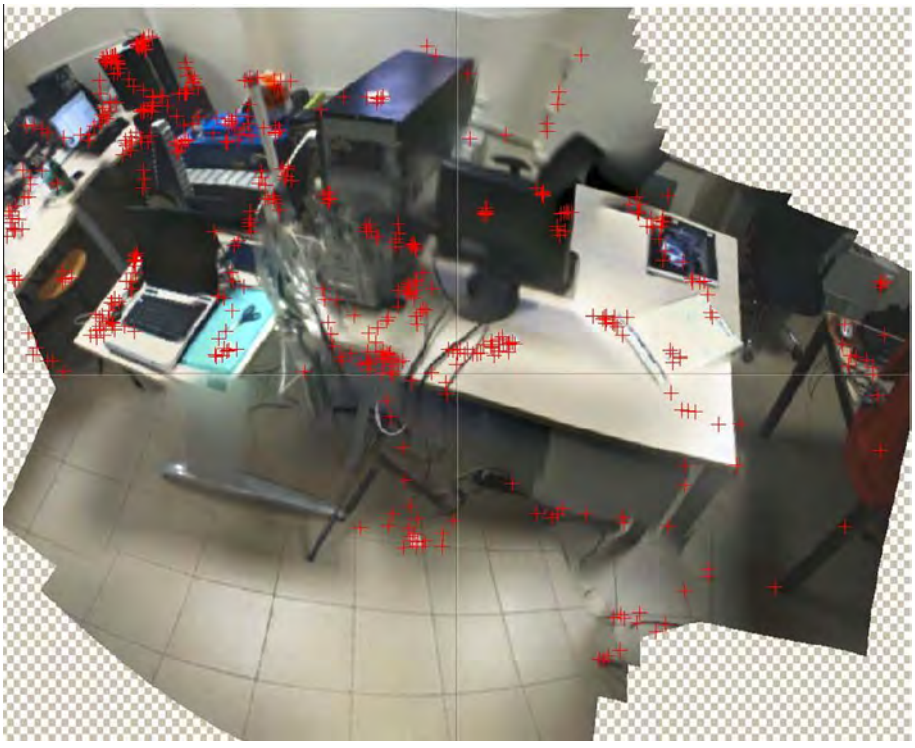

Fig. 5 A panoramic stitching and the PORs collected in 20 s; the stitching has been realized with 30 images over a collection of 600 left images of the scene. The acquisition of the scene is at 30 Hz while the acquisition of the eye is at 120 Hz. The PORs are measured on the scene via dense structure from motion and further reprojected on the retinal plane (image plane).

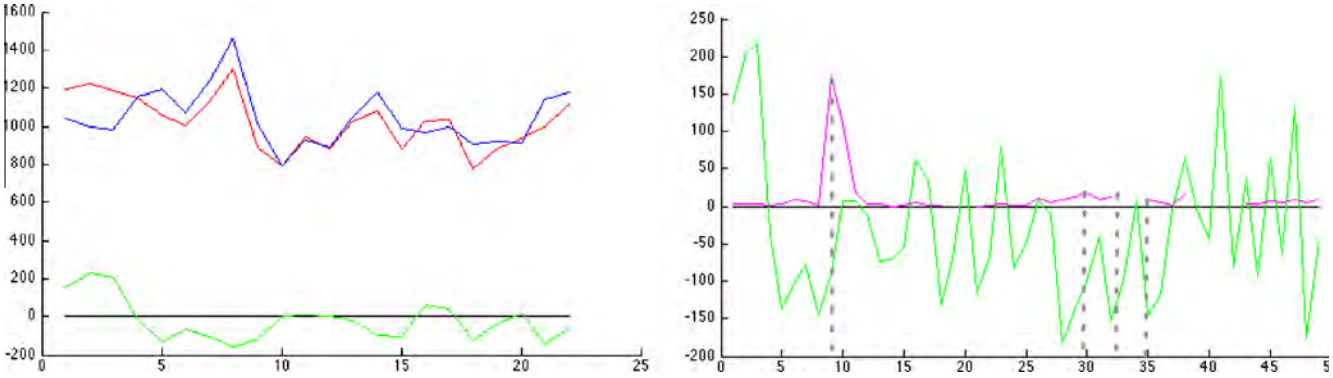

Fig. 6 Keyframe selection criterion. Left: $\Gamma(\mathbf{F})$ (red), $\Gamma(\mathbf{H})$ (blue) and $\Gamma(\mathbf{F}) - \Gamma(\mathbf{H})$ (green). Right: $\Gamma(\mathbf{F}) - \Gamma(\mathbf{H})$ (green) and δ (magenta). Keyframes are selected in correspondence of dashed lines.

treatment). A motion characterized by a small baseline between the current frame and the last keyframe is best described by a plane homography **H**. In contrast, when the subject's head undergoes a translational motion, the fundamental matrix **F** is more suitable to describe a general camera motion.

Building on the Geometric Robust Information Criterion (GRIC, Torr, 1998), a score function is evaluated for both the **F** and **H** motion models at every frame in order to quantitatively measure the fitness of each model to the data. The score function takes into account the n matched features with the last keyframe, the residuals e_i, the number k of model parameters, the error standard deviation σ, the dimensions r of the data and q of the model:

$$\Gamma = \sum_{i=1}^{n} \rho(e_i^2) + [nq \ln(r) + k \ln(rn)], \tag{5}$$

where

$$\rho(e_i^2) = \min\left(\frac{e_i^2}{\sigma^2}, 2(r-q)\right). \tag{6}$$

Eq. (5) returns the lowest score for the model that best fits the data. Once the motion model has been selected, it is used to evaluate the gaze variation, see Fig. 6. According to the selected motion model, changes in the subject's vantage point involving the gaze projections γ and γ' can be detected and new keyframes are instantiated on the basis of the following criterion, balancing between the choice of an homography **H** and of the fundamental matrix **F**:

$$(\Gamma(\mathbf{F}) - \Gamma(\mathbf{H})) \cdot \delta < 0, \quad \delta = \begin{cases} \gamma'^{\top}\mathbf{F}\gamma & \text{if } \Gamma(\mathbf{F}) < \Gamma(\mathbf{H}) \\ \|\mathbf{H}\gamma - \gamma'\| & \text{otherwise.} \end{cases} \tag{7}$$

Upon the instantiation of a new keyframe at time t, the following steps are performed:

- *Subsequence Optimization*. Let $\mathscr{X}$ be the set of unoptimized points, then this set is optimized by Sparse Bundle Adjustment (SBA) (Lourakis & Argyros, 2009) on the sequence of the last k camera poses, using a reprojection error ϵ_{ij} as objective function

$$\min_{\mathbf{R}_i, \mathbf{t}_i, \mathbf{X}_j} \sum_{ij} \epsilon_{ij}, \tag{8}$$

With

$$\epsilon_{ij} = d\left(s\mathbf{K}^l\mathbf{R}_i(\widetilde{\mathbf{X}}_j + \mathbf{t}_i), \mathbf{x}_{ij}^l\right)^2 + d\left(s\mathbf{K}^r\mathbf{R}^s[\mathbf{R}_i(\widetilde{\mathbf{X}}_j + \mathbf{t}_i) - \mathbf{t}^s], \mathbf{x}_{ij}^r\right)^2. \tag{9}$$

Here $i = t-1, \ldots, t-k$, $\widetilde{\mathbf{X}}_j \in \mathscr{X}$ and $\mathbf{x}_{ij}^c$, $c \in \{l, r\}$ is the point $\widetilde{\mathbf{X}}_j$ imaged by the i-th left or right camera respectively.

- *Map Upgrade*. Let $\mathscr{M}$ be the global 3D map, built so far, then $\mathscr{M}$ is updated with the new set of optimized points $\mathscr{X} : \mathscr{M} = \mathscr{M} \cup \mathscr{X}$.
- *Subsequence Initialization*. The set of optimized points is emptied and the number k of camera poses is set to 0.

When a new keyframe is selected, the previous subsequence is terminated, the correspondent points and cameras are optimized and the resultant structure is added to the global map. Each subsequence as defined above is a *coherent subsequence* as it collects a coherent set of PORs, on a specific region in space.

Fig. 7 illustrates the head pose and the PORs related to the scanpath elicited during the search task *looking for J* (see the *Experimental validation* section).

3. Coherent features for point saliency

In the previous section we illustrated how to compute the head scanpath, leading to coherent subsequences of head poses and gaze directions. Once the head poses are retrieved, retrieving the scene structure can be done using the computed camera poses. The scene structure, even if partial, is needed to collect the features of the attended regions. For example, a crucial feature is the space range of PORs, and this is available only if the scene structure is available. Note that by estimating the scene depth, using the computed cameras, a point cloud of the scene structure is obtained.

In this section we illustrate how the coherent subsequence of frames, the point of regard in space and the fixations on the retinal plane can contribute to the definition of the set of features that best specify the visual search task. Though we remark that each search task experiment cleaves the feature set into some unknown prior component; this prior component cannot be recovered

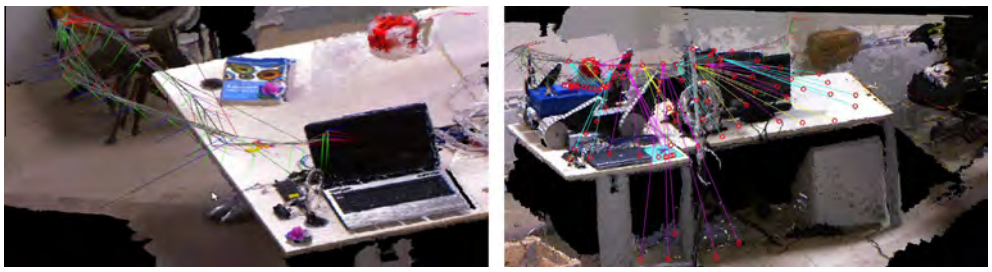

Fig. 7 Head poses of the subject during the experiment *searching for the J*, computed with the described localization algorithm, and the rays joining the head pose with the PORs (the red circles) projected on the scene point cloud. The lines represent, ideally, the intersection of the visual axes.

experimentally from the PORs data, as it is embedded into some prior knowledge the subject has about the shape, dimension and color of both the environment and the object, while she is performing the search.

Now, in our experimental approach, we build an inverse problem, namely given the PORs, the head scan path and the points in the image, we want to determine the properties that are common to all of the experiments. Once these properties are identified then, as described in the following section, we can use them to attempt to define a forward model.

Here we want to recover the features that elicited the PORs, from the scene structure, as computed from different experiments. Features are specific for both the space geometry, such as position on a surface and orientation, and the image such as color and intensity variation. Slightly changing the notation adopted in the previous section, in the following we shall denote a non-homogeneous point in space or on the retinal plane as $\mathbf{X}$ and $\mathbf{x}$, respectively, while in the previous section they were denoted by $\bar{\mathbf{X}}$, and $\bar{\mathbf{x}}$. On the other hand, when a homogeneous point is needed we shall denote it $\widehat{\mathbf{X}}$ or $\hat{\mathbf{x}}$.

Let us consider a coherent subsequence of frames in terms of the set of collected PORs $\mathscr{X} = \{(\mathbf{X}_1, t_0), \ldots, (\mathbf{X}_m, t_q)\}, \mathbf{X}_j \in \mathbb{R}^3$, labeled with the time stamp of their acquisition. It is easy to show that two PORs, even if the same region has been observed at time t and t', cannot coincide, as none is able to observe exactly the same point in space twice. Therefore given the camera $\mathbf{P}_j = \mathbf{K}[\mathbf{R}_j | \mathbf{t}_j]$, there is only one retinal plane $\mathscr{I}_h$ where the POR $\mathbf{X}_h$ is imaged. However if we consider the region around the POR then the points in the region can be imaged into different retinal planes.

Now, for each coherent subsequence, define a monotonic grid of about 12×10^3 nodal points $n_{\mathbf{X}} = (X, Y, Z)^\top$; then we approximate the point cloud with a thin plate surface $S : V \mapsto \mathbb{R}^3, V \subset \mathbb{R}^2$ minimizing the energy functional:

$$\mathscr{M}_\alpha(S) = \sum_{i=1}^{n} (S(X_i, Y_i) - \widehat{Z}_i)^2 + \eta \int_\Omega S_{XX}(X,Y)^2 + 2S_{XY}(X,Y)^2 + S_{YY}(X,Y)^2 \, dX\, dY \tag{10}$$

Here $S(X, Y) = Z$, and $\widehat{Z}_i$ is the depth of the $i-th$ point in the point cloud, $S_{XX}(\mathbf{v}), S_{YY}(\mathbf{v}), S_{XY}(\mathbf{v})$ are the second order derivatives of S, η is a stabilization parameter, and $\Omega \subset \mathbb{R}^2$ is the surface domain; the first term in the rhs of (10) is the penalty term and the second one is the stabilizing functional, for the energy functional, see Hegland, Roberts, and Altas (1997).

A ray $\mathbf{X}(\lambda) = \mathbf{P}^+\mathbf{x} + \mathbf{C}\lambda$ backprojecting a point $\mathbf{x} = (x, y, 1)^\top$, where $\mathbf{P}^+$ is the pseudo inverse of the current camera matrix, and $\mathbf{C}$ its center, shall intersect the surface S into a point $\mathbf{p} = (X, Y, S(X, Y))^\top$, when this point is a POR, it is denoted $\mathbf{p}^\star$. The surface patch around such a point $\mathbf{p}^\star$, is defined according to a distance threshold a; this surface patch is reprojected on the retinal planes of the subsequence, and forms a patch on the retinal planes which is defined the *coherent region*. Therefore a coherent region is the foveated area in the image surrounding a gaze direction. Coherent regions in images are illustrated in Fig. 8.

Given the surface approximating the point cloud, we can sample from the whole data set, retrieved from an experiment, two different set of points: the points on the surface patches centered at $\mathbf{p}^\star$, the pixels on the coherent regions on the retinal planes, and those points, on S and on the retinal planes, who have never been observed, according to the current subsequence. Once these points have been transformed into a feature space, we can obtain a training set $(\mathbf{W}, h)$ such that $h = 1$ if the back transformed item comes from a POR region and $h = -1$ otherwise.

Given a coherent subsequence $\mathscr{I}_1, \ldots, \mathscr{I}_q$ in a time interval $(t_0, t_0 + \Delta t)$, and its associated collection of PORs $\mathscr{X} = \{(\mathbf{X}_1, t_0), \ldots, (\mathbf{X}_m, (t_o + \Delta t)\}, \mathbf{X}_j \in \mathbb{R}^3$, labeled with their time stamp, a surface S, and a region $\mathbf{s}_P = \{\mathbf{p} \in S | \|\mathbf{X} - \mathbf{p}\| \leqslant a\}$, with a the distance threshold indicated above, then for each point in $\mathbf{s}_P$ there is a pixel $\mathbf{x}$ and a retinal plane $\mathscr{I}_s$, $1 \leqslant s \leqslant q$ imaging it. Therefore the set of data, obtained from the POR regions, given a coherent subsequence, in a time interval $(t_0, t_0 + \Delta t)$ and the surface S, is:

$$\{(\mathbf{p}, (\mathbf{x}_1, \ldots, \mathbf{x}_m)) | \mathbf{p} \in S,\ \|\mathbf{X}^P - \mathbf{p}\| < a,\quad \hat{\mathbf{x}}_j = \mathbf{P}_j \widehat{\mathbf{X}}(\lambda), 1 \leqslant j \leqslant m,\ \text{with } \mathbf{x}_j \text{ on some retinal plane } \mathscr{I}_j \text{ in the subsequence}\} \tag{11}$$

Here $\hat{\mathbf{x}}$ and $\widehat{\mathbf{X}}$ are the homogenized version of $\mathbf{x}$ and $\mathbf{p}$, respectively. Points not in this set are the non-observed ones, and are sampled uniformly on the surface and projected on to the corresponding retinal planes points.

Given the above sample set, it is possible to introduce a set of functions mapping points $\mathbf{p} \in S$ and points $\mathbf{x} \in \mathbb{R}^2$ to a suitable feature space. In feature space it is then possible to

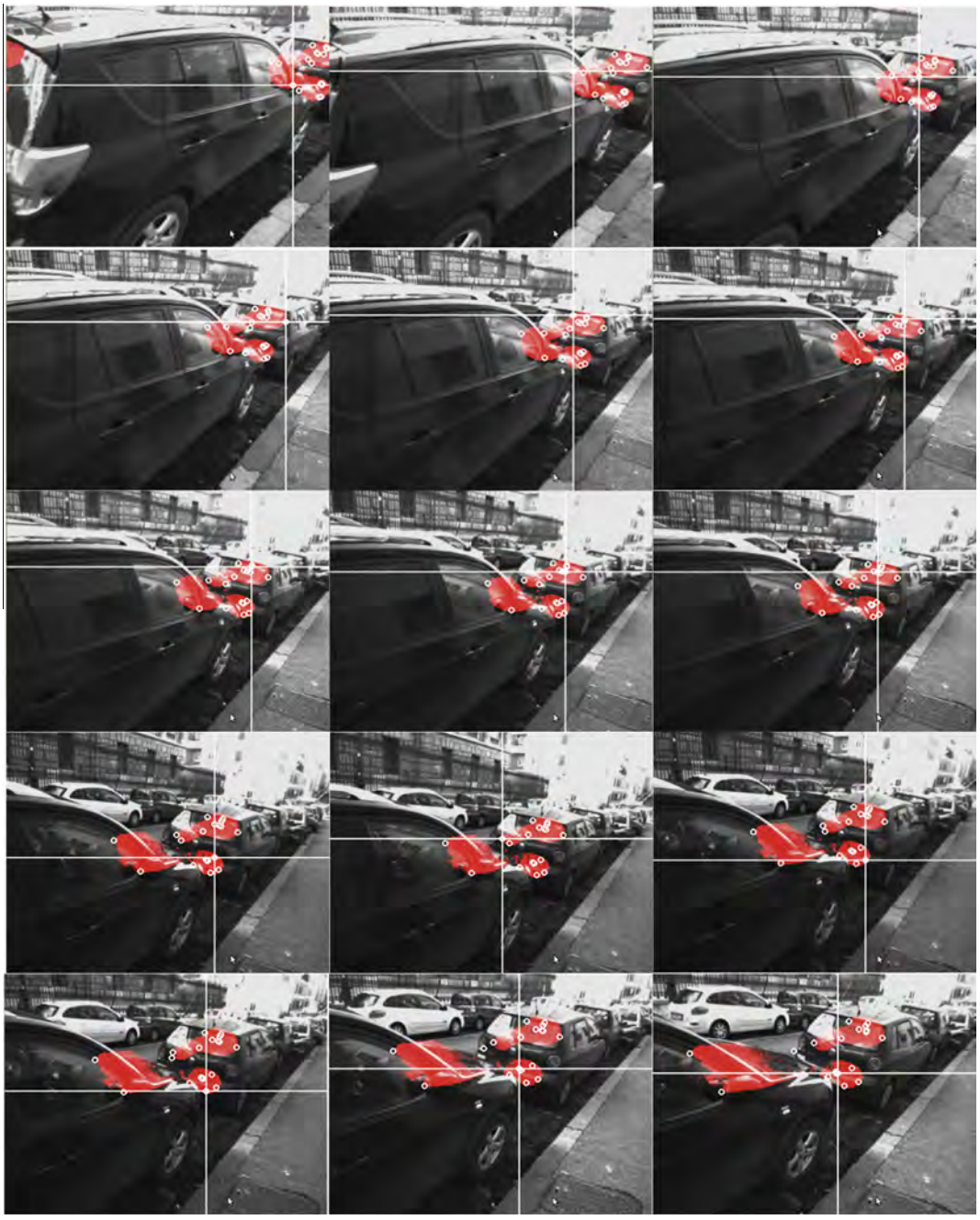

Fig. 8 The sequence of images illustrates the notion of *coherent region*. Here the coherent regions induced by a subsequence of PORs are highlighted in red. They are identified among the frames collected during a search experiment with the GM on the street. In this case the experiment was ''looking for a fine''. The PORs are shown as white circles, while the current POR is shown as a white cross.

learn the function f separating points belonging to salient regions from all the other ones. More precisely, we introduce a set of transformations $\mathscr{F}$ mapping $\mathbf{p} \in \mathbb{R}^3$ and $\mathbf{x}_j \in \mathbb{R}^2$, $j = 1, \ldots, m$, into a feature space, then the learned function f is such that $f(\{\mathscr{F}\} \cdot (\mathbf{p}, (\mathbf{x}_1, \ldots, \mathbf{x}_m))) = h$, $h \in \{1, -1\}$. Here the $\cdot$ indicates that a transformation in $\mathscr{F}$ is applied to the specific set of points, as specified below. We aim at: (1) identify the optimal set of features characterizing a search task and (2) define the function f that separates regions that can/should be attended, according to the search task, from the not attended ones.

A large amount of literature on feature selection (see for example Guyon & Elisseeff, 2003 and references therein) uses a discriminative model, based on the well known family of Support Vector Machines (SVMs) Vapnik (1995), to select the most significant features among a starting base set. Given the set of all possible separating hyperplanes, there are two main optimality criteria for identifying the best one: ℓ_1 and ℓ_2-norm. In the former case, the 1-norm SVM (Mangasarian, 2005) with the ℓ_1-norm, known as lasso penalty is obtained. In the latter case, standard SVM (Cristianini & Shawe-Taylor, 2004; Smola, Bartlett, Schölkopf,

& Schuurmans, 2000) is obtained and the ℓ_2-norm is indicated as ridge penalty. In Zhu, Rosset, Hastie, and Tibshirani (2003) it is argued that 1-norm SVM have advantages over the standard 2-norm, when there are redundant features. The simplest method for achieving feature selection is recursive feature elimination Guyon and Elisseeff (2003), assigning a relative importance to a feature, according to its weight vector within the SVM classifier (see below Eq. (17)). This method allows to remove more than a single feature at a time, once a threshold has been identified.

A first observation for feature selection is that the data collected by the Gaze Machine are available only for training and feature selection, while in general data are taken with a freely moving camera, maybe mounted on a robot pan-tilt head. In general we expect that visual search is performed by a single moving camera, the camera localization and the camera parameters are available during search, a surface patch S for each coherent subsequence is available, though obviously the PORs are available only for the training dataset. Therefore no data specific of the GM can be selected.

Given the surface S, a point $\mathbf{p} = (X, Y, S(X, Y))^\top$ on it and its projection $\mathbf{x}$, we consider different surface parameters that can be obtained from the first and second derivatives of S, in space, and of the image intensity L. The surface $S(X, Y) = Z$ is parametric; let S_X, S_Y be the first order partial derivatives and S_{XX}, S_{YY}, S_{XY} be the second order ones. In the following we identify the surface S with its parametrization.

Let $\mathbf{p}$ be a point on S, the normal N at $\mathbf{p}$ is:

$$N = \frac{S_x \times S_y}{|S_x \times S_y|} \tag{12}$$

Let $\mathbf{v}$ be a vector on the tangent plane at $\mathbf{p}$, the matrices of first and second form for S are:

$$g = \begin{bmatrix} S_x^\top \cdot S_x & S_x^\top \cdot S_y \\ S_x^\top \cdot S_y & S_y^\top \cdot S_y \end{bmatrix}, \quad H = \begin{bmatrix} S_{xx}^\top & S_{xy}^\top \\ S_{xy}^\top & S_{yy}^\top \end{bmatrix} \cdot \begin{bmatrix} N & \mathbf{0} \\ \mathbf{0} & N \end{bmatrix} \tag{13}$$

The above matrices are both symmetric and $det(g) > 0$. Then we consider the Gaussian curvature $K_G = det(H)/det(g)$, namely:

$$K_G = \frac{H_{11}H_{22} - H_{12}^2}{g_{11}g_{22} - g_{12}^2} \tag{14}$$

Actually we considered also the mean Gaussian curvature. Namely, let the best values for $H(\mathbf{v})$ be obtained by $\|\mathbf{v}\| = 1$ and by maximizing the quadratic form $\mathbf{v}^\top H\mathbf{v}$, under the constraint that $\mathbf{v}^\top g\mathbf{v} = 1$. Call these maximal values κ_1 and κ_2. Then the mean Gaussian curvature is:

$$K_M = \frac{\kappa_1 + \kappa_2}{2} \tag{15}$$

We have verified that K_G is more influential than K_M, we indicate the Gaussian curvature of the surface S as σ_S.

Similarly, consider the patches with points $\mathbf{x} = (x, y)^\top$, corresponding to the surface patch with each $\mathbf{x}$ the projection of $\mathbf{p}$ according to the current camera. The Gaussian curvature for the RGB surface is specified as:

$$\sigma_L = \eta_1\eta_2 \tag{16}$$

Here η_1 and η_2 are obtained as κ_1 and κ_2 considering the RGB surface. Therefore also for the intensity surface we have considered the principal curvatures. Both σ_S and σ_L are invariant to rotation.

The last feature that turned out to be important is the task domain, namely the range of the values $\mathbf{p}$ corresponding to PORs. Their importance, as gathered above, is quite intuitive, since we do not search in general an item in the sky unless we know in advance that it can challenge gravity. Clearly the constraints on the range can be given only on S. We define $\mathscr{R}_\tau$ to be the plausibility interval $((X_{min}, X_{max}), (Y_{min}, Y_{max}), (Z_{min}, Z_{max}))$ for a search task τ.

We can now list the features we have inferred. For the scene structure:

- $\mathscr{F}_1$: the surfaces points on S_i, given in global coordinates, whose center $\mathbf{0}$ is the search task starting point; the surfaces are matrices $n \times 3$;
- $\mathscr{F}_2$: σ_S for each patch corresponding to nodal points $\mathbf{p}$ on the surface;
- $\mathscr{F}_3$: the plausible interval $\mathscr{R}_\tau$ on the surface domain;
- $\mathscr{F}_4$: the timestamp.

For the image structure, for each point $\mathbf{x}$, image of $\mathbf{p}$ in frame $\mathscr{I}$, the features are defined as follows:

- $\mathscr{F}_7$: the contrast sensitivity function (see Watson & Ahumada (2005)).
- $\mathscr{F}_6$: σ_L for each image patch;
- $\mathscr{F}_5$: an image patch, centered at $\mathbf{x}$ and having size consistent with a meaningful distance Z of the projected point $\mathbf{p}$. Namely we fix the maximum depth to 3m. and the acute vision angle to about 15 degrees.

This concludes the set of feature operators. We consider a feature point $\mathbf{W} = \{\mathscr{F}\} \cdot \{(\mathbf{p}, (\mathbf{x}_1, \ldots, \mathbf{x}_m))\}$. Following the approach of Schölkopf, Platt, Shawe-Taylor, Smola, and Williamson (2001), we map this set into the vector space defined by a kernel function and set a maximum margin classification problem to separate the data from the origin. Let $\Phi : \mathbb{D}^n \to \mathscr{V}_k$ represent a mapping to the vector space $\mathscr{V}_k$ corresponding to the kernel function $\mathscr{K}$. The separating hyperplane in $\mathscr{V}_k$ space is computed by solving the quadratic program

$$\min_{w \in \mathscr{V}_k, \xi \in \mathbb{R}^+, \rho \in \mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{\upsilon l}\sum_i \xi_i - \rho \tag{17}$$

$$\text{s.t. } (w\Phi(\mathbf{W})) \geqslant \rho - \xi_i, \quad \xi_i \geqslant 0, \tag{18}$$

Here ξ_i are slack variables, while υ is a regularization parameter controlling the trade-off between the goals of maximizing the width of the margin and minimizing the training error at the points $\{\mathscr{F}\} \cdot (\mathbf{p}, (\mathbf{x}_1, \ldots, \mathbf{x}_m))\}$, which takes value 1. So for a new point $\mathbf{W}$ the side of the hyperplane it falls on in $\mathscr{V}_k$ can be determined by evaluating

$$f(\mathbf{W}) = \text{sgn}((w\ \Phi(\mathbf{W})) - \rho). \tag{19}$$

The learned function, in principle, separates salient regions from nonsalient ones. More precisely, given a set of corresponding points $\{(\mathbf{p}, (\mathbf{x}_1, \ldots, \mathbf{x}_m))\}$, according to some cameras $\mathbf{P}_1, \ldots \mathbf{P}_m$ mapping $\hat{\mathbf{p}}$ into a point $\hat{\mathbf{x}}$ in different scene images of the same bundle; given that $(X, Y, S(X, Y))^\top$ is the point on the surface corresponding to $\mathbf{X}(\lambda)$, and given the feature transformations set $\mathscr{F}$, then $f(\mathscr{F} \cdot (\mathbf{p}, (\mathbf{x}_1, \ldots, \mathbf{x}_m))) = 1$ if this is a point in a possible salient region and -1 otherwise.

Results on the classification performed on the above devised feature set are illustrated in the section on *Experimental validation*. We can note that for a 50 s search experiment we collect about 1500 frames, since each image has dimension 480×640, then we have a number of points of the order of $10^{8.5}$. On the other hand as at most 7 PORs are gathered in a single frame and for each POR we collect a surface of about 31×31 pixels then we have positive examples of the order of 10^7, since PORs are often in the same region. Therefore we have rather sparse matrices. The outcome of these experiments is to validate the feature set across different search tasks and to understand what is missing, what is actually part of a prior ability of the searcher and cannot be recovered from the data.

4. Generating proto-objects

In the previous sections we have illustrated a model for head and point of regard localization in space for a gaze machine that can be worn by a subject looking for specific objects in the environment. Using the model we have identified several features, among which we sorted out the most relevant ones for learning a function that can separate the attended regions from the unattended ones, given a specific search task. Note that the function needs to be learned for each task, to cope with the PORs elicited during the specific visual search experiment, though the set of features remain fixed: it is like a continuous recalibration process.

This lack of generalization is to be expected, human visual-search relies on an inner model able to generalize search abstracting from the context and the specific task. We argued in the introduction that this might be a consequence of the way features are aggregated into a coherent structure, that is, a proto-object.

If the unknown function to be learned has to be one generalizing all the learned functions for all the search tasks, then it should be a function minimizing a distance from all the learned functions, for all the experimented tasks. This function u should be one minimizing the following functional:

$$E(u) = \int_{\mathscr{L}} \int_{\Omega} w(\mathbf{X}) \| u_X(W) - f(W) \|^2 d\mathbf{X}\, df \tag{20}$$

Here f is any function learned for the task of visual search, with $\mathscr{L}$ its domain, w is a weight given to the features selected within classification, and $\mathbf{X}$ the observations. In other words, given a search task, the observations, the models specified by the features and the learned function space, $E(u)$ returns the function u which is as close as possible to the value of any possible function selected by the learning process, where the distance is weighted by the features

Here, however, rather than deriving the function u we propose a forward model, based on the previously selected features, which generalizes the learning results. The model is based on wave motion, more specifically it is governed by the equations of a vibrating membrane, with the membranes distributed on the surface S and having an initial displacement induced by the selected features at the specific location.

The main idea of the model is to mimic the stimulus activation, during search, by integrating the features into a vibrational energy. Indeed, due to the initial displacement, the vibration model returns a vibrational energy that is higher where proto-object are expected to be generated and lower or null elsewhere.

In the following, after recalling the model of the finite circular membrane we show how its motion is determined by its initial displacement, induced by the features integration strength. Note that here we do not consider possible interferences between two or more membranes. This will be considered in future works. In Fig. 9 we illustrate the underlying structure of the proposed model.

The general equation for a vibrating circular membrane, occupying a finite region, is the following:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} \right) \quad 0 \leqslant r < a, \theta \leqslant 2\pi, t > 0 \tag{21}$$

This admits a solution by separating variables, and using the positive roots of the Bessel functions of first and second kind. In particular, if the membrane is finite, as in our case, the Bessel functions of the second kind, of any order, are excluded from the solution. Indeed, the general solution of (21), for a membrane that is held fixed at the boundary, $r = a$, and it is finite, is obtained using the Bessel function of the first kind of any order as follows:

$$u(r,\theta,t) = \sum_{m=0,1,,n=1,2,}^{\infty} \sum^{\infty} \{\alpha_{mn} \sin(j_{mn} t) + \beta_{mn} \times \cos(j_{mn} t)\} \{\alpha^{\star}_{mn} \sin(m\theta) + \beta^{\star}_{mn} \cos(m\theta)\} J_m(j_{mn} r) \tag{22}$$

Here J_m is the Bessel function of the first kind of order m, j_{mn} is the nth root of J_m and α, β, $\alpha^{\star}$ and $\beta^{\star}$ are constants that can be determined by the initial conditions of the membrane. We recall that the Bessel functions are the solutions of the second order differential equation

$$z^2 \frac{d^2 y}{dz} + z\frac{dy}{dz} + (z^2 - m^2) y = 0 \tag{23}$$

With two classes of solution, the J_m of the first kind and the Y_m of the second kind. Though, as observed above, here the Bessel of the second kind is disregarded.

The interest of the membrane is in its vibration modes, they provide a plausible model for integrating features and, accordingly, they release energy via their displacement, and because of the Bessel function the energy vanishes in time.

The main aspect of the model is to provide the right initial displacement so that a solution is found in closed form, for up to a certain order, and the energy induced pulls attention or it fades away, as suggested in the coherence theory.

Let (r, θ, Z) be the cylindrical coordinates of a nodal point $\mathbf{X}$ on the surface. Let $\mathbf{c}$ be the contrast sensitivity, and let $\sigma = \sigma_S + \sigma_L + \epsilon$ be the surface variations introduced in the previous section (see Eqs. (14), (16)). We assume that the initial velocity is zero, namely $\partial u/\partial t|_{t=0} = 0$ therefore the general solution becomes:

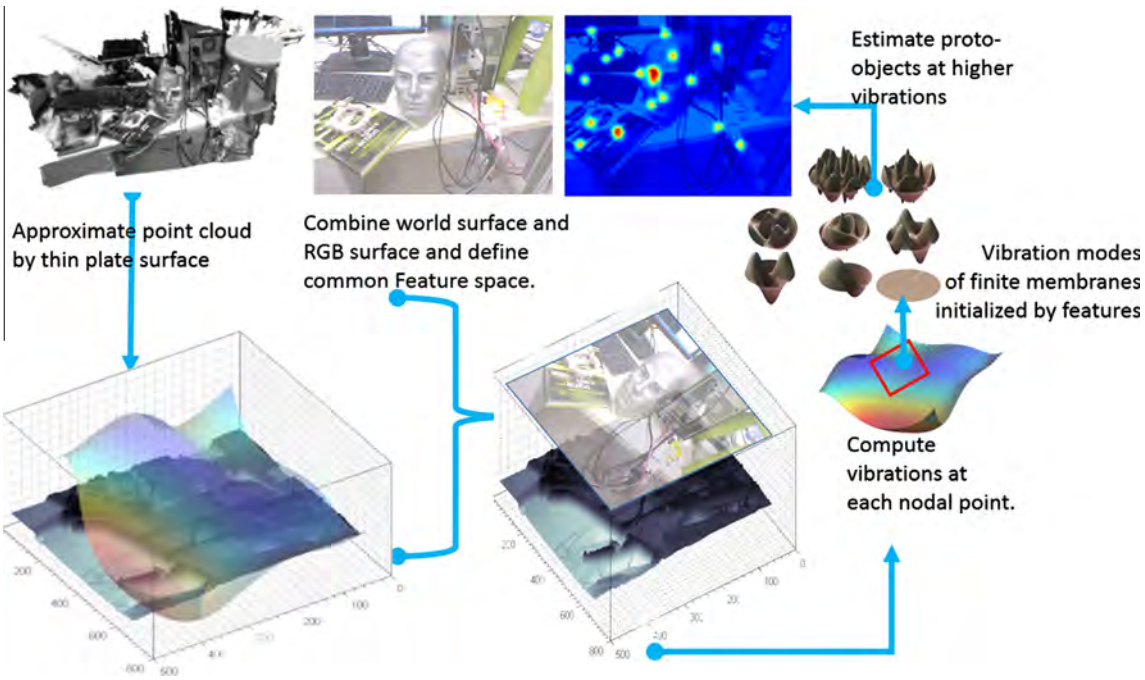

Fig. 9 The figure above illustrates the model for generating proto-object based on wave motion. The model generates vibration at nodal points where, according to the integrated features a stimulus should occur.

$$u(r,\theta,t) = c\sum_{m=0,1}^{\infty}\left(\left[\sum_{n=1,2,}^{\infty}\alpha_{mn}J_m(j_{mn}r)\right]\sin(m\theta) + \left[\sum_{n=1,2,}^{\infty}\beta_{mn}J_m(j_{mn}r)\right]\cos(m\theta)\right)\cos(cj_{mn}t) \tag{24}$$

Using the initial condition $\gamma(r,\theta,0)$, we can separate the inner summations of the above Eq. (24), for $t=0$ as follows:

$$\begin{aligned} C_m &= \sum_{n=1,2,}^{\infty}\alpha_{mn}J_m(j_{mn}r) \\ D_m &= \sum_{n=1,2,}^{\infty}\beta_{mn}J_m(j_{mn}r) \end{aligned} \tag{25}$$

and by Fourier series obtain:

$$\begin{aligned} C_m &= \begin{cases} \frac{1}{\pi}\int_0^{2\pi}\gamma(r,\theta,0)\cos(m\theta)d\theta, & \text{for } m \geqslant 1 \\ \frac{1}{2\pi}\int_0^{2\pi}\gamma(r,\theta,0)d\theta, & \text{for } m=0 \end{cases} \text{ and } D_m \\ &= \frac{1}{\pi}\int_0^{2\pi}\gamma(r,\theta,0)\sin(m\theta)d\theta, m \geqslant 1 \end{aligned} \tag{26}$$

Now, we let the initial displacement be given by the following equation:

$$\gamma(r,\theta,0) = \left(\frac{1}{2}\right)r\sigma\exp\left(\frac{-z^2}{2\sigma^2}\right)\sin\left(\frac{1}{z}\theta\right) \tag{27}$$

This initial displacement ensures that where the surfaces variations σ increase the energy increases too, while the frequency at which the energy is released depends on the radius and the θ values, in such a way that distant points, namely for increasing values of Z, on the surface are penalized. Using Eq. (26) we obtain:

$$\begin{aligned} C_m &= \frac{4zr\sigma\exp\left(\frac{-z^2}{2\sigma^2}\right)\left(-1+\cos(2m\pi)\cos(\frac{2\pi}{z})+mz\sin(2m\pi)\sin(\frac{2\pi}{z})\right)}{\pi(m^2z^2-1)}, \ m>0 \\ C_0 &= \frac{8zr\sigma\exp\left(\frac{-z^2}{2\sigma^2}\right)\sin(\frac{\pi}{z})^2}{\pi} \end{aligned} \tag{28}$$

and

$$D_m = \frac{4zr\sigma\exp\left(\frac{-z^2}{2\sigma^2}\right)\left(\cos(\frac{2\pi}{z})\sin(2m\pi) - mz\cos(2m\pi)\sin(\frac{2\pi}{z})\right)}{\pi(m^2z^2-1)}, \quad m \geqslant 1 \tag{29}$$

Finally the coefficients α_{mn} and β_{mn} are obtained as follows:

$$\begin{aligned} \alpha_{mn} &= \frac{2}{\pi a^2 J_{m+1}(j_{mn}a)^2}\int_0^a rJ_m(j_{mn}r)C_m \\ &= \frac{2^{2-m}\sigma z\Gamma(\frac{m+3}{2})\exp\left(-\frac{z^2}{2\sigma^2}\right)(j_{m,n})^m\left(mz\sin(2\pi m)\sin(\frac{2\pi}{z})+\cos(\frac{2\pi}{z})-1\right)K}{\pi(m^2z^2-1)J_{m+1}(j_{m,n})^2} \end{aligned} \tag{30}$$

Here Γ is the Gamma function, $K = {}_1\widetilde{F}_2\left(\frac{m+3}{2};\frac{m+5}{2},m+1;-\frac{1}{4}(j_{m,n})^2\right)$, where ${}_p\,\mathrm{F}q(a;b;z)$ is the regularized generalized hypergeometric function. And the second parameter β_{mn} is given below:

$$\begin{aligned} \beta_{mn} &= \frac{2}{\pi a^2 J_{m+1}(j_{mn}a)^2}\int_0^a rJ_m(j_{mn}r)D_m \\ &= \frac{2^{2-m}\sigma z\Gamma(\frac{m+3}{2})\exp\left(-\frac{z^2}{2\sigma^2}\right)(j_{m,n})^m\left(mz\cos(2\pi m)\sin(\frac{2\pi}{z})-\sin(2\pi m)\cos(\frac{2\pi}{z})\right)K}{\pi(m^2z^2-1)J_{m+1}(j_{m,n})^2} \end{aligned} \tag{31}$$

Analogously, here Γ is the Gamma function, $K = {}_1\widetilde{F}_2\left(\frac{m+3}{2};\frac{m+5}{2},m+1;-\frac{1}{4}(j_{m,n})^2\right)$, where ${}_p\,\mathrm{F}q(a;b;z)$ is the regularized generalized hypergeometric function. Noting that the roots

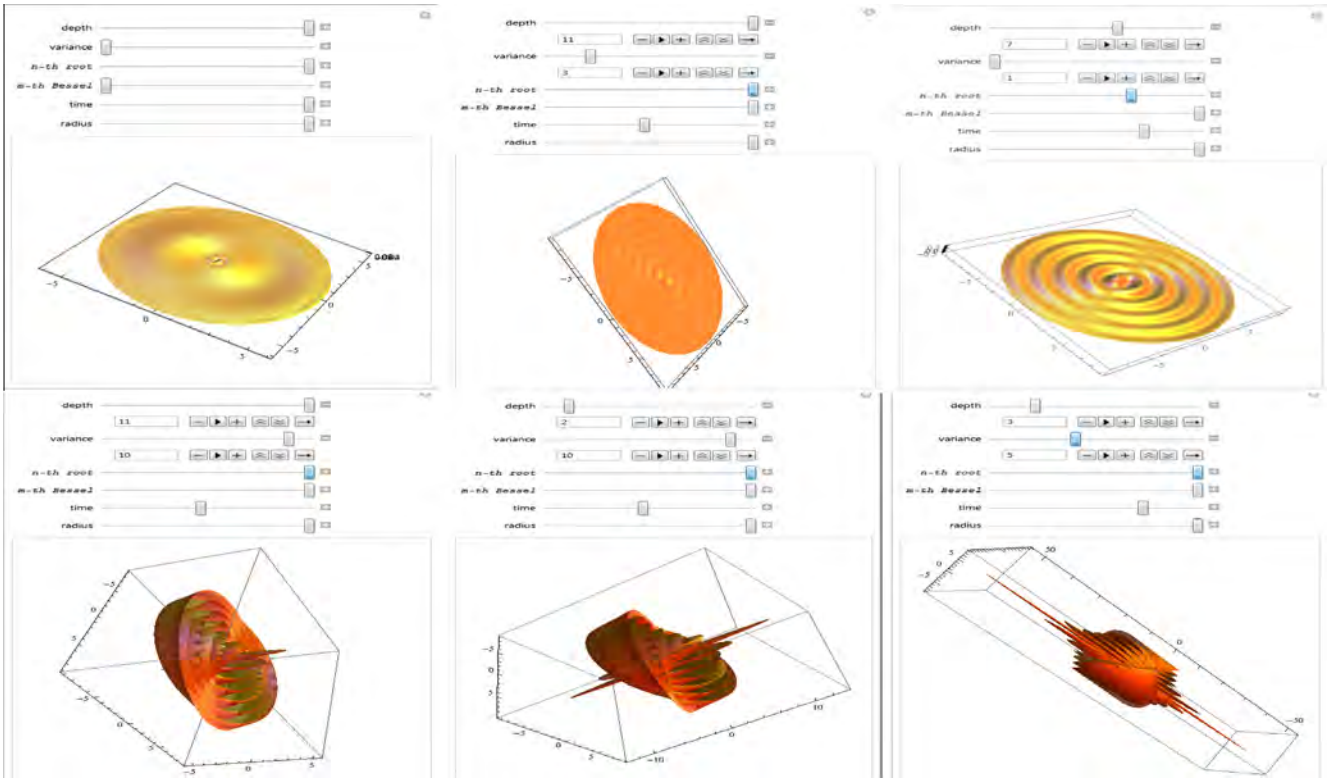

Fig. 10 Vibrations generated by different initial displacements, according to the initial feature values. The interface made in Mathematica, allows to understand the influence of the Gaussian Curvature σ_S and σ_L, for *S* and *L*, specified in the GUI as *variance*, and the distance *Z*, on the vibration frequency.

of the Bessel J_m are easily computed with Mathematica, Matlab or Maple, it follows that up to a given order and to a given root, the vibrating membrane takes a solution for varying features values in closed form. Some of the computed membranes with vibrations varying according to the features, inducing the initial displacement $\gamma(r,\theta,0)$ are illustrated in Fig. 10 showing some of the vibration modes.

The full algorithm to compute the energy elicited by the features structured by the vibrating membrane and to generate proto-object is as follows.

First of all let us define $\mathbb{D} = \bigcup S_{\setminus\mathscr{R}}$ be the domain of all the experiments, in terms of the plausible regions $\mathscr{R}$. Let *Q* be a coherent subsequence of frames, and $\{\widehat{Z}\}_{i=1,\ldots,n}$ the point cloud for *Q*, note that a coherent subsequence includes no more than 15 frames, hence it is labeled by a time interval $(t_0, t_0 + \Delta t)$ of less than half second. Let $\mathbf{K}[\mathbf{I}|\,\mathbf{0}]$ be the reference camera and $\mathbf{R}|\,\mathbf{t}]_1, \ldots,$ $[\mathbf{R}|\,\mathbf{t}]_m$ the poses of the other views with respect to the reference one.

1. For each nodal point $\mathbf{p}$ of *S*, such that $\mathbf{p} \in \mathbb{D}$, and for each projected pixel, according to the camera poses, select the regions generated by the points $(\mathbf{p}, (\mathbf{x}_1, \ldots, \mathbf{x}_m))$ restricted to the domain $\mathbb{D}$.
2. Compute the feature set *W* for the sampled set.
3. Using the above equations, and the obtained features $\mathbf{W}$ at each nodal point, compute the vibrating membrane, allowing the radius *r* to vary about the membrane distribution on *S*, between 1 and 5. Here we exploit the pre computation of the Bessel roots in a lookup table.
4. Compute Eq. (24) for each $0 \leqslant m \leqslant 12$ and for $1 \leqslant n \leqslant 9$. Define the membrane surface as:

$$(rm\cos(\theta), rm\sin(\theta), u(r,\theta,t)) \tag{32}$$

with *t* varying from zero to the maximum time lapse of the subsequence interval. Some examples with varying σ, z, and *r* are illustrated in Fig. 10. Sum the membrane surface absolute values for each time $t \in (t_0, t_0 + \Delta t)$ and using gradient descent, find the membranes that have maximal energy at $t_0 + \Delta t$.
5. The nodal points with maximal energy are generators of proto-objects.
6. Consider the energy of all the neighbor these selected nodal points, according to the maximal radius *a*, and identify these patches in *S* and their projection on the retinal planes of the subsequence as the proto-objects predicting saliency.

Results of this algorithm, for the indoor experiments *looking for J* and *looking for the pink elephant* are illustrated in Fig. 11.

5. Experimental validation

Experiments are at the basis of our experimental model of saliency, whose main stages are shown in the left panel of Fig. 12.

An experiment, begins with a calibration phase, in which the subject moves her/his eyes, head and body while

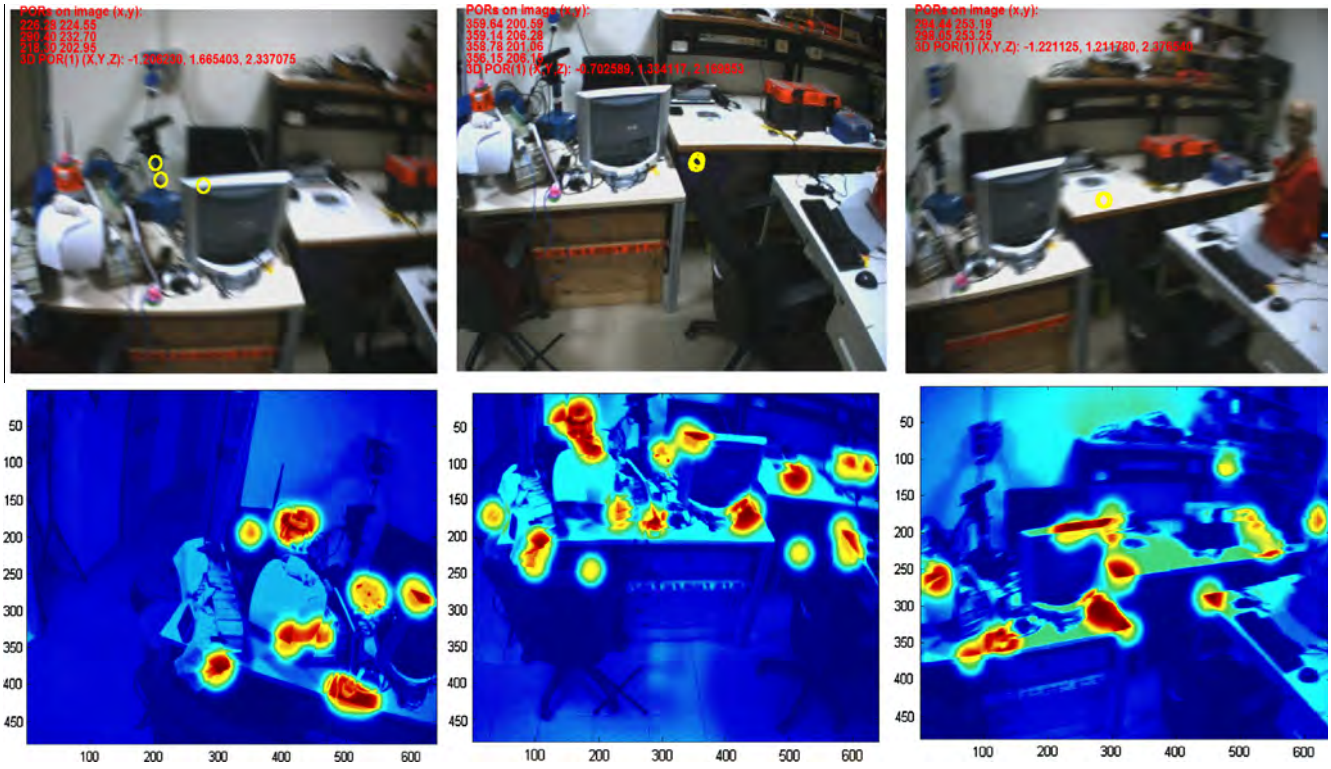

Fig. 11 Comparison between PORs taken from a coherent subsequence and the inferred proto-objects. We can see that in the whole the generated proto-objects are plausible.

fixating a specified target. This phase is needed to calibrate the wearable device with the subject eye motion manifold and scene cameras, as illustrated in Pirri et al. (2011). Thereafter, according to the search task, the search experiment lasts a certain amount of time T, $120\,\text{s} \leqslant T \leqslant 180\,\text{s}$ and it collects the frame sequence F, of the left and right images, at a frequency of $f_T \in [15, 30]$ Hz; frames are gathered in bundles specifying the local coherence of the gaze motion. Further it collects the pupil sequence P at a frequency $f_t \in [120, 180]$ Hz and the head motion H via a compact inertial device part of the acquisition device. Data are processed off-line and the following set of data is returned together with a synchronization of images, visual axes and head poses: the head pose in global coordinates $\mathscr{H}$ via the localization, Pizzoli et al. (2011), the point cloud $\mathscr{M}$ in global coordinates, the visual axes of the eye manifolds, namely the PORs directions, projected as point in the global coordinates of the scene $\mathscr{P}$, the reprojection of the PORs in the images $\mathscr{R}_{POR}$, synchronized, so that in each image a certain amount of PORs, between 7 up to 15 is reprojected. Finally, $\mathscr{B}$ are the relative positions of the observer with respect to the scene.

An experiment, therefore, comes with the following formal structure:

$$E = \langle \mathscr{H}, \mathscr{M}, (\mathscr{B}, \Delta T), (\mathscr{P}, \Delta t), \mathscr{R}_{POR} \rangle \tag{33}$$

Here ΔT is the time lapse between two measurements of the scene, $\Delta T \approx 60$ ms; Δt is the time lapse between two measurements of the PORs direction in the scene, $\Delta t \approx 8$ ms exploiting the scene constancy — namely, the speed of the eyes is faster than any meaningful motion in the scene and of the head and body motion. To these data we add the membrane structures to support the proto-objects. The principal outcomes of an experiment E are the PORs and their localization in the 3D space together with the localization of the head pose in the dense map reconstruction of the scene. These are illustrated in Figs. 3, 7 showing the dense map, the path of the head poses, together with PORs as located in the natural scenes, and in Fig. 5, showing a meaningful part of an experiment, via a stitched panorama, with the PORs reprojected on the images. A typical dataset with the tracked head poses, a dense point cloud with the projected PORs is illustrated in Fig. 13.

5.1. Experimental validation of the acquisition model

Investigating the accuracy of the proposed acquisition model involves different aspects. Localization and mapping of the POR in the 3D scene rely on the estimation of the POR relative position and the localization of the subject in the reference frame of the experiment. In addition, the identification of coherent regions depends on the effectiveness of the keyframe-based mechanism to detect changes in the POR sequence.

A first evaluation focuses on investigating the accuracy of the proposed method in localizing and mapping the PORs. The ground truth has been produced as follows: five visual landmarks have been placed in the experimental scenario and their position has been measured with respect to a fixed reference frame; six subjects have been instructed to fixate the visual landmarks while freely moving in the scenario,

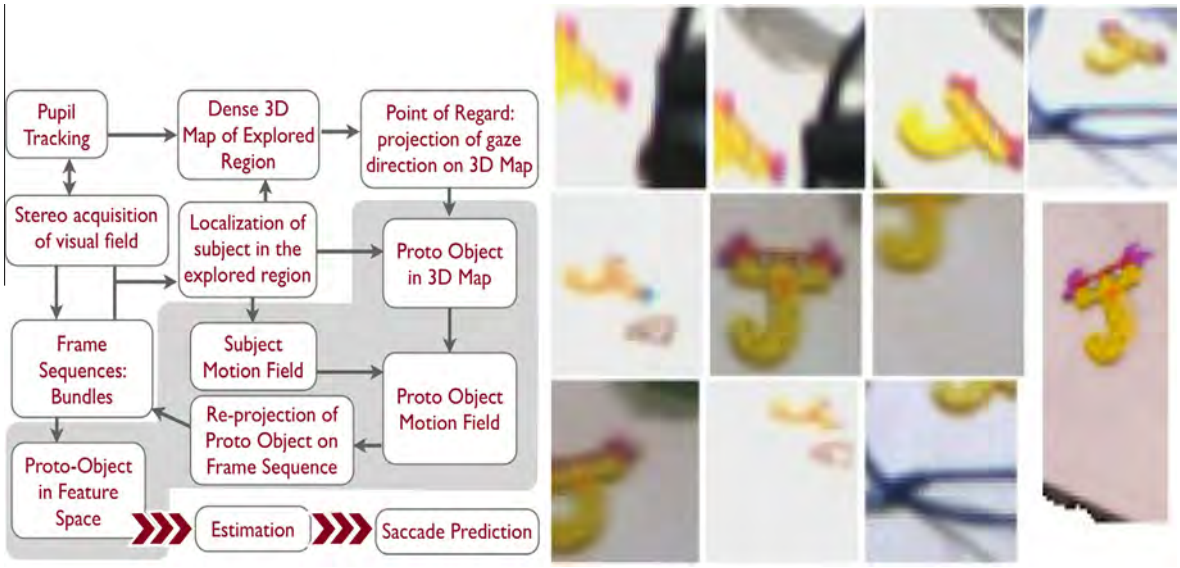

Fig. 12 The left panel shows the stages of saliency prediction according to our *experimental saliency model*. We use the term *experimental* as it is based on 3D measurements of the gaze in natural scenes and of its motion field. The model copes with the coherence theory of attention with respect to the interpretation of Proto-Objects in early attention stages. On the right the backprojection of proto-objects during the task *looking for* **J**, the last image in the right panel is a proto-object in the 3D dense map.

annotating (by voice) the starting and ending of the landmark observations. In each sequence, an average of 60 PORs were produced for each landmark. The validation sequences comprise about 6000 frames each. After registration of the subject initial pose with the fixed reference system, the PORs in the annotated frames were computed and compared with the ground truth, producing a Root Mean Square (RMS) value of 0.094 m.

For a quantitative analysis of the keyframe selection strategy we relied on a manual coding to produce ground truth data: after the acquisition, subjects were shown the scene sequence overlapped with the POR projection on the image plane and used their innate human pattern recognition skill to select coherent subsequences, annotating for each one the starting keyframe. The performance measure is the *agreement*, defined as the ratio between the number of keyframes recognized by the system over the number of keyframes identified by the subject. Experiments on sequences characterized by a number of frames in the range 4000—6000, yielding a number of keyframes in the range 120—200 produced an average agreement of 85%.

Validation of the coherent subsequence Coherent regions constitute the support for the attended proto-objects during an experiment. Each coherent region also selects, in the related sequence of frames, the appearance of the attended structure that is used to train the saliency model. To validate the method introduced in Section *Coherent features for point saliency*, we quantified the extent of the coherent region projections in each of the related bundle images. The result for an experiment producing 16 regions, with centroid distances ranging from 1.8 and 8 m from the observer, is shown in Fig. 14. For each region, the extent of its projection to the frames of the sequence is evaluated as percentage of the total number of pixels in a frame. Scene frames have size 640×480 pixels in the experiments. Fig. 14 shows the median values, the boxes representing the 25th and 75th percentiles, the minimum and maximum values. The validation confirms that the extent of the projections is mostly confined between 1% and 10% of the image area, and is thus suitable for the proposed feature model.

5.2. Validation of the features model

Given a visual search task, we have implemented both a slight varied version of Mangasarian and Wild (2007) and the easier selection addressed in Guyon and Elisseeff (2003). Focusing on sets of features we obtain the *balanced error rate* as follows:

$$ber = \frac{1}{2}\left(\frac{wp_+}{|D|_+} + \frac{wp_-}{|D|_-}\right). \tag{34}$$

Here $|D|_+$ are the positive instances and $|D|_-$ are the negative ones, while wp_+ and wp_- are, respectively, the false negatives and false positives. In the case of the approach of Mangasarian and Wild (2007), to keep trace of the decrease of the objective function on feature groups, we generate $k!/(k-m)!m!$ m-tuples of even features, up to $k = 5$, so as to assign a *ber* value to each feature group.

A model trained on the complete set of features selected as described in Section *Coherent features for point saliency*, is able to predict if a new sample point is likely to be attended, i.e., if it belongs to a coherent region, when the experiment is fixed. To validate this assumption, we ran maximum margin classification experiments. A *K-fold cross-validation* strategy has been followed: we divided the available data comprising more than 6 million points in 3 subsets; in turn, 2 of the three subsets have been used to train the classifier and the remaining one for validation.

The process is iterated until every subset is used for validation. As expected, classification accuracy is very high, as reported in Table 1.

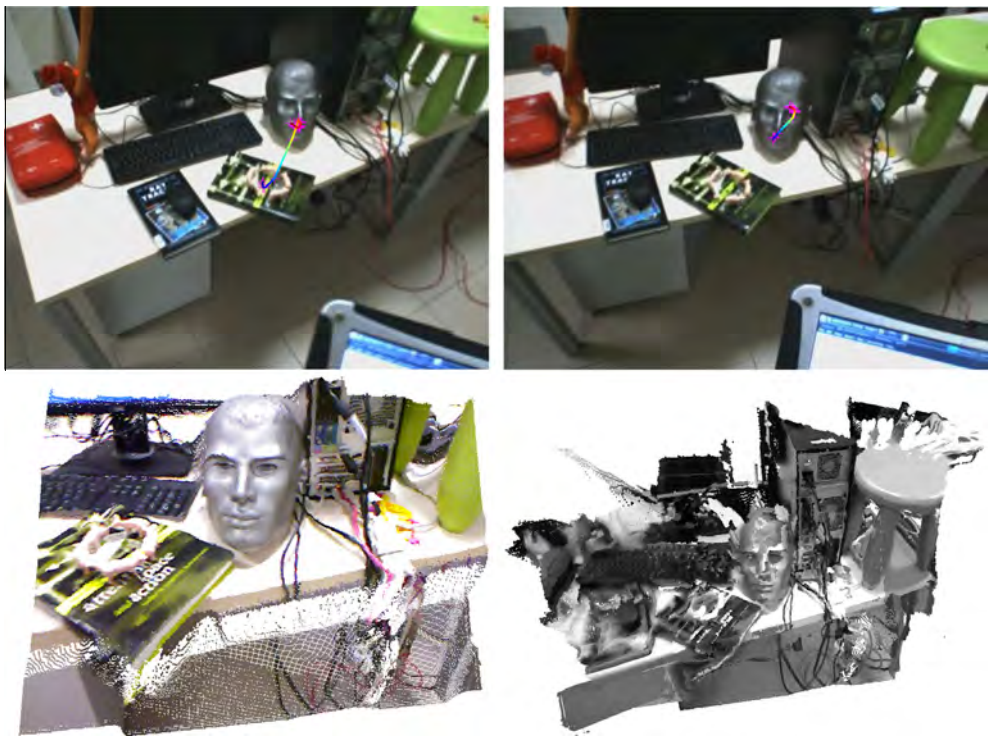

Fig. 13 Dataset *E* of a typical visual search experiment with the GM device; the dataset includes: point cloud, head scan-path, projection of PORs in space and on the retinal planes.

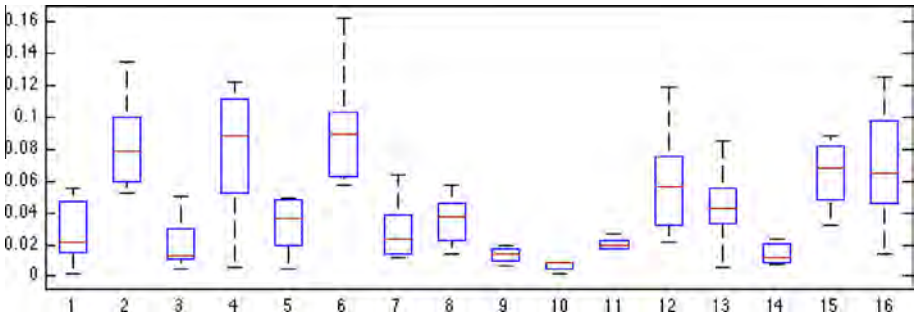

Fig. 14 Box plot for the extent of 16 coherent regions identified in a GM experiment on the street. The extent of the coherent regions is in percentage with respect to the frame dimension in pixels.

Table 1 Results from the *k*-fold cross validation of the maximum margin classification using the complete image+bundle feature set. Here wp^{+} *and* wp^{-} *are, respectively, the false negatives and false positives.*

Iteration	Number of positives	$wp_{+}/\lvert D\rvert_{+}$	$wp_{-}/\lvert D\rvert_{-}$	Accuracy (%)
1	44,707	0.0127	0.0318	95.334
2	46,881	0.01883	0.0206	93.591
3	420,034	0.0093	0.0157	93.019

The accuracy is, in particular illustrated in the tables of Fig. 15, where the outcome of the classification and the measured PORs is highlighted, the first in green and the second in red.

5.3. Validation of the vibration model

To validate the vibration model we have tested the algorithm described in Section *Generating Proto Objects*. The

Fig. 15 Results of features and classification validation for the outdoor experiment *looking for car fines*. In red the PORs, and the coherent patches, in green the estimated point saliency, for the specific task.

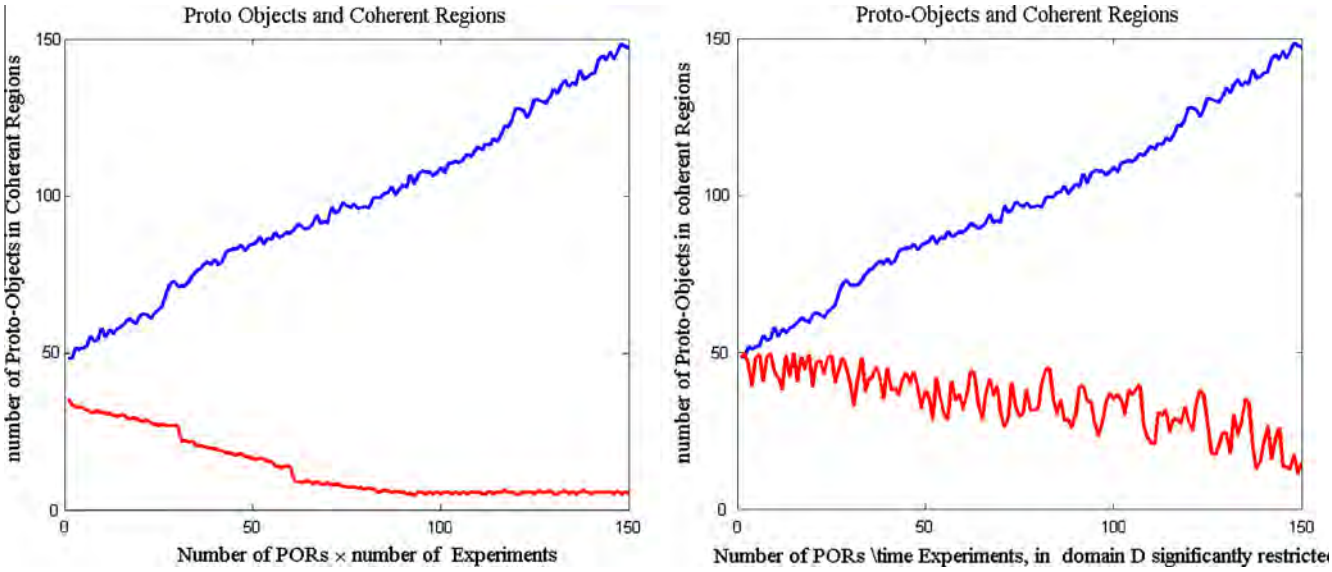

Fig. 16 Results for computed POR as functions of energy vibration at time $t_0 + \Delta t$, given the domain of the specified experiments, and given the limited domain of selected experiments.

implementation of the membrane has been done in both Mathematica, where a GUI is implemented to study the variations according to the initial displacement conditions, see Fig. 10, and in Matlab, exploiting a look up table of the Bessel roots computed in Mathematica. We used also the implementation of *gridfit* by D'Errico (2013) for surface approximation. After classification, we have collected the domain elicited by the learned function. And we have generated two sets. The first set with free domains, namely the range of the **p** values was given by the domains of all experiments. In the second set we have limited the range to similar domains. The results are illustrated in Fig. 16. Here the number of PORs per experiments, indicates the $\mathbf{p}^\star$ collected by the GM, with varying experiments, both indoor and outdoor. The number of proto-objects in coherent regions indicates the regions of maximal energy at $t_0 + \Delta t$, computed at the time steps given for the end of a coherent subsequence.

6. Conclusions

The computational theory of visual attention aims at mimicking the human capability to select, among stimuli acquired in parallel, those that are relevant for the task at hand. Similar to the biological counterpart, artificial systems can accomplish this by orienting the vision sensors toward regions of space that are more promising. 3D saliency prediction resides in defining a quantitative measure of how attention should be deployed in the three-dimensional scene. Current state-of the art does not model the integration of features in space and time, which is required when dealing with a three-dimensional, dynamic scene. In the coherence theory of attention, as introduced in Rensink et al. (2000), the concept of proto-object emerged to explain how focused attention collects features to form a stable object that is temporally and spatially coherent. In this work we address the problem of modeling the process of formation of proto-objects and their relative spatial and temporal coherence according to a double process. At first a pure experimental setting allows as to identify the best features, which are stable across different experiments and different contexts. We show their stability using a classifier that has been exploited also to select the best features. Further we define a forward model based on the selected features. The forward model define a vibrational

energy capturing coherent proto-objects. These encapsulate the information about the search task and we show that some good approximation results are possible. We have thus shown a whole process which, starting from three-dimensional gaze tracking experiments, extract features that are relevant to predict saliency and introduce a novel energy based model to indicates the salient regions in space.

A drawback of the proposed method is the lack of motion features. We intend to address these aspect in future research, note that for an experimental method as the one proposed here it is required to deal with the reconstruction of motion, which is still a hard problem.

Acknowledgments

The research has bee funded by EU-FP7 NIFTI Project, Contract No. 247870.

References

Ackerman, C., & Itti, L. (2005). Robot steering with spectral image information. *IEEE Transactions on Robotics, 21*(2), 247—251.

Bahill, T., Bahill, K. A., Clark, M., & Stark, L. (1975). Closely spaced saccades. *Investigative Ophtalmology, 14*(4), 317—321.

Bahill, T., & Stark, L. (1979). The trajectories of saccadic eye movements. *Scientific American, 240*(1), 1—12.

Belardinelli, A., Pirri, F., & Carbone, A. (2007). Bottom-up gaze shifts and fixations learning by imitation. *IEEE Transactions Systems, Man and Cybernetics B, 37*, 256—271.

Butko, N.J., Zhang, L., Cottrell, G.W., & Movellan, J.R. (2008). Visual saliency model for robot cameras. In *ICRA* (pp. 2398—2403).

Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46*(26), 4333—4345.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. In *NIPS*.

Cristianini, N., & Shawe-Taylor, J. (2004). Kernel methods for pattern analysis. CUP.

D'Errico, J. (2013). Surface fitting using gridfit. Tech. rep., Matlab File Exchange. <http://www.mathworks.com/matlabcentral/fileexchange/authors/679>.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433—458.

Faugeras, O., Luong, Q., & Papadopoulou, T. (2001). *The geometry of multiple images*. MIT Press.

Fiore, P. (2002). Efficient linear solution of exterior orientation. *IEEE PAMI, 23*(2), 140—148.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*(6), 381—395.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *JMLR, 3*(7-8), 1157—1182.

Hartley, R., & Zisserman, A. (2000). Multiple view geometry in computer vision. CUP.

Hegland, M., Roberts, S., & Altas, I. (1997). *Finite element thin plate splines for surface fitting*. Tech. Rep. TR-CS-97-20, Department of Computer Science, Faculty of Engineering and Information Technology, The Australian National University Canberra, ACT 0200.

Hügli, H., Jost, T., & Ouerhani, N. (2005). Model performance for visual attention in real 3d color scenes. In *IWINAC-2005* (pp. 469—478).

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI, 20*(11), 1254—1259.

Julesz, B. (1986). Texton gradients: The texton theory revisited. *Biological Cybernetics, 54*, 245—251.

Koch, C., & Ullman, S. (1985). Shifts in selective visual-attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219—227.

Kowler, E. (2011). Eye movements: The past 25years. *Vision Research*, 1—27.

Lourakis, M., & Argyros, A. (2009). Sba: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software (TOMS), 36*(1), 2.

Mahadevan, V., & Vasconcelos, N. (2010). Spatiotemporal saliency in dynamic scenes. *IEEE PAMI, 32*, 171—177.

Mancas, M., Pirri, F., & Pizzoli, M. (2011). From saliency to eye gaze: Embodied visual selection for a pan-tilt-based robotic head. In *ISVC (1)* (pp. 135—146).

Mangasarian, O. L. (2005). *Exact 1-norm support vector machines via unconstrained convex differentiable minimization*. Tech. rep., Data Mining Institute TR 05-03.

Mangasarian, O. L., & Wild, E. W. (2007). Feature selection for nonlinear kernel support vector machines. In *IEEE-ICDM workshops* (pp. 231—236).

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 200*, 269—294.

Minato, T., & Asada, M. (2001). Image feature generation by visio-motor map learning towards selective attention. In *Proc. of IROS 2001* (pp. 1422—1427).

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology, 7*(4), 480—494.

Orabona, F., Metta, G., & Sandini, G. (2008). A proto-object based visual attention model. In L. Paletta & E. Rome (Eds.), *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint* (pp. 198—215). Berlin Heidelberg: Springer-Verlag.

Pichon, E., & Itti, L. (2002). Real-time high-performance attention focusing for outdoors mobile beobots. In *Proceedings of AAAI spring symposium (AAAI-TR-SS-02-04)* (p. 63).

Pirri, F., Pizzoli, M., Rudi, & A. (2011). A general method for the point of regard estimation in 3d space. In *In (CVPR)* (pp. 921—928).

Pizzoli, M., Rigato, D., Shabani, R., & Pirri, F. (2011). 3d saliency maps. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on* (pp. 9 — 14).

Rensink, R. (2000). The dynamic representation of scenes. *Visual Cognition, 7*, 17—42.

Rensink, R. (2002). Change detection. *Annual Review of Psychology, 53*, 245—277.

Rensink, R., O'Regan, J. K., & Clark, J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 368—373.

Rensink, R., O'Regan, J., & Clark, J. (2000). On the failure to detect changes in scenes across brief interruptions. *Visual Cognition, 7*, 127—145.

Sala, P. L., Sim, R., Shokoufandeh, A., & Dickinson, S. J. (2006). Landmark selection for vision-based navigation. *IEEE Transactions on Robotics, 22*(2), 334—349.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*, 1443—1471.

Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences, 10*(1), 38—45.

Smola, A., Bartlett, P., Schölkopf, B., & Schuurmans, D. (2000). *Advances in large margin classifiers*. Cambridge, MA: MIT Press.

Torr, P. H. S. (1998). Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society of London, Series A, 356*(1740), 1321—1340.

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision Graphics Image Processing, 31*(2), 156—177.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97—136.

Triggs, B., McLauchlan, P.F., Hartley, R., & Fitzgibbon, A. (2000). Bundle adjustment — A modern synthesis. In *ICCV*.

Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artifical Intelligence, 78*, 507—547.

Vapnik, V.N. (1995). The nature of statistical learning theory.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395—1407.

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision, 5*(9), 717—740.

Wolfe, J. M. (1992). The parallel guidance of visual attention. *Current Directions in Psychological Science, 4*, 124—128.

Wolfe, J. M. (1994). Guided search 2.0. a revised model of visual search. *CPsychonomic Bulletin and Review, 2*, 202—238.

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. L., & O'Connell, K. M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 34—49.

Zhou, W., Chen, X., & Enderle, J. (2009). An updated time-optimal 3rd-order linear saccadic eye plant model. *International Journal of Neural Systems, 19*(5).

Zhu, J., Rosset, S., Hastie, T., & Tibshirani, R. (2003). 1-norm support vector machines. *Neural Information Processing Systems*, 16.