

# From Saliency to Eye Gaze: Embodied Visual Selection for a Pan-Tilt-Based Robotic Head

Matei Mancas<sup>1</sup>, Fiora Pirri<sup>2</sup>, and Matia Pizzoli<sup>2</sup>

<sup>1</sup> University of Mons, Mons, Belgium

<sup>2</sup> Sapienza Università di Roma, Rome, Italy

**Abstract.** This paper introduces a model of gaze behavior suitable for robotic active vision. Built upon a saliency map taking into account motion saliency, the presented model estimates the dynamics of different eye movements, allowing to switch from fixational movements, to saccades and to smooth pursuit. We investigate the effect of the embodiment of attentive visual selection in a pan-tilt camera system. The constrained physical system is unable to follow the important fluctuations characterizing the maxima of a saliency map and a strategy is required to dynamically select what is worth attending and the behavior, fixation or target pursuing, to adopt. The main contributions of this work are a novel approach toward real time, motion-based saliency computation in video sequences, a dynamic model for gaze prediction from the saliency map, and the embodiment of the modeled dynamics to control active visual sensing.

## 1 Introduction

The question of determining where to look in front of a scene is the most relevant when designing vision architectures. Real sensors and actuators are characterized by physical limitations, constraining the field of view and motion capabilities. Moreover, computational power should be preserved to focus on crucial aspects for the task at hand. Computational visual attention [1–3] is thus emerging as an appealing component for embodied vision architectures.

To support active vision, control models of eye movements have been designed to compute the optimal sequence to move the visual sensors in order to achieve a given task [4], usually modeled as the problem of maximizing a certain measure of information or reward. In contrast, saliency models [5, 6] provide a biologically consistent prediction of eye movements, in the sense of the frequency with human beings attend regions in the observed scene. The past 25 years of investigations on eye movements have brought a more thorough understanding of how attention works with the oculomotor system in order to control the sensory data collection and extract the interesting information from visually rich environments [7]. Eye movements are explained by the need of keeping the interesting target in the fovea. Similarly, for a robotic system, many active vision problems take advantage of related capabilities: even for non-foveated cameras, tracking by an active vision system requires the target to be in the centre of the field of

view in order to minimize the chances to lose it or to extract stereo information; many robot platforms are endowed with omni-directional vision, coupled with active Pan-Tilt-Zoom or RGBD cameras, which resemble the human subdivision in peripheral and foveated vision.

The present work bridges the gap between control and saliency models, as it combines a motion-based saliency map with a model of eye movements, based on a non-linear Bayesian state-space process, in order to predict the focus of the attention accordingly. A motion-based saliency model provides us with a dynamic, near real-time saliency map. According to this, the next attended location and gaze behavior are selected and used to provide control to the active vision system, thus implementing an embodied visual selection mechanism.

The remainder of the paper is the following: Section 2 describes the computation of the motion-based saliency model; the dynamic prediction of gaze movement is addressed in Section 3, while Section 4 reports the analysis of the experimental results. Finally, in Section 5, the conclusions are drawn.

## 2 Motion-based saliency model

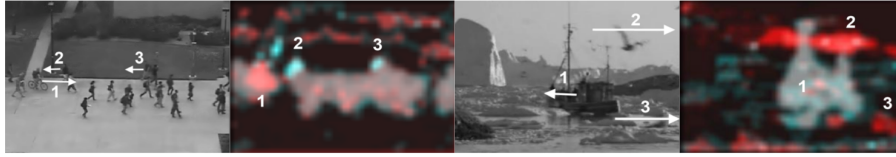
This section describes the main aspects of the motion-based saliency model, which ensures real-time rendition of salient-motion and features. The algorithm is based on three main steps: motion feature extraction, spatio-temporal filtering, and rare motion extraction. The resultant rarity pattern, although computed in an instantaneous context within the same frame, provides information of a short time history induced by spatio-temporal filtering.

As a first step, features are extracted from the video frames. For the purpose of modeling eye motion dynamics, speed and motion direction were computed. The method can be generalized to any other dynamic or static physical feature such as acceleration, rotational motion, color mean, color variance. The motion vector is computed making use of Farneback’s algorithm for optical flow [8], which is quite fast if compared to other techniques. The frame is divided in cells and features are extracted from non overlapping cells for speed. The chosen cell size is 3 or 5 pixels wide in order to take into account small motions. The features are then discretized into 4 directions (north, south, west, east) and 5 speeds (very slow, slow, mean, fast, very fast).

A spatio-temporal low-pass filter is designed to cope with the discretized feature channels, having 4 directions and 5 speeds. The filter first separates the space and time dimensions. Frames are first spatially low-pass filtered; then, a weighted sum is carried out on time dimension by using a loop and a multiplication factor  $0 \leq \beta < 1$ . So, given a feature channel  $F$ , the filtered value in position  $(i, j)$  at time  $t$  is given by:

$$\hat{F}(i, j, t) = \alpha \sum_{n=1}^N \beta^n \sum_{h=-\frac{m}{2}}^{\frac{m}{2}} \sum_{k=-\frac{m}{2}}^{\frac{m}{2}} A(h, k) F(i-h, j-k, t-n) \quad (1)$$

where  $A$  is an  $m \times m$  Gaussian smoothing kernel and  $\alpha$  a normalization term. This process will tend to provide lower weight to those frames entering the



**Fig. 1.** Annotated frames and corresponding saliency maps from the UCSD dataset [6]. The contribution of the speed feature to saliency is indicated by the red color, while cyan indicates directions.

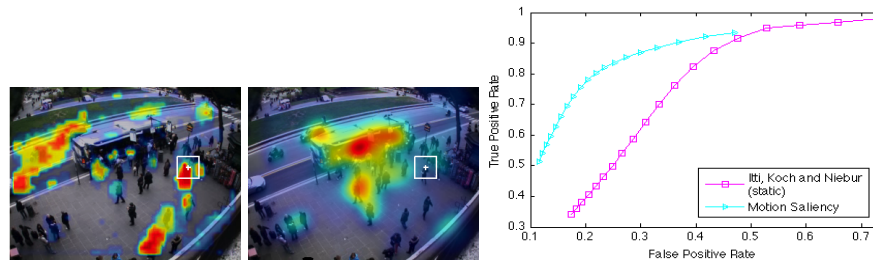
loop several times (the older ones) because of the  $\beta^n$  in (1) which decreases as the loop iteration  $n$  increases. Our approach only takes into account frames from the past (not from the future). This approximation of a 3D convolution provides increasing spatial filtering through iterations and it is suitable for on line processing. The neighborhood of the filtering is obtained by changing the size of the spatial kernel and by modifying the  $\beta$  parameter for the temporal part. If  $\beta$  is closer to 0, the weight applied to the temporal mean will decrease very fast, so the temporal neighborhood will be reduced, while a  $\beta$  closer to 1 will let the temporal dimension be larger. Filtering is implemented at two different scales using  $m \in \{3, 9\}$  and  $\beta \in \{0.9, 0.8\}$ .

After the filtering of each of the 9 feature channels (4 directions, 5 speeds), a histogram with 5 bins is computed for each resulting image and the self-information  $I(b_i)$  of the pixels for a given bin  $b_i$  is computed as

$$I(b_i) = -\log(H(b_i)/\|B\|). \quad (2)$$

Here  $H(b_i)$  is the value of the histogram  $H$  at the bin  $b_i$ , indicating the frequency counts of a video volume, resulting from the 3D low-pass filtering, within the frame;  $\|B\|$  is the cardinality of the frame, namely, the size of the frame in pixels.  $H(b_i)/\|B\|$  is simply the occurrence probability of a pixel of bin  $b_i$ . Self-information is a pixel saliency index. The matrices containing self-information, thus the saliency of the pixels at the two scales, one for each different 3D filters, are added. Then self-information is summed at the different scales. Once a saliency map is computed for each feature channel, a maximum operator is applied to gather the 4 directions into a single saliency map and the 5 speeds into a second saliency map. Rare motion is salient. The two final maps specify the rarity of the statistics of a given video volume at two different scales for a given feature. The two final conspicuity maps represent the amount of bottom-up attention due to speed and motion direction features. Figure 1 illustrates results on groups of moving objects with complex backgrounds. Further details can be found in [9].

To validate the predictiveness of the motion-based saliency model, we chose not to rely on any existing corpus of analyzed data and we collected evidence from gaze tracking experiments making use of a wearable device during natural observation tasks. The collected sequences of eye movements refer to the obser-



**Fig. 2.** Comparison of the *dynamic* saliency model (right) described in Section 2 against the *static* model (left) in [3] in terms of gaze prediction during a natural observation task, in the form of ROC analysis; The ground truth provided by gaze tracking is also shown.

vation of a bus stop (Figure 2). The complete dataset, comprising 20000 frames collected at 30 fps and labeled with the corresponding Point of Regard from gaze tracking, was divided into two sets, one used for deriving the dynamic model and one used for validation.

The static saliency model [3] has been compared to the described one, which is dynamic as it takes into account motion cues in the computation of the saliency map. The ground truth is represented by the collected sequence of points of regard. Figure 2 summarizes the result in the form of ROC analysis. Gazed locations corresponding to maxima in the saliency map are considered good predictions. Those maxima that are not attended by gaze are false positives. Not surprisingly, the dynamic saliency model outperforms the prediction capability of the static model in case of dynamic, natural stimuli. Still it is important to note the high false positive rate, due to the substantial number of un-attended maxima in the saliency map. The reason behind this high number of maxima resides in the integration of appearance and motion cues in the saliency model. Different features compete to gain the focus of attention at each time instant. As a consequence, a way is needed in order to select the next attended region among those that are emerging as salient. In addition, a further choice is requested to select the *gaze behavior*. The dynamic model that controls gazing is responsible of *what* to look, that is what region among the saliency maxima and *how*, that is, if performing fixations, smooth pursuit or saccades.

### 3 From saliency map to gaze prediction

Given a dynamic saliency map, the local maxima at each frame (at each time step  $t$ ), give information about the possible gaze localization. The problem of gaze prediction is the following: *the gaze location at time step  $t$  is observable only via the saliency map, estimating all points of interest for the gaze at  $t$ .* Therefore a usually adopted solution for gaze location prediction is to define a gaze scan path as a path through the local maxima of the saliency map. However, within dynamic environments, the dynamic saliency maps register motion of

different elements in the scene, like people, cars, flickering of several objects, clouds, and mainly ego motion; therefore several local maxima are returned by the saliency map, see Figure 2 and the experiments made available by Itti at <https://crcns.org/data-sets/eye>. This implies that the likelihood of any of the selected scan-paths would be approximatively the same; furthermore the number of possible paths through the local maxima is exponential in the length of the path.

Despite the fact that the saliency map takes care of motion and motion innovation, and accounts for the different gaze behaviors, not being a process itself, it cannot model perspicuously the dynamic of horizontal eye movements such as smooth pursuit, saccades and fixations, and the way these movements are constrained by the embodiment. Therefore it is necessary to reproduce the processes underlying the saliency maps and use the saliency maps as a collection of observations, as we shall specify in the following.

A saccade is a fast eye movement that can reach peak velocities of  $800^\circ/s$ . A saccadic eye motion has no feedback, meaning that the vision system is turned-off while the movement, and correction, via a further saccade, is achieved only when the target is reached. A saccade is a gaze motion from one target to another that can range from less than a degree to  $45^\circ$ . Because saccades follow an exponential function, peak velocity is reached early, and so far it cannot be fully simulated by a controlled mechanical motion, but under severe constraints. Smooth pursuit is a slow movement of the eye exhibited during tracking. Like saccade is a voluntary eye movement, but while saccades are elicited under different stimuli, smooth pursuit is elicited only under the stimulus of a moving target. Smooth pursuit reaches peak velocities of about  $60^\circ/s$  in order to keep the target position in the center of the fovea. As opposite to saccades, during smooth pursuit the vision is clear as it is foveated. Indeed, when the target is in the fovea, then the retinal image is no more in motion, this fact induces the ability of predicting the motion and thus to return to the target using the memorized stimulus. Finally, fixational eye movements are the result of micro saccades, ocular drifts and ocular microtremors related to the prevention of perceptual fading.

Several studies have simulated smooth pursuit, likewise the sudden change induced by saccadic generation (see for example [10]), and microsaccades [11], but always using as observations the current eye position via eye tracking, that is, on the basis of motion identification via the eye. Other studies have modeled the mechanical structure of eye muscles (see for example [12]) to predict eye motion. Despite the interest of these studies for a single step eye motion prediction, these cannot be used to predict gaze localization for embodied systems.

To estimate the gaze scan path we have to consider that the target to be tracked is precisely the gaze and not what the gaze is observing. The gaze, as target, is effectively a point, whose scan path is its projection on the 2D image. The process to be reproduced is based on the following information, at time step  $t$ :

1. The saliency map.
2. The updated eye position at time step  $t - 1$ .

3. The underlying gaze processes induced by the the horizontal eye movements.

The upshot of the above discussion is: the gaze is observed only through the saliency map, and its scan path is induced by at least three types of motion, namely saccade, smooth pursuit and fixation. Accordingly, there are at least three hidden real stochastic processes for the gaze and each of these process needs to be represented via a precise motion model. We note also that, at time step  $t$ , each local maxima of the saliency map is, in principle, the current state of some of the  $k^t$  possible scan paths started at time step  $t_0$ , with  $k$  the number of local maxima.

An important technique for managing multiple interacting dynamic models relies on the Markovian switching systems, also known in general as Interacting Multiple Model (IMM)[13]. Markovian switching systems model a process that changes in time by providing different models for each of the underlying processes. The state of each process is estimated under a bank of  $r$  filters. Each filter serves to estimate the state of the process and a mixing distribution establishes which process is effectively active at time  $t$ .

For the gaze scan path, the state of each process, a 7-dimensional random variable, is estimated by a non linear Bayes filter using the unscented Kalman transform, the transition between the three models has been estimated via a non parametric mixture model, as described in Section 4.

For each process, underlying the gaze scan-path, the estimation of the state  $\mathbf{x}_t$ , involves the following two equations:

$$\begin{aligned}\mathbf{x}_t &= \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t) \\ \mathbf{y}_{t-1} &= \mathbf{h}(\mathbf{x}_{t-1}, \mathbf{w}_t)\end{aligned}\tag{3}$$

The above estimation requires a transformation problem which can be stated as follows;  $\mathbf{x} \sim \mathcal{N}(\mu_x, \Sigma_{xx})$ , statistics  $\mathbf{y}$  are related to  $\mathbf{x}$  by a non linear function  $\mathbf{g}$ , with  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{y} \in \mathbb{R}^m$ ,  $\mathbf{g} : \mathbb{R}^n \mapsto \mathbb{R}^m$ . The unscented Kalman filter [14] is an optimal filtering approximating the filtering distribution of state  $\mathbf{x}$ , based on the unscented transform of the variable  $\mathbf{y}$ . The idea of the unscented transform is to approximate the filtering distribution, by approximating its Gaussian distribution [15]. Namely, a set of  $\sigma$  points are chosen so that their sample mean and covariance are the estimated mean and covariance of the state  $\mathbf{x}$ , then the non-linear function  $\mathbf{g}$  is applied to the  $\sigma$ -points to yield the mean and variance of the looked for distribution. The unscented transform is as follows. Let  $\bar{\mathbf{x}}$  and  $\Sigma_{xx}$  be the mean and covariance of the state  $\mathbf{x}$  of dimension  $n$ , approximated by  $2n + 1$   $\sigma$ -points, whose matrix  $\mathbf{X}$  is formed by  $2n + 1$   $\sigma$ -vectors as follows:

$$\mathbf{X} = [\bar{x} \cdots \bar{x}] + (\sqrt{n + \lambda}) \begin{bmatrix} \mathbf{0} & \sqrt{\Sigma_{xx}} & -\sqrt{\Sigma_{xx}} \end{bmatrix}\tag{4}$$

Here  $\lambda = \alpha^2(n + k) - n$  is a scaling parameter,  $\alpha$  determines the spread of the  $\sigma$ -points around  $\bar{x}$ ,  $k$  is a scaling parameter and  $\beta$  is used to incorporate prior knowledge of  $\mathbf{x}$ .  $\sqrt{(n + \lambda)\Sigma_{xx_i}}$  is the  $i$ -th row of the the matrix square root (see [16]). The  $\sigma$ -vectors are associated with weights  $W_i^{(m)}$  and  $W_i^{(c)}$  defined as

follows:

$$W_0^{(m)} = \frac{\lambda}{n + \lambda}, W_0^{(c)} = \frac{\lambda}{n + \lambda} + (1 - \alpha^2 + \beta), W_i^{(m)} = W_i^{(c)} = \frac{1}{2(n + \lambda)} \quad (5)$$

The transformed points  $\mathbf{Y}_i$  are obtained by applying the non linear function  $\mathbf{g}$  to the sigma points:  $\mathbf{Y}_i = \mathbf{g}(\mathbf{X}_i), i = 0, \dots, 2n$  and the new mean and variances are given as follows:

$$\bar{\mathbf{y}} = \sum_{i=0}^{2n} W_i^{(m)} \mathbf{y}_i \quad \text{and} \quad \Sigma_{yy} = \sum_{i=0}^{2n} W_i^{(c)} (\mathbf{Y}_i - \bar{\mathbf{y}})(\mathbf{Y}_i - \bar{\mathbf{y}})^\top \quad (6)$$

Then the transformation process yielded by the filter consists of the following two steps of prediction and update:

1. Prediction: (a) Compute the matrix of  $\sigma$  points  $\mathbf{X}_{t-1}$ . (b) Propagate the  $\sigma$ -points using the dynamic model  $\mathbf{f}$  of the specific gaze process in so obtaining  $\hat{\mathbf{X}}_t$ . (c) Compute the predicted mean and covariance of the state.
2. Update (a) Compute the matrix of  $\sigma$  points  $\mathbf{X}_t$ . (b) Propagate the  $\sigma$ -points using the observation model  $\mathbf{h}$  of the saliency map  $\mathbf{S}_t$ , in so obtaining  $\hat{\mathbf{Y}}_t$ . (c) Compute the predicted mean and covariance of the observations and the cross-covariance of state and observations. (d) Compute the filter gain, and the state mean and covariance, conditional to the observations.

For the gaze scan path estimation, the state is a 7 dimensional variable  $\mathbf{x}$  specified by the location of the gaze, in the image, the velocity, the acceleration and the saliency value of the point on the saliency map:

$$\mathbf{x} = (x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}, s_{map}) \quad (7)$$

The dynamic model specified by  $\mathbf{f}$ , for each process, is defined as follows. For the characterization of  $\mathbf{f}$  we shall consider position, velocity and acceleration along a single spatial dimension, respectively. For the saccade the dynamic model is the Singer model [17]. In this model the target acceleration is correlated in time, with correlation given by  $\sigma^2 \exp(-\alpha|\tau|)$ ,  $\alpha > 0$ , where  $\sigma^2$  is the variance of the saccade acceleration and  $\alpha$  is the reciprocal of the saccadic behavior, in so depending on the milliseconds needed to accomplish a saccade by the embodied system. Therefore it can be determined by  $\alpha_{max}^2 / 3(1 + 3p_{max} - p_0)$ , where  $\alpha_{max}$  is the maximum rate of acceleration, with probability  $p_{max}$  and  $p_0$  is the probability of no acceleration. The discrete time representation of the continuous model, see [17], is

$$x_t = \Psi(T, \alpha)x_{t-1} + \mathbf{u}_{t-1} \quad (8)$$

Here  $T = 1/30$  is the sampling rate, given by the framerate,  $\mathbf{u}_t$  is the inhomogeneous driving input whose variance is derived in [17], and  $\Psi(T, \alpha)$  is, under the saccadic process, the state transition matrix:

$$\Psi(T, \alpha) = \begin{bmatrix} 1 & T & (1/\alpha^2)[-1 + \alpha T + e^{-\alpha T}] \\ 0 & 1 & (1/\alpha)[1 - e^{-\alpha T}] \\ 0 & 0 & e^{-\alpha T} \end{bmatrix} \quad (9)$$

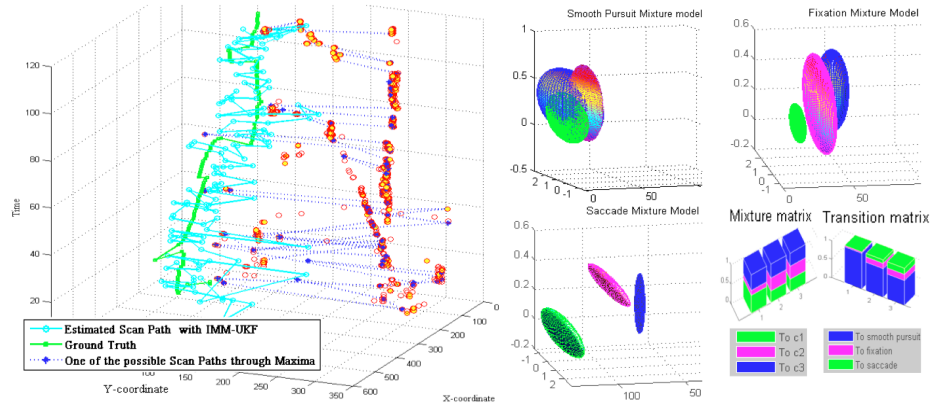
The dynamic model for the fixation motions (micro saccades, ocular drifts and ocular microtremors) is taken to be a Gaussian random walk. Therefore

$$x_t = \varphi x_{t-1} + \sigma_u \mathbf{u}_{t-1} \sim N(\varphi x_{t-1}, \sigma_u^2) \quad (10)$$

Finally the smooth pursuit is modeled by a Wiener process velocity model. The discrete time representation of the model is:

$$x_t = \begin{bmatrix} 1 & \Delta T & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x_{t-1} + \mathbf{u}_{t-1} \quad (11)$$

The measurement model returns only the updated location of the gaze, using the saliency map. To simplify the measurement only the first  $n$  local maxima are used, ordered according to the saliency value; further, among these ones the local maximum, having both highest saliency value, and whose speed and bearing from the predicted gaze location is minimal, is chosen. Clearly this characterization of the measurement is adapted to the particular process. More specifically, let  $S^P = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ , be the vector of positions of the  $n$  local maxima selected, and let  $S^S = (s_1, \dots, s_n)^\top$  be the vector of the saliency values, associated with the local maxima chosen. Furthermore let us denote by  $V_P = (\Delta \mathbf{x}_1, \dots, \Delta \mathbf{x}_n)^\top$  the displacement of the local maxima in  $P$ , measured as the flight distance of the gaze to that position, and let  $\Theta = (\theta_1, \dots, \theta_n)^\top$  be the bearing of the local maxima given the predicted gaze location.



**Fig. 3.** Left: the scan-path estimated with the IMM-UKF in cyan, the Ground Truth scan path in light green and in blue the scan path through the local maxima. Local maxima are the blue-yellow dots. Right: the Gaussian Mixture model and the continuous observation HMM used to estimate the transition kernel and the parameters of the IMM processes.

For the saccade, the measurement is defined as follows:

$$y_t = C(\delta V_P)_t + H x_k + \mathcal{N}(0, q_t); \quad (12)$$

Here  $\delta$  and  $C$  are matrices of dimension  $n \times n$ , whose elements  $\delta_{ij}$  and  $c_{ij}$  are defined as:

$$\delta_{ij} = \begin{cases} \delta_{i+1,j+1} = 1 & \text{if } s_j = \arg \max_j(SS) \\ 0 & \text{otherwise} \end{cases} \quad c_{ij} = \begin{cases} c_{i+1,j+1} = 1 & \text{if } v_j = \arg \min_k(V(\Delta \mathbf{x}_t)) \\ & \text{and } \theta_k = \arg \min_k(\Theta(v_j)) \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

Here  $\Delta \mathbf{x}_t$  is the  $n \times 1$  vector of the displacements, selected by  $\delta$ , to the local maxima,  $v_j$  is the vector of selected displacements  $c_{i,j}$  and

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (14)$$

The model for the fixation simply maintain the state unchanged and the model for the smooth pursuit inverts the ordering specified for the saccade, by first choosing the closest among the selected local maximal value, further it chooses the angle and finally the saliency value. It is easy to see that the three models obey a parsimonious principle to optimizing gaze use of resources.

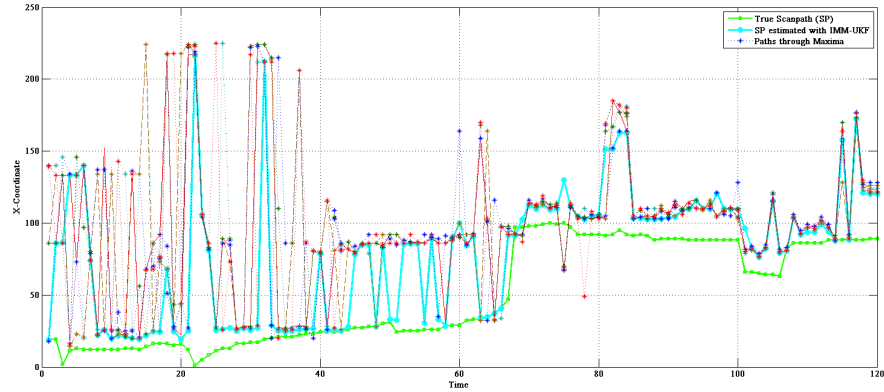
Given the above defined models of the three processes, different transition kernels have been estimated, and are reported in Section 4. The IMM model is finally computed by estimating the mixing probabilities  $\mu_t^{i|j} = (\pi_{ij} \mu_{t-1}^{i|j}) / (\sum \pi_{ij} \mu_{t-1}^{i|j})$ , at each estimation step, made of prediction and update as previously described, the mean, covariance and likelihood of each process is estimated and, finally the combined estimates is computed as:

$$\begin{aligned} \bar{x}_t &= \sum_{i=1}^k \mu_t^i \bar{x}_t^i \\ \Sigma_{xx} &= \sum_{i=1}^k \mu_t^i \times (\Sigma_t^i (\bar{x}^i - \bar{x})(\bar{x}^i - \bar{x})^\top) \end{aligned} \quad (15)$$

In Section 4 we discuss the experiments and the evaluation of the model with respect to gaze scan-paths recorded live in outdoor and dynamic scenes.

## 4 Experiments

Experiments have been performed at the different layers of the system. Experiments and comparisons with other approaches concerning the saliency map have been illustrated in Section 2. Here we are mainly concerned with the estimation of the parameters of the IMM via a continuous observations HMM, with mixtures of Gaussians observations, with the mean squared error of the IMM computed with respect to the *ground truth*, obtained by a wearable gaze tracking device, and finally with the results obtained with a pan-tilt. As test data we have collected the *bus stop* dataset (Figure 2), an outdoor scan-path comprising over 20 thousand gaze tracking frames, with the task of counting the people moving in the area. We have used 4 sets of data, using as features the velocity magnitude and the acceleration to estimate (via the Expectation Maximization algorithm) a continuous observation HMM with three states each accounting for observations sampled from a Gaussian mixture with three components. The estimated

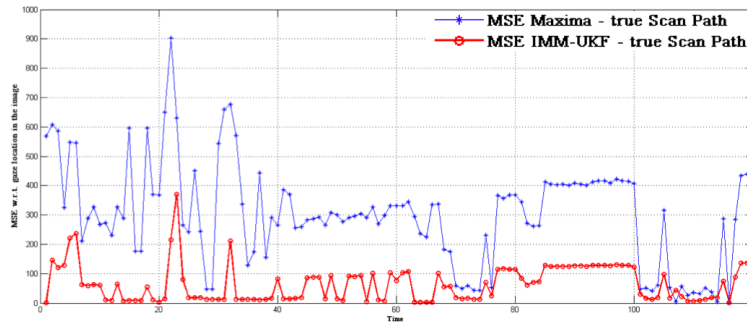


**Fig. 4.** The scan-paths of ten local maxima chosen deterministically at each step, with location determined only with respect to the  $x$ -coordinate, and compared with the IMM-UKF estimated one (in cyan) and the ground truth in light green.

model with the reduced set of features is illustrated in Figures 3. Having set the dimension of the space to three models the best transition kernel, namely the one maximizing the posterior probability of the scan-path, turned out to be the following, likewise the means for the three processes:

$$\mathcal{T} = \begin{bmatrix} & \text{Sacc} & \text{Fix} & \text{SP} \\ \text{Sacc} & 0.1 & 0.45 & 0.45 \\ \text{Fix} & 0.48 & 0.48 & 0.04 \\ \text{SP} & 0.57 & 0.05 & 0.38 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \bar{x}_{\text{sacc}} \\ \bar{x}_{\text{fix}} \\ \bar{x}_{\text{SP}} \end{bmatrix} = \begin{bmatrix} x & y & 1.63 & 9.43 & 1.46 & 17.53 & 0.23 \\ x & y & 0.01 & -0.09 & -0.23 & -0.13 & 0.28 \\ x & y & 0.04 & -0.79 & 1.18 & -1.12 & 2.3 \end{bmatrix} \quad (16)$$

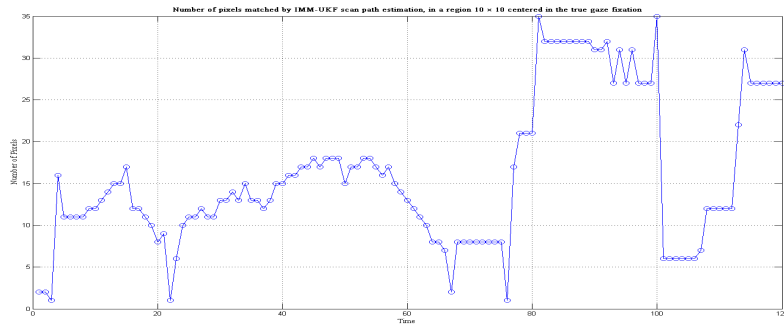
On the left the transition kernel, and on the right the estimated means of the three processes, for the gaze location we have left  $x$ ,  $y$ , since it was set to the center of the image. The mean squared error has been computed both for the



**Fig. 5.** Mean squared error, between the estimated scan path and the ground truth, and between a randomly selected scan-path across all the local maxima.

scan-path obtained by selecting the local maximum at each time step  $t$  and for the scan-path estimated via the IMM-UKF, with respect to the ground truth. The results are illustrated in Figure 5, while in Figure 4 the gaze location only with respect to the  $x$ -coordinate has been taken to illustrate that the local maxima can be hardly chosen, as they are exponential in the length of the scan-path and also only the best one can approximate the scan-path estimated by the switching processes IMM-UKF.

Finally to effectively understand, in terms of observed objects, whether the estimated scan-path would be able to attend, at least partially, the elements in the scene effectively observed by the subject (from which the Ground Truth has been extracted), we have computed the number of pixels that have been in common between a region  $10 \times 10$  around the gaze location, and the point estimated by the IMM-UKF. The outcome is illustrated in Figure 6.



**Fig. 6.** The number of pixel that fall into a region  $10 \times 10$  centered in the gaze location as effectively obtained via the ground truth.

Experiments for the embodied system were performed using a Directed Perception PTU D46 17 pan-tilt. The maximum angle which can be mechanically achieved is  $9^\circ$ , meaning that over this amplitude the delay would be unacceptable for a saccade. The pan-tilt is directly controlled by the position elicited by the Bayesian process.

## 5 Conclusion

We presented a dynamic system to control active vision based on computational attention. The underlying dynamic saliency model integrates motion cues in the computation of the saliency map and its prediction performance has been experimentally evaluated against static approaches using gaze tracking sequences as ground truth. An IMM-UKF takes care of selecting the next attended location among the multiple maxima which are generated by the saliency map, and decides which gaze behavior to implement. The model is trained on evidence collected from humans by means of gaze tracking experiments and provides smooth

scan-paths among saliency maxima which are likely to be attended by a human observer, also exhibiting frequencies of transition between gaze behaviors that are consistent to the biological counterpart. The resultant attentive control is suitable for robot active vision and demonstrates that considerations related to the embodiment naturally lead to a gaze prediction mechanism which seems better correlated to real gaze than the classical use of the saliency map maxima.

## References

1. Treisman, A., Gelade, G.: A feature-integration theory of attention. *Cognitive Psychology* **12** (1980) 97–136
2. Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N., Nufflo, F.: Modeling visual attention via selective tuning. *Artificial Intelligence* **78** (1995) 507 – 547
3. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI* **20** (1998) 1254–1259
4. Butko, N., Movellan, J.: Infomax control of eye movements. *Autonomous Mental Development, IEEE Transactions on* **2** (2010) 91 –107
5. Koch, C., Ullman, S.: Shifts in selective visual-attention: towards the underlying neural circuitry. *Hum. Neurobiol* **4** (1985) 219–227
6. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32** (2010) 171–177
7. Kowler, E.: Eye movements: The past 25years. *Vision Research* (2011) 1–27
8. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: *SCIA*. (2003) 363–370
9. Mancas, M., Riche, N., Leroy, J., Gosselin, B.: Abnormal motion selection in crowds using bottom-up saliency. In: *Proc. of the ICIP*. (2011)
10. Sauter, D., Martin, B., Di Renzo, N., Vomscheid, C.: Analysis of eye tracking movements using innovations generated by a kalman filter. *Medical and Biological Engineering and Comp.* (1991)
11. Engbert, R., Kliegl, R.: Microsaccades uncover the orientation of covert attention. *Vision Research* **43** (2003) 1035 – 1045
12. Komogortsev, O., Khan, J.I.: Eye movement prediction by kalman filter with integrated linear horizontal oculomotor plant mechanical model. In: *ETRA*. (2008) 229–236
13. Blom, H., Bar-Shalom, Y.: The interactive multiple model algorithm for system with markovian switching coefficients. *IEEE Trans. on Automatic Control* **33** (1988) 780–783
14. Julier, S.J., Jeffrey, Uhlmann, K.: Unscented filtering and nonlinear estimation. In: *Proceedings of the IEEE*. (2004)
15. Julier, S.J., Uhlmann, J.K.: A new extension of the kalman filter to nonlinear systems. In: *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*. (1997) 182–193
16. Wan, E., van der Merwe, R.: The unscented kalman filter for nonlinear estimation. In: *Proc. of the Symposium on Adaptive Systems for Signal Processing, Communication and Control*. (2000)
17. Singer, R.A.: Estimating optimal tracking filter performance for manned maneuvering targets. *IEEE Transactions on Aerospace and Electrictronic Systems* (1970)