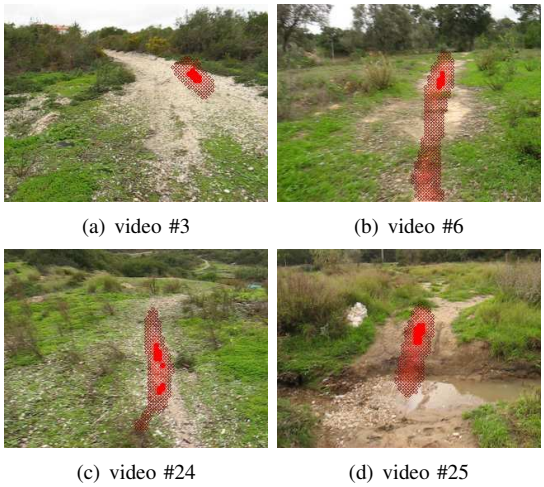
The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Swarm-Based Visual Saliency for Trail Detection

Pedro Santana, Nelson Alves, Luís Correia and José Barata

Abstract— **This paper proposes a model for trail detection that builds upon the observation that trails are salient structures in the robot's visual field. Due to the complexity of natural environments, the straightforward application of bottom-up visual saliency models is not sufficiently robust to predict the location of trails. As for other detection tasks, robustness can be increased by modulating the saliency computation with top-down knowledge about which pixel-wise visual features (e.g., colour) are the most representative of the object being sought. This paper proposes the use of the object's overall layout instead, as it is a more stable and predictable feature in the case of natural trails. This novel component of top-down knowledge is specified in terms of perception-action rules, which control the behaviour of simple agents performing as a swarm to compute the saliency map of the input image. For the purpose of multi-frame evidence accumulation about the trail location, a motion compensated dynamic neural field is used. Experimental results on a large data-set reveal the ability of the model to produce a success rate of 91% at 20Hz. The model shows to be robust in situations where previous trail detectors would fail, such as when the trail does not emerge from the lower part of the image or when it is considerably interrupted.**



(a) video #3 (b) video #6

(c) video #24 (d) video #25

Fig. 1. Typical trail detection results (red overlay) obtained with the proposed model. These results show that the proposed model is able to localise the trail even when it is highly interrupted, blends itself with the background, or does not start from the bottom of the image.

I. Introduction

Trails are usually safe pathways and also free of dead-lock situations. A robot following a trail is thus able to traverse large distances in off-road environments in an effortless way. On the one hand, computation for obstacle detection and trajectory/path planning is saved. On the other hand, fewer are the chances of getting lost or incurring into collisions.

Most of the challenges of trail detection relate to their lack of a well defined morphology or appearance. This hampers a straightforward learning of trail models. In addition, they exist in environments that are unstructured themselves. This in turn complicates the learning of background models. Moreover, the problem of supervising the learning process remains an open issue. This is aggravated by the fact that trails change over time, thus rendering hand-labelling unsuited for the task at hand. Hence, model-free solutions are essential for robust trail detection.

Typical solutions either assume that the robot is already on trail [1], [2] or that strong edges segment it from the background [3]. However, these two assumptions often fail to occur on realistic situations. An alternative is to segment the image, group some of the segments to build trail hypotheses, and then score these hypotheses against a model of the trail [4], [5]. However, contemporary models for robust image segmentation and subsequent grouping are computationally intensive and consequently unsuitable for real-time requirements. Moreover, grouping tends to fail in the presence of interrupted trails. The observation that trails are typically conspicuous in the robot's visual field led us to propose the use of visual saliency to focus the detection process [6]. This approach does not impose any hard constraint on the appearance or shape of both trail and background, nor it requires learning. Moreover, since it is rather common the use of saliency for other tasks in cognitively rich robots [7], [8], the overhead of its computation is diluted over all modules using it. This paper extends considerably this concept by recurring to the swarm-based collective behaviour metaphor and by exploiting evidence accumulation across frames for improved robustness. See Fig. 1 for typical results obtained with the extended model.

In a parallel study, Rasmussen et al. [9] proposes the use of appearance contrast for trail detection, which resembles to some extent the concept of visual saliency. However, they assume that trails are imaged as perfect triangles and both their left and right sides share the same appearance. Natural trails not always possess these properties. Additionally, the extensive use of 3-D information to bias the detection process in their model complicates the assessment of the role played by the appearance-based component. Finally, their method generates several trail hypotheses, whose contrast

P. Santana is with LabMAg, Computer Science Department, University of Lisbon, Portugal, Pedro.Santana@di.fc.ul.pt. The author's work was supported by FCT/MCTES grant No. SFRH/BD/27305/2006.

N. Alves is with IntRoSys, S.A. and UNINOVA, New University of Lisbon, Portugal, nelson.alves@introsys.eu

L. Correia is with LabMAg, Computer Science Department, University of Lisbon, Portugal, Luis.Correia@di.fc.ul.pt

J. Barata is with UNINOVA, New University of Lisbon, Portugal, jab@uninova.pt

is computed afterwards. Conversely, our model only generates hypotheses whose potential for high contrast is known beforehand, thus saving computation.

This paper is organised as follows. Section II overviews the proposed model. The way conspicuity maps are computed is summarised in Section III, which is followed by the detailed description of the swarm-based system in Section IV. Section V shows how the evidence about the trail location is improved across time. Experimental results are presented in Section VI. Finally, some conclusions are drawn and future work is proposed in Section VII.

II. System Overview

We showed in previous work [6] that the saliency map of a given image corresponds itself to an efficiently computed segmentation of the latter. That is, the segmentation of the input image, which can be a computationally intensive task, can be obtained as a by-product of determining which regions of the visual field detach more from the background. Furthermore, the obtained segments are already prioritised by their conspicuity level. We also showed that visual saliency and trail location in the input image are positively correlated.

From these findings it should follow that the highest priority segment in the saliency map matches the location of the trail in the input image. In practise, this is a brittle assumption in the face of not so well behaved saliency maps, which may occur in the presence of distractors or when the trail is considerably heterogeneous. This difficulty can be diminished with top-down boosting of visual features (e.g., colour) that are known to describe the object being sought [10], [11]. However, these visual features are considerably unpredictable in the case of trails in natural environments. In opposition, trails' overall layout is a much more predictable feature. For example, the projection of trails onto the input image typically converges towards a vanishing point. This novel use of top-down knowledge was embedded in our previous work in the form of behaviours ruling the motion of simple agents inhabiting the saliency and its intermediate conspicuity maps. The motion paths of these agents were then taken as the skeleton of a set of trail hypotheses, which were then scored, and three of them selected as the output of the system.

Despite its overall good results, our previous work was unable to reduce the ambiguity of three trail hypotheses, it was brittle in the presence of interrupted trails, and it was unable of exploiting historical information to improve its robustness. Fig. 2 depicts the model proposed in this paper, which extends our previous work to overcome its limitations: (1) by allowing the agents to exhibit collective behaviour through pheromone-based interactions, and (2) by allowing the system to accumulate evidence about the most likely trail location across multiple frames through the use of a dynamic neural field.

In short, two conspicuity maps, $\mathbf{C^C}(t) \in [0,1]$ for colour and $\mathbf{C^I}(t) \in [0,1]$ for intensity information, are computed from the input image $\mathbf{I}(t)$ [6]. A set of agents is then

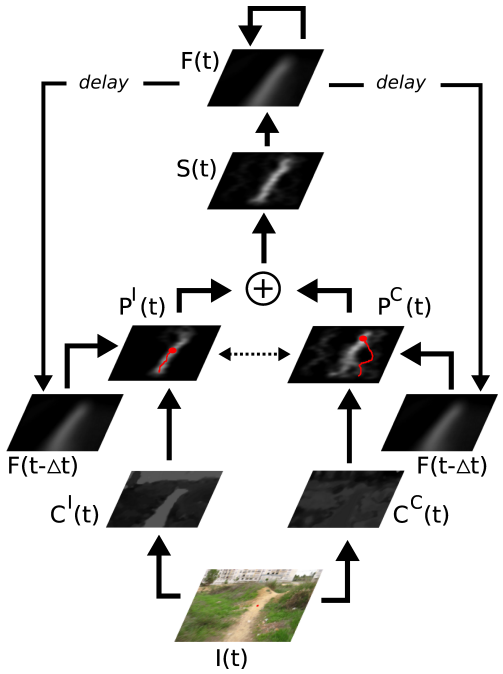


Fig. 2. System overview. The red overlays in both pheromone fields, $\mathbf{P^C}(t)$ and $\mathbf{P^I}(t)$, are two illustrative agent paths. For the sake of clarity, motion compensation aspects are not represented.

deployed on each map. These agents interact with the corresponding conspicuity map according to their perception-action rules, which embed the trail-specific top-down modulation process [6]. During the process, pheromone is deployed and sensed by the agents in two pheromone fields, $\mathbf{P^C}(t) \in [0,1]$ and $\mathbf{P^I}(t) \in [0,1]$, according to the ant foraging metaphor. An additional perception-action rule is introduced to make the agents' behaviour sensible to the pheromone deployed by the swarm, and thus enabling coherent collective behaviour to emerge. This way, agents help each other on the task of perceptual completion, resulting in a global behaviour that is robust to the local variations inherent to trails.

Being the deployed pheromone a function of agents' sensations across their trajectories on the corresponding conspicuity maps, it is influenced by the activity occurring in distant regions of the map. This long-range spatial connectivity allows handling the potentially large size of trails in a robust and parsimonious way.

Rather than blending both conspicuity maps to generate the final saliency map $\mathbf{S}(t) \in [0,1]$, as typically done [12], [10], in this work $\mathbf{S}(t)$ is obtained by blending both pheromone fields. The final saliency map $\mathbf{S}(t)$ feeds a dynamic neural field [13], [14], $\mathbf{F}(t) \in [0,1]$, which integrates pheromone (i.e., evidence) across frames and also implements both lateral excitation and long-range inhibition. This neural field allows the system to maintain a coherent focus of attention across time [14]. Motion compensation is also implemented so that the dynamics of the neural field can be decoupled from the dynamics of the robot. The neural field's state feeds back both pheromone fields so that history influences agents' activity. The output of the system is given

by the current state of the neural field, where the higher the activation of a given neuron the higher its chances of being associated to a trail's pixel.

III. Conspicuity Maps Computation

Conspicuousness computation is about determining which regions of the input image detach from the background at several scales and feature channels. In this paper only intensity and colour channels are used.

Shortly, one dyadic Gaussian pyramid with eight levels is computed from the intensity channel. Two additional pyramids also with eight levels are computed to account for the Red-Green and Blue-Yellow double-opponency colour feature channels. The various scales are then used to perform centre-surround operations [12]. The resulting centre-surround maps have higher intensity on those pixels whose corresponding feature differs the most from their surroundings. An example is a bright patch on a dark background (on-off), as well as the other way around (off-on). On-off centre-surround operations are performed by across-scale point-by-point subtraction, between a level with a fine scale and a level with a coarser one. Off-on maps are computed the other way around, i.e., subtracting the coarser level from the finer one. Then, the centre-surround maps are blended to produce a colour conspicuity map, $\mathbf{C}^{\mathbf{C}}(t) \in [0,1]$, and an intensity conspicuity map, $\mathbf{C}^{\mathbf{I}}(t) \in [0,1]$. The width, w, and height, h, of both maps is 80 and 60, respectively.

When blending maps, the most discriminant ones are promoted by recurring to a normalisation operator. Here we follow the normalisation operator previously proposed by us [6], which was shown to outperform other known models [12], [10] in trail detection. Please refer to [6] for further details and to Fig. 2 for examples of conspicuity maps.

IV. Collective Behaviour

A. Agent Behaviours

This section describes how an agent deployed on a given conspicuity map $m \in \{\mathbf{C}^{\mathbf{C}}(t), \mathbf{C}^{\mathbf{I}}(t)\}$ behaves in order to generate, in cooperation with other agents, a pheromone field $p \in \{\mathbf{P}^{\mathbf{C}}(t), \mathbf{P}^{\mathbf{I}}(t)\}$ whose activity level is correlated with the localisation of the trail. If the agent is allocated to the colour conspicuity map, $\mathbf{C}^{\mathbf{C}}(t)$, then it contributes to the colour pheromone field, $\mathbf{P}^{\mathbf{C}}(t)$. The same process for the intensity conspicuity map.

At the onset of each frame, both pheromone fields are zeroed and subsequently affected by a small ratio λ of the robot motion compensated neural field's previous state, $\mathbf{F}'(t-\Delta t)$, $\mathbf{P}^{\mathbf{C}}(t) = \mathbf{P}^{\mathbf{I}}(t) = \lambda \mathbf{F}'(t-\Delta t)$. In this study $\lambda = 0.1$. Refer to Section V for details on the computation of $\mathbf{F}'(t-\Delta t)$. This pheromone level offset allows agents' activity to be affected by history. This induces stability, robustness to noise and across-frames progressive improvement.

For a given number $n_{max} = 50$ of iterations, whose index is represented by n, the agent builds up a trail hypothesis by updating its position, $o(n)$, according to a set of behaviours B, which are sensible to the level of conspicuity in the agent's surroundings. These behaviours embed top-down information on the object being sought, such as its approximate shape. The agent's motion is also affected by other agents' activity according to the ant foraging metaphor, i.e., via *stigmergy*. That is, agents interact with each other through a pheromone field built by them while moving. Conspicuity-based behaviours and pheromone influence contribute to the agent's motion according to the following voting mechanism,

$$a^+(n) = \arg\max_{a \in O} \left(\sum_{b \in B} \alpha_b f_b(m,a,n) + \beta g(p,a,n) + \gamma q \right) \quad (1)$$

$$\dot{o}(n) = \Gamma\Big(a^+(n)\Big) \quad (2)$$

where: O is the set of possible agent motor actions (e.g., "move to the right"); $\Gamma(.)$ transforms a motor action, $a \in O$, onto pixel coordinates centred on the current agent's position; β is the weight accounting for the contribution of pheromone, which is described by the motor action evaluation function $g(p,a,n) \in [0,1]$; α_b is the weight accounting for the contribution of behaviour $b \in B$, which is described by the motor action evaluation function $f_b(m,a,n) \in [0,1]$; and γ is the weight accounting for stochastic behaviour, being $q \in [0,1]$ a number sampled from a uniform distribution each time the action is evaluated.

The following describes which regions in the local neighbourhood of the current agent position are selected as its next position by each of the five behaviours composing B, and thus embody top-down knowledge about trails,

1) Regions of higher levels of conspicuity, under the assumption that trails are salient in the input image;
2) Regions whose average level of conspicuity is more similar to the average level of conspicuity of the pixels visited by the agent, under the assumption that trails' appearance is somewhat homogeneous;
3) Regions that maintain the agent equidistant to the boundaries of the trail hypothesis being pursued;
4) Upwards regions under the assumption that trails are often vertically elongated;
5) Region targeted by the motor action at the previous iteration, under the assumption that trails' outline is somewhat monotonous.

The newly proposed evaluation function $g(p,a,n)$ greedily provides higher score to the motor actions that take the agent to regions of higher level of pheromone. By making the score proportional to the level of pheromone, this evaluation function guides the agent towards regions recurrently visited by other agents. The outcome is coordinated collective behaviour. By the end of each iteration, the agent contributes to pheromone field p by deploying an amount of pheromone ε in its current position, $o(n)$, and to the other conspicuity map p' a small portion of ε, υ. That is, if $p = \mathbf{C}^{\mathbf{C}}(t)$ then $p' = \mathbf{C}^{\mathbf{I}}(t)$, and the other way around. This process enables loosely coupled cross-modality influence, thus allowing each agent to exploit multiple cues indirectly, and therefore to maintain their simplicity. In this study $\varepsilon = 0.008$ and $\upsilon = 0.3$.

The ratio $\beta/(\sum_{b\in B}\alpha_b+\gamma)$ controls the importance of the collective over the individual experience. In this study $\beta=1.0$ and $\gamma=0.8$. Please refer to [6] for further details on the agent motor actions set O, on the behaviour set B, on its associated weights α_b, and on how the agent's local surroundings is segmented into regions.

B. Agents Recruitment

A set of agents, E_m, is deployed at each conspicuity map $m\in\{\mathbf{C^C}(t),\mathbf{C^I}(t)\}$. The chances of deploying an agent on a given location of conspicuity map m depends on the level of conspicuity at that location and on the level of pheromone at the same position of the corresponding pheromone field p. The following describes in detail the deployment process.

To avoid any noise potentially present at the map's boundaries, agents are deployed with a small offset of the bottom of the conspicuity map in question, i.e., at row $r=h-5$, where h is the height of the conspicuity maps.

To determine the column where each agent is deployed, the unidimensional vector $\mathbf{v^m}=(v_0^m,\ldots,v_w^m)$ is first computed, where w is the width of the conspicuity maps. The element v_k^m of $\mathbf{v^m}$ refers to the average conspicuity level of the pixels in column k, contained between row r and row $r-\delta$, where $\delta=5$ to avoid deploying agents in columns with spurious highly conspicuous pixels. Formally, $v_k^m=\sum_{l\in[r,r-\delta]}m(k,l)/\delta$, where $m(k,l)$ is the conspicuity level at pixel in column k and row l. The same process is repeated to build a vector for the pheromone field in question, $\mathbf{v^p}=(v_0^p,\ldots,v_w^p)$, where $p(k,l)$ is the pheromone level at pixel in column k and row l. In this case, $v_k^p=\sum_{l\in[r,r-\delta]}p(k,l)/\delta$. Then, the test $z<(v_{j\cdot w}^m+\max(v_{j\cdot w-4}^p,v_{j\cdot w+4}^p))$ is repeated until it succeeds, where $z\in[0,1]$ and $j\in[0,1]$ are numbers sampled from a uniform distribution each time the test is performed. At that time, the agent is deployed in column $j\cdot w$. With this test, the chances of deploying an agent in a randomly selected column $j\cdot w$ is as high as the conspicuity and pheromone levels at the deployment region. This sampling process is repeated until $|E_m|=20$ agents are deployed per map m.

V. Evidence Accumulation

To integrate evidence across time, to consider competition between multiple focus of attention, and to promote perceptual grouping, the fusion of both pheromone fields, $\mathbf{S}(t)=\frac{1}{2}\mathbf{P^C}(t)+\frac{1}{2}\mathbf{P^I}(t)$, feeds a 2-D dynamic neural field $\mathbf{F}(t)$. Note that this process only occurs after the agents' activity has ceased, and therefore the pheromone fields have been fully updated.

The dynamical characteristic of the neural fields [13], [14] is what enables their ability to integrate information across time. To avoid the blurring of the neural field when the robot moves, the following three steps explicitly compensate the neural field for the camera motion engaged between the previous and current frames:

1) Estimate the homography matrix $\mathrm{H}(t)$ that describes the perspective transformation between the current frame, $I(t)$, and the previous one, $I(t-\Delta t)$. This step is further detailed in Section V-A.
2) Obtain a perspective compensated version of the previous neural field's state by using the estimated homography matrix, $\mathbf{F}'(t-\Delta t)=\mathbf{H}(t)\mathbf{F}(t-\Delta t)$.
3) Obtain $\mathbf{F}(t)$ by updating the perspective compensated neural field $\mathbf{F}'(t-\Delta t)$ with the pheromone field $\mathbf{S}(t)$. This step is further detailed in Section V-B.

A. Homography Matrix Estimation

To estimate the perspective transformation, a set of Shi and Tomasi [15] corner points are first detected in the previous frame, $\mathbf{I}(t-\Delta t)$. These points are then tracked in the current frame, $\mathbf{I}(t)$, with a pyramidal implementation of the Lucas-Kanade feature tracker [16]. The resulting sparse optical flow is then used to estimate the perspective transformation relating both frames, i.e., the 3×3 homography matrix H,

$$\mathbf{u}'_i=\mathrm{H}(t)\mathbf{u_i} \tag{3}$$

where $\mathbf{u_i}$ is a local feature found in $\mathbf{I}(t-\Delta t)$ and $\mathbf{u}'_\mathbf{i}$ its correspondence in $\mathbf{I}(t)$. Due to noise in the tracking process, the homography matrix is calculated as the least-squares solution that minimises the back-projection error [17]. This process assumes that distortion introduced by the camera lens into the input images has been corrected. It also assumes that either: (1) the terrain in front of the robot is planar or (2) the camera was only rotated, and not displaced, between frames. None of these two constraints can be strictly ensured in off-road environments. Still, in most situations the terrain is somewhat planar and the attitude of the camera changes more significantly than its position. Experiments have shown that the co-occurrence of these two relaxed constraints is sufficient to maintain a robust operation. If a minimum of four correspondences is not found, the homography matrix is set to the identity matrix, $\mathrm{H}(t)=diag(1,1,1)$.

B. Neural Field Update

The neural field $\mathbf{F}(t)$ is a 2D lattice of $w\times h$ neurons with "Mexican-hat"-shaped lateral coupling. This property helps in the formation of a coherent focus of attention [14]. On the one hand, activated neurons excite their neighbours, thus promoting perceptual grouping. On the other hand, activated neurons tend to inhibit distant ones, thus reducing ambiguities in the focus of attention. Formally, the connection's weight between a neuron in position $\mathbf{x}$ and a neuron in position $\mathbf{x}'$ is given by a Difference of Gaussians (DoG), function of the Euclidean distance between both, $w(\mathbf{x},\mathbf{x}')$.

In addition to lateral connectivity, the neural field also has afferent interactions with pheromone field $\mathbf{S}(t)$. The weight of a connection between an element of $\mathbf{S}(t)$ in position $\mathbf{y}$ and a neuron of $\mathbf{F}(t)$ in position $\mathbf{x}$ is given by a Gaussian function of the Euclidean distance between both, $d(\mathbf{x},\mathbf{y})$. This operation enlarges neurons' receptive field to reduce sensitivity to noise.

The average membrane potential of a given neuron at position $\mathbf{x}$ can now be expressed by the following nonlinear integro-differential equation,

$$\tau\frac{\partial \mathbf{F}(\mathbf{x},t)}{\partial t} = -\mathbf{F}(\mathbf{x},t) + \int w(\mathbf{x},\mathbf{x}')\, f\left(\mathbf{F}(\mathbf{x}',t)\right) \mathrm{d}\mathbf{x}' + \int d(\mathbf{x},\mathbf{y})\mathbf{S}(\mathbf{y},t)\mathrm{d}\mathbf{y} + h \quad (4)$$

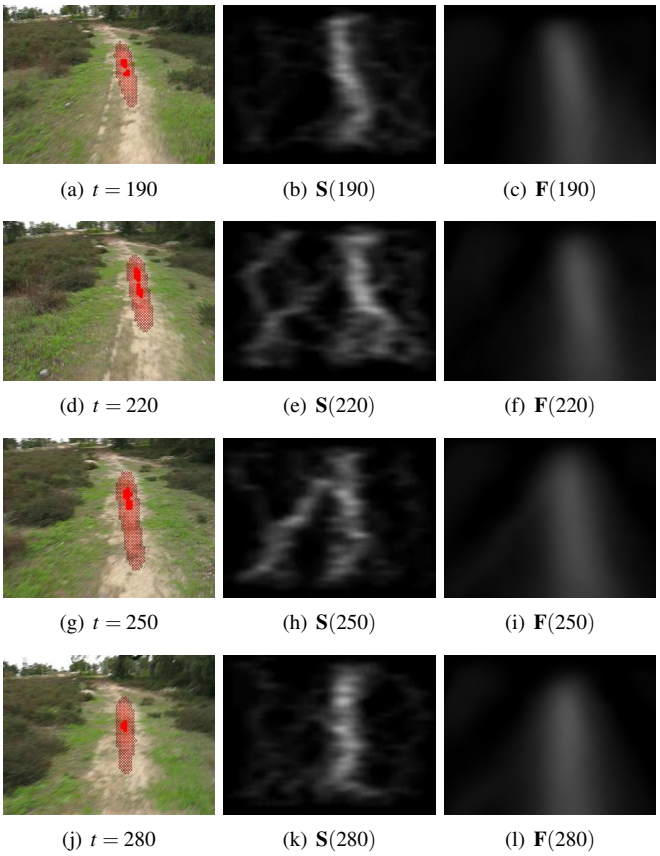
where $f(x) = x$ in this paper, τ is a time constant and $h = 0$ is the neuron threshold. For numerical integration, the Euler forward method is used to obtain an approximation of the neural field, which in matrix form results in the following rearranged expression,

$$\mathbf{F}(t) = \mathbf{F}'(t-\Delta t) + \frac{\Delta t}{\tau}\Big(-a\cdot\left(\mathbf{F}'(t-\Delta t)\right) + b\cdot\left(DoG^{k_1,k_2}_{\sigma_1,\sigma_2} * \mathbf{F}'(t-\Delta t)\right) + c\cdot\left(G^{k_3}_{\sigma_3} * \mathbf{S}(t)\right) + h\Big) \quad (5)$$

where $*$ is the convolution operator, a, b and c are weights defining the contribution of each term, $DoG^{k_1,k_2}_{\sigma_1,\sigma_2} = G^{k_1}_{\sigma_1} - G^{k_2}_{\sigma_2}$, G^{k}_{σ} is a Gaussian kernel of size $k \times k$ and width σ. Note that the neural field's previous state, $\mathbf{F}(t-\Delta t)$, is substituted by its motion compensated counter part, $\mathbf{F}'(t-\Delta t)$. The neural field free parameters have been empirically defined, $\sigma_1 = 4.25$, $\sigma_2 = 14.15$, $\sigma_3 = 2.15$, $k_1 = 25$, $k_2 = 91$, $k_3 = 11$, $a = 2$, $b = 2.5$, $c = 8$, and $\frac{\Delta t}{\tau} = 0.03$. The system showed robustness to small variations around these values as long as the proportions are roughly maintained.

To enable fast computation, the model is synchronously evaluated, meaning that at time t neurons are updated based on the network state at time $t-\Delta t$. Due to robot motion, any potential symmetry at the sensory input does not prevail, making neural field oscillations unlikely to occur over relevant periods of time.

The dynamical characteristic of the model in conjunction with the long-range lateral inhibition results in the following property. The higher the number of frames with the same spot with high activity the more difficult it is, due to lateral connectivity, for other regions to become activated. Hence, transient distractors are actively inhibited once a large evidence on the trail location is accumulated (see Fig. 3).



(a) $t = 190$ (b) $\mathbf{S}(190)$ (c) $\mathbf{F}(190)$
(d) $t = 220$ (e) $\mathbf{S}(220)$ (f) $\mathbf{F}(220)$
(g) $t = 250$ (h) $\mathbf{S}(250)$ (i) $\mathbf{F}(250)$
(j) $t = 280$ (k) $\mathbf{S}(280)$ (l) $\mathbf{F}(280)$

Fig. 3. Example of neural field competition in a situation represented by four ordered frames obtained from video #11 of the tested data-set. The trail is present in the input image for several frames prior to $t = 220$, thus eliciting high level of activity in the neural field, $\mathbf{F}(190)$. Although the transient appearance of a trail-like grass segment in the bottom-left region of the image is felt in the pheromone field, $\mathbf{S}(220)$ and $\mathbf{S}(250)$, this distractor is actively inhibited in the neural field, $\mathbf{F}(220)$ and $\mathbf{F}(250)$.

VI. Experimental Results

An extensive data-set of 25 colour videos encompassing a total of 12023 frames with a resolution of 640×480 has been obtained with a hand-held camera (see Fig. 4). The camera was carried at an approximate height of 1.5 m and at an approximate speed of $1\,\mathrm{ms}^{-1}$. The model's output on these videos is available at the Authors' site[1]. The trail detector was evaluated on a Core2 Duo 2.8 GHz running Linux. OpenCV was used for low-level routines. Table I shows that the model runs on average at 20 Hz, where only 4% refers to the swarm-based activity. The timing reported for the neural field update also includes optical flow computation, homography estimation, and neural field wrapping.

[1]http://www.uninova.pt/∼pfs/iros2010trail.html

The experimental results are twofold. First it is shown that the proposed swarm-based saliency model is more robust than a classical one [12], [10], where conspicuity maps are blended, $\mathbf{S}(t) = \frac{1}{2}\mathbf{C}^{\mathbf{C}}(t) + \frac{1}{2}\mathbf{C}^{\mathbf{I}}(t)$, rather than their corresponding pheromone fields, $\mathbf{S}(t) = \frac{1}{2}\mathbf{P}^{\mathbf{C}}(t) + \frac{1}{2}\mathbf{P}^{\mathbf{I}}(t)$. For the sake of fair comparison, the neural field $\mathbf{F}(t)$, which is fed by $\mathbf{S}(t)$, is used to generate the output in both cases. Then, a qualitative comparison with related trail detectors highlights the advantages of the proposed model. To handle the probabilistic nature of the agents behaviours, a set of 5 runs was performed per video.

The trail is considered correctly localised if the biggest blob of neural field activity above 0.85 (from a maximum of 1) is fully within the trail's boundaries. In cases of ambiguity caused by co-occurrence of two similar blobs, the pheromone field $\mathbf{S}(t)$ is used to assess which blob is being reinforced and consequently should be taken as the output.

Table II shows that the proposed swarm-based saliency model clearly outperforms the classical one. That is, a higher average success rate is obtained along with a smaller standard deviation. It follows from the success rate of $91\% \pm 12\%$ that the proposed model is well suited for off-road autonomous

robots. This result is more stringent if the difficulty of the tested data-set is taken into account. To our knowledge no previous work has been tested against a data-set with trails as narrow, unstructured and discontinuous as the ones herein considered. Moreover, differently from previous works [4], [2], [5], [9], the model succeeds in situations where the trail is not starting from the bottom of the image (see Fig. 1(a)).

It is also worth noting that in 7 of the 25 videos, the proposed model shows 100% success rate for all the 5 five runs. Video 5 is accounted as a long run with almost 5 minutes length. Besides being often interrupted and highly unstructured, the trail in this video also exhibits a variable width. Moreover, the terrain surrounding the trail is heterogeneous and highly populated with potential distractors, such as trees and bushes. The 85% success rate of the model in this video clearly shows its robustness in demanding situations. About 5% of the fail cases refer to situations where the trail is nevertheless noticeable in the neural field. In this case, as in other lower performance videos, ambiguity between trail and surroundings could be reduced by considering additional perceptual modalities, such as texture and depth.

When the trail is highly conspicuous in the environment, as most often occurs, ambiguity is rarely present. When this assumption fails and distractors are scattered, the model is still able to perform correctly. This robustness owes to the agents' sensori-motor coordination capabilities, which allow an opportunistic exploitation of the trail-background segmentation present in the conspicuity maps.

VII. Conclusions

A swarm-based model for top-down modulation of visual saliency with the goal of localising unstructured trails in natural environments was proposed. The model has been successfully validated against a highly demanding and diverse data-set by exhibiting 91% success rate at 20 Hz. These results due to large extent to the swarm-based design, which enabled a robust self-organisation of visual search, perceptual grouping, and multiple hypotheses tracking. All these characteristics allow the system to perform in situations where previous trail detectors fail, such as when the trail does not emerge from the lower part of the image or when it is considerably interrupted. To our knowledge, this work is the most complex application of the agent-based sensori-motor coordination approach to object detection.

The high success rate across the diverse data-set shows that the selected parametrisation is not over-fit to a specific environment, thus highlighting its robustness. Nevertheless, a more extensive sensitivity analysis of the model still needs to be addressed in future work. Other perceptual modalities, such as texture and depth, will be included to further increase the robustness of the model. Finally, we plan to test the swarm-based saliency model to other visual search tasks.

	neural field	**conspicuity maps computation**	**swarm computation**	**total**
time (ms)	12	36	2	**50**

TABLE I

Average computation times.

Acknowledgments

This work was partially supported by FCT/MCTES grant No. SFRH/BD/27305/2006. The authors wish to thank the fruitful comments provided by the anonymous reviewers.

References

[1] C. Rasmussen and D. Scott, "Terrain-based sensor selection for autonomous trail following," in *Proc. of the 2nd Intl. Workshop on Robot Vision (Robvis 2008)*, 2008, pp. 341–355.

[2] D. Fernandez and A. Price, "Visual detection and tracking of poorly structured dirt roads," in *Proc. of the Intl. Conf. on Advanced Robotics (ICAR)*, 2005, pp. 553–560.

[3] A. Bartel, F. Meyer, C. Sinke, T. Wiemann, A. Nchter, K. Lingemann, and J. Hertzberg, "Real-time outdoor trail detection on a mobile robot," in *Proc. of the 13th IASTED Intl. Conf. on Robotics, Applications and Telematics*, 2007, pp. 477–482.

[4] C. Rasmussen and D. Scott, "Shape-guided superpixel grouping for trail detection and tracking," in *Proc. of the 2008 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2008, pp. 4092–4097.

[5] M. Blas, M. Agrawal, K. Konolige, and A. Sundaresan, "Fast color/texture segmentation for outdoor robots," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Nice, France, 2008, pp. 4078–4085.

[6] P. Santana, N. Alves, L. Correia, and J. Barata, "Fast trail detection: A saliency-based approach," in *Proc. of the Intl. Conf. on Robotics and Automation (ICRA 2010)*, Anchorage, Alaska, 2010.

[7] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2008, pp. 962–967.

[8] J. Moren, A. Ude, A. Koene, and G. Cheng, "Biologically based top-down attention modulation for humanoid interactions," *International Journal of Humanoid Robotics*, vol. 5, no. 1, pp. 3–24, 2008.

[9] C. Rasmussen, Y. Lu, and M. Kocamaz, "Appearance contrast for fast, robust trail-following," in *Proc. of the IEEE Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2009.

[10] S. Frintrop, G. Backer, and E. Rome, "Goal-directed search with a top-down modulated computational attention system," *Lecture Notes In Computer Science*, vol. LNCS 3663, p. 117, 2005.

[11] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, 2005.

[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1254–1259, 1998.

[13] S. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," *Biological Cybernetics*, vol. 27, no. 2, pp. 77–87, 1977.

[14] N. Rougier and J. Vitay, "Emergence of attention within a neural population," *Neural Networks*, vol. 19, no. 5, pp. 573–581, 2006.

[15] C. Tomasi and J. Shi, "Good features to track," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[16] J. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker description of the algorithm," *Intel Corporation, Microprocessor Research Labs, OpenCV Documents*, 1999.

[17] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.

Fig. 4. Data-set representative frames. Each image corresponds to one video whose ID is given by increasing order from left to right and top to bottom. The overlaid red blobs represent the model's estimate of the trail location, which corresponds to an activity of the neural field above 0.85.

		Classic Saliency Computation		Proposed Swarm-Based Saliency Computation	
Video ID	**Nr. of Frames**	**Nr. of Correct Frames**	**% of Correct Frames**	**Average Nr. of Correct Frames**	**Average % of Correct Frames**
1	278	124	44.60	278.00 ± 0.00	100.00 ± 0.00
2	204	126	61.76	204.00 ± 0.00	100.00 ± 0.00
3	422	20	4.74	362.40 ± 9.45	85.88 ± 2.24
4	135	0	0.00	135.00 ± 0.00	100.00 ± 0.00
5	2854	927	32.48	2457.40 ± 24.91	86.10 ± 0.87
6	186	52	27.96	185.80 ± 0.45	99.89 ± 0.24
7	121	0	0.00	121.00 ± 0.00	100.00 ± 0.00
8	124	0	0.00	124.00 ± 0.00	100.00 ± 0.00
9	309	58	18.77	277.40 ± 4.51	89.77 ± 1.46
10	147	73	49.66	138.40 ± 1.14	94.15 ± 0.78
11	386	0	0.00	386.00 ± 0.00	100.00 ± 0.00
12	158	0	0.00	108.20 ± 15.71	68.48 ± 9.94
13	134	54	40.30	132.60 ± 1.67	98.96 ± 1.25
14	676	299	44.23	669.60 ± 0.55	99.05 ± 0.08
15	683	181	26.50	559.60 ± 8.38	81.93 ± 1.23
16	770	35	4.55	592.60 ± 11.93	76.96 ± 1.55
17	403	141	34.99	380.40 ± 3.71	94.39 ± 0.92
18	335	325	97.01	331.80 ± 0.45	99.04 ± 0.13
19	230	195	84.78	225.20 ± 2.28	97.91 ± 0.99
20	439	28	6.38	244.20 ± 1.64	55.63 ± 0.37
21	490	18	3.67	479.60 ± 2.30	97.88 ± 0.47
22	230	25	10.87	230.00 ± 0.00	100.00 ± 0.00
23	600	36	6.00	560.20 ± 4.55	93.37 ± 0.76
24	802	0	0.00	683.80 ± 7.05	85.26 ± 0.88
25	907	0	0.00	710.40 ± 9.13	78.32 ± 1.01
	$\sum = 12023$	$\sum = 2717$	$(\mu \pm \sigma) = (\mathbf{23.97 \pm 27.73})$	$\sum = (\mathbf{10577.60 \pm 109.80})$	$(\mu \pm \sigma) = (\mathbf{91.32 \pm 1.01})$

TABLE II

TRAIL DETECTION RESULTS. CLASSIC SAL. COMP.: $\mathbf{S}(t) = \frac{1}{2}\mathbf{C}^{\mathbf{C}}(t) + \frac{1}{2}\mathbf{C}^{\mathbf{I}}(t)$. PROPOSED SAL. COMP.: $\mathbf{S}(t) = \frac{1}{2}\mathbf{P}^{\mathbf{C}}(t) + \frac{1}{2}\mathbf{P}^{\mathbf{I}}(t)$.