Neural-Swarm Visual Saliency for Path Following

Pedro Santana
CTS-UNINOVA, New University of Lisbon
pfs@uninova.pt

Ricardo Mendonça
CTS-UNINOVA, New University of Lisbon

Luís Correia
LabMAg, University of Lisbon

José Barata
CTS-UNINOVA, New University of Lisbon

Last revision: November 10, 2012

Abstract

This paper extends an existing saliency-based model for path detection and tracking so that the appearance of the path being followed can be learned and used to bias the saliency computation process. The goal is to reduce ambiguities in the presence of strong distractors. In both original and extended path detectors, neural and swarm models are layered in order to attain a hybrid solution. With generalisation to other tasks in mind, these detectors are presented as instances of a generic neural-swarm layered architecture for visual saliency computation. The architecture considers a swarm-based substrate for the extraction of high-level perceptual representations, given the low-level perceptual representations extracted by a neural-based substrate. The goal of this division of labour is to ensure parallelism across the vision system while maintaining scalability and tractability. The proposed model is shown to exhibit, at 20 Hz, a 98.67 % success rate on a diverse data-set composed of 39 videos encompassing a total of $29,789$ 640×480 frames. An open source implementation of the model, fully encapsulated as a node of the Robotics Operating System (ROS), is available for download.

1 Introduction

This paper addresses the problem of vision-based path detection and tracking in cross-country environments. A major challenge of this domain relates to the wide variety of paths that must be covered, ranging from engineered beaten paved paths to natural hiking trails with varying shape and tread materials, which defies the devision of a simultaneously robust and computationally efficient detector. This challenge is closely related to the speed-accuracy trade-off humans solve when searching for objects in complex environments, in part by recurring to visual attention mechanisms (Wolfe et al., 2011). The same reasoning applies to object detection in computer vision systems (Elazary and Itti, 2010). Visual attention is commonly thought as of being composed of a sensory-driven bottom-up pre-attentive component (Treisman and Gelade, 1980; Koch and Ullman, 1985; Itti et al., 1998; Palmer, 1999; Corbetta and Shulman, 2002; Hou and Zhang, 2007), which is modulated by top-down context aware pathways (Yarbus, 1967; Wolfe, 1994; Tsotsos et al., 1995; Corbetta and Shulman, 2002; Torralba et al., 2003; Frintrop et al., 2005; Navalpakkam and Itti, 2005; Walther and Koch, 2006; Neider and Zelinsky, 2006; Rothkopf et al., 2007; Hwang et al., 2009), as it has been shown by recent neurophysiological studies (Egner et al., 2008). The outcome is a saliency map that signals the regions of the visual field that are simultaneously conspicuous and share the general properties of the object of interest (e.g., colour).

Recently, a neural-swarm layered model of visual saliency has been proposed and validated in the context of path detection and tracking (Santana et al., 2010a, 2011, 2012). This paper extends this model so that the appearance of the path being followed can be learned and used to bias the saliency computation process in a top-down manner. This extension enhances the system's ability to operate under the presence of strong distractors. Concretely, a 98.67 % success rate on a diverse data-set was obtained at 20 Hz (see Fig. 1 for typical results). These results show the ability of the model to attain accuracy and computational parsimony simultaneously.

With generalisation to other tasks in mind, both original and extended path detectors are presented as instances of a generic neural-swarm architecture for visual saliency computation.

(a) (b)

Figure 1: Representative results obtained with the proposed model. The red overlay corresponds to the the model's estimate about the path location. In (a), it is possible to see the ability of the model to handle sudden changes in both path's appearance and outline. In (b), the model's robustness in the presence of narrow, off-centre, and shadowed paths is highlighted.

As it will be argued, the neural-based layer is suited to realise low-level feature extraction, as it demands mostly for massively parallel application of local image operators, whereas the swarm-based layer is suited for high-level feature extraction from the low-level features, as it demands complex spatio-temporal data integration. Moreover, with this division of labour, parallelism across the vision system is assured without hampering scalability and tractability.

From a cognitive systems perspective, the layered approach results in a system with high local connectivity and global sparse connectivity. These are common properties of small world networks, which are becoming widely recognised as structural in the human brain (Sporns et al., 2004; Bassett and Bullmore, 2006). In line with this coincidence, it is the growing evidence about the similarities between the self-organising properties of the brain processes and the swarm cognition exhibited by social insects (Passino et al., 2008; Couzin, 2009; Marshall and Franks, 2009; Turner, 2011). This connection is triggering the interest in the modelling of robot cognitive behaviour based on the swarm cognition metaphor (Santana and Correia, 2010, 2011; Trianni et al., 2011). By proposing a neural-swarm layered architecture, this paper contributes to bridging the gap between the newborn swarm cognition framework and well-stablished cognitive frameworks of neural basis.

This paper is organised as follows. Section 2 overviews related work in path detection models and swarm-based computer vision in general. Then, after presenting the the neural-swarm architecture in Section 3. Subsequently, in Section 4, a previous model for path following is presented as an instance of the neural-swarm architecture. Then, in Section 5, a set of extensions to this model are proposed for improved performance and robustness. Finally, Section 6 reports a thorough set of experimental results, and Section 7 draws some conclusions and relevant future work avenues.

2 Related Work

Current path detection methods rely considerably on work developed for ill-structured unpaved rural roads. The typical solution in this case is to segment the road region from its surroundings by considering the aggregate of pixels whose likelihood of belonging to the road surface is above a given threshold. This likelihood can either be learnt off-line (Chaturvedi and Malcolm, 2005; Alon et al., 2006) or on-line in a self-supervised way (Thorpe et al., 1988; Fernandez and Casals, 1997; Fernandez and Price, 2005; Song et al., 2007; Thrun et al., 2006; Tue-Cuong et al., 2008; Lookingbill et al., 2007). In general, a simplified model of the road (e.g., trapezoidal) is fit to the segmented image. Region growing is an alternative to the model fitting process for less structured roads (Ghurchian et al., 2004; Fernandez and Price, 2005; Chaturvedi and Malcolm, 2005). By enforcing a global shape constraint, the model-based approach enables the substitution of the road/non-road pixel classification process by an unsupervised clustering mechanism (Crisman and Thorpe, 1991).

These models are the basis of most work on path detection, being the case of trail detection the most demanding. An example is the use of a priori knowledge about the colour distributions of both trail and surroundings for their segmentation (Bartel et al., 2007). Robustness can be

increased if these a priori models are substituted by models learnt on-line (Grudic and Mulligan, 2006; Rasmussen and Scott, 2008b). In contrast to the road domain, the definition of the reference regions from which it is possible to supervise the learning process is not easy. With varying width and orientation, it is difficult to assure that the robot is on the trail, and from that, which regions of the input image can be used as reference. Second, the trail and its surroundings often exhibit the same height, which hampers a straightforward use of depth information to determine a trail reference patch. The use of a global shape constraint (e.g., triangular) to avoid the learning process has also been tested in the trail detection domain (Rasmussen and Scott, 2008a; Blas et al., 2008). This is done by over-segmenting the image, creating sets of segments, and then scoring them against the global shape constraint. Accurate image over-segmentation is a computationally demanding task and usually requires clear edges segmenting the object from the background. Moreover, global shape constraints limit the type of trails that can be detected.

By exploiting the visual saliency of trails and narrow paths in general, the proposed model circumvents the cost of explicitly over-segmenting the input image and the brittleness of depending on accurate appearance and shape models. Rasmussen et al. (2009) proposed the use of local appearance contrast for trail detection, which only superficially resembles the concept of visual saliency. In fact, robust visual saliency should include contrast information between trail and overall scene, besides local contrast information. This is important because it is not guaranteed that the appearance of trails and their immediate surroundings always exhibit sufficient contrast to be robustly exploited. Furthermore, that work still relies on an explicit global shape constraint.

There is already a long line of research on the use of the social insects metaphor for the design of computer vision systems (Poli and Valli, 1993; Liu et al., 1997; Ramos and Almeida, 2000; Owechko and Medasani, 2005; Antón-Canalís et al., 2006; Mobahi et al., 2006; Broggi and Cattani, 2006; Mazouzi et al., 2007; Zhang et al., 2008; Tian et al., 2008; Ma et al., 2010). The most related to ours is the work of Broggi and Cattani (2006), which detects the edges of ill-structured desert roads with a swarm-based system. However, paths in natural environments seldom are delimited from the background by strong edges, which is the reason why a region-based approach is preferred. Furthermore, operating on the appearance space directly, and not on the conspicuity space, the work of Broggi and Cattani (2006) does not exploit the observation that narrow paths are conspicuous structures in the environment. By following a region-based approach and by operating on the conspicuity space, our approach is better suited for the problem of path detection in natural environments.

In parallel with our preliminary swarm-based model for saliency computation (Santana et al., 2010a), an ACO model for edge-based saliency computation was proposed by Ma et al. (2010). Two major differences exist between the two models. First, differently from our model, the model proposed by Ma et al. is focused on edge detection, which is of little use for the problem of path detection in uneven natural terrain. Rarely these paths are surrounded by strong edges. The second major difference is magnified with the extensions proposed in this paper. Namely, while the the work of Ma et al. is purely bottom-up, our model exploits a priori and learned top-down knowledge of the object being sought to bias the bottom-up saliency computation process. Without this ability, which is widely recognised as key for visual attention

deployment, appearance and shape information cannot be used to discriminate the object being sought from distractors scattered in the environment.

3 Neural-Swarm Architecture

This section proposes a neural-swarm architecture for visual saliency computation. Rather than static structures, like neurons, agents composing a swarm are better viewed as active information particles that flow and change in the system (Santana and Correia, 2011). Hence, using agents, the design focus is on the process and not so much on its supporting substrate. These information particles are sensorimotor coordinated units and, thus, capable of exploiting the recognised benefits of active perception (Ballard, 1991), such as actively selecting and shaping their sensory input to increase pose invariance, signal-to-noise ratio, and discriminative power (Beer, 2003). Furthermore, agents' activity is a function of all the local context sensed by them along their trajectories. Hence, the agent design allows an active and, thus, parsimonious, spatio-temporal anisotropic integration of local contexts. To attain such a complexity, a neural-based counterpart would require highly dense and intricate connectivity. The outcome would be a system with low modularity and, consequently, of limited scalability. Conversely, when information processing requires highly dense isotropic local connectivity, then the modularity offered by a neural-based design is adequate.

In the context of computer vision systems, the above considerations result in the following conclusions. The neural-based paradigm is better suited to realise low-level feature extraction (e.g., conspicuity information) as this demands mostly for massively parallel application of local image operators. Conversely, the swarm-based paradigm is well suited to realise high-level feature extraction (e.g., shape information) from the low-level features, as this demands mostly complex spatio-temporal integration of local perceptual information. As a result, we propose that these complementary properties of both computational paradigms are better exploited when layered.

Fig. 2 illustrates the proposed architecture in the context of visual saliency computation. In this case, we propose the existence of several environment-swarm tuples, one per visual feature (e.g., colour or intensity). The goal of having several pathways is to focus the processing elements on a reduced perceptual space. By allowing lateral connectivity between pathways, cross-feature influences are maintained in the system, which is important to cope with noisy sensory input and feature-specific perceptual aliasing.

The environment on which the swarm operates is composed of two main lattices (see Fig. 2), both locally available to the virtual ants. The first lattice is a conspicuity map pin-pointing the regions of the visual field that are more likely to belong to the object being sought, given a set of low-level feature maps. The second lattice is implemented with a dynamical neural field, which is a two-dimensional (2-D) lattice of neurons with "Mexican-hat" shaped lateral coupling (Amari, 1977). This neural field serves the purpose of simulating the physical medium on which pheromone is deposited. With leaky integrator neurons, this neural network is able to emulate pheromone evaporation. With local excitatory connections and regional inhibitory

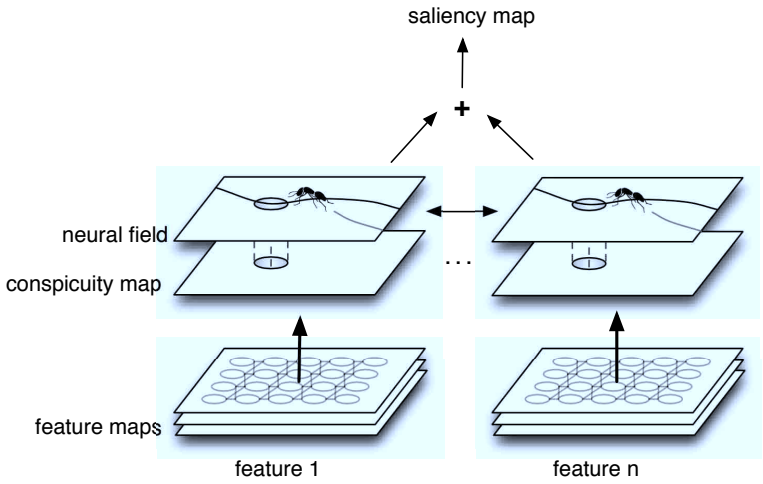


Figure 2: The neural-swarm visual attention architecture. Per visual feature, a set of feature maps are aggregated to produce a feature conspicuity map. This plus a dynamical neural field operate as the environment for the swarm. That is, the swarm senses both conspicuity and neural field activity. It also affects the neural field by depositing pheromone on it. The neural-based nature of the feature maps is represented by the connected circles. The black and the grey curves in the neural field represent a strong pheromone trail deposited by the swarm and the weak pheromone trail deposited by the single virtual ant represented in the picture, respectively. The circles in both neural field and saliency map represent the virtual ant's sensory input. The two-way arrow represents the lateral connectivity between pathways, which can be implemented by means of partial pheromone sharing.

connections, this inter-neuron coupling helps in the formation of a coherent focus of attention (Rougier and Vitay, 2006). In a sense, this lateral connectivity implements isotropic spatial interactions among the pheromone deposited by the several virtual ants. In sum, it helps on the formation of coherent pheromone trails.

The virtual ants in the swarm operate according to a set of stochastic perception-action rules embodying a priori knowledge of the expected object's shape. Each of these perception-action rules has the purpose of inducing the virtual ants to produce trajectories that approximate the object's skeleton, given the low-level features locally available to them and a priori knowledge of the object's shape or appearance. These rules are detailed in Section 4.1. That is, each virtual ant moves with the purpose of grouping a set of pixels that approximates the object's skeleton. This way, the skeleton is defined in a non-parametric way, which is key to deal with not so well behaved objects. Pheromone-based interactions are included to allow a progressive convergence of the virtual ants' semi-exploratory behaviour.

The motion of the virtual ants reflects their history, that is, the local context sensed across their trajectories. This context encompassing both sensory input and other virtual ants activity, through the deposited pheromone, implicitly represents a highly intricate connectivity. As already mentioned, explicitly handling such a connectivity in a neural network would be cum-

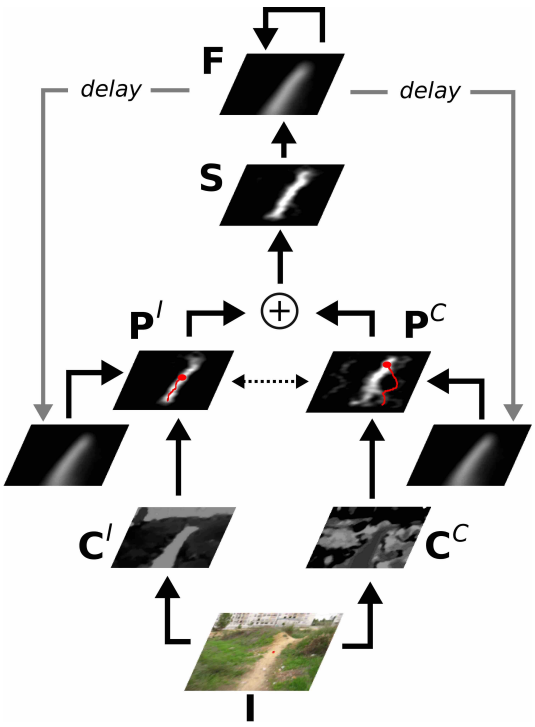


Figure 3: Major structures produced within the neural-swarm path detector (Santana et al., 2010a, 2012)

bersome. Finally, the superposition of all neural fields results in a saliency map, whose highest activity spot corresponds to the most likely location of the object being sought.

4 A Realisation of Path Detection and Tracking

To show the practical plausibility of the architecture proposed in Section 3, this section overviews the neural-swarm model proposed by Santana et al. (2010a, 2011, 2012) as an instance of the architecture for the problem of path detection and tracking.

The model (see Fig. 3) starts by computing feature-dependent conspicuity maps, based on the well-known bio-inspired model of Itti et al. (1998). Concretely, a colour (C) conspicuity map, $\mathbf{C}^{\mathbf{C}}$, and an intensity (I) conspicuity map, $\mathbf{C}^{\mathbf{I}}$, are created based on a set of centre-surround feature maps obtained from the input image at various spatial scales (see Section 5.3 for details). These maps signal the regions of the visual field that detach more from the background on the visual feature in question, such as colour or intensity. Exhibiting high density local isotropic connectivity, they correspond to a realisation of the lower layer of the neural-swarm architecture.

The regions of the conspicuity maps with highest activity should in principle signal the presence of the path. In practise, this is a brittle assumption in the face of not so well behaved

conspicuity maps, which is the case in the presence of distractors or when the path is considerably heterogeneous. Hence, determining the location of the path is much more complex than a simple winner take all process could explain. That is, it is necessary to extract the path's skeleton by aggregating pixels in a non-trivial spatio-temporal way. This process exhibiting global spatio-temporal complex connectivity, it corresponds to a realisation of the upper layer of the neural-swarm architecture. For this purpose, the conspicuity maps are used as sensory input for two swarms of virtual ants that collectively create a colour pheromone map, $\mathbf{P}^C$, and an intensity pheromone map, $\mathbf{P}^I$. These maps represent the self-organised consensus reached by the virtual ants about the best approximation to the actual path's skeleton. To improve robustness, pheromone is accumulated across frames in a dynamical neural field $\mathbf{F}$, which implicitly implements pheromone evaporation and spatial competition. According to the neural-swarm layered architecture, this neural field implementing high density local isotropic connectivity, it is also a realisation of the lower layer of the architecture. To decouple the dynamics of the neural field and the one of the robot, the homography matrix that describes the projective transformation between the current and the previous frames is estimated and used to motion compensate the neural field.

4.1 Pheromone Maps Creation

To build the pheromone maps at each new frame, n virtual ants are sequential deployed and associated, alternatively, to the two visual features. The following overviews the execution of a given virtual ant p_m, associated to one of the visual features, $m \in \{I, C\}$. The other visual feature is represented by m'.

While being iterated for η times, p_m will move on $\mathbf{C}^m$, influenced by the pheromone present in $\mathbf{P}^m$ (see below). While moving, this virtual ant deploys pheromone in each position visited in $\mathbf{P}^m$ with a magnitude $\Phi(p_m, \mathbf{h_{ref}})$, and a small portion of it, υ, in $\mathbf{P}^{m'}$:

$$\Phi(p_m, \mathbf{h_{ref}}) = \varepsilon + \beta \cdot p(T|V_{p_m}, \mathbf{h_{ref}}) \tag{1}$$

where β is a weighting factor, ε is a pheromone level baseline, both empirically defined, and $p(T|V_{p_m}, \mathbf{h_{ref}})$ is the probability of the virtual ant's path, V_{p_m}, to belong to the path (T), given a learned online appearance model $\mathbf{h_{ref}}$. Rather than having virtual ants to deploy a constant level of pheromone along their paths, this approach compels virtual ants deploying higher doses of pheromone on regions of the visual field whose appearance is more similar to the one of the path.

The probability $p(T|V_{p_m}, \mathbf{h_{ref}})$ is approximated by the average probability of pixels, visited by the virtual ant, belonging to the path. These pixels are represented by the set V_{p_m}, and their individual probabilities are obtained directly from the normalized histogram $\mathbf{h_{ref}}$ learned online (see Section 5.4), according to a technique known as histogram back-projection.

The chances of creating a virtual ant p_m on a given location of the visual input is defined as a function of the levels of conspicuity and pheromone at that location. As a consequence, virtual ants are progressively and probabilistically deployed where there are more chances of

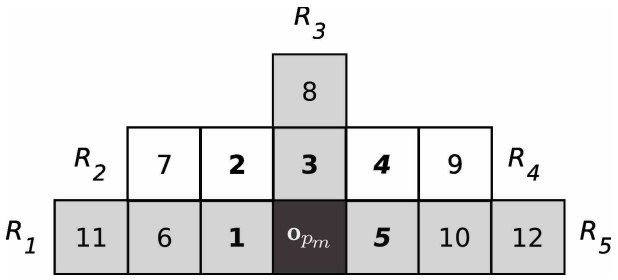


Figure 4: Virtual ant's sensory and action spaces. Space is discretised in pixels and only the ones that the ant is able to perceive are represented. Regions surrounding current virtual ant's position, $\mathbf{o}_{p_m}$, are segmented into a set of receptive fields, $R_1 = \{1,6,11\}, R_2 = \{2,7\}, R_3 = \{3,8\}, R_4 = \{4,9\}, R_5 = \{5,10,12\}$, whose composing pixels are numbered as in the figure. If a given action $a \in A$ is selected, then the next virtual ant's position will be the closest pixel to the virtual ant, represented by the pixels in bold.

being a path, under the assumptions that: (1) paths tend to be conspicuous; (2) the path has been successfully detected in the previous frame (represented by the feedback provided by the delayed neural field state); and (3) that the pheromone accumulated by virtual ants deployed in the current frame builds-up mostly around the actual path's location. Nevertheless, virtual ants are compelled to start their execution from the bottom of the visual input.

The virtual ant is controlled by a set of behaviours that exploit a priori top-down knowledge about paths' overall structure, such as the fact that they converge towards a vanishing point. Before specifying its behaviours, it is necessary to specify its sensory and action spaces. To reduce both sensitivity to noise and computational cost, the sensory input is defined by five coarse receptive fields disposed around the virtual ant's current position, $R_1 \ldots R_5$ (see Fig. 4). Let us now define $\mathbf{C}^m(R_k, \mathbf{o}_{p_m})$ and $\mathbf{P}^m(R_k, \mathbf{o}_{p_m})$ as returning the average conspicuity and pheromone levels of the pixels constituting receptive field R_k, respectively. Parameter $\mathbf{o}_{p_m}$, which represents the virtual ant's position, is used to transform the virtual ant's centred receptive field onto the map's frame of reference. An action $a \in A$ moves the virtual ant to one of the five neighbour pixels not behind the current virtual ant's position. The action space is thus defined by the set $A = \{1,2,3,4,5\}$ (see Fig. 4).

The goal of a virtual ant's behaviour, from the set of behaviours B, is to distribute votes in the interval $[0,1]$ for each possible action $a \in A$, according to a set of local perceptual cues. For instance, the assumption that paths are usually conspicuous structures in the environment is exploited with the following behavioural function:

$$f_{greedy}(p_m, a) = \mathbf{C}^m(R_a, \mathbf{o}_{p_m}). \tag{2}$$

This behavioural function is maximal for the action that takes the virtual ant towards the immediate region with highest conspicuity level. That is, maximising this function over the action set leads to a greedy exploitation of the conspicuity cue to reconstruct the path's skeleton. Other behaviours, not herein formalised for the sake of space, are responsible for driving the

virtual ants towards regions of high conspicuity under the assumption that paths are salient in the input image, regions whose average level of conspicuity is more similar to the average level of conspicuity of all the pixels visited by the virtual ant under the assumption that paths' appearance is homogeneous, regions that maintain the virtual ant equidistant to the boundaries of the path hypothesis being pursued, and regions targeted by the motor action at the previous iteration under the assumption that paths' orientation tend to be monotonous.

The actual action to be engaged by virtual ant p_m, a_{p_m}, is obtained with the following utility function, which incorporates behaviors' votes, pheromone-based interactions, and random fluctuations:

$$a_{p_m} = \arg\max_{a \in A} \left(\sum_{b \in B} \alpha_b f_b(p_m, a) + \mathbf{P}^m(R_a, \mathbf{o}_{p_m}) + \gamma q \right), \tag{3}$$

where α_b is a user defined weight accounting for the contribution of behavior $b \in B$ and γ is the weight accounting for stochastic behavior, being $q \in [0,1]$ a number sampled from a uniform distribution each time the action is evaluated.

The self-organisation of virtual ants trails around the path's skeleton is a result of the trade-off between their exploratory behaviours, their pheromone-based interactions, presence of random fluctuations at their actions, and the structure of the environment. Basically, virtual ants deployed on path regions tend to converge as a result of a successful exploitation of the local structure by the behaviour set. Conversely, random fluctuations tend to dominate when virtual ants are deployed off path and, thus, compel them to disperse. The presence of the path tends to be a global constraint which is only felt by the virtual ants deployed on it. In other words, the path operates as an attractor for the self-organising system. By making virtual ants attracted by high pheromone concentration regions, the difference between diverging and converging virtual ants is reinforced, which helps breaking symmetries. This ensures that, along time, the structure imposed by the presence of the path on the behaviours is stronger than the effects of random fluctuations.

The collective operation of the virtual ants implements a shape-based filter over the conspicuity maps, whereas the neural field implements a spatio-temporal smoothing filter. The outcome is a method that detects paths according to the expectation that they are salient and exhibit a given approximate shape. The synergistic interaction between both bottom-up and top-down pathways, or between neural-based and swarm-based connectivity, reduces the dependency on accurate path appearance and shape models without hampering computational parsimony. As a result, robustness to sudden changes in the path is higher and computation is saved.

The just described system's basic structures and processing flow, represented in Fig. 3, denote its cascaded nature, similar to the ones found in typical computer vision pipelines. Concretely, the cascade is composed of low-level features processing, implemented by the conspicuity maps, shape-based higher-level features extraction, implemented by the swarms, and temporal filtering, implemented by the dynamic neural field. By modulating the swarms with the neural field's state, tracking is also included in the processing flow. Therefore, this model shows the functional equivalence between the neural-swarm architecture and typical computer

vision pipelines. That is, the neural-swarm architecture is able to provide the same level of functionality as a typical computer vision solution, with the advantage of being able to deliver it on a fully parallel and self-organising way.

5 Extending Top-Down Influence

This section proposes a set of extensions to improve the accuracy and the robustness of the neural-swarm path detector described in the previous section.

Fig. 5 illustrates the model with the herein proposed extensions related to the top-down modulation of the saliency computation process. As mentioned, these help the virtual ants disambiguate in situations where the conspicuity information is insufficient by itself. In this work, shape, appearance, and contrast models are used as top-down knowledge of the path. As results will show, the joint operation of these three sources of information is sufficient to ensure safe tracking of the path.

5.1 Shape-based top-down influence

The shape-based top-down influence is, as explained in the previous section, implicitly specified in the form of behavioural rules executed by the virtual ants (Santana et al., 2010a, 2012). Hence, this knowledge is taken as innate and, thus, not affected by learning. Conversely, the appearance model (see Section 5.2) and the contrast model (see Section 5.3) are learned on-line (see Section 5.4).

5.2 Appearance-based top-down influence

As in previous work (Santana et al., 2011, 2012), the appearance model is defined by a three-dimensional normalised 8-bit $16 \times 16 \times 16$ colour histogram, $\mathbf{h_{ref}}$ (see Section 4.1). The model is used to promote the deployment of pheromone on regions of the image whose appearance is more likely to belong to the one of the path. This paper extends the use of this model in two ways. First, the $c_1c_2c_3$ colour space is used, instead of the HSV colour space. This option follows the well-known shadow and illumination invariance of the $c_1c_2c_3$ colour space in outdoor environments (Gevers and Smeulders, 1999; Song et al., 2007). Fig. 6 shows an example in which this property is evident. Fig. 7 shows that, even in the absence of shadows, the $c_1c_2c_3$ colour model is more robust for the task at hand. Second, besides helping virtual ants to know where to deploy more pheromone, the back-projection of the appearance model is also directly superposed onto the conspicuity maps. The outcome are cleaner conspicuity maps, which helps virtual ants to track the path's skeleton in a more robust way (see Fig. 8). Although useful, the appearance model alone is unable to segment the path when its appearance suffers a sudden change, as in the situation depicted in Fig. 9.

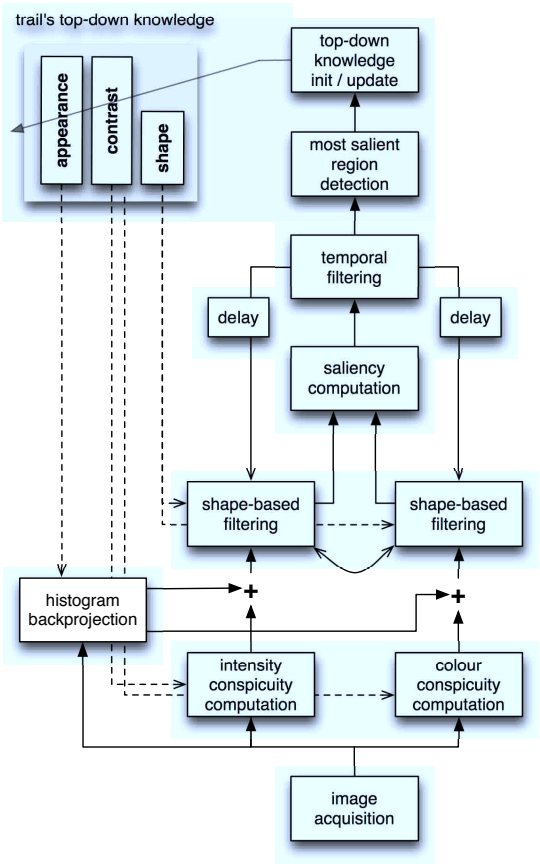
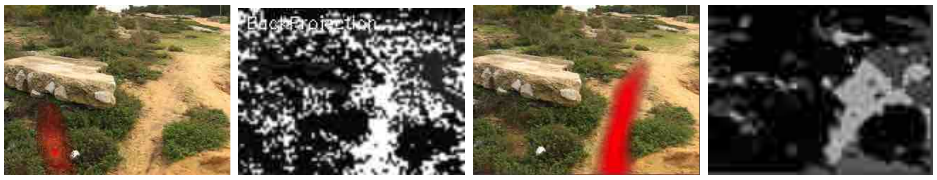


Figure 5: Proposed model's pipeline. The model starts by computing two conspicuity maps from the input image, one encompassing intensity information and another encompassing colour information. After the initialisation phase, these maps are biased by the top-down contrast model and fused with a probability map obtained by back-projecting an on-line learned appearance model. The resulting maps are subsequently shape-based filtered by two swarms of virtual ants, whose behavioural rules represent the top-down knowledge of the overall path's typical shape. These swarms operate over pheromone maps, which are initialised with the neural field's activity. Cross-modality, represented by the two-way arrow, are maintained by allowing swarms to share pheromone. Top-down appearance knowledge is used in this phase by modulating the level of pheromone deployed by the virtual ants. The output of the shape-based filter is a set of two maps that, when fused together, generate a saliency map. This latter map feeds the dynamical neural field, which performs temporal filtering. Finally, the most salient region is obtained as a mask to constrain the learning of both appearance and contrast models.

(a) RGB channels (b) c_1 channel (c) c_2 channel (d) c_3 channel

Figure 6: *RGB* and $c_1c_2c_3$ colour models in the presence of strong shadows.

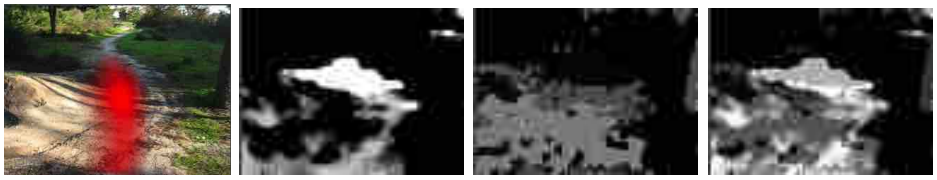


(a) (b) (c) (d)

Figure 7: Comparison between back-projections computed from *RGB*-based and $c_1c_2c_3$-based appearance models. (a) Input image with the system's output overlaid, based on the *RGB*-based appearance model, whose back-projection is depicted in (b). (c) Input image with the system's output overlaid, based on the $c_1c_2c_3$-based appearance model, whose back-projection is depicted in (d).

5.3 Contrast-based top-down influence

Typically, visual saliency maps are obtained by aggregating several maps of contrast visual features (Itti et al., 1998). The presence of a dark region on a bright background at a given scale is an example of a contrast visual feature. However, contrast alone is ambiguous when the environment is populated with several distractors. To reduce ambiguity, the relative importance of each contrast visual feature to the final saliency map can be made a function of a priori knowledge about the object being sought (Frintrop et al., 2005; Navalpakkam and Itti, 2005) (e.g., the object is supposedly brighter than the background). The following describes an extension to



(a) (b) (c) (d)

Figure 8: Typical situation in which the superposition of the conspicuity map and the appearance model's back-projection produces a map with a reduced number of gaps, thus facilitating the swarm operation. (a) Input image with system's output overlaid in red. (b) Conspicuity map of (a). Appearance model's back-projection of (a). Superposition of (b) and (c).

(a) input (b) back-projection (c) conspicuity

Figure 9: Typical situation in which the back-projection of the appearance model, depicted in (b), is insufficient to segment the path from the background. Conversely, the combination of both top-down modulated conspicuity maps, depicted in (c), is able to accurately localise the path in the input image. The failure of the appearance model owes mostly to the sudden appearance of a bridge along the path.

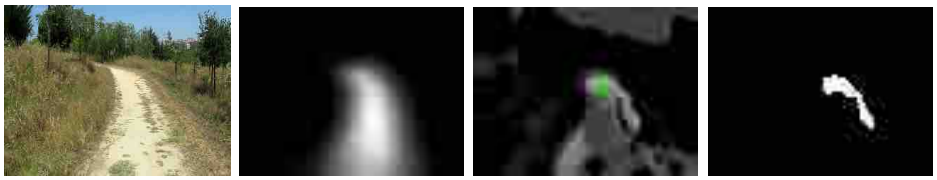
the basic neural-swarm path detector (see Section 4) so that a priori contrast knowledge can be learned and used online.

The conspicuity maps computation process (Santana et al., 2010a, 2012) is based on the bio-inspired model proposed by Itti et al. (1998). This method starts by computing, from the intensity channel, one dyadic Gaussian pyramid (Burt and Adelson, 1983) with eight levels. Two additional pyramids, also with eight levels, are computed to account for the Red-Green and Blue-Yellow double-opponency colour feature sub-channels. Each level corresponds to a given scale. Various scales are then used to create a set of on-off and off-on centre-surround maps per pyramid (Itti et al., 1998). These have higher intensity on those pixels whose corresponding feature differs the most from their surroundings. On-off centre-surround maps are built 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, all centre-surround maps of a given kind, i.e., on-off or off-on, built from the intensity pyramid are resized to a common size, independently scaled in magnitude with the method proposed in Santana et al. (2010b), and finally averaged together to produce an aggregate intensity centre-surround feature map. Then, both on-off and off-on aggregate intensity centre-surround feature maps are scaled and averaged together to produce an intensity conspicuity map. The same process applies to create Red-Green and Blue-Yellow conspicuity maps, which are then averaged together to produce a single colour conspicuity map.

For the extended neural-swarm path detector, the contrast model is defined as a vector of six elements $\mathbf{w}$, each representing the importance weight a given aggregate centre-surround feature map has to the detection of the path. These weights are updated (description in Section 5.4) and used in each frame to override the scaling function applied by default to the aggregate centre-surround maps. Fig. 10 depicts a typical situation in which top-down contrast modulation is key for a proper path detection.

(a) (b) (c) (d)

Figure 10: Typical situation where top-down contrast modulation is key for a proper path detection. (a) Input image with most likely path location overlaid, using top-down contrast modulated conspicuity maps (b). (c) Input image with most likely path location overlaid, using conspicuity maps without top-down contrast modulation (d). The specific environmental configuration results in an inversion of the conspicuity maps, attracting virtual ants to the sides of the path. By including top-down contrast knowledge, this problem is overcome.



(a) input image (b) neural field (c) combined conspicuity maps (d) learning mask

Figure 11: Definition of input image's mask for appearance and contrast models learning. The mask is defined with a region growing process that proceeds on the map combining both conspicuity maps. The process starts on the pixel associated to the highest neural field's activity, represented by the green diamond in (c).

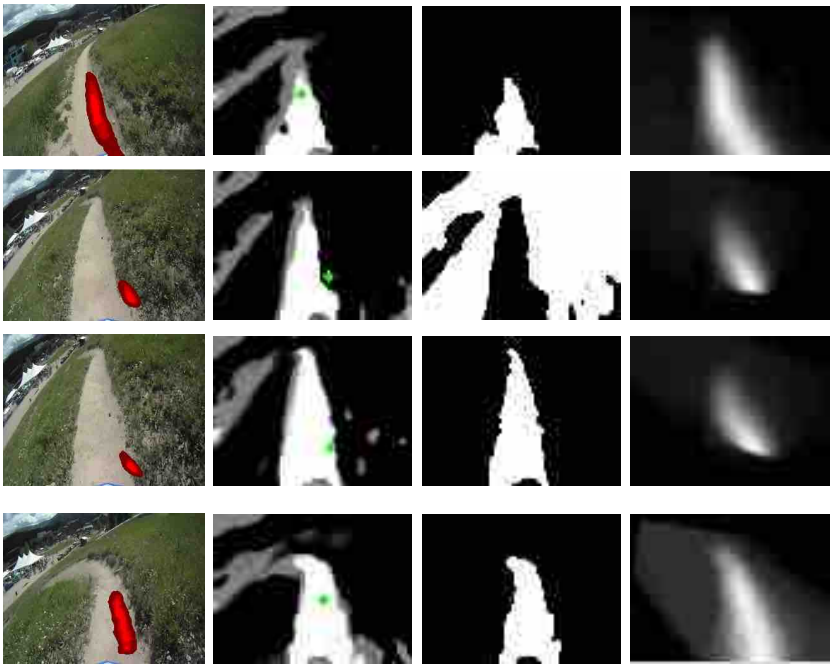
5.4 Learning contrast and appearance models

To learn both appearance and contrast models, it is necessary to specify, in each frame, the region of the input image that corresponds to the most likely location of the path. This is done by searching for the most salient region in the neural field, whose location is used as a seed for a region growing process. Region growing proceeds on a temporary map computed as the average of both colour and intensity conspicuity maps (see Fig. 11). The result is a mask that is provided to both appearance and contrast learning processes.

To update vector the contrast model $\mathbf{w}$ (see Section 5.3), a weight vector $\mathbf{w}'$ is learned from the current frame's masked region, according to the method proposed by Frintrop et al. (2005). Basically, the method sets higher weights to maps that positively correlate with the most likely location of the object. This way, maps are promoted according to their relevance to the object. To cope with noisy data, the adaptation of $\mathbf{w}$ is formulated as

$$\mathbf{w}(t) = (1-\beta)\,\mathbf{w}(t-1) + \beta\,\mathbf{w}'(t), \tag{4}$$

where β is the learning rate. This equation implements a temporal smoothing filter, which is im-



(a) (b) (c) (d)

Figure 12: Momentary path mis-tracking as a result of a faulty motion compensation. (a) Input image with most likely path location overlaid in red. (b) Map combining both conspicuity maps. The diamond corresponds to the location associated to the neural field's highest activity. (c) Learning mask obtained by a region growing process, whose seed is represented by the diamond in (b). (d) Neural field. The second row shows the neural field being erroneously projected off path, which causes a mis-localisation of the seed and, as a result, of the mask for the contrast model update. The inertia of the model attenuates the effect of this erroneous update, which allows the swarm to rapidly converge back to the path.

portant to avoid learning the background in cases of faulty neural field's motion compensation. One of these situations is depicted in Fig. 12. The figure also shows the ability of the system to rapidly recover, i.e., to adjust the neural field. A similar smoothing filter is used to update the normalised histogram $\mathbf{h_{ref}}$ (see Section 5.2) from the histogram of the current frame's masked region (Santana et al., 2010a, 2012).

To avoid using the appearance and contrast models while these are not statistically relevant, an initialisation phase is considered. During this phase the system operates only under shape-based top-down influence. In the current implementation, this phase lasts for 50 frames, which is empirically shown to be sufficient for the system to converge. Fig. 13 illustrates the benefit of using top-down modulation to enforce the pop-out of the path region, even in the presence of strong distractors.

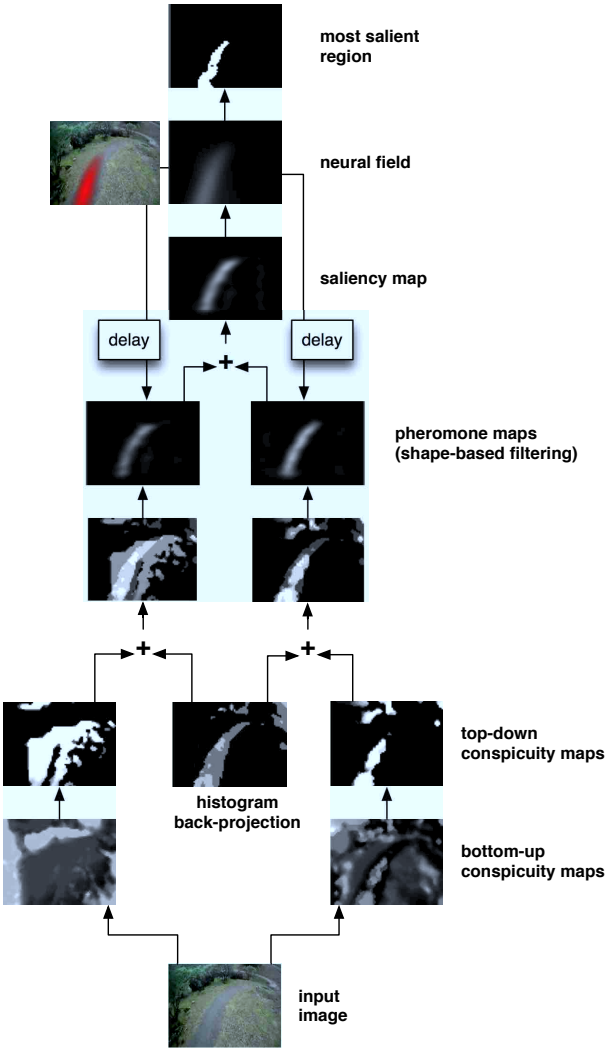


Figure 13: System's snapshot on a typical situation. Due to the environment's specific configuration, the bottom-up conspicuity maps are unable to unambiguously signal the presence of the path. By modulating the conspicuity process with the contrast model learned so far, path-background discrimination improves considerably. It is also possible to see that the back-projection of the histogram used as appearance model provides a discriminatory map that helps the swarms in performing the shape-based filtering. The good quality of the appearance model reflects the good correlation between the most salient region of previous frames and the actual location of the path. The image on the left of the neural field corresponds to its overlay on the input image.

	Conspicuity maps computation	Swarms execution	Neural field update	Total
Time (ms)	18	33	2	53

Table 1: Average computation times. The timing reported for the neural field update also includes motion estimation and neural field's corresponding transformation.

6 Experiments

To study the impact of the proposed extensions to the neural-swarm path tracker, a comprehensive data-set of 39 colour videos, encompassing a total of 29,789 frames with 640×480 resolution, was used. A sub-set of 28 of these videos were recorded with a hand-handled camera at an approximate height of 1.5 m and carried at an approximate speed of $1\,\mathrm{ms}^{-1}$. The remaining 11 videos were acquired from bike-mounted cameras moving at various speeds. In opposition to the first sub-set, the second one was downloaded from YouTube, meaning that, the videos were acquired from cameras with different sensors and field of views. Furthermore, the data-set includes both natural and engineered paths in a wide variety of backgrounds. Hence, to be successful, the proposed method needs to be able to perform in a wide range of camera-environment configurations.

The path is considered correctly detected if the biggest blob of neural field activity (above 85 % of its maximum) is fully localised within the path's boundaries and roughly aligned with its orientation. Such an output should be sufficient for visual servoing a robot at moderate speeds along the path. Complementarily, the whole neural field's activity can be taken as an approximation of the path/background segmentation. Such representation may be useful for detailed motion planning. Although a quantitive analysis is not provided, the obtained qualitative results support this possibility.

The proposed model was implemented as single-core C++ code and tested on a Pentium(R) Core2 Duo 2.53 with 4 Gb of RAM, running a 64-bit Linux Ubuntu 10.10 (Maverick Meerkat). OpenCV 2.3[1] (Bradski and Kaehler, 2008) was used for low-level computer vision routines. An open source implementation of the model, fully encapsulated as a node of the Robotics Operating System (ROS)[2] (Quigley et al., 2009), is available for download from the authors website (Santana, 2011). The model's free parameters presented in Section 4.1 were set as in (Santana et al., 2012). As reported in Table 1, with this configuration, the model performs at 20 Hz. At this processing rate, the model shows to be computationally efficient and, thus, sufficient for smooth path following.

Table 2 presents the detection rate for each video, averaged over five independent runs. Multiple runs are required only for the sake of the statistical analysis of the model, and not

[1] OpenCV: `http://opencv.willowgarage.com`

[2] ROS: `http://www.ros.org`

for on-line execution. The table shows that the proposed extensions improve the model in all videos, resulting on an overall success rate of 98.67 %. The 5 % improvement is significant given the difficulty of improving a system already performing near the optimal. Fig. 14 illustrates the model's output in key frames of the tested data-set. Note the ability of the model to handle interrupted paths, path intersections, and paths whose orientation does not change monotonically. These good results owe considerably to the option of not trying to fit parametric path models to the data; instead, the system produces in a bottom-up way non-parametric path hypotheses, which are progressively refined by means of self-organising behaviour. The videos with the results overlaid are available for download from the authors website (Santana, 2011).

Video ID	Nr of Frames	Original model's detection rate (%)	Extended model's detection rate (%)
1	278	100.00 ± 0.00	100.00 ± 0.00
2	204	100.00 ± 0.00	100.00 ± 0.00
3	422	100.00 ± 0.00	100.00 ± 0.00
4	135	100.00 ± 0.00	100.00 ± 0.00
5	2854	96.29 ± 0.03	100.00 ± 0.00
6	186	100.00 ± 0.00	100.00 ± 0.00
7	121	100.00 ± 0.00	100.00 ± 0.00
8	124	100.00 ± 0.00	100.00 ± 0.00
9	301	93.33 ± 0.39	97.35 ± 0.13
10	147	99.18 ± 1.09	100.00 ± 0.00
11	386	100.00 ± 0.00	100.00 ± 0.00
12	158	95.44 ± 0.25	100.00 ± 0.00
13	134	100.00 ± 0.00	100.00 ± 0.00
14	676	98.02 ± 0.07	100.00 ± 0.00
15	683	93.82 ± 0.06	100.00 ± 0.00
16	770	90.99 ± 0.10	97.45 ± 0.06
17	403	81.04 ± 0.20	100.00 ± 0.00
18	335	100.00 ± 0.00	100.00 ± 0.00
19	230	98.70 ± 0.27	100.00 ± 0.00
20	439	95.81 ± 0.18	100.00 ± 0.00
21	490	100.00 ± 0.00	100.00 ± 0.00
22	230	100.00 ± 0.00	100.00 ± 0.00
23	600	100.00 ± 0.00	100.00 ± 0.00
24	802	99.05 ± 0.10	100.00 ± 0.00
25	907	93.65 ± 0.05	100.00 ± 0.00
26	1553	90.96 ± 0.05	100.00 ± 0.00
27	3011	97.51 ± 0.02	99.47 ± 0.04
28	1288	96.77 ± 0.06	100.00 ± 0.00
29	267	78.95 ± 0.15	96.25 ± 0.24
30	440	82.18 ± 0.18	84.82 ± 0.36
31	1027	67.19 ± 0.12	99.51 ± 0.06

Continued on next page

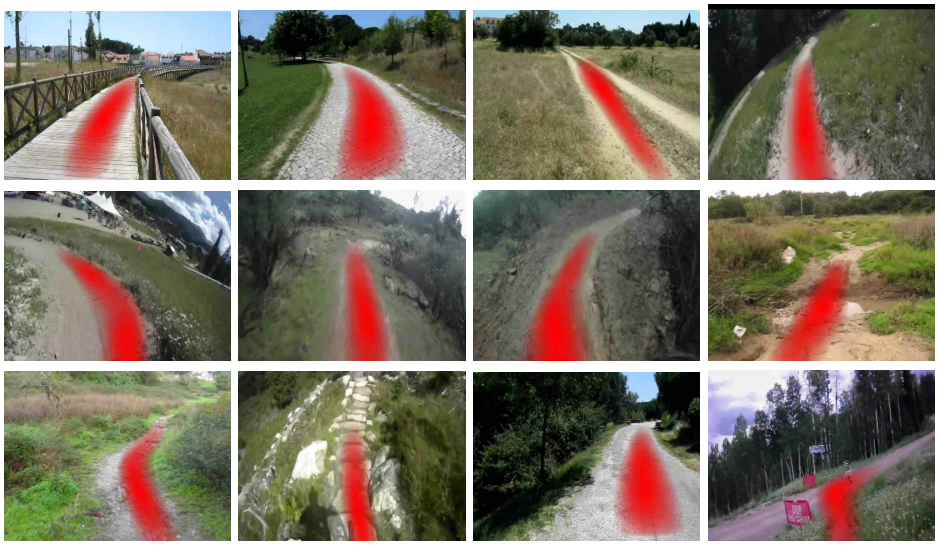


Figure 14: Representative frames of the tested data-set with most likely path location overlaid in red.

Video ID	Nr of Frames	Original model's detection rate (%)	Extended model's detection rate (%)
32	1083	91.51 ± 0.00	89.33 ± 0.07
33	1649	95.17 ± 0.08	92.15 ± 0.10
34	591	57.09 ± 0.08	97.33 ± 0.20
35	388	95.41 ± 0.10	97.42 ± 0.23
36	2515	89.69 ± 0.06	100.00 ± 0.00
37	429	85.78 ± 0.15	100.00 ± 0.00
38	829	95.34 ± 0.06	96.91 ± 0.10
39	2696	94.24 ± 0.03	100.00 ± 0.00
Overall	29,789	93.67 ± 0.10	98.67 ± 0.04

Table 2: Comparison between the original (Santana et al., 2010a, 2012) and the extended path detectors, over 39 videos. Results per video (mean $\pm$ standard deviation) refer to the average of 5 runs. The results in the last row refer to the aggregate detection rate (mean $\pm$ standard deviation), computed as the average of the detection rates obtained over all videos.

To the best of our knowledge, no previous work has been tested against a data set with paths simultaneously as narrow, unstructured, and discontinuous as the ones herein considered. It is worth noting that in 28 of the 39 videos, the proposed model shows 100 % success rate across the 5 five runs. Videos 5, 27, 36, and 39 are accounted as long runs, above 4 minutes length, and are composed of more than 2500 frames. Along these videos, the paths are often interrupted, highly unstructured, and of variable width. Moreover, the terrain surrounding the pay is heterogeneous and highly populated with strong distractors, such as trees and bushes.

The 100 % success rate of the model in these videos (except for video 27, which has 99.47%) clearly shows its robustness in demanding situations.

7 Conclusions

Envisioning robust and efficient robot vision systems, this paper proposes a neural-swarm layered architecture for visual saliency computation. The architecture encompasses a swarm-based substrate for the extraction of high-level perceptual representations, given the low-level perceptual representations extracted by a neural-based substrate. This task allocation follows the rationale that neural networks exhibit a granularity level adequate for dense, local, and isotropic spatio-temporal processing, whereas swarms are more adequate for sparse, global, and anisotropic spatio-temporal processing. That is, swarms are interesting to integrate the local contexts described by the low-level perceptual representations. The outcome is a scalable parallel architecture that relies on self-organising principles to promote robustness and computational parsimony, both key assets of any autonomous robot.

The bottom-up nature of swarms plus their ability to handle dynamic environments - via pheromone propagated in time through a dynamic neural field - allows the extraction and tracking of high-level representations without having to rely on parametric models. This property, which still needs further formalisation, is particularly advantageous when the environment does not afford simplified, i.e., tractable, parametric models.

A previously published vision-based path detector was presented as an instance of the neural-swarm layered architecture. Furthermore, a set of extensions to the detector were proposed. Using a neural-swarm approach, this detector exploits the fact that paths are conspicuous structures in the environment by considering visual saliency as the primary visual feature for their detection. With an overall success rate of 98.67 %, at 20 Hz, it becomes clear that the neural-swarm layered architecture suits well the purpose of supporting the development of robust and efficient vision systems. This success rate is 5% higher than with the previous version of the model. The overall improvements can be credited to the addition of on-line learned contrast and appearance models for top-down modulation of the saliency computation process. This promotes the path's saliency in the presence of strong distractors.

Although the model is logically parallel, its current implementation is sequential. As future work, we expect to produce an implementation that is deployable in parallel hardware. For improved robustness, we expect in the near future to include 3-D perceptual data into the model. Finally, we expect to validate the model on a closed-loop scenario, i.e., on an autonomous robot.

Acknowledgements

This work was partially supported by CTS multi-annual funding, through the PIDDAC Program.

References

Y. Alon, A. Ferencz, and A. Shashua. Off-road path following using region classification and geometric projection constraints. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 689–696. IEEE, 2006.

S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977.

L. Antón-Canalís, M. Hernández-Tejera, and E. Sánchez-Nielsen. Particle swarms as video sequence inhabitants for object tracking in computer vision. In *Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 604 – 609. IEEE Computer Society, Washington, DC, 2006.

D. H. Ballard. Animate vision. *Artificial Intelligence*, 48(1):57–86, 1991.

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 *Proceedings of the 13th IASTED International Conference on Robotics, Applications and Telematics*, pages 477–482, 2007.

D. Bassett and E. Bullmore. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006.

R. D. Beer. The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4):209–243, 2003.

M. Blas, M. Agrawal, K. Konolige, and A. Sundaresan. Fast color/texture segmentation for outdoor robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4078–4085. IEEE Press, Piscataway, 2008.

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

A. Broggi and S. Cattani. An agent based evolutionary approach to path detection for off-road vehicle guidance. *Pattern Recognition Letters*, 27(11):1164–1173, 2006.

P. Burt and E. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

P. Chaturvedi and A. Malcolm. Real-time road following in natural terrain. In *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems*, volume 2, pages 815–820. IEEE, 2005.

M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3):201–215, 2002.

I. Couzin. Collective cognition in animal groups. *Trends in Cognitive Sciences*, 13(1):36–43, 2009.

J. Crisman and C. Thorpe. Unscarf-a color vision system for the detection of unstructured roads. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2496–2501. IEEE Press, Piscataway, 1991.

T. Egner, J. M. P. Monti, E. H. Trittschuh, C. A. Wieneke, J. Hirsch, and M. Mesulam. Neural integration of top-down spatial and feature-based information in visual search. *Journal of Neuroscience*, 28(24):6141, 2008.

L. Elazary and L. Itti. A bayesian model for efficient visual search and recognition. *Vision research*, 50(14):1338–1352, 2010.

D. Fernandez and A. Price. Visual detection and tracking of poorly structured dirt roads. In *Proceedings of the International Conference on Advanced Robotics (ICAR)*, pages 553–560. IEEE, 2005.

J. Fernandez and A. Casals. Autonomous navigation in ill-structured outdoor environment. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 1, pages 395–400. IEEE Press, Piscataway, 1997.

S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *Proceedings of the DAGM 2005, Lecture Notes on Computer Science, 3663*, pages 117–124. Springer-Verlag, Berlin, Germany, 2005.

T. Gevers and A. Smeulders. Color-based object recognition. *Pattern Recognition*, (32):453–464, 1999.

R. Ghurchian, S. Hashino, and E. Nakano. A fast forest road segmentation for real-time robot self-navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 406 – 411 vol.1. IEEE Press, Piscataway, 2004.

G. Grudic and J. Mulligan. Outdoor path labeling using polynomial mahalanobis distance. In *Proceedings of Robotics: Science and Systems*, pages 16–19. MIT Press: Cambridge, MA, 2006.

X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE Computer Society, Washington, DC, 2007.

A. D. Hwang, E. C. Higgins, and M. Pomplun. A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, 9(5):1–18, 2009.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985.

J. Liu, Y. Tang, and Y. Cao. An evolutionary autonomous agents approach to image feature extraction. *IEEE Transactions on Evolutionary Computation*, 1(2):141–158, 1997.

A. Lookingbill, J. Rogers, D. Lieb, J. Curry, and S. Thrun. Reverse optical flow for self-supervised adaptive autonomous robot navigation. *International Journal of Computer Vision*, 74(3):287–302, 2007.

L. Ma, J. Tian, and W. Yu. Visual saliency detection in image using ant colony optimisation and local phase coherence. *Electronics Letters*, 46(15):1066–1068, 2010.

J. A. R. Marshall and N. R. Franks. Colony-level cognition. *Current Biology*, 19(10):395–396, 2009.

S. Mazouzi, Z. Guessoum, F. Michel, and M. Batouche. A multi-agent approach for range image segmentation. In *Proceedings of the 5th international Central and Eastern European conference on Multi-Agent Systems and Applications (CEEMAS), LNAI 4696*, volume 4696, pages 1–10. Springer-Verlag, Berlin, Germany, 2007.

H. Mobahi, M. N. Ahmadabadi, and B. N. Araabi. Swarm contours: A fast self-organization approach for snake initialization. *Complexity*, 12(1):41–52, 2006.

V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45 (2):205–231, 2005.

M. B. Neider and G. J. Zelinsky. Scene context guides eye movements during visual search. *Vision Research*, 46(5):614–621, 2006.

Y. Owechko and S. Medasani. A swarm-based volition/attention framework for object recognition. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 91–98. IEEE Computer Society, Washington, DC, 2005.

S. E. Palmer. *Vision science: Photons to phenomenology*. MIT Press Cambridge, MA., 1999.

K. M. Passino, T. D. Seeley, and P. K. Visscher. Swarm cognition in honey bees. *Behavioral Ecology and Sociobiology*, 62(3):401–414, 2008.

R. Poli and G. Valli. Neural inhabitants of MR and echo images segment cardiac structures. In *Proceedings of the Computers in Cardiology*, pages 193–196. IEEE Computer Society, Washington, DC, 1993.

M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng. Ros: an open-source robot operating system. In *Proc. of the ICRA Open-Source Software Workshop*, 2009.

V. Ramos and F. Almeida. Artificial ant colonies in digital image habitats - a mass behavior effect study on pattern recognition. In *Proceedings of the 2n International Workshop on Ant Algorithms - From Ant Colonies to Artificial Ants (ANTS)*, pages 113–116, Belgium, 2000.

C. Rasmussen and D. Scott. Shape-guided superpixel grouping for trail detection and tracking. In *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4092–4097. IEEE Press, Piscataway, 2008a.

C. Rasmussen and D. Scott. Terrain-based sensor selection for autonomous trail following. In *Proceedings of the 2nd International Workshop on Robot Vision (Robvis 2008)*, pages 341–355, 2008b.

C. Rasmussen, Y. Lu, and M. Kocamaz. Appearance contrast for fast, robust trail-following. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE Press, Piscataway, 2009.

C. Rothkopf, D. Ballard, and M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14)(16):1–20, 2007.

N. Rougier and J. Vitay. Emergence of attention within a neural population. *Neural Networks*, 19(5):573–581, 2006.

P. Santana. Trail detection supporting material. `http://www.uninova.pt/~pfs/traildetection.html`, 2011.

P. Santana and L. Correia. A swarm cognition realization of attention, action selection and spatial memory. *Adaptive Behavior*, 18(5):428–447, 2010.

P. Santana and L. Correia. Swarm cognition on off-road autonomous robots. *Swarm Intelligence*, 5(1):45–72, 2011.

P. Santana, N. Alves, L. Correia, and J. Barata. Swarm-based visual saliency for trail detection. In *Proceedings of the IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems (IROS)*, pages 759–765. IEEE Press, Piscataway, 18-22 October 2010a.

P. Santana, N. Alves, L. Correia, and J. Barata. A saliency-based approach to boost trail detection. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 1426–1431. IEEE Press, Piscataway, 2010b.

P. Santana, R. Mendonça, L. Correia, and J. Barata. Swarms for robot vision: The case of adaptive visual trail detection and tracking. In *Proceedings of the European Conference on Artificial Life (ECAL)*, pages 712–719. MIT Press, Cambridge, 2011.

P. Santana, L. Correia, R. Mendonça, N. Alves, and J. Barata. Tracking natural trails with swarm-based visual saliency. *To appear in Journal of Field Robotics (early view online)*, 2012.

D. Song, H. Lee, J. Yi, and A. Levandowski. Vision-based motion planning for an autonomous motorcycle on ill-structured roads. *Autonomous Robots*, 23(3):197–212, 2007. ISSN 0929-5593.

O. Sporns, D. Chialvo, M. Kaiser, and C. Hilgetag. Organization, development and function of complex brain networks. *Trends in cognitive sciences*, 8(9):418–425, 2004.

C. Thorpe, M. Hebert, T. Kanade, and S. Shafer. Vision and navigation for the Carnegie-Mellon Navlab. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(3):362–373, 1988.

S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney. Stanley: The robot that won the darpa grand challenge. *Journal of Field Robotics*, 23(9): 661–692, 2006.

J. Tian, W. Yu, and S. Xie. An ant colony optimization algorithm for image edge detection. In *Proc. of the IEEE Congress on Evolutionary Computation*, pages 751–756, 2008.

A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 273–280. IEEE Computer Society, Washington, DC, 2003.

A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

V. Trianni, E. Tuci, K. Passino, and J. Marshall. Swarm cognition: an interdisciplinary approach to the study of self-organising biological collectives. *Swarm Intelligence*, 5(1):3–18, 2011.

J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1-2):507–545, 1995.

D.-S. Tue-Cuong, G. Dong, Y. C. Hwang, and O. S. Heng. Extraction of shady roads using intrinsic colors on stereo camera. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 341–346. IEEE, 2008.

J. Turner. Termites as models of swarm cognition. *Swarm Intelligence*, 5(1):19–43, 2011.

D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19: 1395–1407, 2006.

J. Wolfe, M. Vo, K. Evans, and M. Greene. Visual search in scenes involves selective and nonselective pathways. *Trends in cognitive sciences*, 15(2):77–84, 2011.

J. M. Wolfe. Guided search 2. 0. a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.

A. L. Yarbus. *Eye movements and vision*. Plenum Press, New York, 1967.

X. Zhang, W. Hu, S. Maybank, X. Li, and M. Zhu. Sequential particle swarm optimization for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE Computer Socienty, Washington, DC, 2008.