UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

VISUAL ATTENTION AND SWARM COGNITION FOR OFF-ROAD ROBOTS

Pedro Figueiredo Santana

DOUTORAMENTO EM INFORMÁTICA
Especialidade em Engenharia Informática

2011

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

VISUAL ATTENTION AND SWARM COGNITION FOR OFF-ROAD ROBOTS

Pedro Figueiredo Santana

DOUTORAMENTO EM INFORMÁTICA
Especialidade em Engenharia Informática

Tese orientada pelo Prof. Doutor Luís Miguel Parreira e Correia

2011

Abstract

This thesis addresses the problem of modelling visual attention in the context of autonomous off-road robots. The goal of using visual attention mechanisms is to allow robots to focus perception on the aspects of the environment that are more relevant to the task at hand. As this work will show, this capability is a promoter of robustness and computational parsimony in both obstacle and trail detection. These features are key enablers of fast and energetically efficient field robots.

One of the major challenges in modelling visual attention emerges from the need to ensure that the model is able to manage the speed-accuracy trade-off in the face of context and task changes. This thesis shows that this trade-off is handled if the cognitive process of visual attention is modelled as a self-organising process, whose operation is modulated by the robot's action selection process. By closing the loop from the action selection process to the perceptual one, the latter is able to perform on a by-need basis, anticipating actual robot moves.

To endow visual attention with self-organising properties, this work gets inspiration from Nature. Concretely, the mechanisms underlying the ability that army ants have of foraging in a self-organising way are used as metaphor to solve the task of searching, also in a self-organising way, for obstacles and trails in the robot's visual field. The solution proposed in this thesis is to have multiple covert foci of attention operating as a swarm via pheromone-based interactions.

This work represents the first embodied realisation of swarm cognition, which is a new born field that aims to uncover the basic principles of cognition by inspecting the self-organising properties of the collective intelligence exhibited by social insects. Hence, this thesis contributes to robotics as an engineering discipline, and to robotics as a modelling discipline supporting the study of adaptive behaviour.

KEY WORDS: Swarm Cognition, Visual Attention, Autonomous Robots, Off-Road Local Navigation, Stereo-Based Obstacle Detection, Monocular Trail Detection.

Resumo

Esta tese aborda o problema da modelação de atenção visual no contexto de robôs autónomos todo-o-terreno. O objectivo de utilizar mecanismos de atenção visual é o de focar a percepção nos aspectos do ambiente mais relevantes à tarefa do robô. Esta tese mostra que, na detecção de obstáculos e de trilhos, esta capacidade promove robustez e parcimónia computacional. Estas são características chave para a rapidez e eficiência dos robôs todo-o-terreno.

Um dos maiores desafios na modelação de atenção visual advém da necessidade de gerir o compromisso velocidade-precisão na presença de variações de contexto ou de tarefa. Esta tese mostra que este compromisso é resolvido se o processo de atenção visual for modelado como um processo auto-organizado, cuja operação é modulada pelo módulo de selecção de acção, responsável pelo controlo do robô. Ao fechar a malha entre o processo de selecção de acção e o de percepção, o último é capaz de operar apenas onde é necessário, antecipando as acções do robô.

Para fornecer atenção visual com propriedades auto-organizadas, este trabalho obtém inspiração da Natureza. Concretamente, os mecanismos responsáveis pela capacidade que as formigas guerreiras têm de procurar alimento de forma auto-organizada, são usados como metáfora na resolução da tarefa de procurar, também de forma auto-organizada, obstáculos e trilhos no campo visual do robô. A solução proposta nesta tese é a de colocar vários focos de atenção encoberta a operar como um enxame, através de interacções baseadas em feromona.

Este trabalho representa a primeira realização corporizada de cognição de enxame. Este é um novo campo de investigação que procura descobrir os princípios básicos da cognição, inspeccionando as propriedades auto-organizadas da inteligência colectiva exibida pelos insectos sociais. Logo, esta tese contribui para a robótica como disciplina de engenharia e para a robótica como disciplina de modelação, capaz de suportar o estudo do comportamento adaptável.

PALAVRAS CHAVE: Cognição de Enxame, Atenção Visual, Robôs Autónomos, Navegação Todo-o-Terreno, Detecção de Obstáculos Binocular, Detecção de Trilhos Monocular

Resumo Alargado

Esta tese aborda o problema geral da navegação local em robôs móveis a operar em ambientes todo-o-terreno. A resolução deste problema é um passo essencial para permitir a utilização de robôs autónomos em tarefas perigosas ou fisicamente exigentes para os seres humanos, tais como monitorização ambiental. Um dos principais obstáculos à concretização deste objectivo é a dificuldade inerente à implementação de mecanismos percepcionais que sejam simultaneamente rápidos e robustos. Este problema é mais notório em robôs que se baseiam em visão como modalidade sensorial principal, cuja riqueza vem a par de uma elevada complexidade de processamento computacional. Esta tese estuda a possibilidade de se atenuar este problema através do uso de mecanismos de atenção visual. Desta forma, os recursos computacionais do robô são concentrados na percepção dos aspectos do ambiente mais relevantes à sua operação.

Num primeiro estudo, mostramos como a atenção visual pode ser usada para tornar a detecção de obstáculos em terrenos irregulares mais rápida e robusta. Neste trabalho assume-se que o robô está equipado com um sistema de visão binocular capaz de fornecer nuvens de pontos tridimensionais (3-D) do ambiente em seu redor. Estas nuvens, regra geral, são muito densas e ruidosas, o que dificulta o seu processamento. A arquitectura proposta explora o melhor de duas técnicas de detecção complementares. Uma das técnicas, que é rápida e portanto ideal para uma primeira análise, considera obstáculos aqueles pontos 3-D que se destacam do plano do chão a uma dada altura mínima (Konolige et al., 2009). Para evitar um grande número de falsos positivos em terrenos irregulares, onde não existe um plano do chão dominante, o limiar de detecção desse detector deve ser elevado de modo a que somente obstáculos de grande dimensão possam ser detectados. No que respeita à navegação local, a detecção de um obstáculo grande numa dada região do ambiente torna irrelevante a detecção de obstáculos pequenos nessa mesma região. Logo, só se torna necessário a aplicação de um detector de obstáculos pequenos nas regiões onde nenhum obstáculo grande tenha sido encontrado. Para relaxar a restrição de chão planar, que é essencial remover com vista à detecção de obstáculos pequenos, as regiões livres de obstáculos grandes são então analisadas por um detector que opera com base em restrições

geométricas entre pontos 3-D vizinhos (Manduchi et al., 2005). O elevado custo computacional associado à precisão deste segundo detector é compensado pelo facto de poder ser focado pelo primeiro, que sendo mais rápido é menos preciso. Contudo, o seu custo computacional ainda permanece demasiado elevado para permitir uma navegação rápida. Ao adaptar um modelo bio-inspirado, conhecido na área da saliência visual computacional (Itti et al., 1998), para o problema da detecção de obstáculos em ambientes todo-o-terreno e usando-o para focar ainda mais o detector de obstáculos pequenos, a computação total é reduzida em cerca de 20 vezes. O mapa de saliência visual permite o detector de obstáculos focar melhor a sua operação nas regiões do ambiente que se destacam do fundo e portanto têm maior propensão de pertencerem a um obstáculo. Consequentemente, uma vantagem adicional de focar o detector é que evita, em grande medida, a detecção de falsos positivos. Isto é particularmente importante dada a sensibilidade da técnica original de detecção de obstáculos pequenos a este aspecto.

Um dos maiores desafios na modelação de atenção visual advém da necessidade de gerir o delicado compromisso entre velocidade e precisão, inerente às tarefas de pesquisa visual. Por exemplo, no caso de detecção de obstáculos, uma análise detalhada do campo visual do robô pode dificultar uma cobertura completa, ao passo que uma análise grosseira pode tornar a detecção de pequenos obstáculos insuficiente. Um desafio adicional resulta do facto de que, quando o ambiente ou a tarefa mudam, a importância relativa de um obstáculo também pode mudar. Portanto, a obtenção da solução óptima capaz de satisfazer todos estes requisitos é um problema complexo, que dificilmente pode ser resolvido de forma eficiente. Alternativamente, num segundo estudo, complementar ao primeiro, mostramos que se o processo de atenção visual for modelado como um processo auto-organizado, o compromisso velocidade-precisão é resolvido de forma parcimoniosa e insensível a mudanças de contexto. Para além disso, ao modular a alocação de atenção visual com a saída do processo de selecção de acção, a primeira é capaz de antecipar as acções do robô e, consequentemente, de induzir a percepção a operar nas regiões do espaço visual mais relevantes.

Para fornecer o mecanismo de atenção visual com propriedades auto-organizadas, este trabalho obtém inspiração da Natureza. Concretamente, os mecanismos responsáveis pela capacidade que as formigas guerreiras têm de procurar alimento de forma auto-organizada são usados como metáfora na resolução da tarefa de procurar, também de forma auto-organizada, obstáculos no campo visual do robô. A metáfora é usada para modelar vários focos de atenção encoberta como agentes virtuais simples que habitam o espaço visual do robô e operam como um enxame através de interacções baseadas em "feromona". Os agentes procuram na imagem, de forma coordenada, obstáculos relevantes à navegação. Basicamente, os agentes realizam

ciclos locais de atenção visual encoberta, enquanto que o comportamento auto-organizado do colectivo mantém coerência espaço-temporal global. Resultados experimentais demonstram a capacidade do modelo em controlar de forma robusta um robô todo-o-terreno, equipado com um sensor de visão binocular, na tarefa de navegação local, com menos de 1 % do seu campo visual a ser analisado. Um subproduto do sistema é a manutenção de memórias espaciais paralelas, activas e esparsas. Estas memórias são constituídas pelos agentes que deixam o campo de visão do sensor ao tentarem manter-se junto dos obstáculos.

Num terceiro estudo, o uso de conhecimento a priori sobre os objectos a pesquisar no campo visual é explorado. Usar este conhecimento é importante quando a saliência visual é insuficiente para focalizar a atenção na presença de outros objectos ou "alucinações", que se destacam no ambiente, pelo menos tanto quanto o objecto que está a ser procurado. A detecção de trilhos em ambientes naturais todo-o-terreno foi escolhida como caso de estudo neste trabalho. A motivação para a aplicação de modelos de saliência visual na detecção de trilhos advém da observação de que estes são estruturas que frequentemente se apresentam conspícuas no campo visual do robô. Normalmente, o conhecimento a priori do objecto a procurar toma a forma do conjunto de características (e.g., cor) que melhor descrevem o objecto. Em vez disso, a forma aproximada do objecto é utilizada neste trabalho, dado que no caso dos trilhos é uma característica muito mais estável e previsível. Este tipo de conhecimento a priori é especificado indirectamente no modelo proposto, como regras percepção-acção que controlam o comportamento dos agentes que habitam o espaço visual do robô. Tal como no caso de navegação local, esses agentes representam processos locais de atenção encoberta. A auto-organização do seu comportamento colectivo como enxame resulta num mapa de saliência visual e, portanto, num processo de atenção encoberta global. Tanto neste como no estudo anterior, a auto-organização dos enxames de agentes assegura a robustez do sistema com parcimónia de custo computacional.

O método começa por inserir um conjunto de agentes em dois mapas de conspicuidade previamente calculados, um para a cor e outro para a informação de intensidade. Em modelos computacionais típicos de saliência visual (e.g., Itti et al. (1998)), os mapas de conspicuidade são combinados para gerar o mapa de saliência final. Alternativamente, o modelo proposto constrói o mapa de saliência visual a partir da composição de dois mapas de feromona, construídos pelos conjuntos de agentes. Os agentes movem-se por várias iterações de acordo com: (1) a fusão de um conjunto de regras de percepção-acção, (2) um factor aleatório; e (3) um factor de atracção à feromona já depositada. As regras de percepção-acção permitem aos agentes explorar a informação de conspicuidade local, tendo em conta conhecimento a priori da forma

aproximada do trilho, a fim de aproximarem as suas trajectórias ao esqueleto do trilho. As interacções baseadas em feromonas permitem que os agentes superem as variações locais da imagem e, implicitamente, facilitam uma ajuda mútua na construção de um consenso sobre a melhor aproximação ao esqueleto do trilho. Este consenso emerge onde o comportamento semi-exploratório dos agentes é constrangido pela presença do trilho, o *atractor* ambiental. Resultados experimentais num grande conjunto de dados revelam a capacidade do modelo em produzir uma taxa de sucesso de 91 % a 20 Hz.

Todos estes desenvolvimentos contribuem para mostrar que: (1) a robustez, a parcimónia e o desempenho dos robôs todo-o-terreno são melhorados se mecanismos de atenção visual forem empregues nos seus sistemas de controlo e (2) a síntese de comportamento cognitivo auto-organizado em robôs autónomos é facilitada se a metáfora do comportamento colectivo exibido pelos insectos sociais for usada como padrão de desenho. Além disso, esta tese apresenta a primeira realização de cognição corporizada baseada em enxames. Este novo campo de investigação visa a descoberta dos princípios básicos da cognição, inspeccionando as propriedades de auto-organização da inteligência colectiva exibida pelos insectos sociais.

Acknowledgements

First and foremost I want to thank my advisor Prof. Luís Correia. His strong criticism and pragmatism allied with his ability to think outside the box were pivotal (I know he loves this word...) to the development of my work. His friendship and complete availability were also paramount (yet another word he loves...) to the progress and enrichment of this endeavour.

My next acknowledgement goes to all the MSc students that helped me in the implementation work, in particular Magno Guedes, Nelson Alves, Paulo Santos, and Ricardo Mendonça, which gave assistance to the implementation and experimental validation of part of the work reported in Chapters 2 and 5.

I also want to thank the MSc students that contributed to the development of the hardware and software of the robot used in the experiments, namely, Vasco Santos, Carlos Cândido, Mário Salgueiro, and Luís Almeida. This development was done in collaboration with the Portuguese company IntRoSys, S.A.. Hence, I want to thank IntRoSys CEO, Luís Flores, for his commitment and support to this research line. A special thanks to João Lisboa, the man behind the mechanics of the robot, for teaching me so much about mechanics, and most importantly, for his friendship.

A special word to Prof. José Barata. I want to thank him for his understanding when I, already working under him at UNINOVA, needed to take some time off to finish this thesis. The trust and ambition he always puts on me have been important motivational elements in my work.

I also want to thank my colleagues from the research lab GruVA, Carlos Grilo and Pedro Mariano, for their friendship and always rewarding discussions. The same

warm compliment goes to my colleagues and friends from UNINOVA, Luís Ribeiro, Gonçalo Cândido, and Eduardo Pinto.

I also want to thank the fruitful comments and critics provided by Prof. José Santos-Victor and Prof. José Rufino.

I gratefully acknowledge Fundação para a Ciência e Tecnologia for the doctoral grant No. SFRH/BD/27305/2006 and to LabMAg for hosting and funding, which made this work possible.

I want to thank all the unconditional support and motivation given by my family during this enriching but demanding period. Thank you so much: Mãe, Rui, Pai, Joana, Diogo, Avô, Avó, D.ª Fernanda, Sr. Mário, Luís, and Eugénio. A thousand hugs to my sweet wife Ana and to my "suns" Francisco and Rodrigo. They make my life unique and worth living.

Lisbon, May 2011
Pedro Figueiredo Santana

To Ana, Francisco, and Rodrigo

Contents

List of Acronyms

2-D - Two-Dimensional
3-D - Three-Dimensional
DoG - Different of Gaussians
EKF - Extended Kalman Filter
ESalOD - full-fledged Saliency-based small Obstacle Detector
FOV - Field Of View
FPR - False Positive Rate
GPE - Ground-Plane Estimation
GPU - Graphics Processing Unit
LOG - Laplacian Of Gaussian
OOD - Original Obstacle Detector
MCC - Matthews Correlation Coefficient
RANSAC - RANdom SAmple Consensus
ROC - Receiver Operating Characteristic
ROI - Region Of Interest
SalGPE - Saliency-based Ground-Plane Estimation
SalOD - stripped-down Saliency-based small Obstacle Detector
SLAM - Simultaneous Localisation And Mapping
SVD - Singular Valued Decomposition
SVR - Space-Variant Resolution
TPR - True Positive Rate

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Motivation: Robust and Parsimonious Field Robots

The valuable scientific data generated by the Mars rovers in their missions (Squyres et al., 2009), which could not be obtained by any other means, clearly shows the importance autonomous field robots have for the future of humankind. Besides helping us finding a new extra-terrestrial home, autonomous field robots can help us maintaining our current one. Examples of ecological-related applications where field robots are becoming important assets are wildlife monitoring (Tokekar et al., 2010) and water quality monitoring (Sukhatme et al., 2007). As in the extra-terrestrial case, these activities also occur in geographically inconvenient regions, which limits their continuous execution. In this sense, autonomous field robots, as the one depicted in Fig. 1.1, may become the ideal tool for a successful maintenance of Earth ecosystems.

Field robots are also practical in helping us in the prevention of natural disasters; for instance, robots can be used to detect fires early at their onset (Merino et al., 2006). Robots can also be handy in the mitigation of the effects of natural disasters, as it already happens in search & rescue missions (Murphy and Stover, 2008). Human-centred activities can also benefit from including these robots in their operations, such as agriculture (Johnson et al., 2009), humanitarian demining (Santana et al., 2007; Habib, 2007), and patrol/reconnaissance operations (Huntsberger et al., 2011). Common to all these tasks is that they are, to different extents, dangerous or physically demanding for humans. In sum, field robots are tools that can substantially contribute to solve many ecological, social, and economic problems.

Figure 1.1: The off-road robot (Santana et al., 2008a) used in the experiments.

A consequence of performing remote operations is that the repair of a damaged robot is impractical. Therefore, the whole mission may be compromised by the occurrence of a simple malfunction. A straightforward workaround is to avoid single points of failure by exploiting the redundancy inherent to multi-robot systems. The use of several robots being more robust is also more expensive, meaning that robots must be cheap, and consequently, affording only reduced energetic and computational resources. These two constraints are also present in cases where robot size and weight are highly constrained by task requirements; for instance, ground vehicles for operations in confined environments as snake-like robots (Hirose and Fukushima, 2004) or miniature aerial robots with extremely limited payload capacity (Beyeler et al., 2009). It must also be taken into account that in most identified tasks robots must operate for long periods, which poses additional constraints to their efficiency; eventually, these robots will need to perform energy harvesting (Low et al., 2009). In fact, efficiency is important to field robots as it is to any other contemporary industrial equipment, in the sense that ecological sustainability is always an issue.

In the long run, computational parsimony is important for any kind of robot performing interesting tasks in complex environments. Such robots will necessarily engage in demanding multi-faceted cognitive behaviour, and so will require the integration of modules for sensorimotor coordination, high-level abstract reasoning, teamwork, among others. Mostly due to perceptual requirements, robust local navigation is known to be computationally intensive. Abstract symbolic reasoning, either for path planning, object manipulation, or task execution, is also inherently computationally intensive. Teamwork is demanding in terms of perception, planning, and teammate

internal state estimation (e.g., for intentions prediction). If making each of these components run with an acceptable rate is a daunting task with contemporary consumer computers, making them work as a whole to build a cognitively rich robot is still an unfeasible one. As a result, fully autonomous robots currently demand computational parsimony as a key drive in their design.

1.2 The Solution: Visual Attention

An important sensory modality for field robots is vision. Being passive, vision requires low power to operate. Moreover, vision provides the robot with (multi-spectral) appearance, volumetric, and motion data, thus reducing the need for additional sensors. Also important, the acquisition of all this information is synchronised, thus avoiding problems of registration. Finally, vision sensors are small and lightweight. These characteristics are possibly the reason why vision is ubiquitous in Nature and so useful for compact and energetically efficient field robots. However, the richness of vision comes with the price of complex processing. This is particularly challenging in the case of off-road environments, where the absence of an easily observable structure complicates the definition of robust fast to compute heuristics.

The complexity inherent to vision calls for fine and contextualised focus of computational resources on the most relevant aspects of the environment. This process is called visual attention and it is known to be widespread in the animal kingdom (Land, 1999) and it has been extensively studied in humans (Oliva and Torralba, 2007; Wolfe et al., In Press). By focussing perception: (1) computation, and consequently, energy are more efficiently used; (2) the robot becomes less sensitive to noise and erroneous environmental cues (i.e., false positives); and as a consequence of the previous two, (3) faster robot motion, lower cost, and reduced robot size are enabled.

Attention ultimately results in the motion of sense organs, e.g., eyes, towards the relevant stimulus source. This is called overt attention. A faster process is the one of mentally focussing particular aspects of the sensory stimuli. This is called covert attention and its importance to off-road robots is the core of the main hypothesis being addressed in this thesis:

> *parsimony, robustness, and performance of off-road robots is improved if covert visual attention mechanisms are employed in their control systems.*

1.3 The Method: Swarm Cognition

Although a multi-robot system helps to avoid single points of failure, the endurance of the whole system is unavoidably attached to the robustness of each robot. This robustness is connected to the ability of the robot not to collapse in the face of unforeseen situations. With optimality in mind, either a large set (virtually all) of optimal situation specific heuristics would have to be defined, or computationally expensive reasoning mechanisms would have to be running on-board the robot in real-time. The former being unfeasible and the latter being in contradiction with the parsimonious design requirement and with the way Nature has evolved animals, means that the quest for optimality undermines the compliance with both robustness and parsimony requirements. In addition, in dynamic environments, where exceptions rule out regularity, an optimal solution may rapidly become deprecated. Hence, the high computational cost spent on improving a near-optimal solution is better applied in the search of good-enough alternative solutions. These alternative solutions may enable a prompt reaction of the system in the face of change.

For these reasons, this thesis largely considers the design of a control system as the process of creating a stochastic parallel system with self-organising properties. That is, a system capable of dynamically cope with disturbances without explicit context awareness - robust - and capable of generating complexity from the interaction of simple stochastic entities - parsimonious. Hence, besides enabling robustness and parsimony in field robots, this thesis contributes to the wider understanding of self-organising embodied cognition. We think that, while not necessary, the study of embodied cognition gains considerably from the insights obtained from building field robots, in particular those operating in natural environments. This stems from the fact that robots inhabiting the environment where Nature's evolutionary process took place are supposedly subject to similar constraints. Furthermore, by studying embodied cognition in field robots, short-term practical outcomes from the conceptual work are ensured, which would not be obtainable if instead only simulation or toy-world problems would be considered.

Although in line with the dynamical systems approach to human cognition (Beer, 1995; Thelen and Smith, 1996) in what regards its dynamical and self-organising nature, the methods here proposed are not constrained by the existence of supporting analytical tools. We consider that this constraint, although interesting for many ob-

vious reasons, imposes artificial limits on the complexity that is in reach of well designed computational models. Bearing in mind the goal of modelling covert attention under a self-organising perspective with a tractable granularity, the *agent* abstraction is considered. That is, our main computational atom is not a neuron or a well-defined mathematical construct, it is a spatio-temporal dynamical entity that inhabits the sensorimotor space of the robot and moves on it to guide the robot's focus of attention.

Rather than static structures, like neurons, these agents are better viewed as active information particles that flow and change in the system. Hence, using agents, the design focus is on the process and not so much on its supporting substrate. Being sensorimotor coordinated units, these information particles can in the limit exploit the recognised benefits of active perception (Bajcsy, 1988; Aloimonos et al., 1988; Ballard, 1991; Scheier et al., 1998; Nolfi and Marocco, 2002; Beer, 2003; Fend et al., 2003; Balkenius et al., 2004; Floreano et al., 2004; Nolfi, 2005; Suzuki and Floreano, 2006; Pfeifer and Bongard, 2006; Kim and Moeller, 2006; Sporns and Lungarella, 2006; de Croon and Postma, 2007; Choe et al., 2008; Mirolli et al., 2010; Tuci et al., 2010), such as actively selecting and shaping their sensory input to increase pose invariance, signal-to-noise ratio, and discriminative power. Thus, an agent-based design is an interesting alternative to a neuron-based design when it comes to manage complex spatio-temporal problems, as it is the case of shifting covert attention towards the most relevant aspects of the environment.

The selection of the agent construct to model covert attention is further motivated by existing evidence that both covert and overt attention processes share the same neural mechanisms (de Haan et al., 2008), and being the latter a sensorimotor coordinated process we hypothesise so it should be the former. In this sense, agents, which are themselves sensorimotor coordinated entities, become a natural choice to model local loops of covert attention. Then, provided that it is possible to have several agents operating in a self-organising way, we solve the problem of maintaining multiple covert attention processes (Pylyshyn and Storm, 1988; Doran et al., 2009) operating in a spatio-temporal coherent way. To ensure self-organising behaviour, we recur to biological knowledge obtained from similar processes in Nature, such as the self-organising *collective intelligence* (Franks, 1989) exhibited by social insects, i.e., natural agents like bees, ants, and termites. In a sense, these animals embody as a collective a sort of *swarm cognition*, whose considerable similarities with brain cognitive function are becoming widely recognised (Passino et al., 2008; Santana and Correia, 2009; Couzin, 2009; Mar-

shall and Franks, 2009; Trianni and Tuci, 2010; Santana and Correia, 2010a; Turner, 2011; Trianni et al., 2011). This metaphor is particularly powerful due to the easiness of tracking the activity of social insects from the collective down to the individual, at least when compared to the difficulty of inspecting the nervous system activity in a both detailed and comprehensive way. Hence, this metaphor enables a convenient way of transferring knowledge from Nature to engineering. In the specific scope of this thesis, which is visual search, the army ants foraging metaphor is considered.

Following the rationale underlying the power of swarm cognition models for the synthesis of cognition the following hypothesis is also addressed in this thesis:

> *the synthesis of self-organising robot cognitive behaviour is facilitated if the social insects collective behaviour metaphor is used as design pattern.*

1.4 The Impact: A Wider Picture

Being self-organised, the overall system's behaviour emerges from the interaction of simple units, which is often done indirectly by changing and sensing the environment, a phenomenon known as *stigmergy* (Grassé, 1959). High level structures emerge from the bottom-up system operation, meaning that the system is fully specified by the logic ruling the simple homogeneous agents. Consequently, the system's design space is small and fully grounded through the set of sensorimotor rules controlling the agents, which can either be designed by hand, actively learned, or even evolved.

Self-organisation referring to synergistic parallelism, and not to simple concurrence, goes beyond a naive parallelisation by splitting. This might be interesting to better explore the capabilities of parallel hardware. A more radical emerging computational hardware is called *amorphous computing* (Abelson et al., 2000). It envisions the application of micro-fabricated particles or engineered cells for the implementation of massively distributed computation. The distributed, stochastic, and unreliable nature of these particles or cells demands for new computational paradigms capable of coping with these characteristics, which are drastically different from the highly reliable and deterministic computational units currently in use. As self-organising systems exhibit graceful degradation in the presence of unreliable elements and do actually benefit from some level of randomness for a robust operation (Bonabeau et al., 1999; Correia,

2006), *swarm cognition* is a natural candidate for the task at hand.

Finally, studying the synthesis of cognition under a swarm perspective takes a step further in the direction of an unified methodology for the development of robot control systems and multi-robot coordination mechanisms. Basically, it is all about coupling sensorimotor coordination loops. A critical point of swarm research, and adaptive behaviour in general, is the potential for cross-breeding between both natural and artificial sciences. Observations from the Natural world can be used to seed the modelling process in engineering. Conversely, operational engineering models can be used to induce hypotheses regarding specific aspects of the Natural world. The unexpected importance of a given variable in an engineered model may trigger the interest in assessing the relevance of its natural counterpart. Interestingly, new developments, such as bacterial micro-robots (Martel and Mohammadi, 2010), show that the future may bring part natural part artificial *agents*, whose collective cognitive operation will most probably rely on self-organising principles.

1.5 Proposed Models Overview

The phenomenological support of the two research hypotheses will be assessed, in different extents, against three foundational aspects of off-road mobility: obstacle detection, trail detection, and local navigation. The particularities of each case-study elicit different aspects of visual attention, and consequently, enrich the overall argument. In general, the proposed models exhibit methodological novelty and improved performance over existing methods, and so they are contributions by themselves to both robotics and visual attention literature.

The following subsections provide an overview of each contribution, whose details can be found in subsequent independent chapters. First, in Section 1.5.1, a study where visual attention is shown to speed up and augment the robustness of obstacle detection in all-terrain environments is described. Then, in Section 1.5.2, a swarm cognition model capable of integrating visual attention, action selection, and spatial memory to implement a complete local navigation system is presented. Finally, in Section 1.5.3, visual attention is demonstrated to be also useful in the task of trail detection. For this purpose, a swarm-based model for visual saliency computation is shown to enable a robust and fast to compute exploitation of a priori knowledge about the trail's approximate layout.

1.5.1 Visual Attention for Off-Road Obstacle Detection

This first work (Santana et al., 2008b, 2009, 2010d, 2011) is focused on showing the benefits of using visual attention mechanisms to promote robust and fast obstacle detection on uneven terrain.

The proposed architecture exploits the best of two complementary detection techniques. One of the techniques, which is fast and so ideal for a first scan, considers obstacles three-dimensional (3-D) points that stand out from an estimated ground-plane by a given minimum height (Konolige et al., 2009). To avoid a large number of false positives in uneven terrain, where no dominant ground-plane is assured to exist, the detection threshold must be set high so that only large obstacles are detected. In what regards local navigation, the detection of a large obstacle in a given region of the environment turns irrelevant the detection of small obstacles therein. Hence, it is only necessary to search for small obstacles in the regions where no large obstacle has been found. To relax the planar ground constraint, essential to enable small obstacle detection, these regions are analysed by a detector that operates based on geometrical constraints between neighbour 3-D points (Manduchi et al., 2005). The high computational cost associated to the accuracy of this method is compensated by the focus of attention provided by the other faster yet less accurate detector.

Although the small obstacle detection technique is already focused by the large detection technique, its computational cost still remains too high. By adapting a well-known bio-inspired visual saliency model (Itti et al., 1998) to the problem at hand, and by using it to further focus the small obstacle detector, computation is shown to be reduced in about 20 times. The computed visual saliency map allows the obstacle detector to better focus its operation on the regions of the environment that detach more from the background, and consequently are more prone to belong to an obstacle. Another key aspect of the model is that the detector updates the saliency map whenever an obstacle is detected. This closes the loop allowing the saliency map to guide the detector while being opportunistically corrected by it. Besides reducing computational cost, this approach also helps reducing the false positive rate. This is particularly important given that the original small obstacle detection technique suffers considerably from this problem. To further improve the method's robustness, a voting mechanism is included in the small obstacle detector without noticeable computational cost.

1.5.2 Swarm Cognition for Local Navigation

A step further, this section analysis a complementary aspect of visual attention on off-road robots. In particular, the work presented in this section (Santana and Correia, 2009, 2010a, 2011) is about the modelling of covert visual attention as a parallel process that unfolds in synergy with the robot's action selection process. The parallel nature of the proposed model is in accordance to the multiple covert attention hypothesis (Pylyshyn and Storm, 1988; Doran et al., 2009) from cognitive science, and thus, suitable for the concurrent search for multiple objects in the robot's visual field. In this particular study, the action selection process aims at providing the off-road robot with local navigation capabilities, whereas the perceptual process is responsible for the detection of obstacles.

In order to maintain responsiveness in complex environments, perception must be able to deal with the delicate speed-accuracy trade-off inherent to visual search tasks. A detailed analysis of the robot's visual field may hamper a complete coverage, whereas a coarse analysis may render insufficient the detection of small obstacles. Such trade-off is solved in this work by modulating the allocation of visual attention with the output of the action selection process, thus allowing the former to anticipate the robot's actual actions and, as a consequence, to induce perception to operate on a by-need basis. In other words, attention gives priority to the regions of the visual field where the detection of obstacles affects more the action selection process. For instance, if, due to task constraints or current knowledge about the world state, the action selection process outputs a right turn, visual attention increases preference for the right-hand side of the robot's visual field. This preference is set under the rational that an obstacle detected on the left-hand side would have lesser impact on the unfolding of both action selection process and actual robot motion. Another desirable requirement for embodied systems is the ability to produce results on-demand. That is, the perceptual process should be able to progressively improve its results so that when requested it has a good enough solution to provide. An additional challenge comes from the fact that, when either the environment or the task change, the relative importance of an obstacle may also change. Optimally satisfying all these requirements is a complex problem, which can hardly be solved in a computationally efficient way.

Based on the well known formalism that describes the self-organisation of collective foraging strategies in social insects, the joint satisfaction of all these requirements is

solved by recurring to a set of simple virtual agents inhabiting the robot's visual input, called p-ants, which search for obstacles in a collectively coordinated way. Basically, p-ants perform local covert visual attention loops, whereas the self-organised collective behaviour maintains global spatio-temporal coherence in order to robustly maintain a proper speed-accuracy trade-off. The ant foraging metaphor is particularly interesting because of its similarities with the problem tackled in this work. First, ants need to efficiently search (forage) their environment in a parallel manner. Second, ants are also sensorimotor coordinated entities capable of engaging in active perceptual activity.

Experimental results show the ability of the model to robustly control an off-road robot equipped with a stereoscopic vision sensor in a local navigation task with less than 1 % of the robot's visual input being analysed. A by-product of the system is the maintenance of active, parallel, and sparse spatial working memories. These memories are composed of p-ants that leave the sensor's field of view when tracking detected obstacles. Tracking in the absence of visual feedback is carried out by compensating the p-ant's position with ego-motion information obtained from visual odometry. Given the fact that p-ants are deployed based on the action selection process' output, the environment representation is potentially incomplete, yet good enough for navigation purposes. Discarding the typical requirement of completeness follows from the assumption that different purpose-oriented world representations cohabit in the same robot. In addition to the navigation-oriented representation proposed in this work, the robot might accumulate sensor data to generate off-line an accurate and potentially complete map. This latter map could be useful for instance to support tele-operation. In short, the model exhibits the self-organisation of a relevant set of features composing a cognitive system.

Given that the purpose of the work described in this section is to study the interaction between attention, action selection, and memory, the specific obstacle detection technique employed for experimental validation is a simplified one. Concretely, it assumes a mostly planar environment. Nevertheless, with some engineering, this detector could be substituted by the more complex one described in the previous section.

1.5.3 Visual Attention for Trail Detection

Models of visual attention typically assume the existence 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; Hwang et al., 2009). The work described in the two previous sections is consistent with this view: pre-attentive visual saliency obtained directly from the sensory input and top-down context knowledge obtained from the action selection process have been used to guide obstacle detection.

This section describes a third research line (Santana et al., 2010a,b) that complements this analysis by studying the use of object-related a priori knowledge in visual search tasks. The use of object-related a priori knowledge is important when visual saliency is insufficient to focus the attention in the presence of distractors. These distractors are other objects or perceptual aliasing in the environment that happen to detach from the background at least as much as the object being sought. Trail detection in natural environments was selected as case-study for this work. The motivation for the application of visual saliency to trail detection builds from the observation that trails are quite often conspicuous structures in the robot's visual field. Furthermore, the lack of a well defined morphology or appearance of trails limits the application of learning and model-based approaches.

Typically, object-related a priori knowledge is used by top-down boosting of the set of features (e.g., colour) known beforehand to be more representative of the object being sought. Instead, the object's overall layout, which is a more stable and predictable feature in the case of natural trails, whose local appearance often blends with the background, is used in this work. This type of a priori knowledge is specified indirectly in the proposed model as perception-action rules controlling the behaviour of simple agents inhabiting the robot's visual input. Like in the local navigation case, these agents are called p-ants and represent local covert attention processes. Their self-organising collective behaviour results in a saliency map of the input image, and thus, in a global covert attention process.

The method starts by deploying a set of p-ants in two previously computed conspicuity maps, one for colour, and another for intensity information. In typical saliency computational models (e.g., Itti et al. (1998)), conspicuity maps are blended to generate the final saliency map. Alternatively and motivated by the ant foraging metaphor, in the proposed model the saliency map is taken as the blend of the pheromone maps

generated by the set of p-ants. A p-ant starts its motion in the bottom region of a given conspicuity map and moves on it while deploying pheromone on two pheromone maps according to: (1) the fusion of a set of perception-action rules; (2) a random factor; and (3) a pheromone-attraction factor. Perception-action rules allow p-ants to exploit local conspicuity information in order to approximate their trajectory to the actual trail's skeleton. The pheromone-based interactions allow p-ants to overcome the local image variations and implicitly help each other in building a consensus on the best trail's skeleton approximation. This consensus emerges where the semi-exploratory behaviour of the p-ants is consistently constrained by the perceptual footprint of the trail. In a sense, the set of p-ants self-organises around a given environmental attractor.

The final saliency map is integrated across time in a dynamical neural field (Amari, 1977; Rougier and Vitay, 2006), which feeds back to the pheromone maps in the subsequent frame. This process endows p-ants with historical influence, which is key for tracking the trail. A neural field is a two-dimensional (2-D) lattice of neurons with "Mexican-hat" shaped lateral coupling. With local excitatory connections and long-range inhibitory connections, this inter-neuron coupling helps in the formation of a coherent focus of attention (Rougier and Vitay, 2006).

Note that in this trail detection model both neuron-based and agent-based computational paradigms cohabit. Neurons are used in those situations where there is homogeneous and isotropic spatial connectivity, as are the cases of the neural field and of the conspicuity maps. Conversely, agents are useful where connectivity is much more complex and harder to obtain, as it is the case of creating a trail hypothesis over noisy conspicuity maps. Experimental results on a large data set reveal the ability of the model to produce a success rate of 91 % at 20 Hz, showing it 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.

1.6 Organisation

This dissertation is organised as a collection of four journal articles, one per chapter, describing the work presented in the previous sections. The absence of an introductory chapter with literature review owes to the fact that the journal articles are themselves comprehensive on this matter. Moreover, although each journal article has its own

bibliography list in the published version, a document-wise merged version is taken here to avoid unnecessary duplication.

In short, the other chapters of this dissertation are:

Chapter 2: Stereo-Based All-Terrain Obstacle Detection Using Visual Saliency. This chapter is the article published in the *Journal of Field Robotics* on the use of visual attention methods to detect obstacles in uneven off-road environments, based on stereoscopic vision sensors.

Chapter 3: A Swarm Cognition Realisation of Attention, Action Selection and Spatial Memory. This chapter is the article published in the *Adaptive Behavior* journal on swarm cognition for local navigation. This work was only validated in simulation.

Chapter 4: Swarm Cognition on Off-Road Autonomous Robots. This chapter is the article published in the *Swam Intelligence* journal, which extends the swarm cognition model for local navigation to robots equipped with stereoscopic vision sensors operating on mostly planar off-road environments. The focus of this chapter is on the local navigation framework, and not on the specifically employed obstacle detection technique, and so its contributions are orthogonal to the contributions of Chapter 2.

Chapter 5: Finding Natural Trails with Swarm-Based Visual Saliency. This chapter is a work submitted for publication in the *Journal of Field Robotics* on a novel swarm-based method for visual saliency computation and on the first application of visual saliency for the problem of trail detection.

Chapter 6: Conclusions and Future Work. This chapter presents a set of conclusions extracted from the results obtained in this thesis, points the way to future research, and lists the set of publications produced during the PhD period.

Chapter 2

Stereo-Based All-Terrain Obstacle Detection Using Visual Saliency

Pedro Santana, Magno Guedes, Luís Correia, José Barata.
Journal of Field Robotics, 28(2):241-263, 2011.

Abstract

This paper proposes a hybridisation of two well-known stereo-based obstacle detection techniques for all-terrain environments. While one of the techniques is employed for the detection of large obstacles, the other is used for the detection of small ones. This combination of techniques opportunistically exploits their complementary properties to reduce computation and improve detection accuracy. Being particularly computation intensive and prone to generate a high false positive rate in the face of noisy three-dimensional point clouds, the technique for small obstacle detection is further extended in two directions. The goal of the first extension is to reduce both problems by focussing the detection on those regions of the visual field that detach more from the background and, consequently, are more likely to contain an obstacle. This is attained by means of spatially varying the data density of the input images according to their visual saliency. The second extension refers to the use of a novel voting mechanism, which further improves robustness. Extensive experimental results confirm the ability of the proposed method to robustly detect obstacles up to a range of 20 m on uneven terrain. Moreover, the model runs at 5 Hz on 640×480 stereo images.

keywords: all-terrain perception, obstacle detection, stereo-vision, visual saliency, terrestrial robotics.

2.1 Introduction

The unconstrained appearance of obstacles in all-terrain environments results in the essential use of volumetric information for their detection. Recent developments have made practical the use of laser scanners and stereoscopic vision sensors for the acquisition of volumetric data, i.e., dense three-dimensional (3-D) point clouds. A 3-D point cloud is said to be dense when it is not confined to the representation of notable locations in the environment. In the case of stereoscopic vision, this virtually means that the 3-D position of every patch of the environment that is imaged by the sensor needs to be known. With a high resolution vision sensor, one can obtain denser 3-D point clouds than those produced by typical laser scanners. When compared to laser scanners, stereoscopic vision sensors also tend to be less power consuming and to be lighter, smaller, and cheaper. In contrast to laser scanners, stereoscopic vision sensors provide 3-D data whose acquisition is exactly registered and synchronised with visual data. This is an important asset to facilitate both object and place recognition. However, stereoscopic sensors are usually less accurate than laser scanners, more computationally intensive, more dependent on lighting conditions, and less prepared to reconstruct the 3-D structure of textureless surfaces.

Motivated by the previously mentioned benefits of stereoscopic vision, this paper proposes a set of developments to tackle the major challenges associated to its use for the problem of all-terrain obstacle detection. These challenges are mostly related with the ability of efficiently and robustly processing the considerable amount of noisy and inaccurate sensory data that is generated by the sensor. Nonetheless, we believe that the concepts developed in this paper may also be applied to volumetric data generated by 3-D laser scanners, provided that the point cloud is associated with visual data. This can be done by registering a camera to the laser scanner.

Hard assumptions on the environment's topology are usually exploited to enable fast stereo-based terrain assessment. The most common assumption states that the terrain can be reasonably approximated by planar models (Batavia and Singh, 2001; Labayrade et al., 2002; Dang and Hoffmann, 2005; Soquet et al., 2007; Vernaza et al., 2008; Konolige et al., 2009; Poppinga et al., 2008; Hadsell et al., 2009), which is often not the case on all-terrain. A well known alternative defines obstacles according to geometric relationships among neighbour 3-D points (Bellutta et al., 2000; Talukder et al., 2002; Manduchi et al., 2005; Dubbelman et al., 2007; van der Mark et al., 2007),

thus relaxing the planar terrain assumption. However, this alternative technique is computationally intensive and suffers from a considerable sensitivity to both noise and sparsity in 3-D point clouds. This paper proposes and validates a hybrid model for obstacle detection, in which the strengths of both techniques are brought together to attain higher levels of robustness, accuracy, and computational efficiency. In this sense, two key aspects of this work are the use of the planar assumption to detect large obstacles and the use of the geometric relationships among neighbour 3-D points to detect small ones.

Additionally, variable data density, also known as space-variant resolution (SVR), is used in the computations to improve the overall efficiency of the algorithms. Visual saliency is used to determine regions of particular interest where a higher density of data must be processed. This saliency-based modulation of the SVR allows computation to be focused on the most promising regions of the environment while reducing the false positive rate. Finally, a novel voting mechanism is also introduced to augment the robustness of obstacle detection in the presence of noisy 3-D point clouds. Preliminary versions of this saliency-based obstacle detection model can be found in previous publications (Santana et al., 2008b, 2009, 2010d).

This paper is organised as follows. Section 2.2 relates this proposed method to previous work. Section 2.3 overviews the proposed system architecture. Sections 2.4 and 2.5 describe the novel way visual saliency is computed and the ground-plane is estimated, respectively. Then, in Section 2.6, particular focus is given to the small obstacle detector. Parameter selection details are subsequently provided in Section 2.7. Finally, experimental results are described in Section 2.8, followed by conclusions and future work in Section 2.9.

2.2 Related Work

This section describes related work on terrain classification based on volumetric data provided by laser scanners or stereoscopic vision sensors, with particular emphasis on work validated on all-terrain environments.

Only a few assumptions can be made regarding the structure of all-terrain environments. A typical assumption states that the environment's ground can be modelled by planes. In this case, obstacles are considered to be these 3-D points standing out of the estimated ground-planes (Dang and Hoffmann, 2005; Vernaza et al., 2008; Konolige

et al., 2009). Some local statistics regarding the distance 3-D points have to the ground-planes can be used to reduce the method's sensitivity to false positives (Hadsell et al., 2009). This plane-based approach can also be applied indirectly, such as through an estimated homography (Batavia and Singh, 2001) or through a Hough space of planes (Poppinga et al., 2008). Also under the planar assumption, the v-disparity space approach (Labayrade et al., 2002; Soquet et al., 2007) is usually employed for on-road obstacle detection. These plane-based approaches are highly appealing due mostly to their computational efficiency. However, the unevenness of typical all-terrain environments breaks down the planar assumption. When that happens, a terrain's surface variations can be erroneously characterised as obstacles. Nevertheless, compelled by its computational parsimony, we propose in this work to make a contextualised use of the planar assumption, i.e., only for large obstacle detection.

A way of relaxing the planar terrain assumption is through the use of heuristics applied locally to the disparity/range image (Broggi et al., 2005; Schafer et al., 2005; Caraffi et al., 2007; Konolige et al., 2009). Heuristics in the form of local point statistics obtained directly from the 3-D point cloud can also be used to produce accurate results (Wellington and Stentz, 2004; Lalonde et al., 2006). However, this technique is impracticable for stereoscopic vision due mostly to its noisy nature. Heuristics on the residual resulting from a line fitting process can also be applied on a scan-by-scan basis to data generated by two-dimensional (2-D) laser scanners (Moorehead et al., 1999; Batavia and Singh, 2002; Castano and Matthies, 2003; Urmson et al., 2006; Andersen et al., 2008).

Alternative heuristics can be applied when a dense 3-D point cloud is available. One example is traversability estimation in terms of the residuals resulting from several local plane-fitting processes applied to the 3-D point cloud (Moorehead et al., 1999; Gennery, 1999; Singh et al., 2000; Goldberg et al., 2002; Biesiadecki and Maimone, 2006; Ye, 2007). More recently, octree decomposition has been employed to create grid representations of the 3-D point cloud (Rusu et al., 2009). Polygonal models are then fitted to each cell and heuristically labelled as ground, level, vertical obstacles, stairs, or unknown. To determine their traversability, plane-fitting processes are then applied to the models falling in the latter category. In addition to its heuristic definition of obstacle, this method still lacks experimental validation on all-terrain.

Heuristics can ultimately be learned (Wellington and Stentz, 2004; Bajracharya et al., 2008). Linguistic heuristic rules can also be applied to blend several cues when com-

puting traversability indexes (Seraji, 1999, 2006). The major limitation of heuristic-based solutions is the difficulty in defining obstacles in terms of the robot's mobility. A way of circumventing this limitation is through the construction of Digital Elevation Maps (DEM) of the environment, upon which a detailed kinematic and/or dynamic model of the robot can be used for safe motion planning (Kelly and Stentz, 1998; Lacroix et al., 2002; Plagemann et al., 2008; Kolter et al., 2009). However, these solutions tend to be too computationally demanding.

To take into account key mobility properties, such as the clearance height under the robot, obstacles can also be defined in terms of geometrical relationships between neighbour 3-D points. In the case of 2-D laser scanners these relationships can be computed in scan-by-scan fashion (Chang et al., 1999). A similar process can also be applied on a column-by-column way in stereo-based systems (Bellutta et al., 2000; Dubbelman et al., 2007). Probabilistic models have been successfully applied for improved robustness in the integration of evidence across 2-D laser scans (Thrun et al., 2006a). Talukder et al. (2002) and Manduchi et al. (2005) proposed and validated on a stereo-based system a more general geometric approach that can be applied to 3-D point clouds. However, this method is computationally intensive and sensitive to artefacts induced by the 3-D reconstruction process. These limitations have been partially mitigated with the use of look-up tables and explicit handling of uncertainty (van der Mark et al., 2007). In this work, a higher level of robustness is attained with a novel voting filter and an improved posture compensation mechanism. To speed up computation we propose the synergistic use of visual saliency and SVR.

The robot's focus of attention and respective resolution can be defined in terms of the robot's speed so as to avoid holes/overlaps in the analysis of consecutive frames (Kelly and Stentz, 1998). This process is in line with the active vision approach (Bajcsy, 1988; Aloimonos et al., 1988; Ballard, 1991) and it can also be used to reduce the risk of not detecting obstacles (Grandjean and Matthies, 1993). Our method dynamically adapts the resolution and focus of attention based on the contents of the visual input. Hence, despite sharing the same goal of reducing the cost of perception to its minimum, the two approaches show complementary properties. As will be shown, our saliency-based SVR mechanism has the additional advantage of reducing the false positive rate.

Visual saliency has been thoroughly employed for object detection in indoor environments (Vijayakumar et al., 2001; Orabona et al., 2005; Moren et al., 2008; Meger

et al., 2008; Yu et al., 2007). It has also been used to select strong landmarks for visual simultaneous localisation and mapping in urban environments (Newman and Ho, 2005; Frintrop et al., 2007). Laser-based range images have been used to focus the analysis of registered colour images on the task of searching for information placards along dirt roads (Hong et al., 2002). In this case one might say that salient regions in the laser data are used to focus the analysis of a colour-based detector. However, to the best of our knowledge, our proposed method is the first application of visual saliency to modulate all-terrain obstacle detection.

2.3 System Overview

The proposed model (see Fig. 2.1) is characterised in part by the novel integration of two complementary obstacle detection techniques. The two techniques differ mostly in their definition of an obstacle. The first considers as obstacles those 3-D points that stand out from an estimated ground-plane (Dang and Hoffmann, 2005; Vernaza et al., 2008; Konolige et al., 2009). Because most all-terrain environments are not perfectly planar, this detector can search robustly only for large obstacles. Small variations in height on uneven terrains would often be confused with small obstacles. Hence, the obstacle detector is configured in a way that only obstacles above a large distance h off the ground plane are detected. Smaller obstacles are instead considered by the second technique, which relaxes the planar assumption by defining obstacles according to geometrical relationships between neighbour 3-D points. Although this second technique is more robust in uneven terrains, it requires dense 3-D point clouds. The level of noise and sparseness of the 3-D point cloud in large homogeneous objects, due to failure in the stereo reconstruction process, makes this technique less capable of detecting large obstacles. As confirmed by experimental results (see Section 2.8.3), from the strengths of both techniques emerges a more robust all-terrain obstacle detector.

Another relevant contribution of the model is the small obstacle detector itself (see Section 2.6), which departs from previous related work (Talukder et al., 2002; Manduchi et al., 2005; van der Mark et al., 2007) by improving its robustness and computational efficiency. The latter is partially due to the innovative use of visual saliency (see Section 2.4) to modulate all-terrain obstacle detection. The challenges posed by this new domain required some adaptations to the standard way of computing saliency (e.g., (Itti et al., 1998; Frintrop et al., 2005)). The model also exploits for the first time

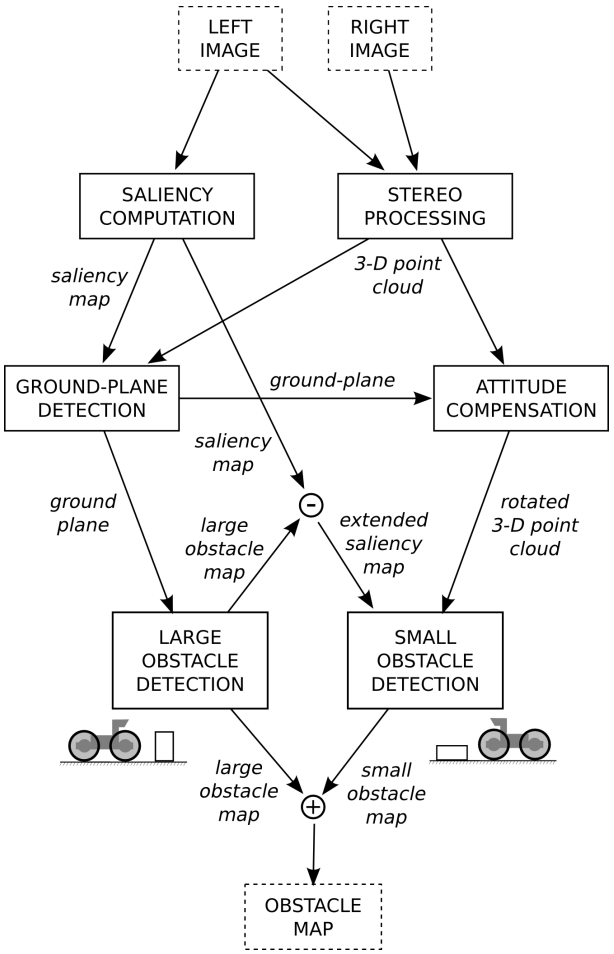


Figure 2.1: Proposed model building blocks.

visual saliency on the task of ground-plane estimation (see Section 2.5). The estimated ground-plane supports the large obstacle detector while it permits the rotation of the 3-D point cloud to compensate for the robot's posture in the small obstacle detector computations (see Section 2.6.5).

To avoid unnecessary and consequently inefficient applications of the small obstacle detector, the output of the large obstacle detector is used to mask the saliency map with negative values. Using the masked saliency map instead of the original saliency map, the small obstacles detector will ignore regions with negative saliency that have already been labelled by the other detector. Finally, the outputs of both detectors are merged in order to produce the final obstacle map.

We now characterise which types of obstacle can be detected with the proposed

model. For this purpose we assume that an environment can be generally classified in terms of its slipperiness, compressibility, permeability, and morphology. With the exception of the fourth property, which can be assessed with volumetric data, these properties require an appearance-based analysis. Consequently, the proposed model is constrained to detect obstacles that can be segmented by the background based solely on their 3-D morphology. Both positive (e.g., rocks/trees) and negative (e.g., ditches/ holes) obstacles are detected by the proposed model provided that they are associated to sufficient 3-D information.

From this analysis it follows that the proposed model is not capable of detecting water bodies nor is it capable of distinguishing between compressible (e.g., wall of tall grass) and noncompressible obstacles. Nevertheless, due to its fragile morphological characteristics, spurious tall grass is filtered out by the system. It follows also that the model is unable to assess the slipperiness of the terrain, as it typically requires the identification of the materials composing it.

2.4 Saliency Computation

The goal of using visual saliency is to determine which regions of the visual field stand out more significantly from the background. The basic idea is that the higher its associated saliency, the more prone a given region of the visual field is to contain an object. Thus, perceptual processes exploiting this input are better fit to handle the speed-accuracy trade-off. The following describes the proposed saliency model, which is an adaptation for all-terrain environments of the biologically-inspired model proposed by Itti et al. (1998).

Let L be the left image of an image pair obtained from a stereoscopic vision sensor. To reduce computational cost, saliency is computed on a region of interest (ROI) of L. The ROI is a horizontal strip between the bottom row and row u. Row u corresponds to the uppermost row containing at least a given percentage ς of pixels with an associated depth within the minimum and maximum considered ranges for obstacle detection, r_{min} and r_{max}, respectively. With this process we guarantee that the definition of the ROI is not affected by spurious pixels whose associated 3-D points are erroneously beyond r_{max}. To further reduce computational cost, all image operators are performed over 8-bit pixels whose magnitude is clamped to $[0, 255]$. All experiments use 640×480 input images.

A dyadic Gaussian pyramid $I(\sigma)$ with six levels $\sigma \in \{0,\ldots,5\}$ is created from the intensity channel of the ROI. The resolution scale of level σ is $1/2^{\sigma}$ times the ROI resolution scale. Intensity is obtained by averaging the three colour channels. Then, four on-off centre-surround intensity feature maps are created to promote bright objects on dark backgrounds. Four off-on centre-surround intensity feature maps are also created to promote dark objects on bright backgrounds. On-off centre-surround operations are performed by across-scale, point-by-point subtraction between level c, with a finer scale, and level s, with a coarser scale linearly interpolated to the finer scale, with $(c,s) \in \Omega = \{(2,4),(2,5),(3,4),(3,5)\}$. Off-on maps are computed the other way around, that is, by subtracting the coarser level from the finer level. On-off, $I^{on-off}(c,s)$, and off-on, $I^{off-on}(c,s)$, centre-surround maps are then combined to generate the intensity conspicuity map

$$C_I = \sum_{i \in \{on-off, off-on\}} \left(\frac{1}{2} \bigoplus_{(c,s)\in\Omega} I^i(c,s) \right), \tag{2.1}$$

where the across-scale addition $\bigoplus$ is performed with point-by-point addition of the maps after being scaled to the resolution of level $\sigma = 3$. Sixteen orientation feature maps, $O(\sigma,\theta)$, are created by convolving levels $\sigma \in \{1,\ldots,4\}$ with Gabor filters tuned to orientations $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. Gabor filters are themselves centre-surround operators and therefore require no across-scale subtraction procedure (Frintrop, 2006). As before, all orientation feature maps are combined at the resolution of level $\sigma = 3$ in order to create the orientation conspicuity map

$$C_O = \sum_{\theta \in \{0^\circ,45^\circ,90^\circ,135^\circ\}} \left(\frac{1}{4} \bigoplus_{\sigma\in\{1,\ldots,4\}} O(\sigma,\theta) \right). \tag{2.2}$$

The saliency map S (e.g., Fig. 2.2) is then obtained by modulating the intensity conspicuity map C_I with the orientation conspicuity map C_O:

$$S = \mathcal{M}\left(\frac{1}{2} \cdot \mathcal{N}(C_I), \frac{1}{2} \cdot \mathcal{N}(C_O) \right), \tag{2.3}$$

where $\mathcal{M}(A,B) = A \cdot sigm(B)$, $sigm(\cdot)$ is the sigmoid operator and $\mathcal{N}(\cdot)$ scales the provided image amplitude between $[0,255]$.

This saliency model is essentially the one proposed by Itti et al. (1998) with one

(a) Input image 5

(b) Saliency map of image 5

(c) Input image 6

(d) Saliency map of image 6

Figure 2.2: Examples of typical saliency maps. ROI with $r_{max} = 10\,\text{m}$. The brighter the pixels in the saliency map, the higher their saliency. From these examples, it is possible to see the concentration of salient pixels on obstacle regions.

general improvement and two adaptations for all-terrain obstacle detection. The improvement was proposed by Frintrop (2006), and it refers to the use of both on-off and off-on feature channels separately. The first adaptation for all-terrain obstacle detection is in the normalisation operator $\mathcal{N}(.)$, which does not try to promote maps according to their number of activity peaks, as is typically done. This is because the frequency at which objects appear in the maps is not necessarily correlated to their proneness of being obstacles to the robot.

The second adaptation for all-terrain obstacle detection is in the way conspicuity maps are blended. Rather than their typical addition, conspicuity maps are non-linearly multiplied. This results in making salient only those regions of the environment that are simultaneously conspicuous in both orientation and intensity channels. The basic idea is that obstacles on all-terrain environments are normally highly textured and at the same time conspicuous in the intensity channel. The non-linearity introduced by the sigmoid operator in Equation 2.3 aims at inhibiting the saliency map in the presence of low orientation conspicuity in order to remove the background noise inherent in textured outdoor terrains. When orientation conspicuity is strong, the sigmoid operator amplifies the saliency map and promotes the orientation conspicuity map over the intensity one.

Although not focused on all-terrain obstacle detection, a parallel and independent study (Hwang et al., 2009) has also reported the benefits of using a weighted product of the conspicuity maps as an alternative to their standard summation. Fig. 2.3 compares

(a) Input image (b) Saliency with proposed model (c) Saliency with Itti's model

Figure 2.3: Comparison between the saliency map generated by the proposed model (b) and the saliency map obtained with the model proposed by Itti et al. (1998) (c), for a typical input image (a). ROI with $r_{max} = 10\,\text{m}$.

a saliency map computed by the proposed model with a saliency map obtained with the model proposed by Itti et al. (1998).

2.5 Ground Plane Estimation

Obstacles can be in part characterised as disturbances occurring at the ground surface. Thus, estimating how the robot relates to the ground-plane based on a dense 3-D point cloud is of extreme importance to support obstacle detection. However, the roughness of outdoor terrain renders highly unlikely the existence of a clear-cut ground-plane. Thus, robust estimation methods, such as random sample consensus (RANSAC) (Fischler and Bolles, 1981), must be employed. In a nutshell, a typical RANSAC procedure for ground-plane estimation in the ROI is composed of the following seven steps:

Step 1. Pick randomly a set R of three non-collinear 3-D points within range $[r_{min}, r_{max}]$ and generate its corresponding ground plane hypothesis, h_R, with some straightforward geometry.

Step 2. The score of the plane hypothesis is the cardinality of the set of its inliers, $score(h_R) = |P_{h_R}|$. An inlier $\mathbf{p} \in P_{h_R}$, where $\mathbf{p} = (p_x, p_y, p_z)^T$, is a 3-D point whose distance to plane h_R, $d(\mathbf{p}, h_R)$, is smaller than a given threshold d_{plane}. P_{h_R} is thus the set of all 3-D points that cope with this condition and that are at the same time located in the ROI.

Step 3. Repeat steps 1 and 2 until n_{hypo} hypotheses, composing a set H, have been generated.

Step 4. Select from H the hypothesis with highest score for refinement, $b = \arg\max_{h \in H} score(h)$.

Step 5. Compute a refined version of b, b', by fitting the inliers set of b, P_b. This fitting is done with weighted least-squares orthogonal regression via the well known Singular Valued Decomposition (SVD) technique. The weight $w_{\mathbf{q}}$ of an inlier $\mathbf{q} \in P_b$, where $\mathbf{q} = (q_x, q_y, q_z)^T$, is given by $w_{\mathbf{q}} = 1 - \frac{d(\mathbf{q},b)}{d_{plane}}$. That is, the farther $\mathbf{q}$ is from b, the less it weights in the fitting process. Compute the inliers set of b', $P_{b'}$, and substitute the current best ground-plane estimate with the refined one. That is, make $b = b'$ and $P_b = P_{b'}$. $P_{b'}$ is computed with the procedure described in Step 2.

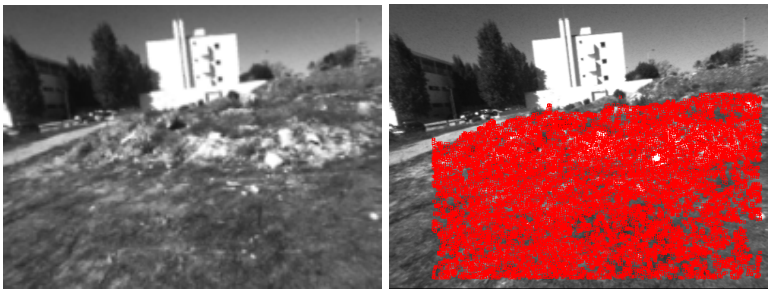
Step 6. Iterate step 5 until $|P_b|$ becomes constant across iterations or a maximum number of iterations, m_{refit}, is reached.

Step 7. Take b as the ground-plane estimate.

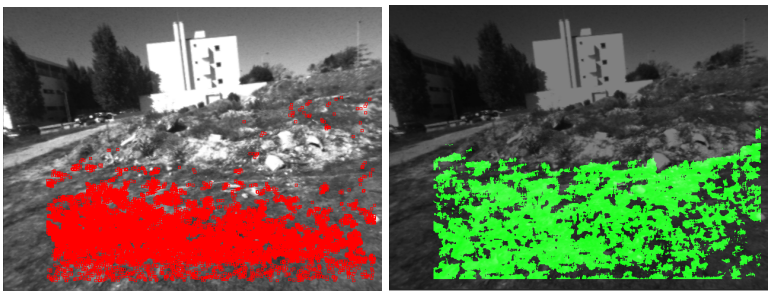
Points belonging to obstacles will inevitably generate poor ground-plane hypotheses in Step 1. This means that in environments cluttered with obstacles, a larger number of hypotheses must be generated to guarantee that a good hypothesis is found. This additional computation can be reduced by rejecting selected points if they are likely to belong to an obstacle. This likelihood is here defined in terms of visual saliency level (see Fig. 2.4(b-c)).

Formally, in Step 1, a randomly selected 3-D point $\mathbf{p}$ is promptly rejected and so not considered to build a plane hypothesis, if $s_{\mathbf{p}'} > \frac{x}{\rho \cdot n_l}$, where $x \in [0, 255]$ is a number sampled from a uniform distribution each time the inequality is tested; $n_l \in [0, 1]$ is the ratio of pixels with saliency below a given threshold l; *local saliency* $s_{\mathbf{p}'} \in [0, 255]$ is the maximum saliency within a given sub-sampled chess-like squared neighbourhood of p', with size $q \cdot n_l$, q being the empirically defined maximum size; and ρ is an empirically defined scaling factor. The goal of using the ratio of pixels with a saliency value under a given threshold is to allow the system to progressively fall-back to a non-modulated procedure as saliency reduces its discriminative power. This happens, for instance, in highly textured terrains, in which the sampling procedure is too constrained as a result of the saliency map's cluttering. See Fig. 2.4(d) for an example of the output generated by the ground-plane estimation process.

Complementary mechanisms could be exploited for improved performance and robustness. In terms of saliency modulation, the score of each ground-plane hypothesis



(a) Left input image (b) Without saliency



(c) With saliency (d) Ground points

Figure 2.4: Typical example illustrating the advantages of using visual saliency to modulate the hypothesis generation step of a RANSAC procedure for ground-plane estimation. (b) Overlaid pixels (in red) correspond to 5000 3-D points randomly sampled without saliency modulation. (c) Same as (b) but with saliency modulation. (d) Overlaid pixels (in green) whose corresponding 3-D points lie on the estimated ground-plane.

could be weighted according to the saliency of its supporting inliers. Additionally, the weight assigned to each 3-D point in the weighted least-squares orthogonal regression process could also be modulated by its associated saliency. Spatio-temporal prior information could also be applied to modulate the RANSAC process (Mufti et al., 2008; Chumerin and Hulle, 2008) with the purpose of increasing the chances of detecting the actual plane given a fixed number of RANSAC iterations. The 3-D points of obstacles detected in the previous frame could also be directly removed from consideration in the estimation process (van der Mark et al., 2007). If ground-plane estimation and detected obstacles removal are done sequentially and in an iterative way per frame, multiple planes could also be detected (Hadsell et al., 2009).

2.6 Small Obstacle Detection

Detecting obstacles based uniquely on the distance between 3-D points and the ground-plane is a brittle procedure on all-terrain. This is mostly because small variations in height on uneven terrains would often be confused with small obstacles. In those situations obstacles are better defined in terms of geometrical relationships between their composing 3-D points. Aligned with this idea, the model proposed by Talukder et al. (2002) and Manduchi et al. (2005), which is summarised in Section 2.6.1 and from now on designated Original Obstacle Detector (OOD), is taken as the starting point for our small obstacle detector. To reduce its computational cost and sensitivity to noise, the OOD is here extended with a SVR mechanism and a voting filter, as described in Sections 2.6.2 and 2.6.3, respectively.

2.6.1 Obstacle Definition

Let $\{x, y, z\}$ be the basis that defines three axes relative to the centre of the left camera and with the z-axis aligned with the sensor's optical axis (see Fig. 2.5(b)). This basis is the sensor's local reference frame, $\mathcal{F}$. Let P be the set of 3-D points, defined with respect to $\mathcal{F}$, computed by the stereo-based 3-D reconstruction process and provided to the obstacle detector. As proposed by Talukder et al. (2002) and Manduchi et al. (2005), a 3-D point is considered to belong to an obstacle if it is *compatible* with any other 3-D point in the point cloud. Two 3-D points obtained from P, $\mathbf{p_a} = (x_a, y_a, z_a)^T$ and $\mathbf{p_b} = (x_b, y_b, z_b)^T$, are said to be *compatible* if

$$H_{min} < |y_b - y_a| < H_{max} \wedge \frac{|y_b - y_a|}{\|p_b - p_a\|} > \sin\theta, \tag{2.4}$$

where θ is the maximum slant angle that the vehicle can safely negotiate, H_{min} is the clearance height under the vehicle, and H_{max} is the maximum height to be considered by the detector.

Fig. 2.5 provides an intuitive geometrical interpretation of this definition of compatibility. That is, all 3-D points that are compatible with a given 3-D point $\mathbf{p} = (x, y, z)^T$ are encompassed by two truncated cones, $U_\mathbf{p}$ and $L_\mathbf{p}$. Although pointing in opposite directions, both upper and lower truncated cones are normal to the xy plane and have their vertexes located in $\mathbf{p}$. Additionally, both cones have an aperture angle of $(\pi - 2\theta)$ and are limited by the planes $y = H_{min}$ and $y = H_{max}$.

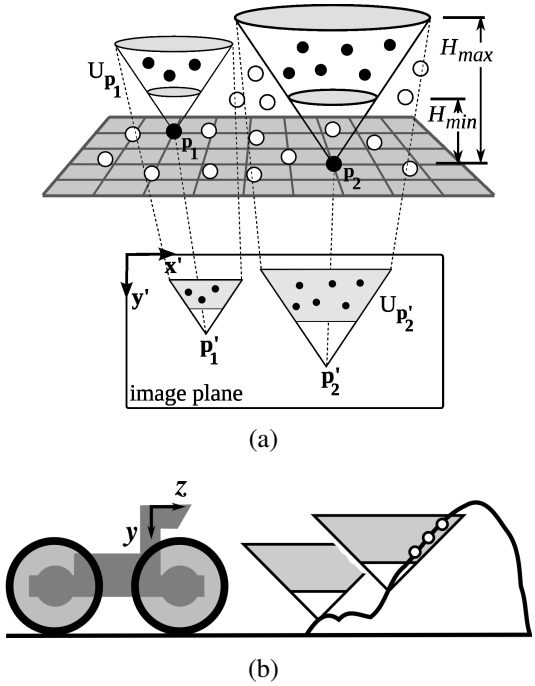


Figure 2.5: Geometric interpretation of the compatibility test (Talukder et al., 2002; Manduchi et al., 2005). (a) Filled and unfilled circles represent 3-D points that are compatible and incompatible, respectively, with points $\mathbf{p_1}$ and $\mathbf{p_2}$. It is also possible to depict the projections of 3-D points $\mathbf{p_1}$ and $\mathbf{p_2}$ onto the image plane, i.e., pixels $\mathbf{p'_1}$ and $\mathbf{p'_2}$, respectively. Note that the farther the points are from the sensor, the smaller the projections of their truncated cones are. For readability reasons, $L_{\mathbf{p_1}}$, $L_{\mathbf{p'_1}}$, $L_{\mathbf{p_2}}$ and $L_{\mathbf{p'_2}}$ are not represented. (b) Diagram illustrating a situation in which the compatibility test allows the detection of an obstacle in front of the robot. The white circles represent the points falling in the upper truncated cone and consequently considered compatible, that is, pertaining to the same obstacle.

Checking the compatibility between all possible 3-D points is limited by real-time requirements. Fortunately, according to Talukder, Manduchi et al., only a reduced subset of all combinations needs to be assessed if the process is carried out in the image space. For this purpose, $\mathbf{p}' = (p'_x, p'_y)^T$ is defined as the projection of the 3-D point $\mathbf{p} \in P$ onto the left camera's image plane. This projection is defined with respect to the local reference frame $\{x', y'\}$ and refers basically to the pixel in the left camera that is the image of the world point $\mathbf{p}$ (see Fig. 2.5(a)). Similarly, the two truncated cones of $\mathbf{p}$ project onto two truncated triangles in the image plane, $U_{\mathbf{p}'}$ and $L_{\mathbf{p}'}$, whose vertexes are both located at pixel $\mathbf{p}'$ (see Fig. 2.5(a)). The height of these truncated triangles is given by $H_{max} f / p_z$, and their width can be approximated by $\frac{2H_{max} f}{\tan\theta_{max} p_z \cos\eta}$, where f is

the camera's focal length and $\eta = \arctan \frac{p_x}{p_z}$. In the image space, the set of pixels that may be compatible with $\mathbf{p}'$, and consequently with $\mathbf{p}$, are now constrained to those encompassed in both truncated triangles. Furthermore, if the image is scanned from bottom to top and from left to right, it suffices to consider only the upper truncated triangles to efficiently detect and label all pixels.

Finally, two points that are compatible to each other are said to pertain to the same obstacle. By transitivity, two points that are linked by a chain of compatible points are also said to belong to the same obstacle. This property will enable the segmentation of obstacles according to their 3-D relationships.

2.6.2 Space-Variant Resolution

Despite the advantages of using a truncated triangle in order to focus the application of the compatibility test, the computation time of the method is still on the order of seconds for 640×480 image-pairs. With quantisation and extensive use of look-up tables it is possible to attain faster processing rates (van der Mark et al., 2007). However, the loss of accuracy is unavoidable with such approximations. This section proposes the use of SVR as a complementary way to reduce the computational cost of the method.

For fast detection, pixels are *coarsely analysed* in a first phase according to the scanning procedure previously described, that is, from bottom to top and from left to right. However, this time the analysis is done with steps of n pixels, which may be skipped based on their visual saliency and two additional constraints (see below). Furthermore, compatibility is tested only against a sub-sampled set of the pixels falling inside the upper truncated triangle of each analysed pixel. This sub-sampling is done in a chess-like pattern with $1/m$ of the image resolution.

Whenever an obstacle is detected, it is foveated by performing a *finer analysis* of the region. Concretely, the pixels within the upper truncated triangle of the pixel just labelled obstacle are re-sampled from $1/n$ of the image full resolution, with $n < m$. This aims at improving the representation of any obstacle that has been detected.

As soon as the scanning procedure is completed, the detector enters in a second phase for *full resolution recovery*. This final phase is important as most morphological filters that might be applied afterwards perform better in high resolution. This second phase is implemented with the following region growing mechanism. Let $\mathbf{p}'_\mathbf{1}$ be a pixel labelled obstacle and $\mathbf{p_1}$ its corresponding 3-D position. Every other pixel $\mathbf{p}'_\mathbf{2}$ whose

distance to $\mathbf{p}'_1$ is smaller than the highest number of skipped pixels, $||\mathbf{p}'_1 - \mathbf{p}'_2|| < m$, is a candidate also to become labelled as obstacle. The final test to determine whether $\mathbf{p}'_2$ is labelled as obstacle is done by checking whether its corresponding 3-D point, $\mathbf{p_2}$, is at a distance from $\mathbf{p_1}$ shorter than an approximation of the maximum allowed distance between 3-D points to be considered compatible, $||\mathbf{p_1} - \mathbf{p_2}|| < H_{max}$.

As mentioned, in order to save computation, the compatibility test is conditionally applied in the coarse analysis phase. Concretely, the compatibility test is applied only if the *local saliency* increases between scanned pixels or at least one of two additional constraints is applicable. Local saliency is preferred over a pixel-wise one to reduce the effects of poor lighting conditions, which in some situations make objects' upper part appear more salient than their lower part. Local saliency is computed by taking the maximum saliency from the set of pixels that share the column of the pixel being analysed and that are contained in its truncated triangle. Owing to the fact that the compatibility test is skipped over non-interesting regions and therefore with low likelihood of containing obstacles, certain regions of the environment are more coarsely analysed and so computational cost is saved.

Let us now describe the two additional constraints that forces the execution of the compatibility test during the coarse analysis. The first constraint aims at performing a fine analysis whenever an obstacle is detected and a *progressive* fine-to-coarse analysis of the obstacle's boundaries. The latter effect is particularly useful in handling noisy data in and around obstacles. For this purpose, the maximum number of pixels that may be skipped before compulsorily engaging in a new compatibility test is progressively increased as the scanning process moves away from a detected obstacle. This number is represented by the variable n_{slide}. This variable is set to n whenever a compatibility test returns positive (i.e., when the scanning process is still on the obstacle) and doubled whenever its value matches the number of skipped pixels (i.e., when the compatibility test has been compulsorily engaged as a consequence of the maximum number of pixels that may be skipped has been reached).

Along the same line of reasoning, instead of analysing every row of the input image, $n + w$ rows are skipped, where w is incremented every time an analysed row does not contain any pixel with a positive compatibility test. To avoid large jumps w is upper bounded by w_{max}. Whenever a compatibility test succeeds, w is zeroed. This procedure intends to reduce the computational load in environments with few obstacles or when these are located mostly in the far-field. Because the truncated triangles associated with

(a) Input image (b) Saliency map (c) Saliency-based SVR

Figure 2.6: Typical example of pixels analysed with saliency-based SVR. (c) Pixels analysed based on the saliency map depicted in (b). White pixels in (c) correspond to points that have been skipped by the detector due to lack of saliency or computed range, whereas grey pixels correspond to points that have been analysed. ROI with $r_{max} = 10\,\text{m}$, $H_{min} = 0.10\,\text{m}$, $H_{max} = 0.4\,\text{m}$, and $\theta = 40^\circ$. Space-variant resolution with $n = 3$, $m = 6$, $n_{max} = 30$, and $w_{max} = 20$. Note the focus on obstacle regions.

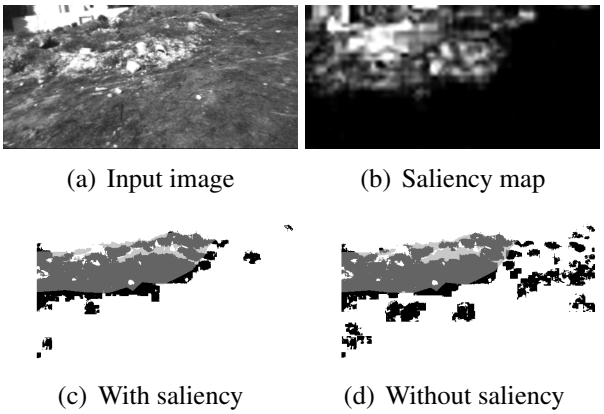
points in the near-field are quite large, skipping rows from the image's bottom when no obstacle is found there greatly reduces the computational load. Fig. 2.6 illustrates the operation of the SVR on a typical input image.

Finally, every time a pixel is labelled obstacle, the saliency of all pixels encompassed by its upper truncated triangle is increased by a given percentage, λ. This reinforcement of the detected obstacle's presence raises the chances of selecting other obstacle's pixels for compatibility testing. The use of saliency to guide a task-specific detector is rather typical (Itti et al., 1998; Frintrop, 2006). However, the proposed model exhibits a novel characteristic to saliency-based systems by allowing the results of the detector to modulate the saliency map.

By not relying on 3-D features, visual saliency is able to guide the detector without being misguided by potential 3-D artefacts. Fig. 2.7 shows that this property helps in the reduction of false positive rate.

2.6.3 Voting Filter

Stereo-based 3-D reconstruction is a rather noisy process. This characteristic is particularly problematic when small distant obstacles must be detected. In this case, the challenge is to devise a set of filters to remove the noise without hampering performance and accuracy. The direct filtering of the 3-D point cloud is a computationally expensive task. The cost comes mainly from the fact that it is not possible to know



(a) Input image (b) Saliency map

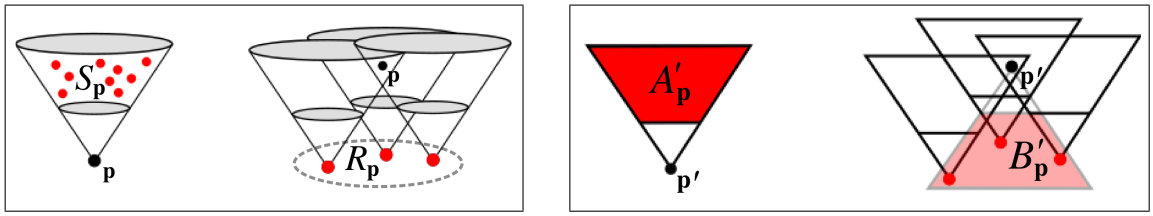
(c) With saliency (d) Without saliency

Figure 2.7: Typical results of the SVR with and without saliency modulation. No additional filters have been turned on. (c)-(d) Black, false positives; dark grey, true positives; bright grey, false negatives. ROI with $r_{max} = 10\,\text{m}$, $H_{min} = 0.10\,\text{m}$, $H_{max} = 0.4\,\text{m}$, and $\theta = 40^\circ$. SVR with $n = 3$, $m = 6$, $n_{max} = 30$, and $w_{max} = 20$. Without the use of saliency, a set of false positives hamper the robot from finding the passage on the right-hand side of the image.

beforehand which regions of the image are important to be handled. An alternative is to perform post-filtering on the segmented obstacles. This focuses the filtering process on those 3-D patches that are known to be relevant for the overall system. However, the higher the noise level the more obstacle segments exist and consequently the more expensive their treatment is. Here, we propose a complementary mechanism: the use of a voting mechanism fully embedded in the obstacle detector. Being embedded in the detection process, the previously mentioned limitations are circumvented. On the one hand, only 3-D points that are relevant to the detector are analysed. On the other hand, removing noise during the detection process reduces the number of obstacle points to be considered by the subsequent segmentation phase.

Let $S_{\mathbf{p}}$ be the set of 3-D points encompassed by the upper truncated cone emanating from the 3-D point $\mathbf{p}$. These points are said to be *voted* by $\mathbf{p}$. Conversely, let $R_{\mathbf{p}}$ be the set of 3-D points whose upper truncated cones encompass $\mathbf{p}$. These points are said to *vote* on $\mathbf{p}$. See Fig. 2.8 for an illustration of these concepts. A direct voting mechanism would be to reject $\mathbf{p}$ as an obstacle if the cardinality of both sets $S_{\mathbf{p}}$ and $R_{\mathbf{p}}$ did not reach a given threshold. However, the effects of projection make the theoretical maximum size of both sets depend on the distance $\mathbf{p}$ is from the camera. In other words, farther obstacles are represented by fewer pixels than closer obstacles. Hence, this approach would result in a scale-variant filtering mechanism and so is inaccurate.

To make it invariant to scale, the voting mechanism applies a threshold to the car-



(a) (b)

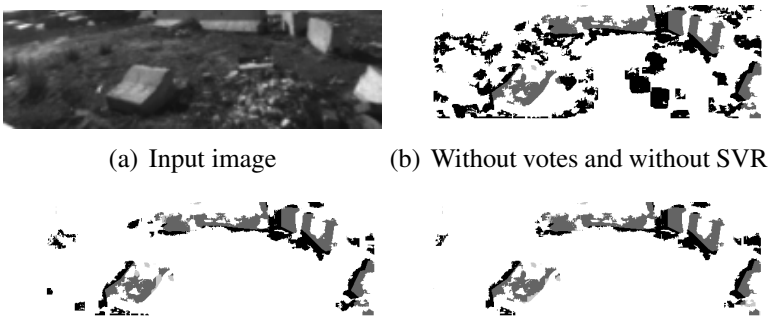
Figure 2.8: Diagrams illustrating the voting mechanism associated with a given 3-D point $\mathbf{p}$ and its corresponding projection on the image plane, $\mathbf{p}'$. (a) Illustration depicting voting relationships between a given 3-D point $\mathbf{p}$ and its neighbours, according to the compatibility test. (b) Projection onto the image plane of the situation depicted in (a). $|A'_{\mathbf{p}}|$ and $|B'_{\mathbf{p}}|$ correspond to the area in pixels of the left and right filled truncated triangles, respectively. If the number of pixels without computed disparity is subtracted from each of these quantities, one obtains the theoretical maximum number of times that $\mathbf{p}$ is able to vote and being voted, respectively.

dinality of the distance-normalised versions of both $R_{\mathbf{p}}$ and $S_{\mathbf{p}}$. As before, let $\mathbf{p}'$ be the projection of $\mathbf{p}$ onto the image plane. Let $A'_{\mathbf{p}}$ be the area in pixels of the upper truncated triangle emanating from pixel $\mathbf{p}'$ (see Fig. 2.8). Let $A''_{\mathbf{p}}$ be the sub-set of $A'_{\mathbf{p}}$ with associated 3-D information. $A''_{\mathbf{p}}$ is said to be the theoretical maximum number of points that may be voted by $\mathbf{p}'$ and consequently by $\mathbf{p}$. Similarly, let $B'_{\mathbf{p}}$ be the area in pixels of the lower truncated triangle emanating from pixel $\mathbf{p}'$ (see Fig. 2.8). Let $B''_{\mathbf{p}}$ be the sub-set of B'_p with associated 3-D information. $B''_{\mathbf{p}}$ is said to be an approximation of the theoretical maximum number of points that may vote on $\mathbf{p}$. This approximation builds on the assumption that the truncated triangles associated to the projections of all 3-D points composing the same obstacle have equal area. Because those points are close to each other, the taken assumption renders a good approximation.

The following test on the number of votes normalised by their theoretical maximum can now be used to determine whether point $\mathbf{p}$ is accepted as an obstacle:

$$\left(\frac{|S_{\mathbf{p}}|}{|A''_{\mathbf{p}}|} > v\right) \vee \left(\frac{|R_{\mathbf{p}}|}{|B''_{\mathbf{p}}|} > v\right), \tag{2.5}$$

where v is an empirically defined threshold. With the voting mechanism, compatibility between two 3-D points is no longer a sufficient condition to consider them as obstacles. Now, a higher level of robustness is attained by defining obstacles with a many-to-many relationship. With this method the detector becomes considerably resilient to the presence of 3-D artefacts (see Fig. 2.9) and even to the type of noise generated by a



(a) Input image (b) Without votes and without SVR

(c) With votes and without SVR (d) With votes and with SVR

Figure 2.9: Typical results of the small obstacle detector without voting mechanism (b), with voting mechanism (c), and with voting mechanism plus SVR (d). In (b) and (c) the SVR has been turned off. (b)-(d) Black, false positives; dark grey, true positives; bright grey, false negatives. ROI with $r_{max} = 10\,\text{m}$, $H_{min} = 0.10\,\text{m}$, $H_{max} = 0.4\,\text{m}$, and $\theta = 40^\circ$. Voting mechanisms with $v = 0.2$ and $a = 25$. SVR with $n = 3$, $m = 6$, $n_{max} = 30$, and $w_{max} = 20$. Note the considerable reduction of false positives when the voting mechanism is turned on. Note also that with SVR, computation time is saved by one order of magnitude for the same detection rate. The lack of data (white pixels) within texture-deprived obstacles is due to failure in the stereo-based 3-D reconstruction process.

partial damage of one of the lenses composing the stereoscopic vision sensor (Santana et al., 2008b).

2.6.4 Area Filter

Although the voting filter is extremely powerful, experimental results will show that its operation can be better exploited when in conjunction with an area filter (see Section 2.8). The area filter comes into play to remove any residual noise left by the voting filters. For this purpose, the obstacle points are first segmented in the 3-D space (Manduchi et al., 2005). The area filter then eliminates those segments whose volume projected onto the image plane is characterised by having a small area. Formally, an obstacle point $\mathbf{p}$ is re-labelled as non-obstacle if

$$|A_{\mathbf{p}}| < \frac{a \cdot 10^2}{p_z^2}, \tag{2.6}$$

where A_p is the set of points composing the segment that encompasses p and a is an empirically defined scalar. This test verifies whether the projected area of the segment is below a pre-specified area $a \cdot 10^2$, properly normalised by the squared distance to the

obstacle. This normalisation procedure introduces scale-invariance into the filter.

2.6.5 Pitch-Roll Compensation

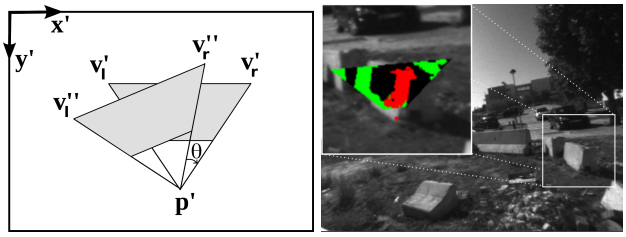
All the above geometrical considerations assume that the camera is not pitched nor rolled with respect to the ground-plane. A possible way of removing this constraint is to compensate for small variations on the camera's attitude by overestimating the truncated triangles' size (Talukder et al., 2002; Manduchi et al., 2005). A lighter and more exact alternative is proposed next.

Let Q be the set of 3-D points obtained for the current scene. Let α and θ be the pitch and roll angles of the sensor with respect to the ground-plane, which is estimated as in Section 2.5. Let $R_{(\theta,\alpha)}$ be a 3-D rotation matrix built upon both pitch and roll angles. The 3-D points composing the set P provided to the obstacle detector now correspond to the elements of Q properly rotated by $R_{(\theta,\alpha)}$:

$$\mathbf{p} = R_{(\theta,\alpha)}\mathbf{q},\ \forall \mathbf{p} \in P, \mathbf{q} \in Q. \tag{2.7}$$

This rotation aligns the stereoscopic vision sensor's local frame of reference with the frame of reference of the ground-plane (van der Mark et al., 2007). This accommodates the 3-D point cloud for the application of the canonical compatibility test. From another perspective, the truncated cone of the compatibility test is implicitly rotated according to the attitude of the vision sensor with respect to the ground-plane. However, the projection of this transformation must also be accounted for. That is, the truncated triangle must also be rotated; otherwise some pixels would be erroneously skipped, whereas others would be unnecessarily analysed in the compatibility test.

A solution to this problem would be to rotate the back-projection of the truncated triangle's vertexes using $R_{(-\theta,-\alpha)}$, which would then be re-projected onto the image plane to become the new vertexes used in the compatibility test. However, the resulting triangle would most probably no longer be isosceles, which would complicate the scanning procedure within it. This plus the additional projective transformations make this solution computationally expensive. For this reason, in practice only the roll angle is taken into account in this operation. This approximation draws from the empirical observation that this rotation is essential for a proper operation of the small obstacle detector, whereas the disregard of the pitch angle only results in missing the top of some obstacles. Bearing this in mind, the vertexes of the rotated triangle associated to



(a) (b)

Figure 2.10: Roll compensation in the image plane. (a) Illustrative diagram, where the truncated triangle is rotated in order to compensate for a roll angle θ. (b) Typical real situation in which the rotation of the truncated triangle is the result of the roll compensation mechanism. The zoomed image depicts the results of the compatibility test associated with the pixel in the truncated cone's vertex, where red (darker), green (lighter), and black pixels are compatible, incompatible, and without associated 3-D information, respectively.

$\mathbf{p}$, $\mathbf{v}''_{\mathbf{l}}$ and $\mathbf{v}''_{\mathbf{r}}$, are obtained by rotating the non-rotated triangle's vertexes, $\mathbf{v}'_{\mathbf{l}}$ and $\mathbf{v}'_{\mathbf{r}}$, so as to compensate for the ground-plane's roll angle:

$$\mathbf{v}''_{\mathbf{i}} = \mathrm{R}_{(-\theta)}(\mathbf{v}'_{\mathbf{i}} - \mathbf{p}') + \mathbf{p}', \tag{2.8}$$

where $\mathbf{i} \in \{l, r\}$ and $\mathrm{R}_{(-\theta)}$ is a 2-D rotation matrix. Note that the ground-plane's roll angle is assumed to project directly onto the image plane, which is an approximation that was found to be sufficient in practice (see Fig. 2.10).

2.7 Parameter Selection

This section provides guidelines regarding the instantiation of every free parameter included in the model. The ROI definition requires the specification of a minimum range, r_{min}, and a maximum range, r_{max}. Whereas the former is constrained by the sensor's field of view, the latter is upper bounded by the point cloud's noise level. For instance, the shorter the baseline of the stereoscopic vision sensor, the smaller the r_{max}. The additional parameter ς is an upper bound in the expected proportion of pixels encompassed in a row that are erroneously beyond r_{max}. A value of 15 % is usually sufficient.

The number of iterations m_{refit} and n_{hypo} in the ground-plane estimation RANSAC procedure could be theoretically defined, provided that the distribution of the sampling space was known. Because this is not the case on all-terrain environments, these

parameters must be defined empirically. A large number of iterations is required only if a large number of obstacles are expected to occur in the scene. The related parameter d_{plane} defines an upper bound for the distance between a point and a plane hypothesis to be accepted as an inlier. On uneven ground, a small value ($< 10\,\text{cm}$) has the negative effect of fitting the ground-plane estimate to small planar patches of the terrain. In practice, larger patches of the terrain can be covered and consequently more robust estimates attained if a larger value is considered.

The number of non-salient pixels is used in the RANSAC procedure to estimate the discriminative power that saliency has in the current scene. In the process, a small threshold l is used to determine which pixels are non-salient. A small value above zero is usually sufficient for a proper instantiation of this parameter. A measure of the local saliency is used in the ground-plane estimation process to determine which points should be taken into account. The larger the width of the window used to compute local saliency, q, is, the more difficult it is to accept points near salient regions. Practice suggests a value below 25 % of the image's width to properly handle cluttered environments. The selection pressure is also controlled by the scaling factor ρ. Higher values reduce the chances of selecting non-salient pixels at the cost of a higher number of required samples. Practice suggests that a scaling factor of around 4 manages the trade-off well.

The small obstacle detector also requires the specification of a set of parameters. H_{min} and θ are obtained directly from the robot's mechanical properties, whereas H_{max} equals h in the hybrid framework. That is, H_{max} is as high as the smallest obstacle to be detected by the large obstacle detector. SVR parameters n_{max} and w_{max} must also be instantiated. Parameter n_{max} trade-offs speed with the risk of failing to detect a non-salient obstacle in the coarse analysis of the image. Because large obstacles are those that may not be salient in the image, n_{max} can be safely instantiated to a large value. Not being sensitive to saliency information, the parameter controlling the maximum number of rows that can be skipped in the coarse analysis of the image, w_{max}, must be at least as small as the expected height in pixels of obstacles detectable at a safe distance. The same reasoning applies to the variable controlling the coarse analysis, m. In practice these variables are tuned to achieve the required computational performance under a given accuracy constraint. An example of such constraint is the expected projected width onto the image plane, $\Delta x'$, of an object with a given width Δx and located at a given distance z from the sensor. This can be expressed as $\Delta x' = \frac{f}{z \cdot o} \Delta x$, where f is

the sensor's focal length and o is the pixel's width. For instance, for a camera with focal length of 4.4 mm and pixel size 0.006 mm, a 20 cm obstacle at 5 m from the robot will have a corresponding size in the image plane of roughly 30 pixels. This means that both w_{max} and m must have smaller values to ensure that the obstacle is sampled. Parameter λ, used to detect ascending saliency variations, is empirically minimised under the constraint that the scanning procedure must not get trapped by small local variations resulting from noise in the input image. A good indicator for its parameterisation is 10% of the saliency's range.

Both voting and area filters are also dependent on two parameters, namely v and a. Their values depend mostly on the level of noise needed to be handled. As experimental results will show, by varying these thresholds it is possible to smoothly move on the trade-off curve relating false positive rate and true positive rate. This robustness will also be confirmed by the ability of using the same thresholds in different environments. In practice, by making $v \approx 0.2$ and $a \approx 25$ most of the noise impinging stereo-vision is filtered without significant signal loss.

2.8 Experimental Results

This section presents the experimental results obtained with two different data sets composed of 640×480 image pairs. The first data set is composed of 36 heterogeneous stereo image pairs (see Fig. 2.18 in Appendix A) that have been acquired with a 9 cm baseline Videre Design STOC sensor under dynamic conditions and at an approximate height of 1.5 m. This set will be used to quantitatively assess the small obstacle detector and the ground-plane estimation technique. For this purpose all images have been hand-labelled (obstacle/non-obstacle pixels) for ground truth. To reduce imprecision in the hand-labelling process, the ROI of this data set has been limited to 10 m and images were selected so that a dominant ground-plane was actually present and vegetation as much absent as possible. This allows an accurate quantitative analysis of the several components of the system.

To test the hybrid detector, a data set with images containing large obstacles and considerable uneven terrain is required. This data set should also allow the test of the method against more distant obstacles. Finally, the data set should be extensive and disparate. All these aspects have been considered in the second data set, which encompasses three long runs with 798, 998 and 600 frames, obtained at 7.5 Hz in off-

road, urban, and mixed environments, respectively. With the stereo head hand-held at an approximate height of 1.5 m, the acquisition process took place by walking in the environments at an average speed of approximately $1\,\mathrm{ms}^{-1}$. To allow the detection of obstacles up to 20 m a sensor with a baseline of 30 cm was used. Every 20 images were hand-labelled to enable a quantitative analysis. The higher complexity of this data set when compared to the previous one results in more ambiguous hand-labels. This explains why the components of the system are first tested individually and thoroughly with the previous data set.

Small Vision System (SVS) (Konolige, 1997; Konolige and Beymer, 2007) and OpenCV (Bradski and Kaehler, 2008) were used for stereo computation and other low-level computer vision routines, respectively. Stereo computation uses an area-based L1 norm (absolute difference) correlation method, operating over Laplacian Of Gaussian (LOG) transformed images. The result is interpolated to a precision of $1/4$ pixel and the correlation window size is 11×11. To increase the amount of information available in the point cloud, the disparity calculation is carried out at the original resolution, and also on images reduced by $1/2$. With this multi-scale approach, the extra disparity information is used to fill in dropouts in the original disparity calculation.

SVS also provides a set of standard filters to reject 3-D points that are potentially erroneous at the cost of reducing too much the density of the 3-D point cloud in poorly textured environments. Briefly, according to a threshold f_c, a confidence filter eliminates stereo matches that have a low probability of success due to lack of image texture. A uniqueness filter performs a consistency check to ensure that the minimum correlation value is lower than all other match values by a threshold f_u. Finally, a speckle filter eliminates small disparity regions that are not correct by imposing a threshold f_s on the minimum region size. The three filters are used in both data sets with $f_c = 12$, $f_u = 10$, and $f_s = 400$. The ability of the model to handle noisier point clouds will be demonstrated in a final experiment, in which the strength of both confidence and uniqueness filters will be reduced, $f_c = 6$, $f_u = 6$. With this reduction, the point cloud is denser, but also noisier, and so the overall results of our model are more significant.

The following summarises the parameterisation of the model used in the experiments, which has been defined according to the guidelines presented in Section 2.7. Robot related parameters, $H_{min} = 0.1\,\mathrm{m}$, $H_{max} = h = 0.4\,\mathrm{m}$, $\theta = 40^\circ$, ground-plane estimation parameters, $m_{refit} = 10$, $n_{hypo} = 40$, $d_{plane} = 0.2\,\mathrm{m}$, $l = 5$, $q = 150$, $\rho = 4$, and large obstacle detection threshold, $h = 0.4\,\mathrm{m}$, are stable across experiments. To account for

the different sensors' baselines, the ROI is parameterised with $r_{min} = 1\,\text{m}$, $r_{max} = 10\,\text{m}$, $\varsigma = 15\,\%$ and $r_{min} = 2\,\text{m}$, $r_{max} = 20\,\text{m}$, $\varsigma = 15\,\%$ for the first and second data sets, respectively. Unless otherwise noted, SVR of the small obstacle detector and related filtering mechanisms are set as $n = 3$, $m = 6$, $n_{max} = 30$, $w_{max} = 20$ and $v = 0.2$, $a = 25$, respectively. The specific values of v and a will be justified according to the results obtained with the small obstacle detector in the first data set.

2.8.1 Ground-Plane Estimation Results

The first set of experiments intends to demonstrate the usefulness of using saliency to modulate the ground-plane hypotheses generation step. Ground-truth is given in terms of obstacle/non-obstacle hand-labels, rather than in terms of ground-plane coefficients. This option results from the fact that the hand-labelling of the obstacles is a much more accurate process than the process of hand-labelling the ground-plane coefficients that better approximate the terrain. This compels us to assess the ground-plane estimation process by indirect means. Concretely, validation is done by comparing the obstacles detected using the plane-based detector with the obstacles present in the ground-truth. The better the estimated ground-plane, the closer the detected obstacles match the ground-truth. This process is repeated with and without using saliency to modulate the ground-plane estimation process.

A large set M of 10000 ground-plane hypotheses per image in the first data set was created with the Saliency-based Ground-Plane Estimation (SalGPE) approach. This large set results in statistics varying $\approx 1\,\%$ across experiments. A set U with the size of M was created for each image as well, but this time without saliency modulation and thus representing the canonical RANSAC-based Ground-Plane Estimation (GPE). The set of obstacles detected using each ground-plane hypothesis is compared against the ground-truth to obtain the True Positive Rate (TPR), the False Positive Rate (FPR), and the two-class Matthews Correlation Coefficient (MCC). The MCC metric is well known for its ability to handle unbalanced data sets. The closer MCC is to 1 the better the hypothesis matches the ground-truth. Obstacles are those points whose orthogonal distance to the plane is above $0.2\,\text{m}$, which is a reasonable upper bound for most wheeled robots. A lower value would be inappropriate for a plane-based detection approach. Then, the mean (μ) and standard deviation (σ) of the above variables over all hypotheses in both M and U are computed.

Image	μ^M_{MCC}	σ^M_{MCC}	μ^U_{MCC}	σ^U_{MCC}	$\ominus_\mu$	$\ominus_\sigma$
00	0.57	0.36	0.23	0.51	**0.34**	-0.15
22	0.68	0.23	0.41	0.37	**0.27**	-0.14
23	0.56	0.26	0.30	0.38	**0.26**	-0.12
21	0.64	0.22	0.4	0.35	**0.25**	-0.13
16	0.75	0.28	0.56	0.41	**0.19**	-0.14
33	0.63	0.32	0.48	0.38	**0.16**	-0.06
06	0.50	0.24	0.38	0.28	**0.12**	-0.04
19	0.79	0.23	0.70	0.34	**0.10**	-0.11
31	0.55	0.27	0.47	0.33	**0.08**	-0.07
02	0.67	0.18	0.60	0.25	**0.07**	-0.08
24	0.23	0.31	0.16	0.29	**0.07**	0.02
20	0.68	0.33	0.62	0.38	**0.06**	-0.05
07	0.40	0.28	0.34	0.31	**0.06**	-0.03
01	0.62	0.16	0.57	0.23	**0.05**	-0.07
08	0.39	0.11	0.35	0.12	**0.04**	-0.02
27	0.78	0.21	0.73	0.25	**0.04**	-0.04
25	0.63	0.28	0.59	0.30	**0.03**	-0.02
11	0.54	0.23	0.51	0.26	**0.03**	-0.03
34	0.60	0.18	0.59	0.20	**0.02**	-0.02
29	0.58	0.29	0.55	0.31	**0.02**	-0.02
05	0.39	0.14	0.37	0.16	**0.02**	-0.01
35	0.32	0.17	0.31	0.17	**0.01**	0.00
26	0.36	0.11	0.35	0.11	**0.01**	0.00
10	0.58	0.15	0.57	0.15	**0.01**	0.00
15	0.17	0.14	0.16	0.15	**0.01**	-0.01
03	0.40	0.19	0.38	0.21	**0.01**	-0.02
14	0.21	0.09	0.20	0.09	**0.00**	0.01
13	0.00	0.00	0.00	0.00	**0.00**	0.00
09	0.02	0.05	0.02	0.05	**0.00**	0.00
18	0.00	0.00	0.00	0.00	**0.00**	0.00
28	0.20	0.32	0.20	0.31	**0.00**	0.00
30	0.27	0.21	0.28	0.20	**-0.01**	0.01
12	0.15	0.09	0.16	0.09	**-0.01**	0.00
04	0.30	0.15	0.31	0.15	**-0.01**	0.00
32	0.50	0.32	0.52	0.30	**-0.02**	0.02
17	0.37	0.13	0.40	0.11	**-0.03**	0.02

(a)

Image	MCC^M_p	MCC^U_p	$\ominus_p$
34	0.38	0.21	**0.17**
19	0.69	0.52	**0.17**
33	0.11	-0.03	**0.14**
29	0.5	0.36	**0.14**
7	0.24	0.11	**0.13**
0	0.74	0.61	**0.12**
2	0.73	0.61	**0.12**
27	0.55	0.44	**0.11**
31	0.6	0.48	**0.11**
22	0.65	0.54	**0.11**
1	0.42	0.33	**0.09**
6	0.41	0.31	**0.09**
26	0.4	0.31	**0.09**
5	0.41	0.32	**0.09**
21	0.72	0.63	**0.08**
12	0.25	0.17	**0.08**
20	0.69	0.61	**0.08**
4	0.5	0.42	**0.07**
25	0.45	0.37	**0.07**
23	0.58	0.51	**0.07**
16	0.77	0.7	**0.07**
28	0.23	0.16	**0.07**
30	0.6	0.54	**0.06**
3	0.5	0.44	**0.06**
35	0.25	0.19	**0.06**
32	0.35	0.29	**0.06**
17	0.06	0.02	**0.04**
11	0.43	0.39	**0.03**
8	0.23	0.2	**0.03**
10	0.14	0.11	**0.03**
9	0.05	0.03	**0.02**
24	0.24	0.24	**0.01**
14	-0.02	-0.02	**0.01**
15	0.29	0.29	**0**
13	0	0	**0**
18	0	0	**0**

(b)

Image	MCC^M	MCC^U	$\ominus$
29	0.5	0.33	**0.17**
34	0.28	0.11	**0.17**
0	0.6	0.45	**0.15**
23	0.57	0.44	**0.14**
7	0.32	0.19	**0.13**
19	0.58	0.45	**0.13**
25	0.26	0.13	**0.13**
2	0.68	0.55	**0.13**
27	0.35	0.23	**0.13**
33	-0.02	-0.15	**0.13**
16	0.68	0.55	**0.12**
30	0.5	0.38	**0.12**
21	0.61	0.49	**0.12**
31	0.46	0.34	**0.11**
28	0.2	0.1	**0.1**
22	0.57	0.47	**0.09**
20	0.63	0.53	**0.09**
3	0.45	0.36	**0.09**
26	0.27	0.18	**0.09**
1	0.39	0.31	**0.08**
5	0.43	0.35	**0.08**
12	0.21	0.14	**0.07**
35	0.2	0.14	**0.06**
32	0.29	0.23	**0.06**
6	0.35	0.29	**0.06**
4	0.32	0.27	**0.06**
11	0.16	0.11	**0.05**
15	0.18	0.15	**0.03**
10	0.09	0.06	**0.03**
24	0.27	0.25	**0.02**
14	-0.02	-0.03	**0.01**
9	0.03	0.02	**0.01**
8	0.24	0.24	**0**
13	0	0	**0**
18	0	0	**0**
17	0.04	0.07	**-0.03**

(c)

Table 2.1: MCC results. (a) Ground-plane estimation with $\ominus_\mu = \mu^M_{MCC} - \mu^U_{MCC}$ and $\ominus_\sigma = \sigma^M_{MCC} - \sigma^U_{MCC}$. (b) Obstacle detection with ground-plane compensation, where $\ominus_p = MCC^M_p - MCC^U_p$. (c) Obstacle detection without ground-plane compensation, where $\ominus = MCC^M - MCC^U$. M = saliency-based RANSAC; U = canonical RANSAC.

According to the MCC results (see Table 2.1(a)), two main image sub-sets emerge. One (grey shaded) aggregates images in which the RANSAC saliency-based hypotheses generation step outperforms the canonical one (refer to the two last columns in Table 2.1(a)). The MCC differences are residual (i.e., $< 5\,\%$) for the remainder images, meaning that saliency is essentially neutral there. Images without obstacles (13 and 18)

have MCC values of 0. Images benefiting from saliency share a characteristic: a considerable presence of objects. In these situations (e.g., Fig. 2.2), saliency easily segments objects from background. Saliency thus operates better in those situations in which it is most required. In the absence of obstacles an uninformed solution suffices.

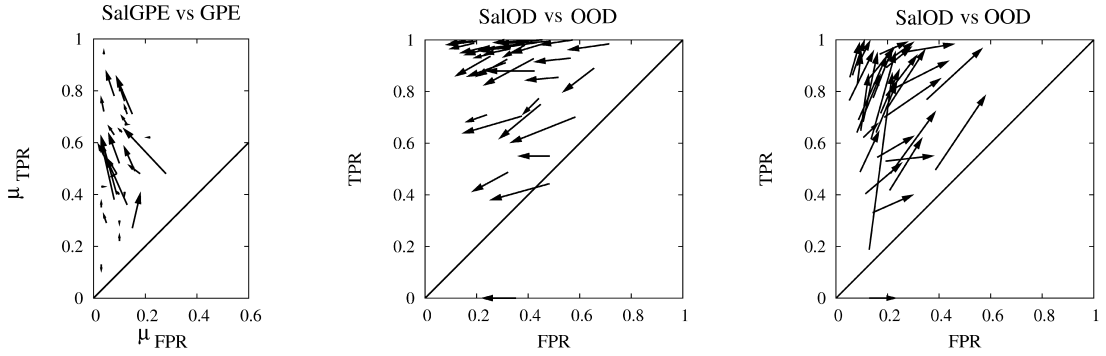
Fig. 2.11(a) depicts the Receiver Operating Characteristic (ROC) results of the experiment. For each image k, an arrow is drawn to connect each without-saliency point, $(\mu^U_{TPR},\mu^U_{FPR})_k$, to its corresponding point with-saliency, $(\mu^M_{TPR},\mu^M_{FPR})_k$. The closer a point is to the upper-left corner of the graph the better the corresponding set of sampled hypotheses matches the ground truth. A clear dominance of arrows heading towards the upper-left corner is observed.

However, the arrow corresponding to image 23 in Fig. 2.11(a) goes in a clearly different direction. The justification for this fact relates to the particular configuration of the environment, whose effects on the analysis are the following. Without saliency modulation, the estimated plane is raised slightly off the ground, influenced by the large obstacle in the image. As a result, by not being fully above the estimated plane, the obstacle induces a low TPR. Moreover, by being below the estimated plane, some false positives do not contribute to the FPR. Conversely, because the obstacle does not affect the fitting process when saliency modulation is employed, the effects are the opposite. That is, the better approximation of the estimated plane to the actual ground-plane results in that surface variations induce a higher FPR, and a bigger portion of the obstacle contributes to an also higher TPR.

2.8.2 Small Obstacle Detector Results

To isolate the saliency modulation capabilities, a stripped down version of the small obstacle detector was tested in an initial experiment. For this purpose, the voting filter was turned off, the maximum number of pixels and rows that can be skipped in the coarse analysis was fixed to 30 and 1, respectively. This means that the skipping procedure was not progressive.

Fig. 2.11(b) shows the results obtained from comparing the OOD (Talukder et al., 2002; Manduchi et al., 2005) with the stripped down, saliency-based detector (SalOD), i.e., with $(n,m,n_{slide}) = (1,2,30)$. When compared to the OOD, the SalOD exhibits a considerably reduced FPR and only a slightly smaller TPR. This combined result shows



(a) (b) (c)

Figure 2.11: Ground-plane estimation and obstacle detection ROC plots. For a given image, each arrow connects the TPR/FPR trade-off point obtained without saliency modulation to the TPR/FPR trade-off obtained with saliency modulation. (a) Ground-plane estimation with (SalGPE) vs. without (GPE) saliency modulation. (b) Obstacle detection with (SalOD) vs. without (OOD) saliency modulation. (c) Same as (b) but skipping n_{slide} pixels for the OOD case. Line $x = y$ displayed for reference.

the benefits of using saliency to selectively discard non-obstacle points. Moreover, it should be noted that the still undesirable level of TPR reduction is owed in part to the absence of progressive skipping in the SalOD. This signal-to-noise ratio improvement is confirmed by the contrast of the MCC with (MCC_p^M) and without (MCC_p^U) saliency modulation (see Table 2.1(b)). Label p means that ground-plane compensation is on. Saliency contributes in the same amount when the ground-plane compensation is off (see Table 2.1(c)), highlighting its resilience.

To reinforce the evidence that the reduction in FPR is due to the saliency's selective nature instead of the reduced number of pixels being analysed, an additional test was carried out. The OOD was configured to systematically skip n_{slide} pixels, rather than n, when displacing the truncated triangle. In this situation, the results (see Fig. 2.11(c)) show that although the OOD now produces a smaller FPR, due to blindly skipping more pixels, a considerable reduction in the TPR is also observed. As for obstacle detection, smaller TPR means higher risk of collision, saliency shows itself to be a useful cue for informed false positive removal.

To test the full-fledged small obstacle detector (ESalOD), all its features were turned on, including both voting and area filters. Fig. 2.12 plots the ROC curves of this experiment. A first analysis shows that ESalOD with $(n \times m) = (3 \times 6)$ exhibits a TPR vs. FPR trade-off at least as good as the trade-off exhibited by the OOD. This stems from

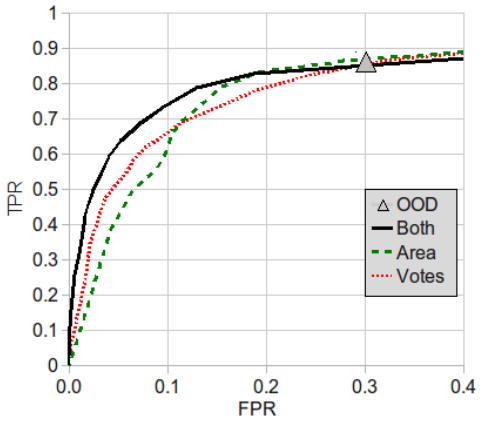


Figure 2.12: Obstacle detection ROC curves. Each plot is the average of the ROC curves built over all images in the first data set. A ROC curve is obtained by varying either the voting or the area filter parameters. Votes, ESalOD with voting filter on, $v \in \{0, 0.05, \ldots, 1\}$, and area filter off, $a = 0$; Area, ESalOD with voting filter off, $v = 0$, and area filter on, $a \in \{0, 5, \ldots, 100\}$; Both, ESalOD with both filters on, $a = 25$ and $v \in \{0, 0.05, \ldots, 1\}$; $\triangle$ OOD, by not having any associated filter is limited to a ROC point.

the observation that the ESalOD ROC curve intercepts the OOD ROC point. The ROC curve also shows that for high values of TPR the EsalOD exhibits a better TPR vs. FPR trade-off than the OOD. This demonstrates how advantageous it is to embed into the detector both visual saliency and voting filter. The curve shows, for instance, that it is possible to reduce the FPR of the OOD by $\approx 70\,\%$ while diminishing the TPR by only $\approx 10\,\%$.

The area under each ROC curve obtained with either the voting or the area filters alone is smaller than the area obtained with both filters operating in conjunction. This shows that the latter configuration exhibits a better overall performance. Note that for low FPR values (i.e., < 0.15) none of the filters alone is capable of approaching the ROC curve obtained with both filters operating simultaneously. This clearly shows their complementary role. This is further confirmed by the ROC curve of the area filter (see Fig. 2.13), where the higher the area under the curve, the more active the area filter is.

From the analysis of the area filter ROC curve (see Fig. 2.13) we conclude that the configuration $a = 25$ (see Equation 2.6) is the most adequate for the first data set. After a thorough inspection of the model's behaviour in the first data set we also conclude

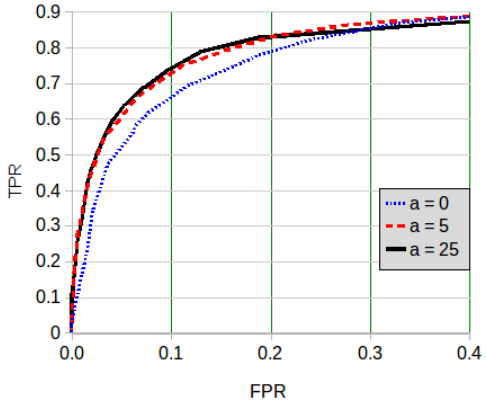


Figure 2.13: Impact of the area filter. Each plot represents the average of the ROC curves obtained over all images in the first data set, for a given area filter configuration. For a given image and filter's parameterisation, $a \in \{0, 5, 25\}$, the ROC curve is built by sliding the threshold of the votes filter over its domain, $v \in \{0, 0.05, \ldots, 1\}$. The absence of the area filter, $a = 0$, results in the lowest area under the curve, which shows the usefulness of the filter. The relative performance of the other two parameterisations, $a \in \{5, 25\}$, swaps at the intersecting point of their ROC curves. Nevertheless, $a = 25$ is shown to be the most interesting configuration as it is the one that performs better for higher TPR values.

that the configuration $v = 0.2$ (see Equation 2.5) for the voting filter shows the best performance. With this configuration, false negatives are mostly absent and the false positives are mostly concentrated around obstacles, resulting only in their enlargement (see Fig. 2.9).

2.8.3 Hybrid Obstacle Detection Results

Having confirmed the performance of the model's individual components, this section tests the hybrid solution as a whole. For this purpose, the more extensive second data set is used. The terrain in this data set is much more uneven, and large obstacles at different ranges are much more frequent, which will help highlighting the advantages of focussing each detection technique on a specific type of obstacle. As will be shown, this specialisation promotes a better TPR vs. FPR trade-off as well as a reduced computational cost. The ROI in these experiments has been changed to $r_{min} = 2\,\text{m}$, $r_{max} = 20\,\text{m}$.

Table 2.2 summarises the quantitative results obtained for the three runs composing the data set. The table's first row shows the ability of the model to maintain a

	Urban Run		Off-Road Run		Mixed Run	
	TPR	FPR	TPR	FPR	TPR	FPR
Base	0.95 ± 0.08	0.11 ± 0.19	0.89 ± 0.16	0.02 ± 0.03	0.93 ± 0.20	0.05 ± 0.05
+ disp.	0.97 ± 0.04	0.11 ± 0.06	0.96 ± 0.08	0.05 ± 0.04	0.98 ± 0.05	0.12 ± 0.08
+ filters	0.91 ± 0.10	0.05 ± 0.03	0.90 ± 0.12	0.02 ± 0.02	0.90 ± 0.16	0.06 ± 0.06
+ n_{max}	0.91 ± 0.09	0.05 ± 0.03	0.90 ± 0.11	0.02 ± 0.03	0.91 ± 0.13	0.06 ± 0.06

Table 2.2: Results obtained with the hybrid detector (mean $\pm$ standard deviation). Each row corresponds to a given configuration. The small obstacle detector in the "base" configuration is parameterised as for the first data set. The "+ disp." configuration takes the same parameterisation but with a much denser/noisier disparity map. The "+ filters" configuration adds to the previous configuration a stronger filtering mechanism, $v = 60$, $a = 70$. The last configuration adds to the previous configuration a larger upper bound to the saliency-based pixel skipping procedure, $n_{max} = 100$.

high TPR with the parameterisation obtained from the first data set. This is an exhibition of the model's robustness to different sensors and environments. The second row shows that with a less filtered disparity map, and consequently with a denser and noisier point cloud, the number of false positives grow for both off-road and mixed environments. But interestingly the TPR grows as well. This shows the robustness of the model to changes in the stereo-based 3-D reconstruction process. If the growth in FPR is nevertheless undesirable, which depends on the model's client, one can reduce it by empirically pushing further the voting and area filters to $v = 60$, $a = 70$. The third row of the table shows exactly this. See for instance that the FPR is reduced on average by 55 % with only a TPR average reduction of 7 %.

To assess the sensibility of the model to the mechanism that most strongly constrains the saliency-based SVR, n_{max} was extended from 30 to 100 pixels. The results in the last row of the table show that no significant difference can be observed with this new parameterisation. This means that the saliency map is accurate enough to guide the detector. This last configuration of the hybrid detector will be used in the remaining experiments. The videos are available in the online version of this article. The videos permit the qualitative verification of the model's ability to detect the vast majority of the obstacles present in the several tested environments. Moreover, the videos also show that most of the false positives are not stable across frames, and consequently they should be easily filtered out under a probabilistic mapping framework.

Table 2.3 summarises the results obtained with both small and large obstacle detectors in isolation. The goal of these experiments is to study the contribution of each

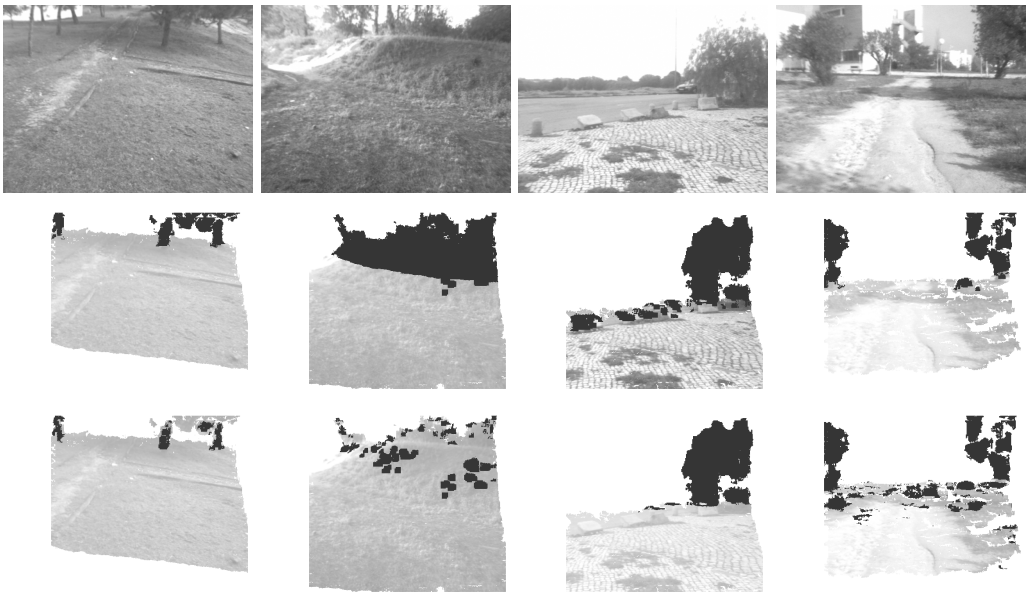
	Urban Run		Off-Road Run		Mixed Run	
	TPR	FPR	TPR	FPR	TPR	FPR
Small	0.63 ± 0.23	0.04 ± 0.03	0.62 ± 0.24	0.01 ± 0.01	0.63 ± 0.28	0.05 ± 0.05
$h = 0.4\,\text{m}$	0.81 ± 0.17	0.04 ± 0.03	0.80 ± 0.21	0.02 ± 0.02	0.79 ± 0.25	0.04 ± 0.06
$h = 0.1\,\text{m}$	0.93 ± 0.06	0.13 ± 0.13	0.95 ± 0.07	0.19 ± 0.16	0.92 ± 0.13	0.17 ± 0.14

Table 2.3: Results obtained with isolated detectors (mean $\pm$ standard deviation). Each row corresponds to a given configuration. The "small" configuration considers only the output of the small obstacle detector as considered in configuration "+ n_{max}" (see Table 2.2) . The "$h = 0.4\,\text{m}$" configuration considers only the large obstacle detector. The last configuration considers only the large obstacle detector, but this time with $h = 0.1\,\text{m}$ and also encompassing negative obstacles.

technique to the overall hybrid model. The first row of the table shows that the small obstacle detector alone is not able to obtain the TPR level of the hybrid solution (cf. bottom row of Table 2.2). This is mostly because it fails to detect some large obstacles whose 3-D point cloud is sparser or noisier (e.g., Fig. 2.14(a)). This is a limitation of methods that identify obstacles based on geometrical relationships between neighbour 3-D points. Sometimes, the small obstacle detector alone also fails to label as obstacle those gentle slopes that due to their height are not navigable (e.g., Fig. 2.14(b)). The second row of the table shows that the large obstacle detector, as parameterised for the hybrid solution, $h = 0.4\,\text{m}$, is also unable to reach the desired TPR (cf. bottom row of Table 2.2). In this case the detector fails to detect small obstacles and the bottom part of large ones (e.g., Fig. 2.14(c)).

In conclusion, the two techniques are shown to perform in a complementary way, thus justifying the benefits of a hybrid solution. By making $h = 0.1\,\text{m}$, the ability of the plane-based solution to detect both small and large obstacles was tested in a final experiment. As this configuration intends to also substitute the use of the small obstacle detector, it must also be able to detect negative obstacles. Therefore, in order to allow that points below the ground-plane would also be considered obstacles in this test, the distance of 3-D points to the ground-plane is considered in absolute terms. The third row of Table 2.3 shows that with this configuration the FPR grows approximately 44 times the growth of TPR (e.g., Fig. 2.14(d)). See, for instance, the case of the off-road run, in which an 850 % increment of FPR is followed by only a 19 % increment of TPR. Therefore, the hybrid solution exhibits the best TPR vs. FPR trade-off.

Typical outputs of the hybrid obstacle detector in the off-road, urban, and mixed environments can be depicted in Fig. 2.15, 2.16, and 2.17, respectively.



(a) Small (b) Small (c) $h = 0.4\,\mathrm{m}$ (d) $h = 0.1\,\mathrm{m}$

Figure 2.14: Comparison between hybrid (middle row) and isolated (bottom row) detectors over a set of typical input images (top row). Results are organised in a columnwise manner, each column being associated to a given configuration (see Table 2.3). Detected obstacles are overlaid in dark grey over the corresponding input images. White pixels are those out of the ROI or without computed 3-D information and are thus discarded by the detection process.

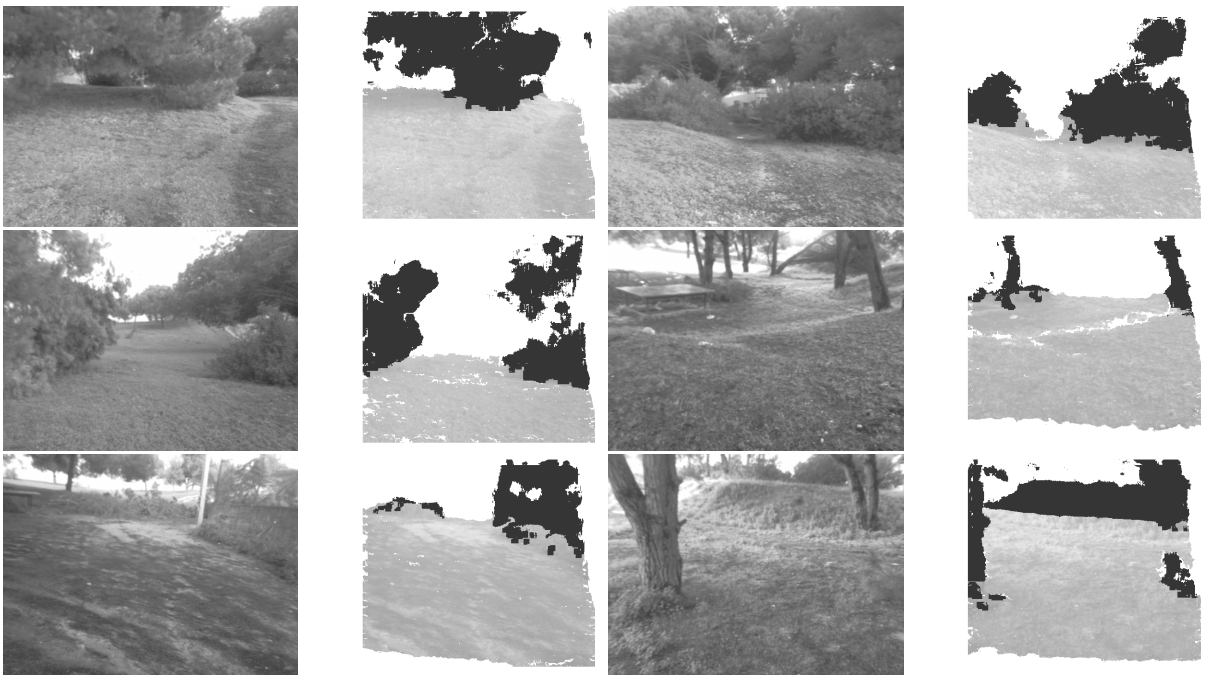


Figure 2.15: Typical results obtained with the proposed hybrid model in the off-road long run.

Figure 2.16: Typical results obtained with the proposed hybrid model in the urban long run.

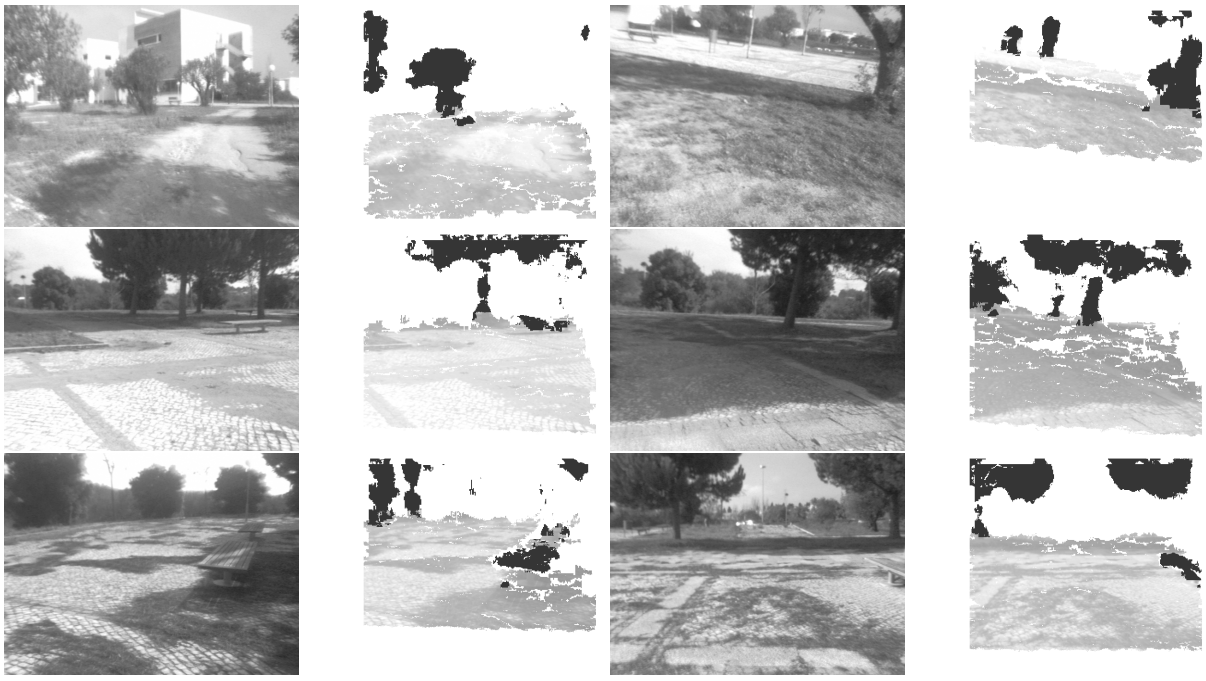


Figure 2.17: Typical results obtained with the proposed hybrid model in the mixed long run.

		Hybrid Detector				OOD
Data set	Stereo	Saliency	Plane	Detection	Total	Total
1 (9 cm)	44 ± 1	51 ± 4	31 ± 7	168 ± 106	294 ± 107	6464 ± 4074
2 (30 cm)	59 ± 1	53 ± 7	25 ± 5	46 ± 25	183 ± 26	3198 ± 909

Table 2.4: Timing information, in milliseconds, for both hybrid detector and OOD (mean $\pm$ standard deviation). Partial timings for stereo processing, saliency computation, ground-plane estimation, and actual obstacle detection are also provided.

2.8.4 Computational Performance

Table 2.4 summarises the computation time spent at each step of the proposed model. It also compares the computation time of the hybrid detector with the computation time taken by the OOD, proposed by Talukder et al. (2002) and Manduchi et al. (2005). These results were obtained with a Linux-based 2.8 GHz Intel Core 2 Duo Laptop equipped with 4 GB of RAM. The model runs on a single core.

The hybrid obstacle detection takes on average 183 ms on the second data set, as opposed to the 294 ms obtained with the first data set. The 61 % increment of computation time relates to the different ROI used in each data set, which is in turn a consequence of using sensors with different baselines. When covering larger areas of the near-field, the truncated triangles considered by the detector are quite large and consequently expensive to analyse. The increment in computation time for the OOD case is higher, 100 %, meaning that its performance decays more drastically for sensors with shorter baselines. Finally, the hybrid detector performs on average 17 and 22 times faster than the OOD for the first and second data sets, respectively.

2.9 Conclusions

A model for stereo-based all-terrain obstacle detection was presented. By hybridising two complementary obstacle detection techniques, the model innovates at the architectural level. A common characteristic of the two techniques is that they perform in the image space, rather than in a digital terrain map.

The presented hybrid system is capable of searching for obstacles with more than 10 cm height up to a range of 20 m on uneven terrain. It performs at 5 Hz on 640×480

images. This has been attained by focussing each detection technique on a particular type of obstacle (i.e., large vs. small) depending on each technique's characteristics. Also novel at the architectural level is the extensive use of visual saliency to guide the detection process. This mechanism was shown to improve robustness, accuracy, and computational efficiency of obstacle detection.

The technique employed for small obstacle detection is known for its ability to perform well on uneven terrain. However, in its classic form, it is also known to be computationally expensive and brittle in the face of noise. These limitations have been overcome in the proposed model with the novel use of saliency-based space-variant resolution, with a mechanism for the camera's pitch-roll compensation, and with a voting mechanism. The latter was shown to be extremely powerful in enabling the operation of the detector in the presence of noisy 3-D point clouds. The fact that the filter is fully embedded in the detector's operation, almost discards the need for time consuming post-filters.

The revealed success of visual saliency in all-terrain obstacle detection is in part due to the novel way conspicuity maps are blended in this work. Another key aspect is that the detector updates the saliency map whenever an obstacle is detected. This allows the saliency map to guide the detector while being opportunistically corrected by it.

With the exception of the compatibility test, all other aspects of the proposed model are easily parallelisable. This opens the door for future Graphics Processing Unit (GPU)-based implementations to further reduce processing cycle. As a further improvement, the number of independent parameters shall also be reduced. We also expect in the future to assess the benefits for obstacle detection of modulating the saliency map with top-down knowledge, such as the expected appearance of most frequent obstacles.

Acknowledgements

This work was partially supported by FCT/MCTES grant No. SFRH/BD/
27305/2006. We thank Paulo Santos for his participation in the early developments of the obstacle detector. We also acknowledge the useful comments provided by Nelson Alves and by the four anonymous reviewers of this paper.

2.10 Appendix A

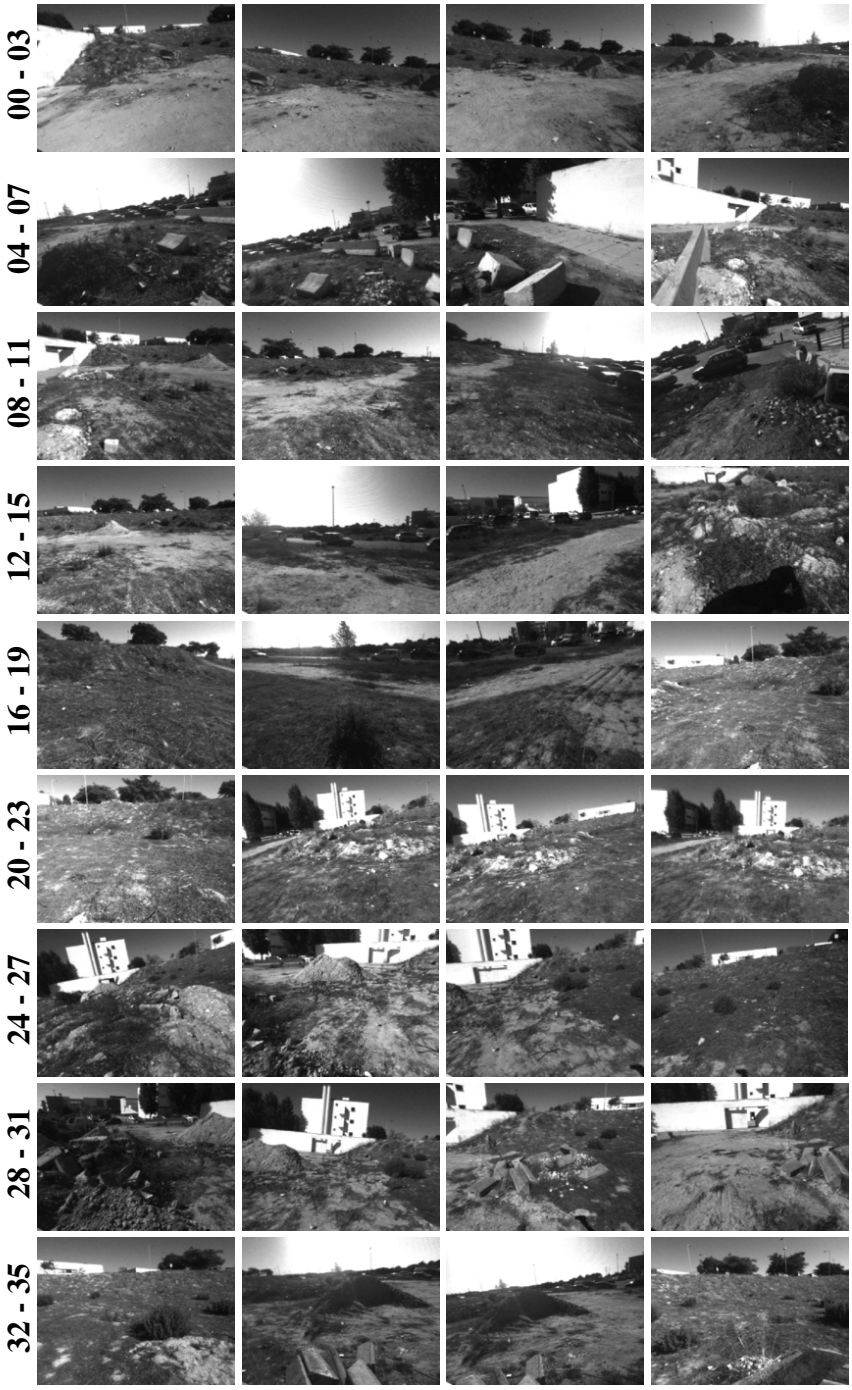


Figure 2.18: Stereo data set (left images only) obtained with a 9 cm baseline configuration.

2.11 Appendix B: Index to Multimedia Extensions

Extension	Media type	Description
1	Video	Off-road environment
2	Video	Urban environment
3	Video	Off-road/urban mixed environment

Table 2.5: The videos are available as Supporting Information in the online version of this article.

Chapter 3

A Swarm Cognition Realisation of Attention, Action Selection and Spatial Memory

Pedro Santana, Luís Correia.
Adaptive Behavior, 18(5):428-447, 2010.

Abstract

This paper reports a study on modelling covert visual attention as a parallel process that unfolds in synergy with the embodied agent's action selection process. The parallel nature of the proposed model is according to the multiple covert attention hypothesis and thus suitable for the concurrent search for multiple objects in the embodied agent's visual field. In line with the active vision approach, the interaction with the action selection process is exploited by the model to deploy visual attention in a by-need way. Additionally, the multiple focuses of attention considered in the proposed model interact in a way that their collective behaviour robustly self-organises for a proper handling of the speed-accuracy trade-off inherent to visual search tasks. Besides the self-organisation of a global spatio-temporal visual attention policy, the model also produces parallel, sparse, and active spatial working memories, that is, local maps of the environment. The underlying mechanisms of the model are based on the well known formalism that describes the self-organisation of collective foraging strategies in social insects. This metaphor is particularly interesting because of its similarities with the problem tackled in this paper, that is, the one of parallel visual attention. This claim is validated by experimental results on a simulated robot performing local navigation, where the ability of the model to generate accurate visual attention and spatial

memories in a parsimonious and robust way is shown.

Keywords: swarm cognition; active vision; visual attention; action selection; artificial life; autonomous robots

3.1 Introduction

More than an information process, perception is a sensorimotor process on its own right (Ballard et al., 1997; Pfeifer and Scheier, 1999; O'Regan and Noe, 2001). This active nature of perception is essential to enable robustness and efficiency (Bajcsy, 1988; Aloimonos et al., 1988; Ballard, 1991). This is in line with the idea that modules within a cognitive system are dynamical systems mutually entrained, as well as they are entrained with the external world through the agent's body (Beer, 1995; Thelen and Smith, 1996). To focus perception in a by-need way, this article addresses the particular entrainment observed between action selection and visual attention.

Visual attention ultimately results in the motion of sense organs towards the relevant stimulus source. This is called overt attention. A faster process is the one of mentally focussing particular aspects of the sensory stimuli. This is called covert attention. Studies on human subjects support the hypothesis that multiple covert attention processes co-exist in the brain (Pylyshyn and Storm, 1988; Doran et al., 2009). Based on this, the model herein proposed considers covert visual attention as a process composed of a set of parallel focuses of attention, which cohabit and interact among themselves and with the action selection process for a better and by-need coverage of the embodied agent's visual field.

One of the most powerful formalisms capable of providing these properties to systems composed of many interacting elements is the one that describes the self-organisation of collective intelligence displayed by social insects (Franks, 1989). In this line, the proposed model uses the ant foraging metaphor to control the collective behaviour of a set simple virtual agents that inhabit the embodied agent's sensorimotor and decision spaces. The behaviour of each of these virtual agents implements a locally coherent covert attention policy. The self-organisation of these virtual agents leads in turn to a global spatio-temporal collective pattern of coherent, robust, and efficient covert attention policy. Parallel, sparse, and active spatial working memories are also maintained by the system, which besides helping the action selection process to

reach better decisions, also improve the visual attention process by endowing it with some predictive power.

The experimental apparatus is composed of a simulated robot that must focus its perceptual resources to enable the detection and retention of memories of the obstacles that are essential to ensure safe local navigation. As argued by Slocum, Downey, and Beer (Slocum et al., 2000), "if we hope to ... analyse in detail model agents exhibiting genuinely cognitive behaviour, it is essential to focus on the simplest possible agent-environment systems that exhibit the cognitive behaviour of interest". It not being the goal of the model to encompass the complexity of object identification and/or recognition, means that the only cognitive behaviour of interest is the one of directing multiple covert visual attention processes in a by-need basis. Hence, the simplest agent-environment system in this case is the one where the perceptual process needs to search a sensory space, whatever its dimensions, and the identification/recognition aspects are reduced to a well defined binary decision process. Bearing this in mind, the simulated robot is provided with a body-centred two-dimensional image of the local environment in which obstacles can be detected by simply checking which image pixels are occupied.

This article is organised as follows. Section 3.2 surveys related work in both natural and applied sciences. Then, in Section 3.3, an overview of the proposed model is presented, as well as a focused inspection on the parallel perceptual system. The experimental results, obtained with a simulated mobile robot, are reported in Section 3.4. A discussion about key aspects of the model is given in Section 3.5. Finally, a set of conclusions is drawn, and pointers to future work are suggested in Section 3.6.

3.2 Related Work

The ability to shift the focus of attention, ubiquitous in animals (Land, 1999), highlights the active nature of perception (Bajcsy, 1988; Aloimonos et al., 1988; Ballard, 1991). From this perspective, body, nervous system, and environment must be seen in an holistic way (Gibson, 1979; Ashby, 1952; Beer, 1995; Thelen and Smith, 1996). An advantage of active vision is the possibility of the embodied agent to act in order to shape its sensory information flow (Sporns and Lungarella, 2006), and consequently make its processing tractable. Sensorimotor coordination (Dewey, 1896) thus plays a key role for adaptive behaviour (Brooks, 1991; Ballard et al., 1997; Pfeifer and Scheier,

1999; O'Regan and Noe, 2001; Pfeifer and Bongard, 2006; Mossio and Taraborelli, 2008).

Models of visual attention, a component of active vision, typically assume the existence 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 of the interplay between bottom-up and top-down processes is typically assumed to be some sort of moving "spotlight" from the most to the least relevant regions of the image. However, it is unlikely that a massively parallel structure, as the human brain is, only runs a single sequential processing stream. In fact, studies with human subjects (Pylyshyn and Storm, 1988; Doran et al., 2009) revealed exactly the opposite. Moreover, in vision in general, interacting perceptual pathways dedicated to different purposes are more likely to exist (Milner and Goodale, 1995; Goodale, 2008), than a global and isomorphic representation of the environment (Marr, 1982).

Overt attention being a sensorimotor process, and supported by the evidence that both overt and covert share the same neural mechanisms (de Haan et al., 2008), we speculate that covert attention can also be modelled as a sensorimotor process. Under this assumption, it is reasonable to consider the focus of attention as a dynamical entity inhabiting the sensory space of the embodied agent. This entity can then proactively move, that is, guide the focus of the embodied agent, towards the most relevant regions of the sensory space in a sensorimotor coordinated way. This entity, which can be modelled as a virtual agent, thus behaves as a locally sequential covert attention process. The feasibility of such agent-based modelling has considerable support from the fields of embodied cognition and active vision, where simple virtual agents performing sensorimotor coordination have been successfully synthesised (Beer, 1996; Scheier et al., 1998; Slocum et al., 2000; Nolfi and Marocco, 2002; Beer, 2003; Balkenius et al., 2004; Floreano et al., 2004; Nolfi, 2005; Suzuki and Floreano, 2006; Pfeifer and Bongard, 2006; Kim and Moeller, 2006; Sporns and Lungarella, 2006; de Croon and Postma, 2007; Choe et al., 2008).

Brain computational modelling with multiple virtual agents is not a new idea (Minsky, 1988; Chialvo and Millonas, 1995). Although the first realisations of the idea to

solve computer vision related problems are also not new (Poli and Valli, 1993; Liu et al., 1997), only more recently it has received considerable attention (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). These parallel computational models are mostly stand-alone engineered parallel perceptual systems, lacking the interaction with action selection and mapping processes. This deficit undermines their explanation power regarding the mechanisms actually building up adaptive behaviour. Conversely, sensor planning, which is a relatively stable field in computer vision and robotics communities, is actually trying to bridge the gap between body motions and information gathering through the sensors (Dickmanns et al., 1990; Nabbe and Hebert, 2003; Kwok and Fox, 2004; Patel et al., 2005; Hernandez et al., 2007; Sprague et al., 2007). However, none of these models considers parallel covert attention operating in an intricate way with the action selection process.

A remarkable metaphor from the natural world encompassing the characteristics of parallel deployment of attention is the foraging behaviour of army ants. These ants are able to leave their nests and cover large areas to search for food items in a purely parallel and robust way (Deneubourg et al., 1989), exhibiting a sort of collective intelligence (Franks, 1989). This mapping between insect-based swarms and neuron-based brains of vertebrate has been also suggested by parallel and independent work (Passino et al., 2008; Couzin, 2009; Marshall et al., 2009; Marshall and Franks, 2009). Our work, instead, approaches the problem of studying cognition through social insects behaviour by building it, following the synthetic approach to embodied cognition (Pfeifer and Scheier, 1999) and artificial life (Bedau, 2003). The advantages of using artificial life models for this purpose, though without the support of any practical realisation, have also been dissected in a parallel study (Trianni and Tuci, 2010). All these accounts can be framed in the emerging multidisciplinary field of swarm cognition (Trianni and Tuci, 2009), which attempts to uncover the basic principles of cognition, that is, adaptive behaviour, recurring to self-organising principles, mainly those exhibited by social insects. Furthermore, conceiving cognition as a self-organising process is essential to understanding how open-ended learning can develop, and consequently, how it can be synthesised.

3.3 Proposed Model

This section starts by providing an introduction to the biological inspiration of foraging strategies in social insects as a model of parallel covert visual attention. This introduction is followed by the description of the algorithm specifying the way perception, which includes visual attention, and action selection interact in the proposed model. Details on both perceptual and action selection processes are provided. The model is described under the context of local navigation. That is, the action selection process aims at deciding which motor action should be engaged by the embodied agent to allow its safe progression in the environment. This decision is influenced by a desired heading of motion and constrained by the obstacles present in the environment, whose detection is responsibility of the perceptual process.

3.3.1 Biological Inspiration

Covert visual attention is mostly related to the parallel search of objects in the embodied agent's visual field. A remarkable metaphor of this process is the one of army ants engaging on foraging behaviour. Assuming that the environment that these ants inhabit corresponds to the embodied agent's visual field, each ant can be seen as an individual covert visual attention process. Their collective behaviour can then be seen as a parallel covert visual attention process.

Following this metaphor from the natural world, the visual process in the proposed model is composed of a swarm of simple homogeneous virtual agents (hereafter p-ants) that inhabit the visual field of the embodied agent (hereafter simply *agent*). These p-ants are probabilistically created (recruited) to search (forage) the agent's visual field along those regions where detected obstacles are likely to affect the action selection process more strongly. Hence, p-ants operate on a by-need basis being affected by the action selection process. As in natural ants, p-ants do this search in a stochastic way. P-ants interact through *stigmergy* for better coverage and tracking of detected objects. It allows the coexistence of positive and negative feedback loops that lead to robust collective behaviour. In conclusion, random fluctuations and both positive and negative feedback, which are necessary ingredients for self-organisation to occur (Bonabeau et al., 1999), are integral part of the model.

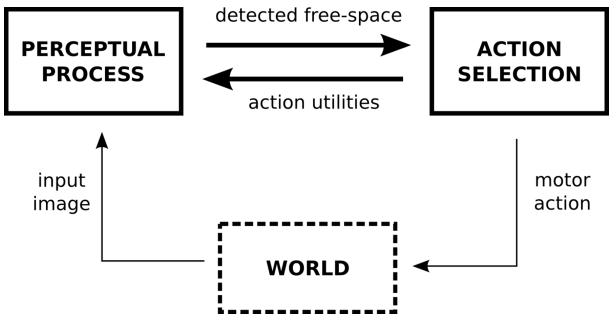


Figure 3.1: Building blocks of the proposed model.

3.3.2 Overall Process

Fig. 3.1 illustrates the connectivity between perceptual and action selection processes, whose control logic is outlined in Algorithm 1. Basically, after receiving a new frame, I, both perceptual and action selection processes interact (thicker arrows) for several iterations before a final motor action decision is reached and eventually engaged. These interactions occurring between both processes allow them to progressively unfold in parallel, and consequently, to enable accurate deployment of visual attention, which is an integral part of the perceptual process. The following describes these interactions for a given frame.

At each iteration, the action selection process sends a message to the perceptual one with an action utility vector, $\mathbf{u} = (u^1, u^2, \ldots, u^k)$, where $u^j \in [0,1]$ is the utility of performing action j, and k is the cardinality of the agent's action repertoire. This vector is computed according to a desired heading of motion, h, and constrained by information about free-space connectivity of the local environment, C, which has been sent by the perceptual process in the previous iteration. In this study, actions u^j are defined as linear trajectories centred on the agent and directed outwards in a radial pattern.

In turn, the perceptual process uses the just received vector $\mathbf{u}$ to iterate its search for obstacles. The set of obstacles detected in the current iteration, O', is appended to the set O, which accumulates obstacles previously detected. This set is then used to compute C, which encompasses range information regarding those radial sectors of the local environment that contain free-space for agent motion. Finally, C is sent back as a message to the action selection process so as to affect its next iteration. Since the perceptual process is affected by the incoming vector $\mathbf{u}$, which states the utility of performing each possible action and consequently which regions of the environment

Algorithm 1: Overall system's execution pseudo-code.

Input: desired heading, h, desired speed, s
Data: i_{max}, i_s are empirically defined constants.

create an empty set of p-ants, $A \leftarrow \emptyset$
while *true* **do**
  reset iterations counter, $i \leftarrow 0$
  reset stable iterations counter, $c \leftarrow 0$
  clear the set of obstacles, $O \leftarrow \emptyset$
  clear free-space connectivity information, $C \leftarrow \emptyset$
  reset perceptual shared mediums, $R \leftarrow -1, P \leftarrow 0$
  reset best action, $v \leftarrow 0, v'' \leftarrow 0$
  get new frame, I
  estimate agent motion, $(\widehat{\Delta x}, \widehat{\Delta y}, \widehat{\Delta \theta})$
  // iterate until best action stabilises for i_s iterations or a maximum nr. of iterations, i_{max}, is reached
  while $(c < i_s \wedge i < i_{max})$ **do**
    $i \leftarrow i + 1$
    // iterate the action selection process
    compute the action utility vector, $\mathbf{u} = (u^1, u^2, \ldots, u^k)$, based on C, h, and v''
    save previous best action, $v' \leftarrow v$
    obtain the highest utility action in $\mathbf{u}$, $v \leftarrow \arg\max_j u^j$
    // iterate the perceptual process
    iterate obst. search,
    $(R, P, A, O') \leftarrow$ **iterateDetection**$(I, (\widehat{\Delta x}, \widehat{\Delta y}, \widehat{\Delta \theta}), R, P, A, \mathbf{u}, s, i, k)$ *// see Algorithm 2*
    update the set of obstacles, $O \leftarrow O \bigcup O'$
    compute free-space connectivity, C, based on O
    // update counter of stable iterations
    if $v = v'$ **then**
      $c \leftarrow c + 1$
    else
      $c \leftarrow 0$
    end
  end
  generate motor action for v and desired speed s
  save motor action sent to the actuators, $v'' \leftarrow v$
end

should be more carefully analysed, it is considered to operate on a by-need fashion.

The mutual influence between action selection and perceptual processes is carried per frame: (1) for a maximum number of iterations, i_{max}; or (2) until the action with highest utility is the same over i_s iterations. Finally, the highest utility action at the time interactions cease is passed to the low-level motion controller through a low-pass filter. In this way, sudden changes at the system's output are smoothed to avoid jitter at the actuators level.

Details regarding both perceptual and action selection processes are provided in the following sections.

3.3.3 Perceptual Process

As mentioned, the iteration of the perceptual process is three fold (see Algorithm 1). First, a set of obstacles, O', is detected in the current iteration. Second, O' is appended to the cumulative set O. Third, O is used to re-compute C, that is, the local environment's free-space connectivity.

The set of obstacles O is used to determine the maximum distance, d^j, the agent is able to travel along each possible linear trajectory, $j \in [1,k]$, without hitting any obstacle. Refer to (Santana and Correia, 2006) for further details. The set C is then the aggregate of all free-space, defined in terms of pairs (j,d^j). With this information the action selection process is able to assess the quality of each possible motor action.

At each iteration, the set O' is created as follows (see Algorithm 2). The perceptual process starts by extending the set of p-ants, A, with a new one, A'. The rate at which p-ants are created is a function of the agent's speed, $s \in [0,1]$. This allocates more perceptual resources when the agent is moving faster and consequently when an action must be selected sooner. The creation rate is also a function of the action utility vector. The higher the utility of a linear trajectory, j, the higher the chances of creating a p-ant, a_j, with the following initial position:

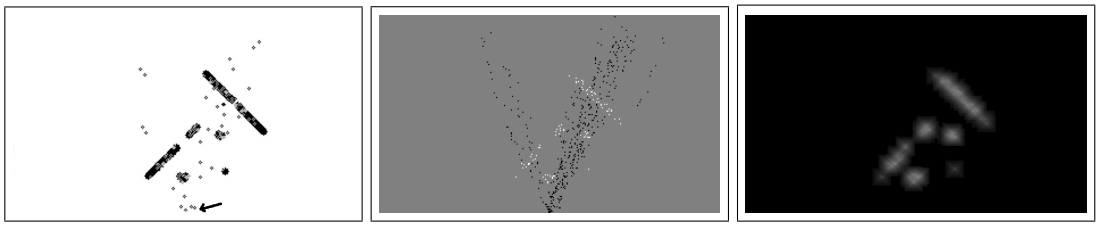
$$p_{a_j} \leftarrow [\angle_j, 0]^T, \tag{3.1}$$

which is defined in polar coordinates in the image plane and where $\angle_j$ is the direction of linear trajectory j. This position corresponds to the bottom-centre of the visual field, so that p-ants can start their search for obstacles in the close vicinity of the agent (see Fig. 3.2(a)).

Algorithm 2: iterateDetection

Input: $I, (\widehat{\Delta x}, \widehat{\Delta y}, \widehat{\Delta\theta}), R, P, A, \mathbf{u}, s, i, k$
Output: R, P, A, O'
Data: $(p_a, z_a, \sigma_a, o_a, b_a, r_a)$ is the set of properties of a given p-ant $a \in A$ (explained in the text).
Data: η, ζ, υ, ρ, r_{max}, and l are empirically defined constants and FOV is the set of pixels composing the agent's visual field (explained in the text).

create an empty set of obstacles, $O' \leftarrow \emptyset$
sample a number from an uniform distribution, $y \in [0,1]$
if $y < s \wedge i > 1$ **then** *// the faster the agent the higher the chances of creating p-ants*
    // create a new set of p-ants, A', to be added to A
    $A' \leftarrow \emptyset$
    foreach $j \in [0,1,\ldots,k]$ **do** *// go over all possible actions*
        sample a number from an uniform distribution, $x_j \in [0,1]$
        if $x_j < u^j$ **then** *// the higher the utility of the action the higher the chances of creating a p-ant*
            create new p-ant a_j with $(p_{a_j}, z_{a_j}, \sigma_{a_j}, o_{a_j}, b_{a_j}, r_{a_j}) \leftarrow ([\angle_j, 0]^T, [\angle_j, 0]^T, 0, \rho, \text{SEARCH}, r_{max})$
            add the new p-ant a_j to set A', $A' \leftarrow A' \bigcup \{a_j\}$
        end
    end
    add the new set of p-ants, $A \leftarrow A \bigcup A'$
end
// iterate all p-ants included in A
foreach $a \in A$ **do**
    reduce p-ant's energy available, $o_a \leftarrow o_a - 1$
    if $i = 1$ **then** *// new frame received*
        use $(\widehat{\Delta x}, \widehat{\Delta y}, \widehat{\Delta\theta})$ in Equation 3.2 to compensate p-ant's position, p_a, for agent motion
    end
    if $o_a \leq 0$ **then** *// p-ant without energy*
        remove p-ant, $A \leftarrow A \backslash \{a\}$
    end
    // iterate search behaviour
    if $b_a = SEARCH$ **then**
        iterate search behaviour, $(R, P, A, O') \leftarrow$ **iterateSearch**(I, R, P, A, O', FOV, a) *// see Algorithm 3*
    end
    // iterate track behaviour
    if $b_a = TRACK$ **then**
        iterate track behaviour, $(R, P, A, O') \leftarrow$ **iterateTrack**$(I, R, P, A, O', FOV, a, \zeta, i)$ *// see Algorithm 4*
    end
    // iterate local search behaviour
    if $b_a = LOCAL\_SEARCH$ **then**
        iterate local search behaviour,
        $(R, P, A, O') \leftarrow$ **iterateLocalSearch**$(I, (\widehat{\Delta x}, \widehat{\Delta y}, \widehat{\Delta\theta}), R, P, A, O', FOV, a, \eta, \upsilon, l, i)$ *// see Algorithm 5*
    end
end
return R, P, A, O'



(a) (b) (c)

Figure 3.2: Snapshot of the visual process activity, in a situation where the action utility vector states that moving to the right is preferred to moving to the left, and both are preferred to moving forward. (a) P-ants (small grey circles) in search (over non-obstacle regions) and tracking (over obstacles, in black) behaviour. Note that the input image is a body-centric two-dimensional Euclidean image of the local environment. The point where p-ants are created is at the bottom-centre of the input image, indicated by the arrow. (b) Shared medium R representing the pixels classified as obstacle (white), as non-obstacles (black), and that have not been tested so far (grey). (c) Shared medium P (in grey level) updated by p-ants as obstacles are detected.

A newly created p-ant, a_j, is endowed with an initial energy level, $o_{a_j} \leftarrow \rho$, which is reduced at each iteration and restored when an obstacle is detected (see below). P-ants with null energy are removed from the system to maintain memory and computation within boundaries. A p-ant initiates its operation with the activation of the *search behaviour* (see below), $b_{a_j} = \text{SEARCH}$. With this behaviour, the p-ant will move on the agent's visual field with the purpose of detecting an obstacle, which is then tracked by the activation of other behaviours.

If the current iteration is the first after obtaining the current frame, $i = 1$, p-ants are not created and the agent's motion since the last frame, $(\widehat{\Delta x}, \widehat{\Delta y}, \widehat{\Delta\theta})$, is estimated recurring to wheel odometry. This estimate is then used to compensate the position of every p-ant $a \in A$, for the agent motion:

$$p_a^* \leftarrow \mathcal{R}(-\widehat{\Delta\theta}) p_a^* - \begin{bmatrix} \widehat{\Delta x} \\ \widehat{\Delta y} \end{bmatrix}, \tag{3.2}$$

where p_a^* corresponds to p_a in Cartesian coordinates, $\mathcal{R}(-\widehat{\Delta\theta})$ is a two-dimensional rotation matrix in order to compensate for the angle $\widehat{\Delta\theta}$. This procedure helps p-ants maintaining their positions with respect to the objects in the environment independently of agent motion. The motion compensated position of each p-ant can also be seen as an interesting point upon which covert attention should be deployed when a new frame arrives. This introduces a prediction component to the covert attention pro-

cess. During the motion compensation process, some p-ants will eventually leave the agent's visual field and so will ultimately implement a sparse local map of the environment. Since the motion compensation is performed by all p-ants, these already out of the visual field will also be affected by the process. This is essential to maintain the local map updated and also to allow some of those p-ants to re-enter the agent's visual field when an environment is revisited.

After these preparatory procedures, the currently active behaviour of each p-ant in A is iterated. That is, each p-ant will perform a motion step according to a given stochastic perception-action rule. The following sections describe the iteration of each of the three possible behaviours, namely *search*, *track*, and *local search*.

3.3.3.1 Search Behaviour

The iteration of the search behaviour is a simple stochastic motion step on input frame I along the preferred search direction defined at p-ant creation time (see Equation 3.1). In the case of a p-ant $a \in A$ with current position p_a the motion step is defined as follows:

$$p_a \leftarrow p_a + \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \begin{bmatrix} N(0,1) & 0 \\ 0 & |N(0,1)| \end{bmatrix}, \tag{3.3}$$

where $N(0,1)$ samples a number from a Gaussian distribution with mean 0 and variance 1 (see Fig. 3.2(a)), λ_1 and λ_2 are empirically defined scalars. High λ_1 and λ_2 values facilitate fast detection of large obstacles, at the cost of missing smaller ones. Small values result on slower, though finer, detection.

After performing a motion step, p-ant a invokes the function $(R,d) \leftarrow detect(I,R,p_a)$, whose result is $d = 1$ if an obstacle is found at position p_a, and $d = 0$ otherwise. Parameter R is a shared medium that the function uses to check whether any other p-ant has already tested the position in question (i.e., if $R(p_a) \neq -1$). If that is the case, then $R(p_a)$ is used as the result. Conversely, if the position is tested for the first time, the function applies the obstacle detector and updates the shared medium accordingly, that is, it does $R(p_a) \leftarrow 1$ if the result is positive and $R(p_a) \leftarrow 0$ if otherwise. This helps saving computation when the cost of applying the obstacle detector is high. In sum, the function *detect* abstracts the detection process as well as the management of the shared medium R. With the exception of this function call, the rest of the perceptual process is about visual attention.

After detecting an obstacle, a set of steps is executed by the p-ant a. First, its position is appended to the set of obstacles O', $O' \leftarrow O' \bigcup \{p_a\}$, which is also updated by other p-ants and eventually used to feed the action selection process. Furthermore, the p-ant's level of energy is restored, $o_a \leftarrow \rho$, and its behaviour changed to *track behaviour*, $b_a \leftarrow$ TRACK, so that the detected obstacle can be followed across frames. Then, a shared medium P is updated to reflect the localisation of the detected obstacle. This is done by adding to P a top-view pyramidal shape of top magnitude 20 and linear decay (0.9) outwards, centred on p_a. This information will then be used by other p-ants to have an idea of the density of p-ants on the region (see below).

Finally, a p-ant in search behaviour that leaves the agent's visual field is removed from the system, $A \leftarrow A \backslash \{a\}$. See Algorithm 3 for further details.

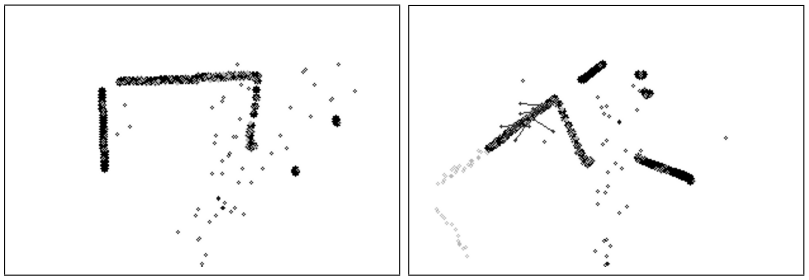
Algorithm 3: iterateSearch

Input: I, R, P, A, O', FOV, a
Output: R, P, A, O'

use Equation 3.3 to update the p-ant's position, p_a, along a preferred search direction
if $p_a \notin FOV$ **then** *// p-ant out of the visual field*
    remove p-ant, $A \leftarrow A \backslash \{a\}$
else *// p-ant in the visual field*
    apply detector at p-ant's position, $(R, d) \leftarrow detect(I, R, p_a)$
    if $(d = 1)$ **then** *// an obstacle has been detected in input image I*
        report the presence of obstacle at p-ant's position, $O' \leftarrow O' \bigcup \{p_a\}$
        raise the obstacles density shared signal at p-ant's position, $P(p_a)$
        restore p-ant's level of energy, $o_a \leftarrow \rho$
        change p-ant to track behaviour, $b_a \leftarrow$ TRACK
    end
end
return R, P, A, O'

3.3.3.2 Track Behaviour

A p-ant engages in *track behaviour* when it finds an obstacle on its current position. Whenever a new frame is acquired, $i = 1$, that p-ant has its position compensated for the agent motion (see above), which ultimately results in the construction of a local map of the environment. However, if the motion estimation is noisy or the obstacle is dynamic, the updated position of the p-ant may no longer be the same as the one of



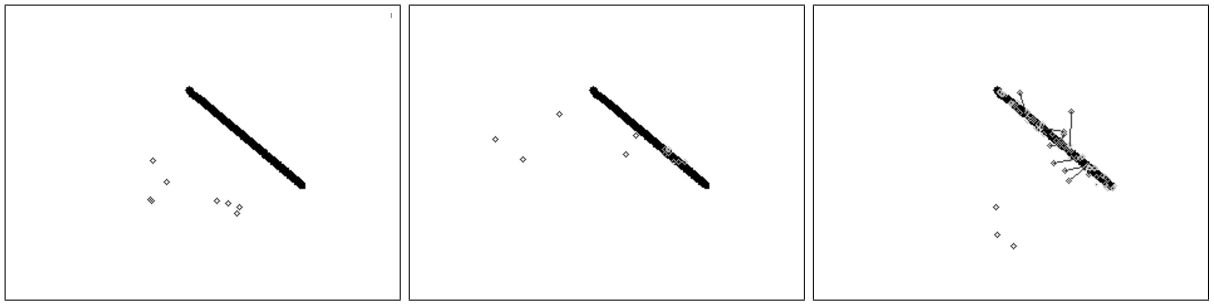
(a) t_1 (b) t_2

Figure 3.3: Graphical output generated by the perceptual process to illustrate the mapping capabilities in two sequential situations. (a) The agent faces a set of obstacles, which are detected and registered in the action selection process. Consequently, the decision to turn to the right is taken. In turn, the decision modulates the visual process, compelling p-ants to flow more on that direction. (b) After a displacement with rotation to the right, some obstacles move out of the agent's visual field (to the left). However, those obstacles remain represented by means of their associated p-ants, and thus still influencing the action selection process. Refer to Fig. 3.5 for the meaning of the lines associated to some p-ants. The grey-level colour coding is as in Fig. 3.2.

the obstacle. Therefore, immediately after obtaining a new frame, every p-ant in track behaviour invokes the *detect* function. In the presence of a positive result, the p-ant's position is appended to the set of obstacles, O', and its energy restored. See Fig. 3.3 for an illustration of the spatial working memory operation. As in the search behaviour, the shared medium P is updated to reflect the localisation of obstacles detected in the agent's visual field.

With the purpose of rapidly covering detected obstacles, p-ants in track behaviour that are in the agent's visual field are allowed to locally replicate themselves. At each track behaviour iteration, a clone a' of a given p-ant a may be created and positioned in a randomly chosen location q, $p_{a'} = q$, provided that this position is not densely populated with other p-ants, $P(q) < \eta$, and it contains an obstacle. q is selected by searching for a pixel at a distance $(\lambda + \beta \cdot y_a)$ of p_a, where $y_a \in [0,1]$ is a number sampled from an uniform distribution. After appending the clone a' to the set of p-ants, $A \leftarrow A \bigcup \{a'\}$, its position is reported as an obstacle to O' and P is updated accordingly.

A p-ant is only allowed to clone for a limited number of times, r_{max}. A clone inherits the number of replications of its ancestor so as to control the diffusion process. This means that the number of descendants of a single p-ant can amount up to $r_{max}!$. In practice the number is much smaller as the cloning process is constrained by the



(a) t_1 (b) t_2 (c) t_3

Figure 3.4: Graphical representation of p-ants in a diffusion process. (a) A set of p-ants in search behaviour. (b) Some p-ants find the obstacle, enter in track behaviour, and start cloning themselves. (c) The diffusion process concludes. Refer to Fig. 3.5 for the meaning of the lines associated to some p-ants.

obstacle boundaries and by the density of p-ants in the location in question. Therefore, a single p-ant is generally not capable of cloning up to the point of fully covering a large obstacle. If the object is in a region of the environment that is really important to the action selection process, its uncovered region will be seen as free space and consequently attract new p-ants towards itself. These new p-ants will eventually reach the obstacle, a new diffusion process will be started, and the obstacle will be fully covered. See Fig. 3.4 for an illustrative example of the diffusion process.

If a given p-ant a in track behaviour happens to be in a region with a high density of p-ants, $P(p_a) > \zeta$, or detects that it is no longer on an obstacle, then it leaves the track behaviour to initiate a local search behaviour, $b_a \leftarrow$ LOCAL_SEARCH. By making $\eta < \zeta$, a sort of hysteresis is implemented, and with that, massive fluctuations of neighbouring p-ants entering and leaving local search behaviour are avoided.

Finally, the energy of p-ants in track behaviour that are out of the agent's visual field receive an extra decrement. As a consequence, p-ants not refreshed with immediate sensory information are maintained in the system for a smaller period of time. See Algorithm 4 for further details.

3.3.3.3 Local Search Behaviour

Local search is modelled as a random walk around an anchor point, whose initial position is defined as the p-ant's position at the time local search was initiated, $z_a \leftarrow p_a$. Hence, with the goal of re-detecting a lost obstacle or of finding a less cluttered region of a tracked obstacle, the local search behaviour randomly changes the position of a

Algorithm 4: iterateTrack

Input: $I, R, P, A, O', FOV, a, \zeta, i$
Output: R, P, A, O'

apply detector at p-ant's position, $(R, d_p) \leftarrow detect(I, R, p_a)$
if $i = 1 \wedge (p_a \notin FOV \vee (p_a \in FOV \wedge d_p = 1)$ **then** *// update obstacles only at frame onset*
  report the presence of obstacle at p-ant's position, $O' \leftarrow O' \bigcup \{p_a\}$
  raise the obstacles density shared signal at $P(p_a)$
  if $p_a \in FOV \wedge d_p = 1$ **then**
    restore p-ant's level of energy, $o_a \leftarrow \rho$
    if $r_a > 0$ **then** *// limit the number of replications*
      randomly generate a neighbour position $q \in FOV$ for p-ant replication
      apply detector at neighbour position, $(R, d_q) \leftarrow detect(I, R, q)$
      if $P(q) < \eta \wedge d_q = 1$ **then** *// new position at not too crowded obstacle region*
        decrease the number of p-ant's available replications, $r_a \leftarrow r_a - 1$
        replicate p-ant, $a' \leftarrow a$, do $p_{a'} \leftarrow q$, and add new p-ant to the p-ants set, $A \leftarrow A \bigcup \{a'\}$
        report the presence of obstacle at new p-ant's position, $O' \leftarrow O' \bigcup \{q\}$
        raise the obstacles density shared signal at $P(q)$
      end
    end
  end
end
if $p_a \notin FOV$ **then** *// p-ant out of visual field*
  extra reduction in p-ant's energy available, $o_a \leftarrow o_a - 1$
else *// p-ant in the visual field*
  if $P(p_a) > \zeta \vee d_p = 0$ **then** *// p-ant at too crowded or on non-obstacle region*
    change p-ant to local search behaviour, $b_a \leftarrow$ LOCAL_SEARCH
    set an anchor point for the p-ant's local search, $z_a \leftarrow p_a$
    set the initial spread of the local search, $\sigma_a \leftarrow 0$
    lower the obstacles density shared signal at $P(p_a)$
  end
end
return R, P, A, O'

p-ant a, around an anchor point, z_a:

$$p_a^* \leftarrow z_a^* + \sigma_a[N(0,1), N(0,1)]^T, \tag{3.4}$$

where $\sigma_a \leftarrow \min(\sigma_a + 1, l)$ and $N(0,1)$ samples a number from a Gaussian distribution with mean 0 and variance 1. By increasing σ_a at each iteration, the local search spreads up to the upper-bound l, which constrains the search to avoid migration of p-ants between obstacles.

The anchor point changes to the p-ant's current position, $z_a \leftarrow p_a$, whenever the density of p-ants there is higher than the one at the current anchor's location, yet not too high, $\eta > P(p_a) > P(z_a)$. This directs the local search towards regions where other p-ants, but not too many, reported the existence of an obstacle. Conversely, if any of these conditions is not met, then the anchor position is not changed and the level of energy of the p-ant in question is dramatically reduced by an amount of υ. This compels p-ants in local search behaviour to remain in the system only for a reduced number of frames. Nevertheless, since p-ants may be in local search behaviour across frames, the anchor points must be compensated for the agent's motion the same way p-ants positions are, at $i = 1$ (see above).

If a p-ant in local search happens to detect a not too cluttered obstacle, it is appended to O', the shared medium P is updated accordingly, the p-ant's energy is restored, and it changes to *track behaviour*. Finally, a p-ant a in local search that leaves the agent's visual field is removed from the system, $A \leftarrow A \backslash \{a\}$. See Algorithm 5 for further details and Fig. 3.5 for an illustrative example of the local search behaviour.

3.3.3.4 Location-Specific P-Ants

The described mechanism for p-ants creation is basically driven by the action selection process output. With it, the perceptual process is capable of adapting to the dynamic nature of the surroundings. However, there are some specific spots in the agent's workspace that demand for invariant attention. To take into account these particular cases, new p-ants are deterministically deployed in each frame at points of interest. In this work we do it for the region right in front of the agent, where obstacles must be detected as soon as possible. These p-ants are setup to immediately engage in *local search behaviour*.

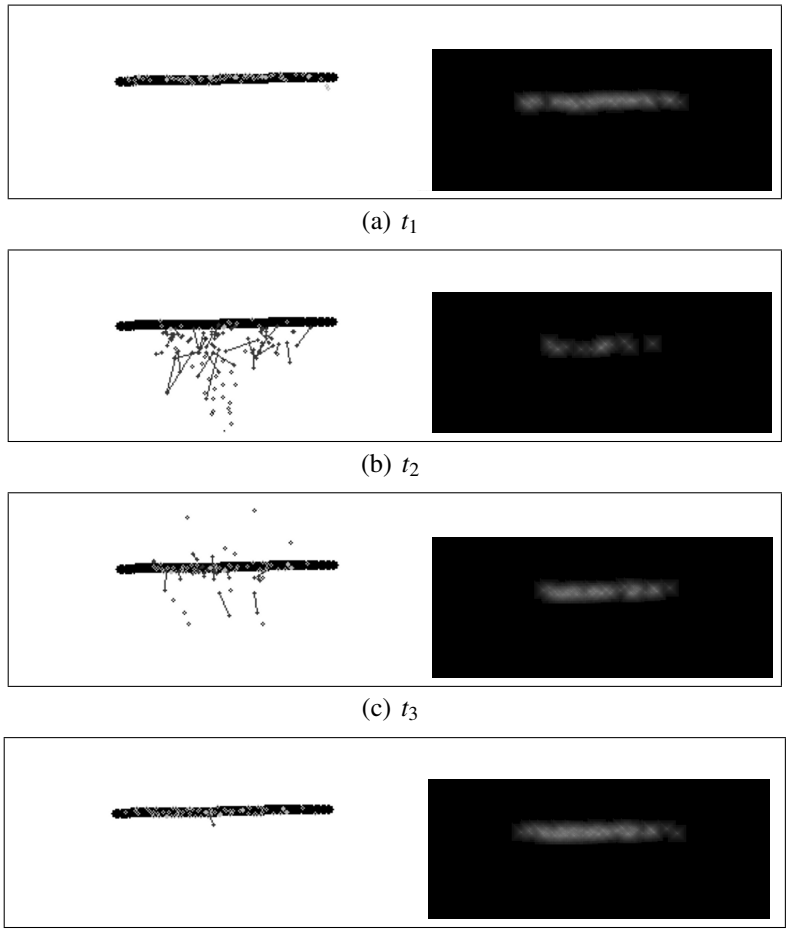
Algorithm 5: iterateLocalSearch

Input: $I, (\widehat{\Delta x}, \widehat{\Delta y}, \widehat{\Delta\theta}), R, P, A, O', FOV, a, \eta, \upsilon, l, i$
Output: R, P, A, O'

if $i = 1$ **then** *// new frame received*
  use $(\widehat{\Delta x}, \widehat{\Delta y}, \widehat{\Delta\theta})$, as in Equation 3.2, to compensate p-ant's anchor point, z_a, for agent motion
end
use Equation 3.4 to update p-ant's position around anchor point, z_a
if $p_a \notin FOV$ **then** *// p-ant out of visual field*
  remove p-ant, $A \leftarrow A \backslash \{a\}$
else
  apply detector at p-ant's position, $(R, d) \leftarrow detect(I, R, p_a)$
  if $d = 1 \wedge P(p_a) < \eta$ **then** *// p-ant on non-crowded region of obstacle*
    report the presence of obstacle at p-ant's position, $O' \leftarrow O' \bigcup \{p_a\}$
    raise the obstacles density shared signal, $P(p_a)$
    restore p-ant's level of energy, $o_a \leftarrow \rho$
    change p-ant to track behaviour, $b_a \leftarrow$ TRACK
  else *// local search must carry on*
    if $P(z_a) < P(p_a) < \eta$ **then** *// anchor point less crowded than p-ant's position*
      update p-ant's anchor point, $z_a \leftarrow p_a$
    else
      reduce dramatically p-ant's energy available, $o_a \leftarrow o_a - \upsilon$
    end
  end
  increment the spread of the local search, $\sigma_a \leftarrow \min(\sigma_a + 1, l)$
end
return R, P, A, O'

3.3.4 Action Selection Process

The action selection process produces at each iteration the multi-valued output **u**. As mentioned, actions in this article are instantiated as linear trajectories centred on the agent and directed outwards in a radial pattern. For this study, $j = 0$ and $j = k$ correspond to $+90^\circ$ and -90° linear trajectories, respectively, both perpendicular to the straightforward motion. Hence, the utility of each possible action is defined as the value of moving the agent along the corresponding linear trajectory. The closer an obstacle is from the agent, the lower the utility of the linear trajectory affected by the obstacle.



(a) t_1

(b) t_2

(c) t_3

(d) t_4

Figure 3.5: Graphical output generated by the perceptual process with erroneous motion estimate to illustrate the local search behaviour. Although the agent does not move, the motion estimate, which is noisy, reports a straight ahead motion. By motion compensation, p-ants move away from the obstacle. As the time unfolds (from t_1 to t_4) they manage to find the way back to the obstacle. Lines connect anchor's and p-ant's current positions. The grey-level colour coding is as in Fig. 3.2.

The specific fast obstacle avoidance method employed in this study (Santana and Correia, 2006) determines which linear trajectory the agent should take to produce faster progress along the direction of the goal heading, h, without hitting any obstacle reported by the visual process. Its geometric nature ensures smooth navigation in cluttered environments. Accordingly, the non-normalised utility of a given linear trajectory, $j \in [0,k]$, represented by the pair $(j,d^j) \in C$, is

$$v^j = \beta_1 \cdot \left(\frac{(d^j/r) \cdot \cos(|\angle_j - h'|) + 1}{2} \right) + \beta_2 \cdot \left(1 - |(d^j/r) \cdot \sin(\angle_j - h')|\right) + \beta_3 \cdot \left(1 - \frac{|j - v''|}{k+1}\right), \tag{3.5}$$

where v'' is the action sent to the actuators in the previous frame, $\angle_j$ is the angle between linear trajectory j and the agent's main axis of motion, and h' is desired heading h transformed to the agent's frame of reference. The utility is higher in directions whose range without obstacles, d^j, projected on the desired direction of motion is larger (first term), whose lateral displacement from that direction is smaller (second term) and that are closer to the last actually sent to the actuators (third term) in order to introduce some inertia. β_1, β_2, and β_3 are empirically defined importance weights set up to avoid oscillatory behaviour.

Then, those trajectories whose utility equals the highest one have their utilities amplified three times. This empirically defined scaling factor allows the best linear trajectory to become even more interesting. The utility normalisation finally occurs by doing $u^j = \max(\frac{v^j}{\arg\max_j v^j}, \varepsilon)$, where ε is an empirically defined minimum residual utility every linear trajectory must contain. In this way all linear trajectories have a non-zero probability of being analysed by p-ants.

3.4 Experimental Results

3.4.1 Experimental Setup

All experiments were carried out on a Centrino Dual Core 2 GHz running Linux/Ubuntu. A set of videos representative of the executed experiments can be found in the authors' Website (Santana and Correia, 2010b). In this study, a wheeled robot simulated in Player/Stage (Gerkey et al., 2003) with a width of $w = 0.3$ m is used as testbed. The robot's sensory input takes the form of 300×150 body-centric Euclidean two-dimensional images of the robot's 16 m $\times$ 8 m frontal environment, constrained by a 90° field-of-view. Obstacles are represented by black pixels, whereas non-obstacle regions are represented by white pixels. To emulate occlusions, only the closest obstacle

along each radial direction outwards the robot is imaged.

The following describes the model's parameterisation. The number of iterations, i_s and i_{max}, described in Section 3.3.2, have been set to 10 and 20, respectively. A higher number of iterations provided no significant added value to the stability of the system. The stochastic motion control parameters from Equation 3.3, λ_1 and λ_2, have been empirically defined as 0.3 and 1.0, respectively. These values produce an adequate speed-accuracy trade-off, for the environments where the simulated robot has been tested. The hysteresis control parameters used to determine when an obstacle is too cluttered with p-ants, η and ζ (Section 3.3.3.2), have been set to 100 and 150, respectively. With these values, p-ants are allowed to densely populate obstacles without exhibiting too much overlap. The diffusion process control parameters (see Section 3.3.3.2), λ and β, have been set to 3 and 7, respectively. These values have been chosen in order to generate a fast diffusion process, and also a relatively dense coverage of the obstacles. The number of times a p-ant is allowed to clone itself is $c = 10$. P-ants top energy is $\rho = 1000$ (see Section 3.3.3), which corresponds roughly to a 5 s lifespan. This is sufficient to maintain a useful short-term memory for local navigation. The control parameter to avoid migration of p-ants between obstacles when performing a local search (see Section 3.3.3.3), l, has been set to 10. Location-specific p-ants (Section 3.3.3.4) are deployed with polar coordinates $[0, 20]^T$ (see Equation 3.1). At this position, the p-ant will guarantee that obstacles in the close vicinity of the robot are surely found, independently of the remaining configuration of the environment. The cardinality of the robot's action repertoire, k, is 80. This action granularity is sufficient to ensure navigation in cluttered environments. The importance weights for the action selection process (see Section 3.3.4), β_1, β_2, and β_3, have been set to 0.35, 0.55, and 0.1, respectively. The residual action utility, ε, has been set to 0.005.

Two 16 m $\times$ 16 m environments have been devised to carry out the experiments. The first (see Fig. 3.6), denominated *lines* environment, contains large (easily detectable) obstacles in the far-field. A proper behaviour in this case requires good look-ahead capabilities. Conversely, the second environment (see Fig. 3.7), denominated *dots*, is cluttered with small (hard to detect) obstacles. In this case the robot should be able to perform thorough near-field obstacle detection. In a given sense, the *lines* environment affords speed whereas the *dots* environment demands accuracy. Using the same parameterisation for both environments, that is, without explicit context awareness, we show the system self-organising in order to handle the different speed-accuracy

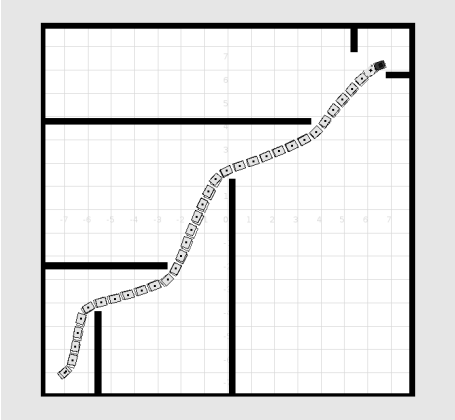


Figure 3.6: Behavioural results in the *lines* environment. Since all runs with active perception produced very similar trajectories, only one is plotted. The robot starts from the bottom-left aligned with the desired heading in the world frame. Filled trails left by the robot correspond to the ground-truth behaviour. Unfilled trails correspond to runs using the proposed active perception model. They are barely visible in this figure as a consequence of being covered by the filled trails. This means that the trajectory executed with the proposed model is near optimal, in the sense that it follows closely the ground-truth.

trade-offs.

The ability of p-ants located outside the robot's field-of-view to report their associated obstacles will be activated only in the last experiment, when explicitly mentioned. In this way, in the remaining experiments, we avoid any influence this feature may have in the results.

3.4.2 Results

During five runs per environment, the robot is asked to travel at a speed of $1.0\,\mathrm{ms}^{-1}$, along the heading coincident with the one of the robot when in its start position. Position (in meters) and heading (in radians) estimates are polluted with additive Gaussian noise, $(\widehat{\Delta x} + 0.1 \cdot N(0,1), \widehat{\Delta y} + 0.1 \cdot N(0,1), \widehat{\Delta\theta} + 0.001 \cdot N(0,1))$, where $N(0,1)$ represents samples from a Gaussian distribution with mean 0 and variance 1. With this error profile, the re-detection of tracked obstacles is required nearly in almost every frame. This is a reasonable situation for robots performing, for instance, in outdoor (unstructured) off-road environments. For comparison purposes, a ground-truth behaviour is generated by providing the action selection process with all sensory information contained in the robot's field-of-view, that is, with a fully-informed passive perception set-up.

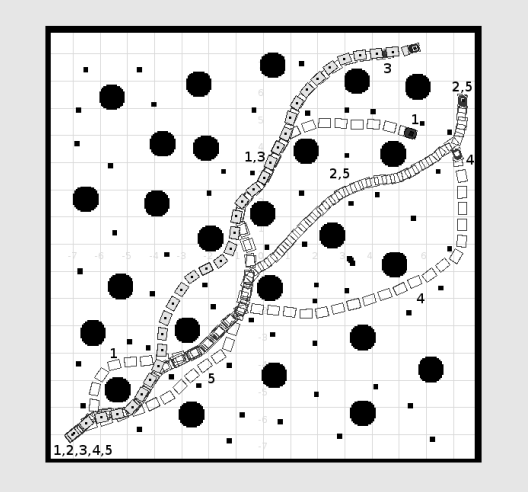
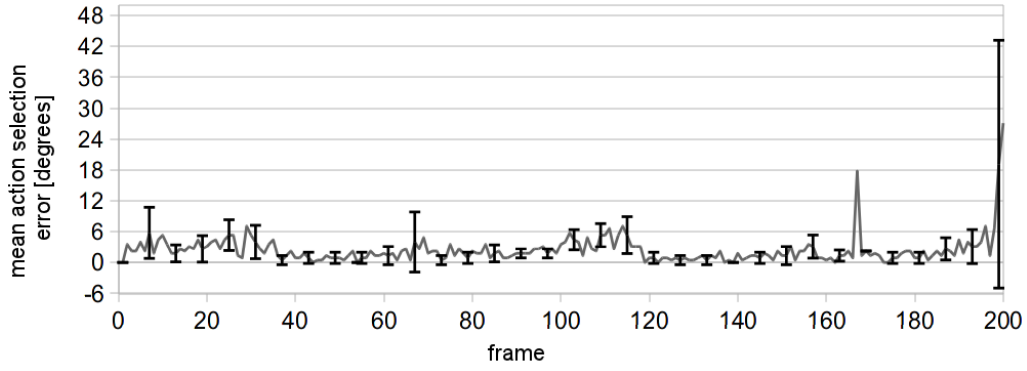


Figure 3.7: Behavioural results in the *dots* environment. The variability exhibited across the five runs (labelled with corresponding numbers) is a result of the environment's complexity. Small variations in the decision at key points cause considerable differences in the path taken. Nevertheless, in qualitative terms, the behaviour of the robot with the active perception set-up follows closely the ground-truth. Same colour code and process of obtaining ground-truth as in Fig. 3.6.

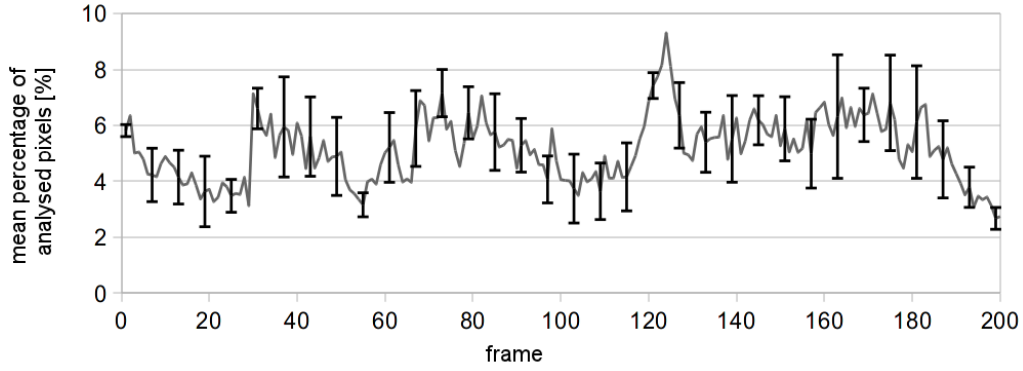
The proposed model generates smooth robot paths that follow closely the ground-truth (see Fig. 3.6 and Fig. 3.7), which is itself near optimal. This is further confirmed by Fig. 3.8(a), in which the steering angle error per frame is plotted. This plot was obtained by frame-wise averaging the error over the five independent runs, in the *lines* environment. The error metric is computed as the absolute difference between the steering angle cast by the action selection process to the actuators and the one associated with the ground-truth behaviour. The small error magnitude confirms the behavioural analysis, revealing the ability of the proposed model to accurately reproduce the activity of a fully-informed (near-optimal) system.

Following the ground-truth is not a big achievement if the percentage of analysed sensor data is high, thus approaching the fully-informed situation. Fig. 3.8(b) illustrates exactly the opposite situation for the proposed model on the five runs in the *lines* environment. Namely, the mean percentage of pixels analysed per frame, that is, in which the obstacle detector is actually applied, is only 5 % of the robot's field-of-view. This shows that the method maintains an adequate speed-accuracy trade-off with parsimonious allocation of resources.

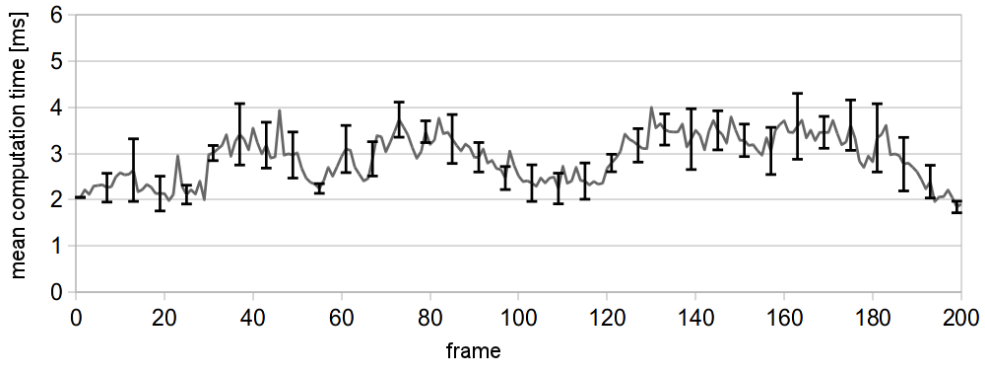
Because of the chaotic nature of the *dots* environment, independent runs on it lead



(a)



(b)



(c)

Figure 3.8: A set of performance metrics per frame. Values are the mean computed over five independent runs in the *lines* environment. Standard deviation represented by the error bars.

inevitably to different paths. Being considerably different, their correlation along time is poor. The quantisation of the performance metrics followed by their frequency analyses, i.e., through histograms, is much more informative in this case. These histograms are built by bin-wise averaging the individual histograms obtained for each of the five independent runs. This frequency analysis highlights the ability of the system to match ground-truth (see Fig. 3.9(a)) and accurately focus attention (see Fig. 3.9(b)) in the *dots* environment. These results are comparable to those obtained in the *lines* environment. The frequency analysis for the *lines* environment are coherent with the time-based analysis (cf. Fig. 3.8).

Fig. 3.8(c) and Fig. 3.9(c) demonstrate the ability of the system to maintain a low computational cost, per frame, in what refers both action selection and visual processes. The mean per frame of 2.9 ms for the *lines* environment, and similarly to the *dots* environment, is residual when compared to typical cost of vision-based obstacle detectors, ≈ 100 ms (Santana et al., 2008b). Thus, the overhead caused by the active vision infrastructure will not be noticed when applied to real-life obstacle detectors. Note that the additional cost of image acquisition and motor actuation, which together take ≈ 100 ms in the Player/Stage simulator, has not been considered. The fast computation obtained is in part due to the small convergence time of the action selection process. This can be observed by the small number of iterations required by the action selection process to stabilise its output (see Fig. 3.10).

As seen above, the system performs near-optimally and parsimoniously in both environments, with the same parameterisation. Although this shows the model's robustness, it is interesting to dissect the underlying causes of such adaptability. Fig. 3.11(a) shows a shift to higher values on the number of agents in local search behavioural mode for the *dots* environment, when compared to the ones in the *lines* environment. Fig. 3.11(b) shows that the same trend is observed for the number of agents in search mode. Refer to Fig. 3.12 for a detailed view on the number of p-ants in each behaviour during the five runs engaged in the *lines* environment. The complexity of the *dots* environment presents a larger number of potential directions of motion, which in turn results in a higher number of p-ants being created. This explains the higher number of p-ants searching in the *dots* environment. The more p-ants are created the higher their concentration around obstacles, thus explaining the higher number of p-ants engaging in local search due to saturation of shared medium P. While trying to find a less saturated spot, p-ants local search increasingly spreads over wider areas around the

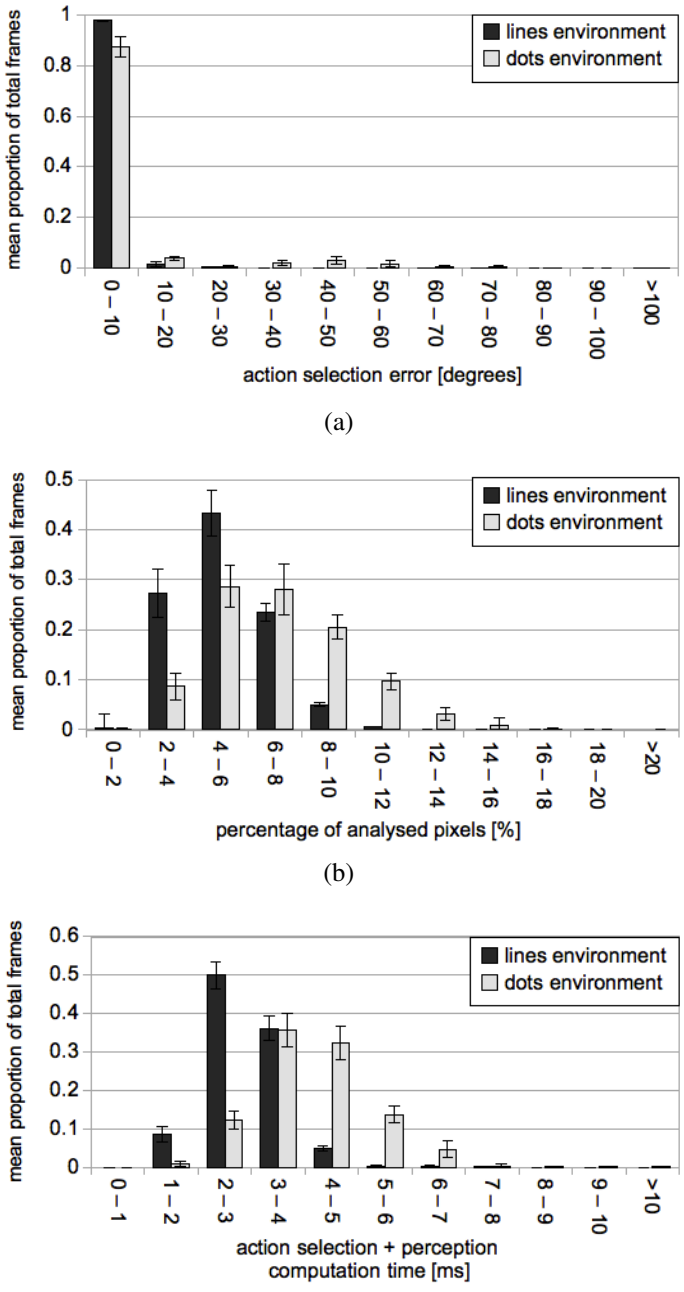


Figure 3.9: Histograms of a set of performance metrics. Values are the mean computed over five independent runs. Histograms normalised by the total number of frames. Standard deviation represented by the error bars.

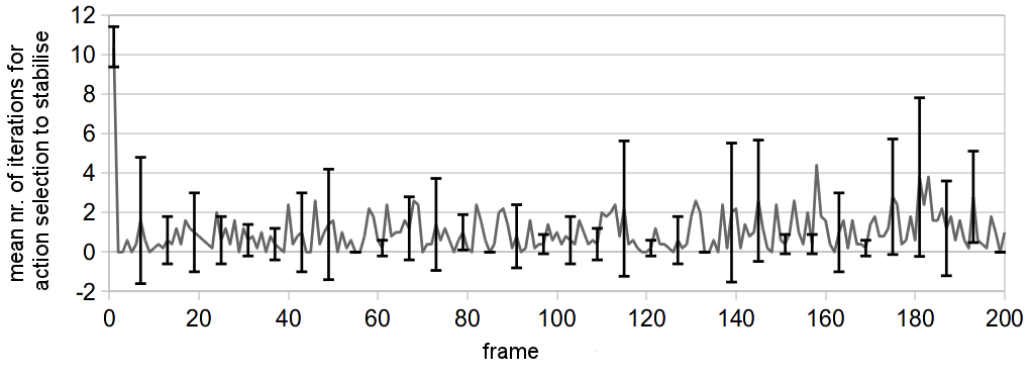
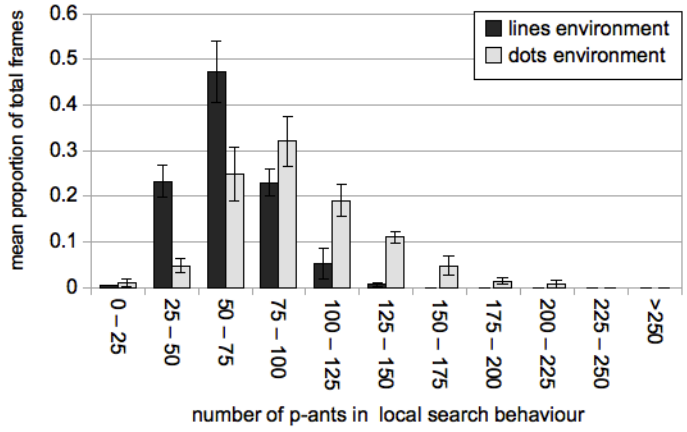


Figure 3.10: Number of iterations per frame, required for action selection to stabilise. Mean computed over five independent runs in the *lines* environment. Standard deviation represented by the error bars.

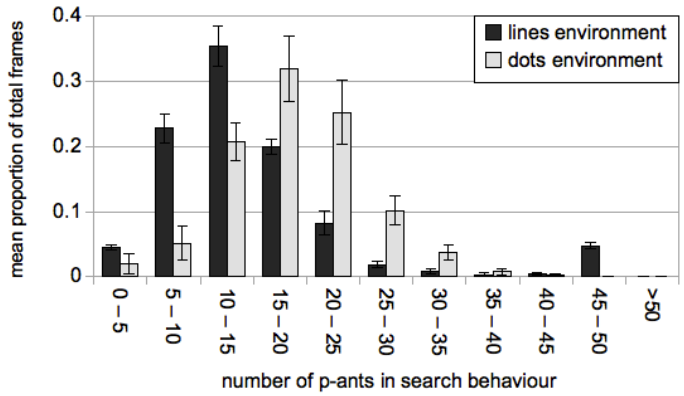
objects, thus resulting in a qualitatively different behaviour from the observed in the *lines* environment. Additionally, in cluttered environments p-ants tend to be locked on obstacles closer to the robot, raising the density of p-ants where it is most needed. In sum, the qualitatively different behaviour raises the chances of detecting relevant obstacles. Hence, the more complex the environment the more p-ants are actively exploring it, and the more focused they are on the vicinities of the robot. This can be observed also in Fig. 3.9(b), where a higher number of analysed pixels in the *dots* environment than in the *lines* one is reported.

As previously argued, the use of parallelism is essential to promote load balancing by enabling some predictive power. To verify this argument the system was tested with and "without" parallelism. To avoid interferences of memory in this test, p-ants were not allowed to survive across cycles. Absence of parallelism occurs when: (1) noise in the stochastic search is not employed, that is, $(\lambda_1, \lambda_2) = (0.0, 0.0)$; (2) p-ants replication is inhibited; and (3) p-ants are deployed only in the direction corresponding to the action with highest utility, that is, v. In these conditions p-ants operate mostly on a greedy manner without massively engaging in self-organising collective behaviour, and consequently without fully exploiting the advantages of parallelism. As expected, Fig. 3.13(a) and Fig. 3.13(b) confirm that in these conditions both ground-truth matching error and time to action selection convergence are higher.

Memory is important because it is what enables p-ants to move across frames and consequently to propagate evidence along time. This is confirmed by the reduced number of iterations required for action selection to stabilise when memory is activated, as



(a)



(b)

Figure 3.11: Histograms of the number of p-ants in search behaviours. Values are the mean computed over five independent runs. Histograms normalised by the total number of frames. Standard deviation represented by the error bars.

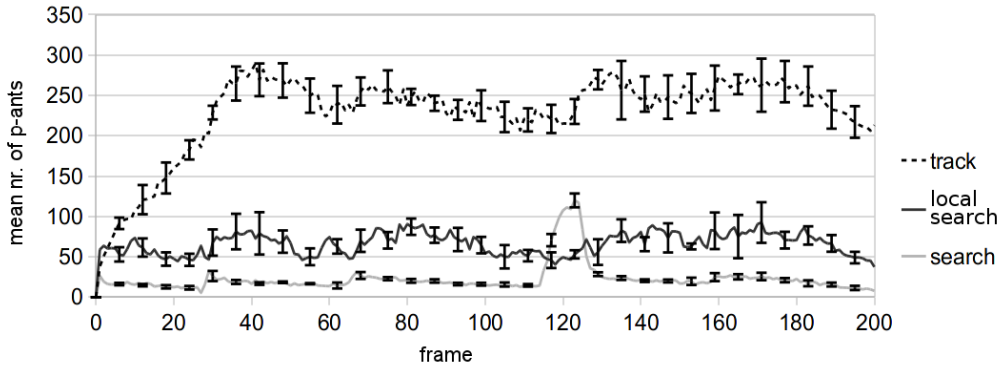
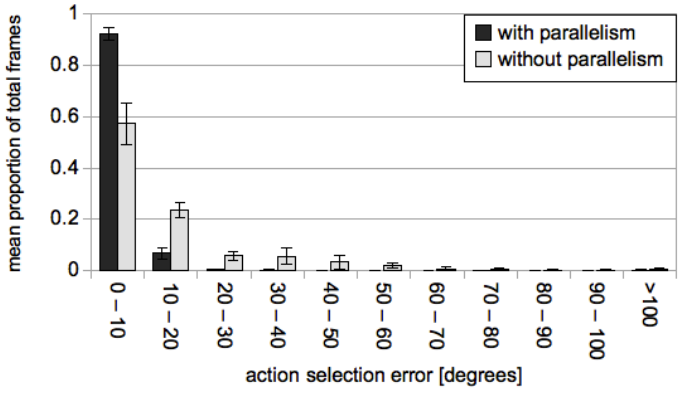
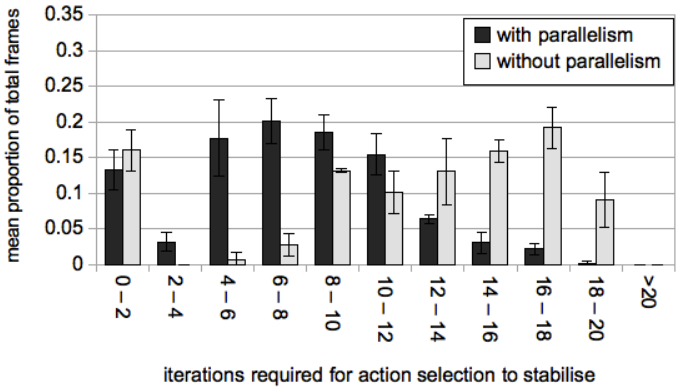


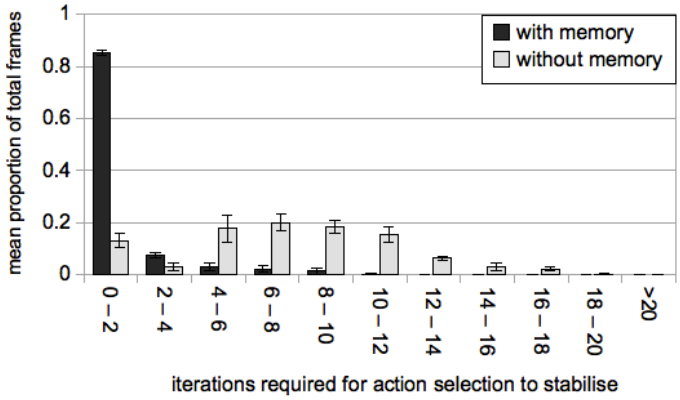
Figure 3.12: Cardinality of p-ants in each behavioural mode per frame. Values are the mean computed over five independent runs in the *lines* environment. Standard deviation represented by the error bars.



(a)



(b)



(c)

Figure 3.13: Results on the influence of parallelism and memory. The histograms were built for the *lines* environment by the same method as in Fig. 3.9. The benefit of using memory and parallelism is reflected in the consistent shift of the corresponding histograms towards the left, in respect to the ones corresponding to the system without these features. (a) Histograms of the ground-truth matching error with and without parallelism. (b) Histograms of the number of iterations required for action selection to stabilise, with and without parallelism. (c) Histograms of the number of iterations required for action selection to stabilise, with and without memory.

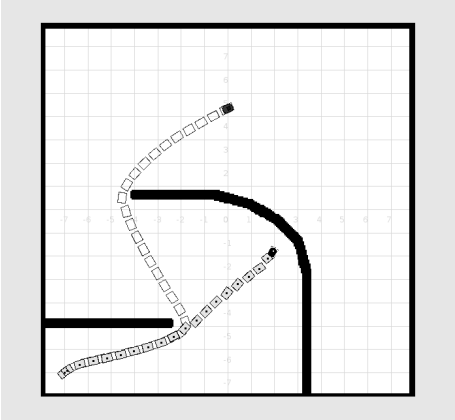


Figure 3.14: Short-term memory results. The robot starts from the bottom-left aligned with the desired heading. The grey robot's path corresponds to a typical situation where only information contained in the sensor's field-of-view (90°) is used. In this situation the action selection is only provided with perceptual information concerning a small region of the dead-end. As a result the action selection process, performing in a greedy way, repetitively switches its output between hard-left and hard-right turns, leading ultimately to a deadlock. With an extended field-of-view (180°), enabled by the short-term memory, the action selection process is already able to better assess the utility of each action (lighter robot's path).

depicted in Fig. 3.13(c). In a last experiment, the ability of p-ants located outside the robot's field-of-view to report their associated obstacles and consequently to affect the action selection process, was activated. Fig. 3.14 shows that this capability is essential for the robot to be able to avoid dead-ends.

3.5 Discussion

3.5.1 Role of Parallelism

Focussing all perceptual resources on the analysis of the environment's region associated to highest utility action is inefficient. Objects detected in other regions may also affect the utility of the most relevant region. The need for perceiving more than one region at a time demands a visual attention policy capable of handling the speed-accuracy trade-off. A detailed accurate analysis of the highest utility region may render impossible the analysis of the other regions. A too coarse analysis will hamper the detection of small obstacles. Defining a optimal visual search policy with this trade-off can only be done for well defined sets of agent-environment configurations. Instead, a

self-organising visual parallel search policy is capable of covering a wider envelope at the cost of doing it sub-optimally. Robustness being more relevant than optimality for embodied agents, self-organisation is paramount.

When either the environment or the task change, the importance of an obstacle for the action selection process may also change. Had the sub-optimal parallel search not been performed while the context was static, the system would be overloaded while trying to stabilise both perceptual and action selection processes to the new context. Hence, parallelism is important to maintain a robust balance of processing load in dynamic environments.

3.5.2 Implicit Context Awareness

In the proposed model there is no explicit control of the maximum number of p-ants that can be deployed. It is fully defined by the utility vector, and consequently it is an emergent property resulting from the immersion of the system in a specific environment. Formally, the average number of p-ants being created per iteration is $\sum_{j\in[0,k]}(u^j s)$. Practically, this means that for a speed of $0.5\,\mathrm{ms}^{-1}$, in an environment where the $k = 81$ possible motor actions exhibit an average utility 0.5, an average of 405 p-ants is created per frame, assuming an execution of $i_{max} = 20$ iterations. In a situation where only three possible motor actions are possible, and with highest utility, the average number of created p-ants drops to 30. That is, the number of p-ants being created varies with the environment. Hence, the system adapts to the context without explicit awareness.

3.5.3 Comparison With Other Population-Based Methods

The fact that p-ants inspect the input frame according to the agent's action selection process, in contrast to previous swarm-based related work for image analysis (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), is a significant difference from an embodied cognition standpoint. The same applies to the ability of p-ants to implement spatial working memories and to be affected by the agent's proprioception for motion compensation.

Another possible alternative to p-ants for obstacle search and tracking are Particle Filters. Although supported by a Bayesian framework, these filters exhibit two undesirable properties for them to be included in a *self-organising embodied cognition* frame-

work. First, particles are sampled by a centralised process. This limits their explanation power with respect to the emergence of cognition in animals and their implementation onto parallel hardware, which is essential to enable robot miniaturisation and energetic sustainability. Second, particles do not communicate with each other, thus neglecting the recognised role of lateral connectivity for sustainable self-organisation in neural structures. In fact, the recently reported out-performance of swarm-based methods over particle filters in visual tracking tasks (Zhang et al., 2008; John et al., 2010) is most often attributed to these two aspects.

Particle filters are recognisably powerful in providing autonomous agents with Simultaneous Localisation and Mapping (SLAM) capabilities (Thrun et al., 2005). The link between particles and p-ants in this case is not straightforward. In general, a particle in SLAM is a sample of the joint distribution robot/map, where a map is composed by a set of Extended Kalman Filters (EKF), each associated to a given landmark of the environment. Conversely, a p-ant is associated to a single landmark of the environment. Hence, a p-ant is more related to the role played by each EKF than with the particles themselves. In contrast to EKFs in SLAM, p-ants perform visual search in addition to tracking, and do it guided by the action selection process. Moreover, while p-ants interact to improve their individual performance, EKFs in particles are independent.

P-ants do not contribute to the estimation of the localisation of the agent, as particles do. Local navigation being the focus of the proposed model, the localisation problem is handled solely with dead-reckoning. While SLAM methods could deliver both localisation and mapping, their high computational and memory cost makes them unsuited for parsimonious local navigation. To reduce computation, usually particle filters are considerably simplified and consequently lose their most important property, that is, their theoretical completeness. From this and the previous paragraph one can conclude that the problem of local mapping in *self-organising embodied cognition* is more convincingly solved by the proposed model than by SLAM models.

3.6 Conclusions

A parallel model of covert visual attention was proposed and validated in a simulated robot performing local navigation. The covert visual attention process has been modelled with swarms of simple virtual agents, named p-ants, based on the social insects

foraging metaphor, motivated by the multiple covert attention hypothesis (Pylyshyn and Storm, 1988; Doran et al., 2009) and in line with the new born swarm cognition framework (Passino et al., 2008). Basically, p-ants perform local covert visual attention loops, whereas the self-organised collective behaviour maintains global spatio-temporal coherence.

The main goal of the model was to focus attention and consequently to reduce the processed image area. Experimental results confirm that this goal was attained, which suggests that the proposed model is able to run in real-time in vision-based robots with modest computational resources. The underlying rationale of the proposal is that the swarm-based fragmentation of the whole behaviour into simple local rules provides a robust system. Results also confirmed this.

Although the use of multiple virtual agents in image analysis is not new, this is the first report on its interaction with action selection. That is, p-ants opportunistically report their results to the action selection process, being at the same time guided by it. In other words, both action and perceptual processes are loosely coupled and unfold in parallel through cross-modulatory signals. This is fundamentally different from previous work in visual attention. For instance, in connectionist models, as typically considered in embodied cognition, both perceptual and action components share, to a large extent, the same computational units, that is, neurons. Being so tightly coupled, those approaches are difficult to modularise and consequently to scale. Conversely, another typical approach is to decouple both processes to the point they only share information in a master-slave way, such as in typical sensor planning strategies. This forces visual attention to operate at longer time-scales, which has little utility for the covert case.

The instantiation of the model onto a vision-based physical robot, to be pursued in future work, will be straightforward. It requires mostly: (1) the conversion between three-dimensional world and two-dimensional image space; and (2) changing the simple black/white discriminator that operates as obstacle detector by an appearance-based template matcher. In the future, we will also extend the model to enable the emergence of hierarchies among p-ants. This would allow hierarchical decomposition of visual attention and sparse spatial working memory. The parallel nature of the model can be further exploited, due to the emergence of new parallel computational platforms.

3.7 Acknowledgements

This work was supported by FCT/MCTES (grant number SFRH/BD/27305/2006). We would also like to thank the fruitful comments provided by Magno Guedes, Nelson Alves and anonymous reviewers.

Chapter 4

Swarm Cognition on Off-Road Autonomous Robots

Pedro Santana, Luís Correia.
Swarm Intelligence, 5(1):45-72, 2011.

Abstract

This paper contributes with the first validation of swarm cognition as a useful framework for the design of autonomous robots controllers. The proposed model is built upon the authors' previous work validated on a simulated robot performing local navigation on a two-dimensional deterministic world. Based on the ant foraging metaphor and motivated by the multiple covert attention hypothesis, the model consists of a set of simple virtual agents inhabiting the robot's visual input, searching in a collectively coordinated way for obstacles. Parsimonious and accurate visual attention, operating on a by-need basis, is attained by making the activity of these agents modulated by the robot's action selection process. A by-product of the system is the maintenance of active, parallel and sparse spatial working memories. In short, the model exhibits the self-organisation of a relevant set of features composing a cognitive system. To show its robustness, the model is extended in this paper to handle the challenges of physical off-road robots equipped with noisy stereoscopic vision sensors. Furthermore, an extensive aggregate of biological arguments sustaining the model is provided. Experimental results show the ability of the model to robustly control the robot on a local navigation task, with less than 1 % of the robot's visual input being analysed. Hence, with this system the computational cost of perception is considerably reduced, thus fostering robot miniaturisation and energetic efficiency. This confirms the advantages of using a swarm-based system, operating in an intricate way with action selection, to

judiciously control visual attention and maintain sparse spatial memories, constituting a basic form of swarm cognition.

keywords: Swarm Cognition, Visual Attention, Spatial Memory, Action Selection, Autonomous Robots, Biological Inspiration.

4.1 Introduction

Almost all embodied agents, either natural or artificial, deeply rely on perception for a proper interaction with their surrounding environment. Thus, understanding how perception self-organises is essential not only to deepen our understanding about the natural world, but also because it facilitates the synthesis of robust robotic systems. Although the function of perception is well understood, its self-organisation is not. That is, we understand what the perceptual system does, but not so much how it builds up from the interaction of distributed simple processing units. This is more striking if the interaction between perception, action selection, and memorisation is also taken into account in perception models. Uncovering these phenomena is key to enable, for instance, the synthesis of developmental mechanisms capable of allowing robotic systems to learn from their interaction with the environment. This is a daunting task, which could profit from a new modelling paradigm.

The most elementary unit model of robot control systems is the artificial neuron. In nature, neurons are the lower level units building up cognition. Being the physical substrate of cognition, neural activity and topological organisation have been the focus of several models (refer to (Gerstner and Kistler, 2002) for a thorough review on the field). With these models it is possible to emulate how neurons affect each other and how they behave collectively. This granularity is, however, often too fine when it comes to handle highly complex spatio-temporal problems, where several brain areas are involved. An example is perception which, as previously mentioned, must interact deeply with action selection and memory processes.

Therefore, a new modelling paradigm would be useful, in particular for the synthesis of embodied cognition systems, such as autonomous robots. A possibility that has been gaining momentum in the last few years is to model cognition in terms of the behaviour exhibited by social animals, with particular emphasis on insects. The swarm intelligence paradigm (Bonabeau et al., 1999) is particularly interesting to model per-

ception. This is because insects are simple entities that need to move and interact so as to actively perceive their environment. Being simple, their abstraction is well suited as the atom of the self-organising cognitive system. The abstraction of their collective behaviour, in turn, is useful to model the interactions occurring between the atoms of the cognitive system. As will be shown, the active nature of the swarm elements can be easily exploited to model the active nature of embodied vision systems.

In previous work (Santana and Correia, 2010a), we have exploited this swarm-based paradigm to synthesise self-organising visual attention, operating in an intricate way with spatial memory and action selection. The result was an accurate and parsimonious perceptual system, performing on a by-need basis and capable of handling the speed-accuracy trade-off in a context-dependent way. However, the model was only validated on a simulated robot performing local navigation on a noise free planar environment.

To assess whether the model is able to exhibit the same interesting properties on a demanding embodied setup, this paper adapts the model to handle its realisation on an all-terrain physical robot equipped with stereoscopic vision. These adaptations are mostly related to the fact that the model must cope with a robot moving on a non-planar terrain, whose sensory input is a noisy three-dimensional (3-D) point cloud. These requirements demand the model to handle the projective nature of the vision sensor and to use evidence accumulation for noise sensitivity reduction. Furthermore, this work deepens the model's biological supporting arguments.

This paper is organised as follows. In Section 4.2, the rationale behind modelling visual attention under the swarm cognition framework is presented. Then, the problem definition is specified in Section 4.3. The proposed model is finally presented in Section 4.4. Next, the results of the validation of the method by a set of experiments in a physical robot is presented in Section 4.5. A discussion on the model's biological plausibility is given in Section 4.6. Finally, a set of conclusions and future work pointers are highlighted in Section 4.7.

4.2 Introducing Swarm Cognition for Attention Modelling

This section describes how this paper builds upon previous work on perception in both natural and applied sciences, as well as how it departs from it.

4.2.1 Active Vision

The act of deploying attention, by means of eyes, head or body motion, is known to be an ubiquitous feature of animals endowed with rich vision (Land, 1999). The ability to shift the focus of attention highlights the active nature of perception, rather than a passive one. This means that perception can only be properly studied when the agent is freely acting on its environment, as proposed by Gibson's ecological (ethological) approach to visual perception (Gibson, 1979). In fact, body, nervous system and environment must be seen in a holistic way (Ashby, 1952; Beer, 1995; Thelen and Smith, 1996).

In the face of the limitations of passive set-ups (Fermuller and Aloimonos, 1995), active vision (Bajcsy, 1988; Aloimonos et al., 1988; Ballard, 1991) has also emerged in the realm of artificial vision systems. Among the advantages of an active vision perspective there is the possibility for the agent to: (1) act in order to shape its sensory information flow (Sporns and Lungarella, 2006) so as to make its processing tractable; (2) increase the sensory input signal-to-noise ratio; and (3) reduce ambiguities and consequently perceptual aliasing. These properties are a product of sensorimotor coordination, a concept that can be traced back to Dewey (1896), related to the enactive approach (Varela et al., 1991), and widely recognised as central to the sustainability of embodied adaptive behaviour (Brooks, 1991; Ballard et al., 1997; Pfeifer and Scheier, 1999; O'Regan and Noe, 2001; Pfeifer and Bongard, 2006; Mossio and Taraborelli, 2008).

Sensorimotor coordination requires a tight coupling between the agent and the environment. Given that the involved timescales of this coupling require fast perceptual processing, the classical view that a global isomorphic representation of the environment is continuously maintained by the agent (Marr, 1982) lacks support. Interacting perceptual pathways, dedicated to different purposes, are more likely to exist (Milner and Goodale, 1995; Goodale, 2008). For instance, pathways associated to motor action control, supposedly the dorsal visual stream, are prone to map directly to motor actions. This is congruent with the idea of the "world as its own best model" championed by Brooks (1991). Being perception so intricately linked to action, it is reasonable to assume that both unfold in parallel.

4.2.2 Parallel Covert Attention

Focussing perception on the most relevant aspects/regions of the environment, taking into account the current context, is advantageous to increase both performance and robustness (Itti and Koch, 2001; Hayhoe and Ballard, 2005; Rothkopf et al., 2007). Furthermore, the intricate relationship between visual attention and other structures, such as bottom-up sensory pathways, categorical reasoning, and action selection, highlights its centrality and consequently its relevance to the understanding of the dynamics involved in an embodied cognitive system. Attention ultimately results in the motion of sense organs, e.g., eyes, towards the relevant stimulus source. This is called overt attention. A faster process is the one of mentally focussing on particular aspects of the sensory stimuli. This is called covert attention and it will be the main focus of the following discussion.

Computational models of visual attention typically assume the existence of a sensor-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 drives attention to salient regions of the visual input. The more a region of the visual input detaches from the background, the higher its level of saliency. Top-down contextual knowledge on the visual search task, i.e., on the object being sought, is also known to play an important role on the modulation of attention (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 has been shown by recent neurophysiological studies (Egner et al., 2008). The outcome of these two processes' interplay is a saliency map whose activity level is higher in the regions of the visual input that are more likely to contain the object being sought. The saliency level is typically assumed to guide the motion of a "spotlight" from the most to the least relevant regions of the visual input. In the case of overt attention, this spotlight guides the motion of the sense organs to centre the visual input on the region of interest. In the covert case, a task-specific object recogniser is applied at each region where the spotlight falls, e.g., (Navalpakkam and Itti, 2005). This approach highlights a dichotomy between the parallel nature of saliency map computation and the sequential nature of the processes operating on its basis. Due to the involvement of mechanical components, this sequencing is unavoidable in overt attention. Conversely, although constrained to pro-

cess sensory input that is sequentially determined by the overt attention, the covert attention process is more likely to operate in a parallel fashion. Two main arguments can be put forth to support this hypothesis.

First, the human brain is a massively parallel structure, which renders unlikely that it would be instantaneously dedicated to a single covert attentional sequential process. That is, the effort of maintaining a unique purely sequential analysis is costly by itself, as it requires complex arbitration mechanisms. Second, studies with human subjects revealed the existence of a unitary spotlight of covert attention to be unlikely (Pylyshyn and Storm, 1988; Doran et al., 2009). The tests involved having subjects paying covert attention to objects that were moving so fast that the chances of effective attention switching were considerably low. The observed successful tracking of these objects supports the multiple spotlights hypothesis, which is also coherent with evidence on the existence of independent tracking mechanisms in the two cerebral hemispheres (Alvarez and Cavanagh, 2005).

There is a considerable bulk of knowledge suggesting that both overt and covert shifts of attention share the same neural mechanisms (de Haan et al., 2008), supporting the pre-motor theory of attention (Rizzolatti et al., 1987). Impressively, this link between response preparation and spatial attention is not limited to the oculomotor system (Eimer et al., 2005). Thus, this theory predicts that common sensorimotor coordination mechanisms are involved in the control of attention and action. We argue that this opens the door for the application of sensorimotor coordination principles, studied mostly on overt control, to better understand and explore the dynamics of covert attention.

4.2.3 Agent Abstraction of Covert Attention

In a more classical view, the control of the attention spotlight would be considered exogenous, in the sense that it is only function of sensory saliency and top-down modulation signals. However, under the assumption that covert attention is a process of sensorimotor coordination, it is more reasonable to consider the spotlight as a dynamical entity inhabiting the sensorimotor space of the embodied agent. This entity can then pro-actively move in this space to guide the focus of the embodied agent in a sensorimotor coordinated way, and thus exploiting all the aforementioned advantages of sensorimotor coordination to shape the sensory input. This entity, from now on called

an agent, thus behaves as a locally sequential covert attentional process.

The feasibility of agent-based modelling for sensorimotor coordination has been extensively validated in the fields of embodied cognition and active vision (Scheier et al., 1998; Nolfi and Marocco, 2002; Beer, 2003; Fend et al., 2003; Balkenius et al., 2004; Floreano et al., 2004; Nolfi, 2005; Suzuki and Floreano, 2006; Pfeifer and Bongard, 2006; Kim and Moeller, 2006; Sporns and Lungarella, 2006; de Croon and Postma, 2007; Choe et al., 2008). A population of agents can generate coherent collective (parallel) behaviour. In perception, the agents' motion corresponds to an attention shift, whereas the collective spatio-temporal self-organised pattern implements a multi-focus attention process. Brain computational modelling with multiple agents is not a new idea (Minsky, 1988; Chialvo and Millonas, 1995). Although the first realisations for computer vision related problems are also not new (Poli and Valli, 1993; Liu et al., 1997), only more recently it has received considerable attention (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). In general, these models exploit the metaphor of swarming behaviour on social insects.

These computational models are mostly stand-alone engineered parallel perceptual systems, lacking a body capable of purposive behaviour. This deficit undermines their explanation power regarding the mechanisms actually building up adaptive behaviour. Conversely, sensor planning, which is a relatively stable field in computer vision and robotics communities, is actually trying to bridge the gap between body motions and information gathering through the sensors (Dickmanns et al., 1990; Sukthankar et al., 1993; Kelly and Stentz, 1998; Behringer and Muller, 1998; Nabbe and Hebert, 2003; Kwok and Fox, 2004; Patel et al., 2005; Bagdanov et al., 2006; Urmson et al., 2006; Tessier et al., 2007; Hernandez et al., 2007; Sprague et al., 2007). However, none of these works models parallel covert attention operating in an intricate way with the action selection process.

4.2.4 Swarm Cognition for Attention Modelling

The previous sections showed the relevance of a parallel model for covert attention, as well as the importance of its tight coupling with motor action preparation. It was also suggested that the multi-agent paradigm is appropriate for this modelling task.

A remarkable metaphor from the natural world encompassing these characteristics,

and consequently exploited in this paper, is the foraging behaviour of army ants. By exploiting pheromone-based local interactions, these ants are able to forage in large areas around their nest in a parallel and robust way (Deneubourg et al., 1989). These foraging ants exhibit a sort of collective intelligence (Franks, 1989), allowing the group to be seen as an individual decision making process.

Rather than static structures, like neurons, these agents are better viewed as active information particles that flow through the system. Hence, using agents, the design focus is on the process and not so much on its supporting substrate. Being sensorimotor coordinated units, these information particles can actively shape their sensory input, use their sensorimotor history to induce long-range influences on other information particles, and in the limit improve their own behaviour. When together, these modular units can exploit the synergy of self-organisation and emergent properties. We argue that reaching the complexity of such a system with a connectionist model is possible but not likely tractable. Although not covered in this article, we foresee the usefulness of connectionist models to implement the agents themselves, as it would enable tractable neuro-evolution (Floreano et al., 2008) to build complex systems. Hence, we suggest that using an agent-based design the system modeller is able to reach higher levels of tractable complexity than when using connectionist models. Although the results reported in this paper show the benefits of an agent-based design, further phenomenological support is still required for a definitive comparison with connectionist models.

The use of the insect swarm metaphor to model vertebrate brain function has been also suggested by parallel and independent work (Passino et al., 2008; Couzin, 2009; Marshall et al., 2009; Marshall and Franks, 2009). Nest site selection in honey bees, for instance, exhibits some of the characteristics of decision making in brain, such as the statistical speed-accuracy trade-off predicted by diffusion models (Ratcliff and Smith, 2004). Our work, instead, approaches the problem of studying cognition through social insects behaviour by building it on an embodied setup, following the synthetic approach to embodied cognition (Pfeifer and Scheier, 1999) and Artificial Life (Bedau, 2003). The advantages of using Artificial Life models for this purpose, though without the support of any practical realisation, were reviewed also in a parallel study (Trianni and Tuci, 2010).

All these accounts can be framed in the emerging multidisciplinary field of Swarm Cognition, which attempts to uncover the basic principles of cognition exploiting self-

organising principles, mainly those exhibited by social insects. However, as we only conceive cognition under the embodiment framework, we consider that a swarm-based model not involved in the embodied agent's sensorimotor coordination loop can hardly be considered as an instance of Swarm Cognition. Note that all these considerations are under the assumption that Swarm Cognition can model and help to understand the behaviour of a multi-cellular individual.

4.3 Problem Definition

This section specifies the objective, hypotheses, and assumptions related to the proposed model. The problem to solve is primarily the one of minimising the perceptual resources required to support the proper execution of the action selection process controlling the embodied agent, hereafter simply called robot. Saving time in perception enables faster robot motion. The ultimate goal is to show that this complex cognitive problem can be easily managed by recurring to the swarm metaphor.

A perceptual resource is said to be expended when a given object detector is applied to a given region of the robot's visual input. Minimising perceptual resources expenditure is achieved by reducing the number of detector applications to the point from which the action selection process would no longer be sufficiently informed for a proper decision making.

Supported by the previous section, the following describes the three main hypothesis we tested in this work. The first hypothesis is that the considered problem can be solved by allowing both perceptual and action selection processes to evolve in synergy, as it should promote the application of the detector, i.e. focus attention, on pixels whose positive detection most impacts the action selection process, and consequently helps it in rapidly stabilising its output. If the robot's goal is to move forwards, detecting an obstacle in its frontal region will compel a sudden change from forward motion to a turn. Conversely, an obstacle detected on the robot's side would effect no change in action selection. Hence, finding first the frontal obstacle has the highest impact on the action selection process and consequently helps the robot reaching the correct decision faster. The second hypothesis is that to ensure that the attentional process is robust, it should be parallel and provided with self-organising properties. Given the unpredictability of the environment, the complexity of taking into account the interaction with the action selection process, and the difficulty of defining a fixed speed-accuracy

trade-off, turns the alternative of using an optimal covert visual attention policy into an infeasible solution. Finally, the third hypothesis is that an adequate formalism to model this parallel covert visual attention process is the one describing the collective behaviour exhibited by social insects.

In previous work (Santana and Correia, 2010a), we have already tested and confirmed these three hypotheses in the context of simulated robots performing local navigation in two-dimensional (2-D) noise free environments. This article tests the three hypotheses on a more generalised setup, namely, on a physical robot performing local navigation in 3-D noisy environments. In turn, the added complexity of such an experimental apparatus demands some adaptations to our previous work (Santana and Correia, 2010a). Concretely, the model has been extended with an evidence accumulation mechanism (see Section 4.4.3.3) to better handle the noisy nature of the sensory input. This is a fundamental mechanism since, without it, noise impinging the robot's sensors would propagate through the system, hampering the maintenance of a coherent cognitive representation. We also needed to handle the projective nature of the employed vision sensor in a rough terrain.

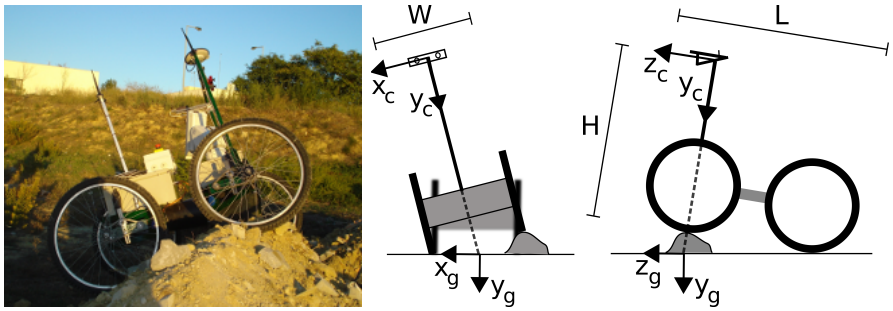
Without loss of generalisation, the robot used to validate the hypotheses is a wheeled robot equipped with a stereoscopic vision sensor that must be able to perform local navigation in rough terrain. Hence, the object detector in this case is a 3-D obstacle detector, whereas the action selection process is a fast obstacle avoidance algorithm. The following sections detail these assumptions.

4.3.1 The Ares Robot

The robot employed in this study, i.e., the Ares robot (Santana et al., 2008a), is a vehicle with four independently steered wheels (see Fig. 4.1). Although it enables the use of several locomotion modes, here only the Double Ackerman mode is considered. In this mode, the robot moves in a car-like way, but with both front and rear wheels steering symmetrically. This gives the robot a considerable manoeuvrability.

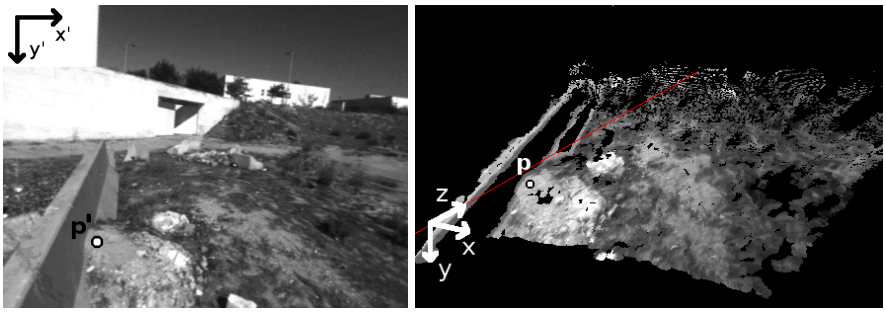
Off-road terrains are uneven and densely populated with protuberances, such as small rocks (see Fig. 4.1). In these situations the robot exhibits considerable tilt levels, which makes the relative position and pose of the sensor with respect to the ground plane to vary as the robot moves.

The robot is equipped with two cameras calibrated to perform stereoscopy. This



(a) Physical robot (b) Front view (c) Side view

Figure 4.1: The Ares robot. In the depicted situation the Ares robot is stepping an obstacle with its frontal left wheel, which results in non-zero pitch and roll angles, with respect to the ground plane.



(a) Left input image (b) 3-D reconstruction

Figure 4.2: 3-D reconstruction process. (a) Left input image obtained with stereoscopic vision, L. (b) Reconstructed 3-D model limited to a depth of 10 m. The 3-D point **p** is projected onto L in pixel $\mathbf{p}'$. The red line corresponds to the optical axis of the left camera. Note the absence of computed depth in many pixels, in particular along the homogeneous vertical structure on the left and frontal walls.

means that each pixel from each camera is associated to a 3-D coordinate in the camera's frame of reference, provided that stereoscopy could be computed for the pixels in question (see below). The left input image, L, is taken as reference and hereafter simply called visual input or frame.

Given that stereoscopy can be computed for the pixel $\mathbf{p}'$ in L, this pixel can be back-projected so as to obtain the corresponding 3-D point **p**. With a projection matrix P obtained from calibration, it is possible to do the inverse operation, i.e., to determine which pixel $\mathbf{p}'$ in L is the projection of the environment's 3-D point **p** (see Fig. 4.2).

The distance between both cameras is 30 cm. This enables sufficiently accurate depth computation between 2.5 m and 20 m. This means that obstacles close to the

robot are not perceivable, making mapping an essential asset for this configuration. Two additional aspects render stereoscopic vision a challenging sensory modality. First, absence of 3-D information in low texture and under illuminated surfaces is rather common (see Fig. 4.2). This owes mostly to the difficulty of stereo processing to match both left and right camera images in those regions. Second, high levels of noise are common with poor lighting condition and depth computation error grows quadratically with range. Hence, in order to operate with a stereoscopic vision setup, a perceptual system must be highly robust.

4.3.2 Mapping Between World and Sensor Coordinate Frames

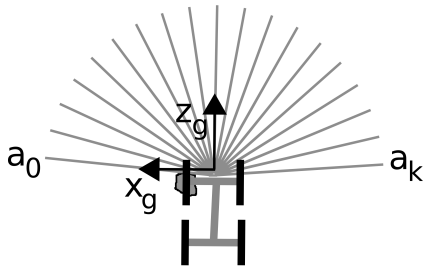
In this study, robot actions are considered to be linear trajectories radially overlaid on the ground in front of the robot, i.e., parallel to plane $z_g x_g$ and with $y_g = 0$ (see Fig. 4.1). To allow obstacle search to be affected by action selection, these linear trajectories must be projected onto L.

This projection requires the estimation of transformation Q, between the ground's frame of reference, $\{x_g, y_g, z_g\}$, and the camera's frame of reference, $\{x_c, y_c, z_c\}$. Due to terrain's roughness, the 4×4 homogeneous transformation matrix Q is computed at each frame by recurring to a robust ground-plane fitting process (Santana et al., 2009). After transforming with Q the linear trajectories to the camera's frame of reference, these can finally be projected onto L with projection matrix P (see Fig. 4.3).

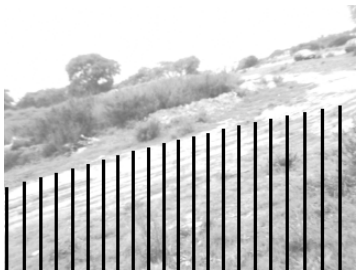
In order to reduce the chances that obstacles in the environment may be confounded with the terrain itself during the ground-plane estimation process, the pixels of the obstacles that have been detected in the previous frame are discarded from the process (van der Mark et al., 2007).

4.3.3 Obstacle Detection

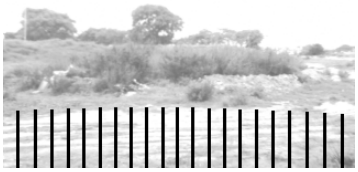
In this study, obstacles are detected in a pixel-wise way over the visual input L. Concretely, a pixel is said to be an obstacle point whenever the pixel's associated 3-D point is at a height from the estimated ground plane that cannot be climbed by the robot (Santana et al., 2009). Formally, the detection result at pixel $\mathbf{p'_a} = (x_a, y_a)$, with the corresponding 3-D point $\mathbf{p_a} = (X_a, Y_a, Z_a)$, is given by



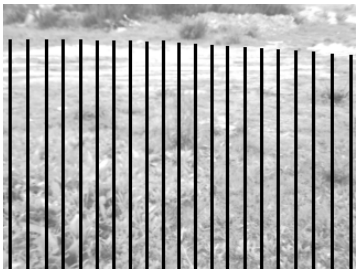
(a) Top view



(b) Projected image of (a)



(c) Projected image with positive pitch angle



(d) Projected image with negative pitch angle

Figure 4.3: Action set parallel to the ground-plane in (a), and respective projection on L in (b). In this situation, the robot is stepping an obstacle with its frontal left wheel (as depicted in Fig. 4.1). Note that the projected action set (black lines) is compensated for the robot's non-zero roll and pitch angles. The central line in all projected images correspond to the straight ahead motion. Note that even with considerably different pitch angles, as in (c) and (d), the projected lines terminate roughly in the same image regions. That is, they are clearly relative to the ground-plane coordinate system defined by $\{x_g, y_g, z_g\}$, and so invariant to the robot's pose.

$$D(\mathbf{p}'_\mathbf{a}) \quad = \quad \begin{cases} 1 & \text{if } d\left(\mathbf{p}_\mathbf{a}, (a,b,c,d)\right) > h_{min} + \beta Z_a \\ 0 & \text{otherwise} \end{cases} , \tag{4.1}$$

where, $d\left(\mathbf{p}_\mathbf{a}, (a,b,c,d)\right)$ is the orthogonal distance between point $\mathbf{p}_\mathbf{a}$ and the estimated ground plane $ax_c + by_c + cz_c + d = 0$, as computed in section 4.3.2. Note that $\mathbf{p}_\mathbf{a}$ is given relative to $\{x_c, y_c, z_c\}$. The term βZ_a makes the detection threshold grow with the range, Z_a, of the point in question. This mechanism compensates for the range-dependent error growth in the 3-D reconstruction process. Thus, it reduces the potential for false positives in the far field. The lack of 3-D information in some pixels results in the impossibility of applying the detector in these cases.

Note that the simplicity of this detection method comes at the cost of only being

applicable to moderately rough terrain. More complex detectors (see, e.g., Santana et al. (2010d)), which are required for more demanding environments, could be used in the model without any specific customisation. In fact, the more complex the detector, the more expensive is its computation, and consequently the more important is to focus the robot's attention.

4.3.3.1 Motion Estimation

A key component of any complex robotic system is motion estimation. With it, the system will be able for instance to produce local maps of the environment.

Frame-to-frame robot motion, i.e., translation and rotation, is computed with a particular model (Santana and Correia, 2008) of visual odometry (Matthies, 1989; Agrawal and Konolige, 2006). Visual odometry basically estimates the rigid body transformation matrix M that explains the change in position of notorious 3-D points in the environment across frames. The major advantage of this method over others, like dead-reckoning, is that it is fully synchronised with the sensor data feeding the perceptual process. This results in very accurate registration of information across frames.

4.4 Model Description

This section starts by mapping the biological metaphor of foraging strategies in social insects to the concepts of parallel covert visual attention to be employed in the proposed model. In short, the metaphor is exploited to provide the model with the self-organising properties essential to maintain a robust parallel focus of attention. After presenting the biological inspiration, the proposed model is described. Since it is built upon authors' previous work, it is not described with full detail. For additional information the reader is referred to the previous publication (Santana and Correia, 2010a).

4.4.1 Biological Inspiration

Covert visual attention is mostly about parallel search of objects in the robot's visual input. A remarkable metaphor for this process is the one of army ants engaging in foraging behaviour (see Section 4.2.4). Assuming that the environment where these ants

inhabit corresponds to the robot's visual input, each ant can be seen as an individual covert visual attention process. As a consequence, their collective behaviour can be considered as a parallel covert visual attention process.

Following this metaphor from the natural world, the visual process in our model is composed of a swarm of simple homogeneous virtual agents that inhabit the visual input of the robot. These agents are probabilistically created (recruited) and therefore they receive hereafter the designation of p-ants. They search (forage) the robot's visual input along those regions where detected obstacles are more likely to stronger affect the action selection process. Hence, p-ants operate on a by-need basis being driven by the action selection process. In the case of local navigation, the utility of perceiving a given region of the visual input is related to the utility of navigating on its corresponding region of the environment. In the proposed model, these regions are associated to the projected linear trajectories, as described in Section 4.3.2.

As in natural ants, p-ants do this search in a stochastic and coordinated way. By not behaving greedily, the system is more robust to unforeseen situations and faster in adapting to contextual changes. P-ants interact through *stigmergy*, i.e., they interact through perceptual shared mediums, for better coverage and tracking of detected obstacles. This allows the coexistence of positive and negative feedback loops that lead to robust collective behaviour. In conclusion, random fluctuations and both positive and negative feedback, which are necessary ingredients for self-organisation to occur (Bonabeau et al., 1999), are included in the model.

4.4.2 Proposed Model

The proposed model is decomposed into two interacting processes, one for perception and another for action selection. The perceptual process includes the parallel covert visual attention aspects. Fig. 4.4 illustrates the connectivity between both processes. Basically, after receiving a new frame, L, both perceptual and action selection processes interact (thicker arrows) for i_{max} iterations before a final motor action decision is reached and eventually engaged. These interactions occurring between both processes allow them to progressively unfold in parallel, and consequently, to enable accurate deployment of visual attention.

At each iteration, the action selection process sends a message to the perceptual process with an action utility vector, $\mathbf{u} = (u^1, u^2, \ldots, u^k)$, where $u^j \in [0, 1]$ is the utility of

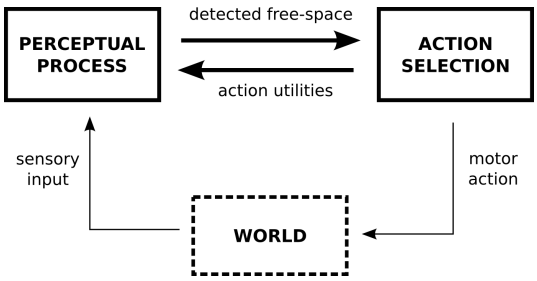


Figure 4.4: Building blocks of the proposed model, adapted from (Santana and Correia, 2010a).

performing action j, and k is the cardinality of the robot's action repertoire. The action utility vector is computed according to a desired heading of motion, h, and constrained by information about free-space connectivity of the local environment, C, which has been sent by the perceptual process in the previous iteration. This set encompasses range information regarding those radial sectors of the local environment that contain free-space for robot motion. In this study, actions u^j are defined as linear trajectories centred on the robot and directed outwards in a radial pattern (see Section 4.3.2). Hence, the utility of each possible action is defined as the value of moving the robot along the corresponding linear trajectory given a goal location, $\mathbf{p_{goal}}$, and obstacles disposition in the environment. The more directed a given linear trajectory is to the goal location, the higher its utility. The closer an obstacle is to the robot along a given linear trajectory, the lower the trajectory's utility. The final utility is a combination of both factors.

At each iteration, the perceptual process uses the just received vector $\mathbf{u}$ to iterate its search for obstacles on a by-need basis (see Section 4.4.3). The positions of obstacles detected since the arrival of the current frame are accumulated in set O. This set is then used to compute C. Finally, C is sent back as a message to the action selection process so as to induce its next move. The action with highest utility at the time the maximum number of iterations is reached is passed to the low-level motion controller through a low-pass filter. This way, sudden changes at the system's output are smoothed to avoid jitter at the actuators level.

4.4.3 Perceptual Process

The perceptual process is composed of a set of p-ants, A, whose elements individually implement a covert attention process. The size of this set is zero at the first frame and

varies along iterations.

At each iteration, the perceptual process starts by appending new p-ants to the set A (see Section 4.4.3.1). Then, each p-ant moves, according to its set of behaviours (see Section 4.4.3.2), updating its 2-D position in the visual input. Possibly, some of the p-ants will find an obstacle to track across iterations and even frames. In between frames, the position of every p-ant is updated to compensate for any robot motion that has occurred (see Section 4.4.3.4). This way, its position is kept relative to the environment and consequently invariant to robot motion. This update is done by applying to p-ant's 3-D position, obtained with stereoscopy from its 2-D position in the visual input, a transformation matrix representing the robot motion (see Section 4.3.3.1). The transformed 3-D point is then projected back onto the visual input with a projection matrix. In the process, due to robot motion, this projection may fall out of the visual input, which means that the p-ant has been moved out of the robot's visual field of view. If at that moment the p-ant was tracking an obstacle it is now said to constitute part of a body-centric local map. To maintain the local map updated (i.e., body centred), p-ants in that situation are also motion compensated in between frames. If a given part of the environment is revisited, the projected positions of some of these p-ants will be again within the visual field of view. These p-ants are then reactivated in order to track their associated obstacles in the visual input. At the end of each iteration, the 3-D positions of all p-ants that are tracking an obstacle, either in or out of the robot's visual field of view, may be used to extend the obstacles set O. Evidence accumulation (see Section 4.4.3.3) on the presence of an obstacle is used by each p-ant to determine whether or not to append its position to set O.

For this process to operate, every p-ant $a \in A$ needs to maintain state information across iterations and frames. This information includes p-ant's body-centred 3-D position, $\mathbf{p_a}$, and its 2-D projection onto a plane parallel to and centred on visual image L, $\mathbf{p'_a}$. With projection matrix P, $\mathbf{p'_a}$ can be computed given $\mathbf{p_a}$. Hence, $\mathbf{p'_a}$ is a 2-D vector defined in the reference frame of L. With stereoscopy, $\mathbf{p_a}$ can be computed given $\mathbf{p'_a}$, provided that $\mathbf{p'_a} \in FOV$, where FOV is the set of possible positions in L. If $\mathbf{p'_a} \notin FOV$ or stereoscopy failed to back-project $\mathbf{p'_a}$, then $\mathbf{p_a}$ can only be estimated, such as through compensating its previous state with the transformation matrix representing the robot motion estimate, M. Since with P it is possible to obtain $\mathbf{p'_a}$ from $\mathbf{p_a}$, one can then assess when the p-ant re-enters the visual input, i.e., when $\mathbf{p'_a} \in FOV$.

The following sections details how the perceptual process works in each iteration.

4.4.3.1 P-ants Creation and Removal

With the exception of the first iteration in each frame, the set of p-ants, A, is extended by creating new p-ants as a function of the incoming action utility vector, $\mathbf{u}$. This set is empty in the first frame. P-ants are not created in the first iteration of each frame because, at that time, vector $\mathbf{u}$ is not yet updated (set C not yet computed).

The higher the robot's speed, $s \in [0,1]$, and the utility of a linear trajectory, j, the higher the chances of creating a corresponding p-ant, a_j, in L. The initial 2-D position of this p-ant, $\mathbf{p}'_\mathbf{a}$, is the point where the linear trajectory j projected onto L intersects the bottom row of L (see Section 4.3.2). This way p-ants start their search for obstacles in the close vicinity of the robot.

The newly created p-ant a_j is endowed with an initial energy level, ρ, which is reduced by one unit in each iteration and restored whenever the p-ant considers to be on an obstacle. With zero energy, the p-ant is removed from the system, $A \leftarrow A \backslash \{a_j\}$, to avoid memory and computation to grow unbounded.

4.4.3.2 P-ants Behaviours

Fig. 4.5 illustrates the finite state machine responsible for the switching of the behaviours ruling each p-ant's activity. The following describes these behaviours as well as their transitions.

After appending new p-ants to A, every p-ant $a \in A$ looks for obstacles at its position $\mathbf{p}'_\mathbf{a}$. To reduce computation, it first looks at a perceptual shared medium S, which represents obstacles detected previously in the current frame by any p-ant. If $\mathbf{p}'_\mathbf{a}$ is not represented in S and there is 3-D information associated to that point, the obstacle detector $D(\mathbf{p}'_\mathbf{a})$ (see Equation 4.1) is called and the result is stored in S. In the absence of 3-D information the point is considered without obstacle.

At creation time, every p-ant a starts in *search behaviour*. As depicted in Fig. 4.6(a), each iteration of this behaviour performs a simple stochastic motion step on L along the associated projected linear trajectory:

$$\mathbf{p}'_\mathbf{a} \leftarrow \mathbf{p}'_\mathbf{a} + \mathbf{v}'_\mathbf{a}, \tag{4.2}$$

where $\mathbf{v}'_\mathbf{a}$ is a velocity vector with angular direction $\lambda_1 \cdot N(0,1)$ and magnitude $\lambda_2 \cdot N(0,1)$, with $N(0,1)$ sampling a number from a Gaussian distribution with mean 0 and

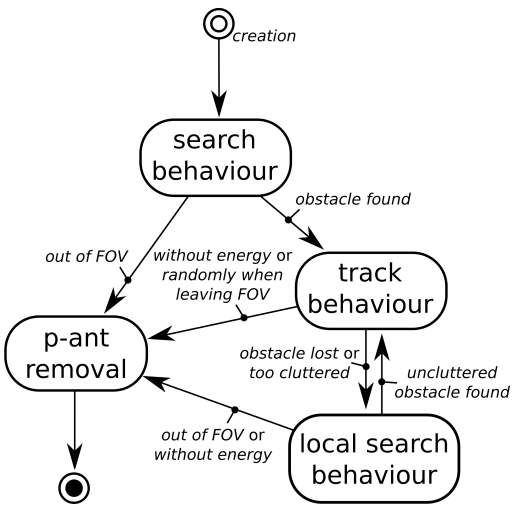


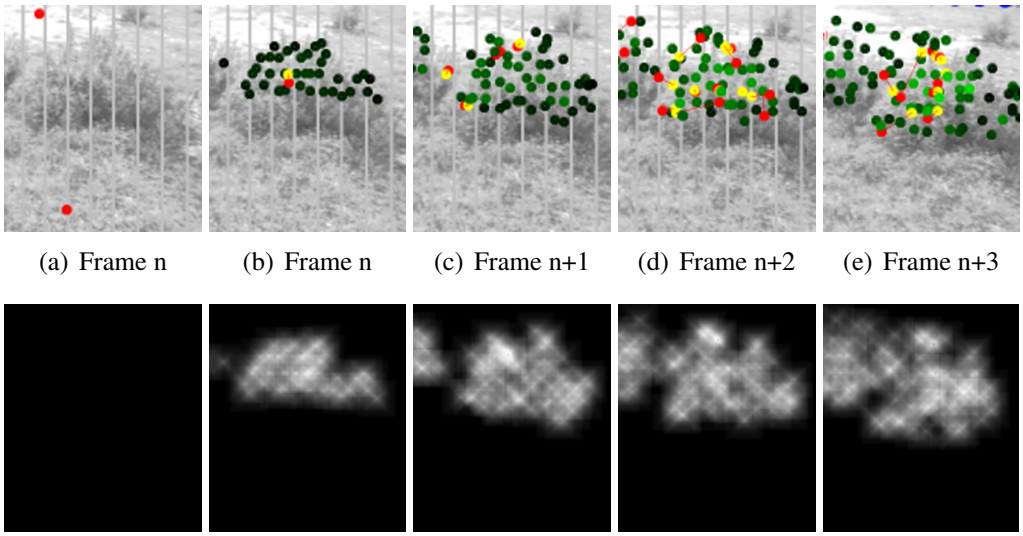
Figure 4.5: Finite state machine each p-ant's activity. Ovals and links represent behavioural states and their transitions, respectively. The labels associated to each transition specify the conditions for the transitions to occur.

variance 1, and λ_1 and λ_2 are empirically defined scalars. High λ_1 and λ_2 values facilitate fast detection of large obstacles, at the cost of missing smaller ones. Small values result in slower, though finer, detection. If while in this behaviour the p-ant moves out of the visual input boundaries, $\mathbf{p}'_{\mathbf{a}} \notin FOV$, the p-ant is removed from the system, i.e., $A \leftarrow A \backslash \{a\}$. This may happen due to p-ant's motion or due to the influence of the robot motion compensation process.

If an obstacle is found by a p-ant while in *search behaviour*, it switches to *track behaviour*. As illustrated in Fig. 4.6(b), in this new behaviour, p-ants are allowed to clone themselves for r_{max} times. A single cloning is attempted per iteration by randomly selecting a position in the neighbourhood (radius between 10 and 30 pixels) of the p-ant's current position. If the randomly selected position contains an obstacle and it is not too cluttered with other p-ants, then a clone is created and initiated on it. A clone inherits the number of replications of its ancestor so as to control the diffusion process, whose goal is to rapidly cover the detected obstacle.

A p-ant assesses the level of cluttering by inspecting the level of activity at its position on a perceptual shared medium, R. This 8-bit structure has the size of L. Therefore, p-ant's position there corresponds to p-ant's position on L. The activity level is a clutter measure because all p-ants in *track behaviour* increase the local activity of the perceptual shared medium R at the corresponding locations. Hence, R represents pheromone

directly emitted by p-ants, and not a chemical they lay in the environment. For a p-ant a, this is done by adding to R a top-view pyramidal shape of top magnitude 20 and linear decay (0.9) outwards, centred on $\mathbf{p'_a}$. If the activity at $\mathbf{p'_a}$ in R is higher than an empirically defined scalar, $R(\mathbf{p_a}) > \eta$, it is said that the p-ant's position is too cluttered. Fig. 4.6(f-j) traces the evolution of shared medium R for a given situation.



(a) Frame n (b) Frame n (c) Frame n+1 (d) Frame n+2 (e) Frame n+3

(f) Frame n (g) Frame n (h) Frame n+1 (i) Frame n+2 (j) Frame n+3

Figure 4.6: Detection, diffusion and evidence accumulation example. The images correspond to a small region of the current frame. Red dots are p-ants in *search behaviour*. Green p-ants refer to those that are in *track behaviour*. The lighter the green the greater the evidence on the presence of the obstacle. The vertical lines correspond to a sub-set of projected linear trajectories with positive utility. Note that only at frame $n+3$ these lines disappear, showing the progressive accumulation of evidence on the presence of the obstacle before reporting it, i.e., appending it to O. It is also possible to depict the diffusion process in b), triggered by a single p-ant (red dot off the obstacle), where p-ants are cloned to better cover the obstacle. Red p-ants on the obstacle are in *local search behaviour* around a given anchor point, represented by the connected yellow dots. This occurs when the region is already too cluttered with other p-ants. The activity level in the perceptual shared medium, R,which is used by p-ants to assess the level of clutter and to help each other on finding the obstacle when in local search, is depicted in the bottom row. Videos with situations as the one depicted in this figure are available as online supplementary material.

Besides cloning, p-ants in *track behaviour* have no additional activity. Nevertheless, robot motion compensation (see Section 4.4.3.4) will maintain their positions relative to the associated obstacles, independently of robot motion. As the robot moves, these p-ants will eventually leave the robot's visual input and contribute to the robot's spatial

working memory with probability ψ. That is, not all p-ants leaving the visual input become part of the spatial memory. This helps on the regulation of the computational load by reducing the sampling of the peripheral environment.

Being inaccurate, the estimated robot motion is usually insufficient to completely cancel the effects of robot motion on the relative position of p-ants. The outcome is that p-ants eventually loose track of the obstacles. To cope with this, any p-ant in *track behaviour* that considers to have sufficiently accumulated evidence of being no longer on an obstacle switches to *local search behaviour*. As will be shown in Section 4.4.3.3, evidence grows and decays as the obstacle detector returns positive and negative, respectively. As a result, if evidence in obstacle presence is high, transient obstacle detection negatives may not be sufficient to engage behaviour switching. This is key to deal with noisy sensory data.

The switch to *local search behaviour* also occurs if the p-ant in position $\mathbf{p}'_\mathbf{a}$ determines that it is on an obstacle too cluttered with other p-ants, $\mathrm{R}(\mathbf{p}'_\mathbf{a}) > \zeta$. By making $\eta < \zeta$, a sort of hysteresis is implemented, and thus massive fluctuations of neighbouring p-ants entering and leaving *local search behaviour* are avoided.

As depicted in Fig. 4.6(c-e), when in *local search behaviour*, a given p-ant a performs a random walk around an anchor point, $\mathbf{z}'_\mathbf{a}$, whose initial position is defined as the p-ant's position at the time *local search behaviour* was initiated, $\mathbf{z}'_\mathbf{a} \leftarrow \mathbf{p}'_\mathbf{a}$. Hence, with the goal of re-detecting a lost obstacle or of finding a less cluttered region, $\mathrm{R}(\mathbf{p}'_\mathbf{a}) < \eta$, this behaviour randomly changes the position of the p-ant around the anchor point as follows:

$$\mathbf{p}'_\mathbf{a} \leftarrow \mathbf{z}'_\mathbf{a} + \sigma_a [N(0,1), N(0,1)]^T, \tag{4.3}$$

where $\sigma_a \leftarrow \min(\sigma_a + 1, l)$ and $N(0,1)$ samples a number from a Gaussian distribution with mean 0 and variance 1. By increasing σ_a at each iteration, the local search spreads up to the upper-bound l, which constrains the search to avoid migration of p-ants between obstacles.

The anchor point changes to the p-ant's current position, $\mathbf{z}'_\mathbf{a} \leftarrow \mathbf{p}'_\mathbf{a}$, whenever the clutter level there is higher than the clutter level at the current anchor's location, yet not too high, $\eta > \mathrm{R}(\mathbf{p}'_\mathbf{a}) > \mathrm{R}(\mathbf{z}'_\mathbf{a})$. This directs the local search towards uncluttered regions where other p-ants reported the existence of an obstacle. If it is not possible to move the anchor point, p-ant's energy level is dramatically reduced by an amount υ. Since p-ants

may be in *local search behaviour* across frames, the anchor points must be compensated for the robot's motion the same way p-ants positions are. If a p-ant in *local search behaviour* happens to detect a not too cluttered obstacle, $\mathrm{R}(\mathbf{p}'_\mathbf{a}) < \eta$, it switches to *track behaviour*. Finally, p-ants in *local search behaviour* that leave the robot's visual input are removed from the system. This may happen due to p-ant's motion or due to the influence of the robot motion compensation process.

4.4.3.3 Evidence Accumulation

In the previous section nothing was said related to the update of the obstacles set O. This is done by having every p-ant a in *track behaviour*, whose *evidence* on the presence of an obstacle on its position is sufficiently high, appending its 3-D position, $\mathbf{p}_\mathbf{a}$, to O. This accumulation of evidence before reporting an obstacle is important to reduce sensitivity to noise in the 3-D point cloud. Otherwise, noise in the sensory input would propagate to the action selection process. The following describes how evidence is integrated by each p-ant.

At the first iteration of each frame, every p-ant a that (1) is in *track behaviour*, (2) is within visual input boundaries, $\mathbf{p}'_\mathbf{a} \in FOV$, and (3) its position has associated 3-D information, updates its evidence on obstacle presence, y_a, according to the following equation:

$$\tau \dot{y}_a = -y_a + \gamma D(\mathbf{p}'_\mathbf{a})\mathrm{R}(\mathbf{p}'_\mathbf{a}), \tag{4.4}$$

where τ and γ are empirically defined scalars, $\mathrm{R}(\mathbf{p}'_\mathbf{a}) \in [0,1]$ is the activity level of the perceptual shared medium R at p-ant's position, $\mathbf{p}'_\mathbf{a}$, and $D(\mathbf{p}'_\mathbf{a})$ returns the result of the pixel-wise obstacle detector (see above).

With this system, evidence decays to zero when an obstacle is not found for several frames, provided that 3-D information is available. This condition ensures that detected obstacles are not forgotten just because stereoscopy is no longer able to compute the obstacle's 3-D position. This allows for instance tracking an obstacle even when it is too close to the robot to be sensed by both left and right cameras.

If the obstacle is found for several frames, then the system converges towards the fixed point $y|_{\dot{y}_a \to 0} = \gamma \mathrm{R}(\mathbf{p}'_\mathbf{a})$. This means that evidence grows faster when the level of activity in the perceptual shared medium R is higher, i.e., when other p-ants also detect the presence of the obstacle. Finally, if $y_a > \alpha$, where α is a confidence threshold, then

the p-ant is confident enough on the existence of an obstacle in its position, and so it can be reported to the action selection process, i.e., appended to O.

To append an obstacle to O, it is necessary to project the corresponding 3-D point, which is in the camera frame of reference, $\{x_c, y_c, z_c\}$, to the ground-plane frame of reference, $\{x_g, y_g, z_g\}$ (see Section 4.3.2). This way the obstacle is represented in the ground coordinate system, and so it affects the action selection process in a robot's pose invariant way.

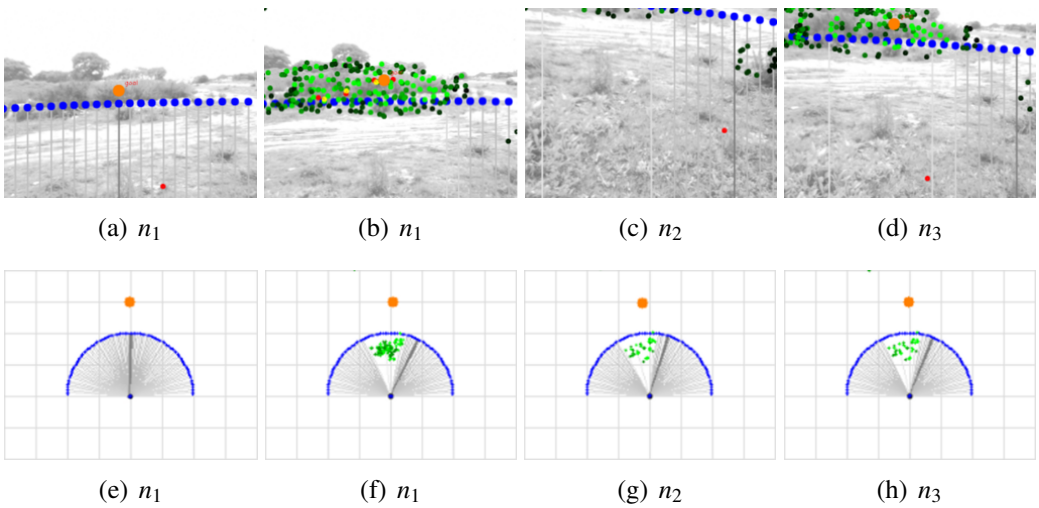
Fig. 4.6 illustrates a situation where: (a) p-ants are guided by the action selection process; (b) they find an obstacle and a diffusion process is initiated; (c), (d) evidence is accumulated for some frames; (e) and finally the obstacle is reported to the action selection process.

A probabilistic framework could also be considered for evidence accumulation, provided that a better noise description would be available. Nevertheless, the presented dynamical system implements a leaky-integrator and consequently a sound mechanism in terms of neural support. Moreover, allowing the collective to influence the individual decision is also sound as it implements a sort of lateral influence between neural structures. In fact, the use of stigmergy allows this to happen without the explicit modelling of a large set of connections.

4.4.3.4 Motion Compensation

Motion compensation is performed in the first iteration of each frame by first transforming the 3-D position of every p-ant according to the robot motion estimate matrix M (see Section 4.3.3.1). Then, this transformed 3-D position is projected onto visual input L according to projection matrix P. The result is the p-ant's motion compensated position in L. See Fig. 4.7 and Fig. 4.8 for an illustration of the motion compensation used to update the spatial memory.

Allowing p-ants to be updated when out of the visual field of view to maintain spatial memories is important so as to avoid that the action selection process operates with a myopic view of the environment. Moreover, it allows the system to reduce perceptual requirements when revisiting a given environment (refer to our previous work on simulation (Santana and Correia, 2010a) for experimental results showing this benefit). The fact that not all p-ants leaving the robot's visual field of view are kept in the system and that p-ants have a limited life span permits execution time to be kept



(a) n_1 (b) n_1 (c) n_2 (d) n_3

(e) n_1 (f) n_1 (g) n_2 (h) n_3

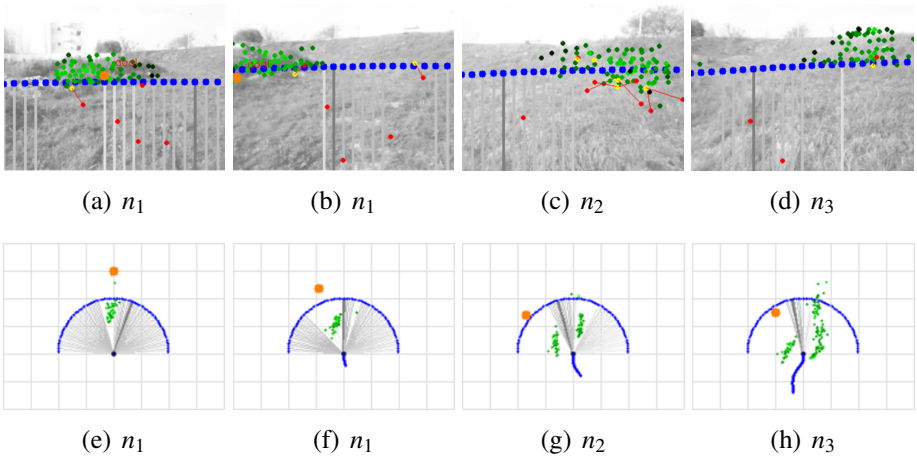
Figure 4.7: Motion compensation example. (a) The robot moves towards the target. (b) A set of p-ants detect a large obstacle. (c) Some frames latter the robot is tilted and these p-ants are obligated to track the obstacle outside the sensor's field of view. (d) When the tilt angle is zeroed these p-ants are once again in the sensor's field of view and consequently projected onto L. At that time, p-ants update their 3-D position and confirm the presence of the obstacle. The smaller number of p-ants, when compared to (b), is a consequence of the probabilistic removal of p-ants when they leave the field of view. The figure at the bottom-right of each image is a body-centred representation of the detected obstacles and linear trajectories. The darker the linear trajectory the higher its utility, as casted by the action selection process. It is possible to see the stability of the representation independently of the robot's current pose. Videos with situations as the one depicted in this figure are available as online supplementary material.

low (see Section 4.5.2). In conclusion, the ability of p-ants to operate as elements of the spatial working memory does not conflict with the idea of minimal perception.

4.5 Experimental Results

4.5.1 Experimental Setup

Small Vision System (SVS) (Konolige, 1997; Konolige and Beymer, 2007) and OpenCV (Bradski and Kaehler, 2008) were used for stereo computation and other low-level computer vision routines on the 320×240 input images, respectively. Stereo computation uses an area-based L1 norm (absolute difference) correlation method, operating over Laplacian Of Gaussian (LOG) transformed images. The result is interpolated to a precision of $1/4$ pixel and the correlation window size is 11×11. To increase the amount of



(a) n_1 (b) n_1 (c) n_2 (d) n_3

(e) n_1 (f) n_1 (g) n_2 (h) n_3

Figure 4.8: Spatial memory operation example (same colour code as in Fig. 4.7). The example given refers to a situation where the robot has to move towards a goal location 30 m ahead of its start position. To reach its goal, the robot circumnavigates the large obstacle area visible in (a) and (b), which as a consequence of the robot's motion leaves the sensor's field of view in (c) and (d). Without the spatial memory, the action selection process would not be able to take into account the obstacles outside the sensor's field of view, and consequently it could lead the robot towards a collision. Videos with situations as the one depicted in this figure are available as online supplementary material.

information available in the point cloud, the disparity calculation is carried out at the original resolution, and also on images reduced by $1/2$. With this multi-scale approach, the extra disparity information is used to fill in dropouts in the original disparity calculation.

SVS also provides a set of standard filters to reject 3-D points that are potentially erroneous at the cost of reducing too much the density of the 3-D point cloud in poorly textured environments. Briefly, according to a threshold f_c a confidence filter eliminates stereo matches that have a low probability of success due to lack of image texture. A uniqueness filter performs a consistency check to ensure that the minimum correlation value must be lower than all other match values by a threshold f_u. Finally, a speckle filter eliminates small disparity regions that are not correct by imposing a threshold f_s on the minimum region size. The three filters are used with $f_c = 12$, $f_u = 10$, and $f_s = 400$.

The following describes the model's parameterisation used in the experiments. With $\gamma = 1$, $\tau = 0.8$, and $\alpha = 0.6$, the dynamics of the confidence level in Equation 4.4 is such that spurious noise could be effectively filtered out, without imposing a compro-

mising latency to obstacles registration. Assuming a 10 Hz frame rate, with a maximum number of iterations per frame $i_{max} = 10$, and a p-ant's top energy $\rho = 5000$, obstacles are maintained in the robot's spatial memory for about 50 s. This timing is sufficient for local obstacle avoidance. The strong energy decay occurring when local search is engaged, i.e., $\upsilon = 100$, guarantees that p-ants in that situation are removed from the system before a new frame is acquired. To produce an adequate speed-accuracy trade-of, the stochastic motion control parameters from Equation 4.2, λ_1 and λ_2, have been empirically defined as 0.3 and 24, respectively. The number of times a p-ant is allowed to clone itself is $r_{max} = 5$. The cardinality of the robot's action repertoire, k, is 80, which is sufficient to ensure navigation in cluttered environments. The control parameters to determine when an obstacle is too cluttered with p-ants, η and ζ, have been set to 40 and 250, respectively. This ensures a good coverage of the obstacle with p-ants without too much overlap. The control parameter to avoid migration of p-ants between obstacles when performing a local search, l, has been set to 20. Finally, the probability of removing a p-ant when it leaves the visual input is $\psi = 0.4$. This value establishes an adequate trade-off between computational efficiency and the density of the robot's spatial memory.

4.5.2 Experiments

To validate the proposed model the robot was asked to perform at a speed of $s = 0.5\,\mathrm{ms}^{-1}$ a set of five runs on the off-road environment depicted in Fig. 4.9. In each run, the robot started from a different initial position and its goal was to reach a location 30 m ahead. Although mostly planar, the environment includes regions of variable slope. This, in addition to the considerable amount of tall vegetation, results in a considerable proneness for false positives occurrence.

Table 4.1 summarises the quantitative results obtained from the experiments. The robot reached the goal location in every run by negotiating all obstacles between itself and the goal. The executed paths were smooth and short, i.e., the robot avoided obstacles soon enough and smoothly followed their contour whenever necessary. This means that the system was capable of rapidly detecting obstacles at distance and kept tracking them in memory. Notice that whenever the robot followed the contour of obstacles outside its field of view we know that it is tracking them in memory.

Visual attention refers to the ability of accurately focussing perceptual resources

(a) (b)

Figure 4.9: Experimental site. (a) The test site overlaid with approximate paths executed by the robot in two runs. In each run, the robot started from a different position and moved autonomously to reach the specified final position, represented by a circle. (b) Situation where the difficult conditions faced by the robot are evidenced: dense and tall vegetation, which often promotes the occurrence of false positives.

Percentage of analysed pixels	$0.9\,\% \pm 0.6\,\%$
Control system computation time	$112.1\,\text{ms} \pm 5.4\,\text{ms}$
Swarm computation time, t_s	$1.7\,\text{ms} \pm 0.6\,\text{ms}$
t_s / full analysis computation time	0.37 ± 0.17

Table 4.1: Results summary (mean $\pm$ standard deviation).

where these are the most needed, according to the action selection process. The fact that the robot exhibits a near-optimal behaviour, i.e., it produces a highly directed motion towards the goal (see Fig. 4.9), with only 0.9 % of analysed pixels on average per frame, is a clear demonstration that the model produces accurate visual attention. Note that accuracy here stands for being able to focus precisely on the regions of the current frame that are the most important for the action selection process.

For the following analysis we split the time spent by the swarm infrastructure from the one spent by p-ants when applying the obstacle detector to the image in their current position. Were the swarm's computation time larger than the one taken by a full image analysis, the benefits of the proposed model would be marginal. This is not the case as it only requires 37 %, on average, of the time spent by the brute force approach.

As the complexity of the obstacle detector grows, the less significant is the computation time of the swarm infrastructure to the overall cost. Note that the obstacle detector used by p-ants is actually the simplest one for a stereoscopic setup, as it only requires the computation of a distance to the ground plane. Hence, the reported results

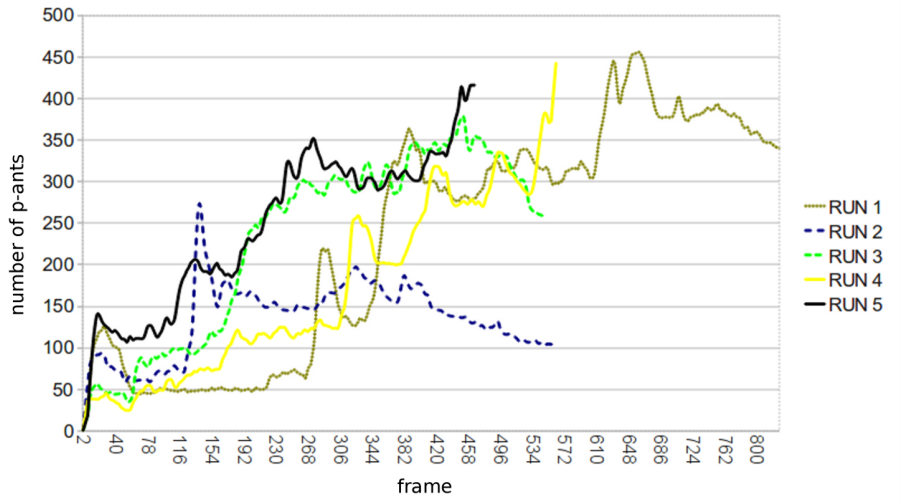


Figure 4.10: Context-dependent load balancing results. Plots refer to the total number of p-ants during each run.

can be understood as the baseline performance of the proposed model. A final remark to say that the whole control system operates almost at 10 Hz, which showed to be sufficient to maintain a stable control of the robot at a speed of $0.5\,\mathrm{ms}^{-1}$. This cost includes swarm update, visual odometry, ground-plane estimation and 3-D reconstruction.

One recognised characteristic of self-organised systems is their robustness to varying contexts. In this work, p-ants self-organise to build visual attention and spatial memories as the environment unfolds in the sensor's field of view. A by-product of this process is the adaptation to the context, which can be observed by the number of p-ants along time (see Figure 4.10). A raise in the number of p-ants corresponds typically to situations where new obstacles are detected. When these are actually noise, their associated p-ants are eventually removed, resulting in a corresponding drop in their cardinality. A smoother decay in the quantity of p-ants usually refers to the removal of p-ants that have reached zero energy. Thus, the variability of the number of p-ants is a sign of system's adaptation to the environmental context and robot-environment interaction history.

Fig. 4.11 shows the results of a comparison between the proposed model for visual attention and a random sampling of the visual input. From the comparison stands out that in 7 % of the tested frames the random policy outperforms the swarm-based policy. However, this percentage corresponds to situations where no obstacles are present in sensor's field of view, meaning that the comparison is made solely on the basis of false

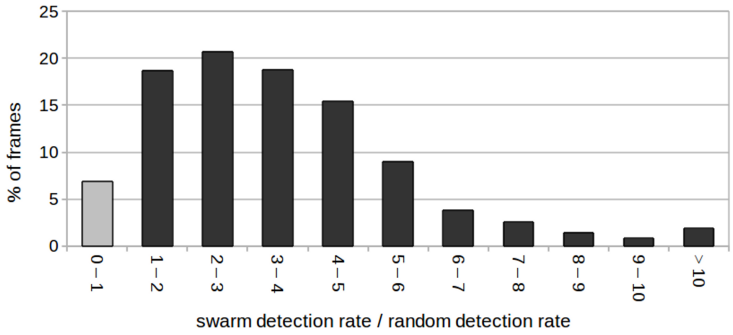


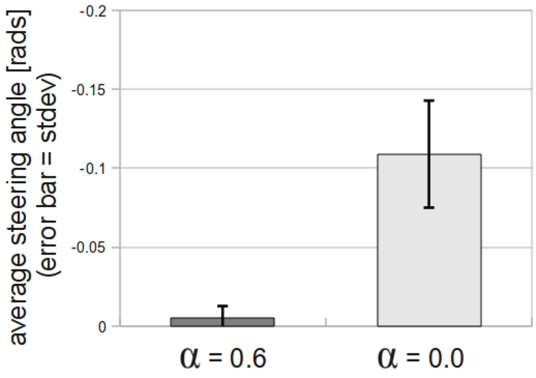
Figure 4.11: Comparison between the proposed model and a random policy. The bins of the histogram refer to the ratio between the detection rate of the proposed method and the detection rate of a random policy, over the set of five runs. Detection rate is defined as the ratio of analysed pixels that are reported as obstacles. The higher the detection rate the more focused and effective is the attentive process. As each frame, the random policy operates by randomly sampling as much pixels as those that have been analysed by the swarm-based one in the same frame. That is, this test compares the effectiveness of each method for the same computational cost. The first bin (light grey) corresponds to the situations where the random policy outperforms the swarm-based one, 7 % overall.

positives and consequently is of little value.

One can then conclude that the results clearly show the benefits of the swarm-based solution guided by the action selection process over a baseline random one. With a larger field of view and a higher resolution, i.e., in a situation where the chances of randomly selecting an obstacle pixel are smaller, it is reasonable to assume that the comparison would further highlight the advantages of the proposed model.

Fig. 4.12 shows that without the evidence accumulation mechanism, all false positives caused by the noisy nature of the sensory input are propagated to the action selection process and consequently to the actuators. Conversely, with evidence accumulation the actuators are barely affected by sensor noise. The inertia introduced by the mechanism to remove noise could cause latency in the sensory information processing. However, the proper behaviour exhibited in the previous experiments allows us to conclude that this is not the case. Note that the evidence accumulation process is affected by the activity of the shared medium R, thus implementing a sort of implicit multi-agent consensus protocol. This is essential to help discriminating actual obstacles from spurious isolated noise, without imposing a slow evidence accumulation dynamics to every situation. Moreover, the use of this form of implicit communication (*stigmergy*) to reach a sort of collective decision is a feed-forward mechanism without

(a)



(b)

Figure 4.12: Comparison between model with ($\alpha = 0.6$) and without ($\alpha = 0.0$) evidence accumulation. The comparison is based on statistics obtained from running each configuration five times on a pre-recorded video sequence. The video sequence was acquired by tele-operating the robot along the straight ahead $\approx 25\,\mathrm{m}$ motion overlaid in (a), whose final point is represented by a circle. To magnify the differences between both configurations, the obstacle detector's threshold is not set to grow with range, i.e., $\beta = 0$ in Equation 4.1. The action selection's goal in both configurations is to move forwards. No obstacles along the path inhibited the forward motion from being the one with highest utility. Therefore, in the absence of false positives influencing the action selection process, the latter would output a zero steering. The bars and error bars in (b) correspond to the mean and standard deviation of the steering angle for each configuration.

computational overhead.

A set of representative videos with the proposed model's output is available as online supplementary material. For complementary experimental evidence on simulation regarding the key role of some aspects of the model, such as stochastic vs. deterministic behaviour, memory-based vs. pure reflexive operation, parallel vs. a pure greedy deployment of p-ants, please refer to our previous work (Santana and Correia, 2010a).

4.6 Discussion

A p-ant represents a covert focus of attention to a given portion of the visual input, upon which a task-specific detector can be applied. Analogously, a natural ant can be seen as a mobile sensor that locally senses the environment for the presence of food items. Natural ants reinforce pheromone trails leading to fruitful locations so as to promote the rate at which food items are returned to the nest. P-ants also use a form

of pheromone (activity in shared medium R) to recruit other p-ants to the location of detected obstacles. A negative feedback mechanism is included under the form of a too high pheromone concentration repelling p-ants. This guarantees an adequate number of p-ants to aggregate around a given obstacle.

Patterns in human brain may break when the visual input changes, but rapidly re-emerge in order to track the dynamics of the environment. Ant colonies are also known to be endowed with similar resilience in what regards tracking environmental changes. Thus, this ability of ant colonies can be used as metaphor to model tracking behaviour in parallel visual processes. Tracking is the action of updating the estimate of a given obstacle's position in the visual input, which can change, for instance, due to motion. In general this requires the tracking entity to perform a local search with the goal of re-detecting the obstacle. Accordingly, if a p-ant determines that it has lost the obstacle it has been focused on, it initiates a local search. As their natural counterparts, p-ants also exploit other p-ants pheromones to boost this search. In addition, p-ants are also sensitive to the robot's motion, allowing them to perform the respective compensation.

This ability of tracking and motion compensation endows the visual system with an intrinsically parallel and active spatial map of the environment. That is, each obstacle is represented by a set of p-ants that actively seek to maintain contact with it.

It is largely recognised that the amplification of random fluctuations is essential to allow the emergence of novel situations in foraging tasks, and in self-organised systems in general (Bonabeau et al., 1999). In this line, p-ants' behaviour also encompasses a probabilistic component. We add the following advantages of using randomness in search tasks. The goal of explicitly using random fluctuations allows p-ants to actively search the surroundings with complex search patterns, without their explicit coding. Moreover, the use of stochastic behaviour has the additional advantage of naturally handling noisy data. A deterministic program would require an extensive exception handling facility, which tends to be error-prone. In short, random fluctuations are essential to foster both cheap and robust design. In fact, the variability in neural response to identical stimulus has supporting evidence to be key in explaining the nearly optimal human brain performance, in a Bayesian sense (Ma et al., 2006). Refer to our previous work on simulation (Santana and Correia, 2010a) for experimental evidence supporting the importance of random fluctuations in the proposed model.

Neurophysiological studies reveal that when human subjects are requested to produce fast response times, in detriment of accuracy, an increase in the baseline activ-

ity is observed in a network of brain areas related to decision, response preparation, and execution, in opposition to sensory related ones (van Veen et al., 2008). The raise in baseline activity allows reaching a decision threshold faster. In our model this is implemented by constraining the time available, i.e., maximum number of iterations per frame, for the action selection process to reach a conclusion. Until the time limit is reached, both action and perceptual processes act the same way for each possible speed-accuracy context. We speculate that this is only possible because the act of perceiving has the goal of pushing action selection forward, while being led by it. By using a parallel approach, our model allows perception to robustly track a good solution right from its onset and without the need of explicit context awareness. This allows both perceptual and action processes to progressively unfold by analysing best possible actions in decreasing order. The result is a good solution whatever the speed-accuracy trade-off. Thus, context modulation is emergent through the decision process, rather than from an explicit context-dependent modulatory signal operating on the perceptual process.

As for the perception-action coupling aspect, the parallel visual process herein proposed also maps seamlessly on the current understanding about the way covert attention operates. First, the use of a parallel system follows the idea of multiple loci of attention (Pylyshyn and Storm, 1988; Doran et al., 2009). Second, being parallel, the approach robustly handles the speed-accuracy trade-off (van Veen et al., 2008). Finally, the proposed model includes a natural way of integrating noise, which is known to be very important to achieve near-optimal brain operation (Ma et al., 2006).

From an engineering standpoint, it is useful to compare the proposed model with particle filters as employed for the problem of simultaneous localisation and mapping (SLAM) (Thrun et al., 2005). In general, a particle in SLAM is a sample of the joint distribution robot/map, where a map is composed by a set of Extended Kalman Filters (EKF), each associated to a given landmark of the environment. Conversely, a p-ant is associated to a single landmark, and so it is more related to the role played by each EKF than with the particles themselves. Differently from EKFs in SLAM, p-ants perform an active search in addition to tracking, and do it guided by the action selection process. Moreover, while p-ants interact to improve their individual performance, EKFs in particles are independent.

P-ants do not contribute to the estimation of the localisation of the robot, as particles do. Local navigation being the focus of the proposed model, the localisation problem is

handled solely with dead-reckoning. While SLAM methods could deliver both localisation and mapping, their high computational and memory cost makes them unsuited for parsimonious local navigation. For a more comprehensive discussion refer to our previous work (Santana and Correia, 2010a).

4.7 Conclusions

The first embodied realisation of swarm cognition was proposed and validated on the local navigation problem for vision-based off-road robots. Although simulations are recognised as important tools to validate minimalist cognition models (Slocum et al., 2000), we believe that only when facing the burdens of the real world, i.e., uncertainty, a model is capable of truly exhibiting its scalability. In this sense, this report is an important contribution to the demonstration of swarm cognition as a promising approach to the synthesis of cognitive systems.

The model considers that perceptual and action selection processes operate in an intricate way to focus attention and consequently reduce the processed image area. The underlying idea of the model is that perception should focus on the regions of the visual input where detected objects most impact the action selection process. This helps action selection to rapidly stabilise its output. Motivated by the multiple covert attention hypothesis (Pylyshyn and Storm, 1988; Doran et al., 2009), covert visual attention is modelled with swarms of simple virtual agents, named p-ants, based on the social insects foraging metaphor. Basically, p-ants perform local covert visual attention loops, whereas the self-organised collective behaviour maintains a robust global spatio-temporal coherence.

The benefits of these properties are confirmed by experimental results obtained with a physical all-terrain robot performing local navigation and equipped with a stereoscopic vision sensor. Concretely, the perceptual process is showed to be capable of providing the action selection process with sufficient information for a proper decision making with only 0.9 % of visual input being analysed. This accuracy of the focus of attention is credited to: (1) the ability of action selection to guide perception on a by-need basis, as verified against a random policy; and (2) the self-organising principles exhibited by the parallel covert visual attention process, which provides the system with robustness and consequently the ability of maintaining an adequate focus of attention across environments. Spatial memory, another relevant asset of a cognitive

system, was also shown to self-organise in the proposed model. The evidence accumulation mechanism added to the model revealed to be essential on the reduction of the sensitivity to false positives, which due to their high quantity would impair cognitive behaviour.

On the one hand, these results show the feasibility and advantages of relying on self-organisation to synthesise robust embodied cognitive systems, by recurring to the social insects metaphor. On the other hand, they show that with such a model, autonomous robots can considerably reduce the cost of perception, which in turn fosters real-time performance, computational parsimony, and energetic efficiency. This is an important asset to enable miniaturisation, long-lasting operation, and higher robot speed. The proposed model is a systematic method to exploit action selection information to maintain an accurate focus of perceptual attention. It allows the robot designer to exploit the most adequate detector for the task at hand without having to customise the method to cope with the computational requirements, and consequently to lose some of its detection capabilities.

As future work, we plan to devise a parallel implementation of the model in dedicated hardware, such as Graphics Processing Units (GPU), thus fully exploiting the advantages of a multi-agent design. In addition, we intend to generalise the model so that it can be applied to other perception tasks.

Acknowledgements

This work was partially supported by IntRoSys, S.A. and by FCT/MCTES grant No. SFRH/BD/27305/2006. We would like to thank the fruitful discussions with Magno Guedes and Nelson Alves, as well as their support to the field experiments. We also thank IntRoSys, S.A. for the availability of the Ares robot.

Chapter 5

Finding Natural Trails With Swarm-Based Visual Saliency

Pedro Santana, Nelson Alves, Luís Correia, José Barata.
Submitted to *Journal of Field Robotics*.

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 based on a priori 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. Bearing in mind both computational parsimony and detection robustness, this 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 20 Hz. 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.

keywords: terrain perception, trail detection, visual saliency, bio-inspired methods, terrestrial robotics.

5.1 Introduction

Autonomous navigation in off-road environments demands for robust perceptual capabilities. An outstanding one is the ability to perform free-space visual assessment, which can be done via the explicit analysis of the volumetric properties of the terrain's surface (Batavia and Singh, 2001; Lacroix et al., 2002; Manduchi et al., 2005; Broggi et al., 2005; Seraji, 2006; Konolige et al., 2009; Kolter et al., 2009; Rusu et al., 2009; Santana et al., 2011). In addition to this direct inference of terrain's state, one can also use indirect cues regarding free-space connectivity, such as the presence of trails.

Trails, such as the ones created for hikers and bikers, are usually safe pathways, free of dead-lock situations. A robot following a trail should thus be able to traverse large distances in off-road environments in an effortless way. On the one hand, computation for obstacle detection and path planning is saved. On the other hand, fewer are the chances of getting lost or of incurring into ground traps. A practical application of robots with the ability of trail following could be natural parks patrolling, in which robots would be engaged in actively maintaining and cataloguing the environment, and possibly providing support to human hikers.

Without disregarding the benefits of using range information for the task of trail detection, provided for instance by stereo-vision or laser scanners, this paper addresses the problem from a complementary two-dimensional (2-D) vision perspective. This complementarity is important because trails not always have a recognisable volumetric signature and because accurate three-dimensional (3-D) point clouds are difficult to obtain off-road. That is, while stereo-vision sensors strongly rely on good lighting conditions to triangulate features in both cameras, laser scanners rely on good ego-motion estimates for the registration of consecutive 2-D scans, which may be hard to achieve when moving fast in bumpy terrains. An alternative would be to use 3-D laser scanners, such as Velodyne. However, trail following being a task more suitable for small/medium robots can hardly be solved with such a bulky and expensive sensor.

Most of the challenges of trail detection relate to their lack of a well defined morphology or appearance. This hampers a straightforward specification or learning of trail models. In addition, trails exist in environments that are unstructured themselves,

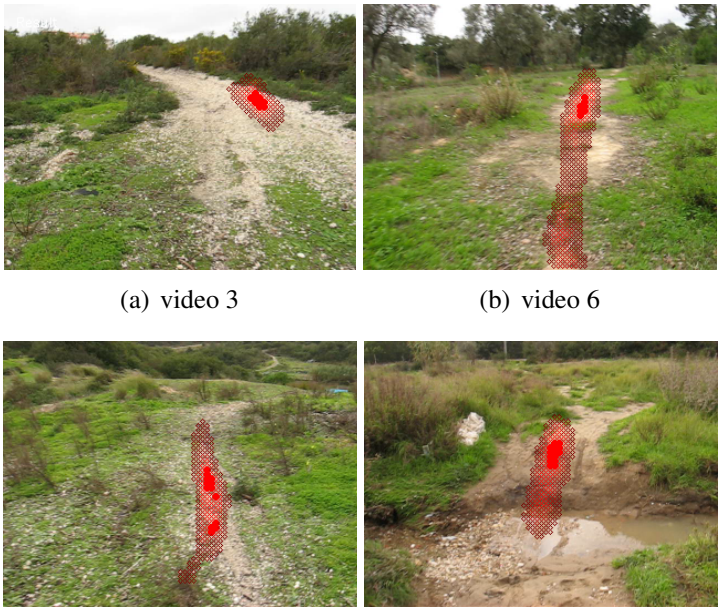
which complicates the learning of background models. Moreover, the fact that trails change over time renders data hand-labelling for the learning supervision unsuited. Hence, model-free (meaning as free as possible) solutions are essential for their robust detection.

This paper exploits the observation that trails are typically conspicuous in the visual field of the robot to propose the use of visual saliency for the purpose of their detection. In other words, this work exploits the overall scene's context to guide the localisation of the trail. This approach does not impose any hard constraints 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 (Ruesch et al., 2008; Moren et al., 2008; Santana et al., 2011), the overhead of its computation is diluted over all modules using it.

Besides confirming the hypothesis that visual saliency and trail location are indeed positively correlated, it will also be shown that the conspicuity maps of a given input image correspond themselves to efficiently computed segmentations 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 at various scales. Furthermore, the obtained segments are already prioritised by their conspicuity level. Given that this saliency-based image segmentation is done "labelling" each region according to its detachment level with respect to the overall scene, it does not require hard edges separating the regions, which is known to be a problem when over-segmenting an image (Cour and Shi, 2007).

From these findings, it should follow that the segment in the saliency map with highest priority matches the location of the trail in the input image. However, in practise, this is a brittle assumption in the presence of not so well behaved conspicuity maps, which may occur in the presence of distractors or when the trail is considerably heterogeneous. This difficulty can be diminished by top-down boosting of the set of features (e.g., colour) known beforehand to better describe the object being sought (Frintrop et al., 2005; Navalpakkam and Itti, 2005). However, these visual features are considerably unpredictable in the case of trails in natural environments. In contrast, 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 type of a priori knowledge is embedded in the model herein proposed in the form of behaviour rules for the motion of simple agents inhabiting the conspicuity



(a) video 3 (b) video 6

(c) video 24 (d) video 25

Figure 5.1: Typical trail detection results (red overlay) obtained with the proposed model. These results show the model's ability to localise the trail even when it is highly interrupted (b)(d), blends itself with the background (c), or does not start from the bottom of the image (a).

maps associated to specific visual features, namely, colour and intensity. The paths executed by these agents are used as skeletons of trail hypotheses. However, a system based solely on the individual behaviour of these simple agents is brittle when presented with less structured (e.g., interrupted) trails. To overcome these limitations, the proposed model exploits: (1) the metaphor of the *collective intelligence* (Franks, 1989) exhibited by social insects to ensure that agents cooperatively build up a robust approximation of the actual trail's skeleton; and (2) dynamical neural fields (Amari, 1977; Rougier and Vitay, 2006) to accumulate evidence about the most likely trail location across frames, extended with a mechanism to compensate for robot motion.

Extensive experimental results show that the proposed model attains a success rate of 91 % at 20 Hz in demanding scenarios, as depicted in Fig. 5.1. This fast computation owes greatly to the extensive use of bottom-up mechanisms, which besides fostering future implementations in parallel hardware, also avoid expensive computations required by model-based solutions.

This paper extends two previous conference papers (Santana et al., 2010a,b) and is organised as follows. Section 5.2 overviews related work. Then, Section 5.3 introduces

the proposed model, which is followed by experimental results in Section 5.4. Finally, some conclusions are drawn and future work is proposed in Section 5.5.

5.2 Related Work

Current trail detection methods rely considerably on work developed for the road detection domain, which for this reason is first surveyed.

While the detection and tracking of paved roads is facilitated by their rather predictable appearance and presence of well delimited limits, this is not the case of ill-structured unpaved rural roads. The typical solution in the latter 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. The likelihood of a given pixel to belong to the road surface can either be learnt off-line from labelled images (Chaturvedi and Malcolm, 2005; Alon et al., 2006) or, for more robust operation, learnt on-line given a set of reference regions in the input image automatically labelled as road/non-road. In the latter case, the labelling process is done assuming that the robot is on the road, and so the region right in front of it can be surely labelled as road (Thorpe et al., 1988; Fernandez and Casals, 1997; Fernandez and Price, 2005; Song et al., 2007), or exploiting short range volumetric information obtained from other sensors (e.g., stereo or laser) to discriminate the road plane from others (Thrun et al., 2006b; Tue-Cuong et al., 2008). To attain depth invariance, locally labelled visual elements can be traced back until the moment they entered the camera's field of view, i.e., when they still were in the far field (Lookingbill et al., 2007).

In general, once the segmentation is concluded, a simplified model of the road (e.g., trapezoidal) is fit to the segmented image. Region growing is an interesting alternative to the model fitting process when it regards handling hard to model roads (Ghurchian et al., 2004; Fernandez and Price, 2005; Chaturvedi and Malcolm, 2005). However, 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). This reduces the burden of maintaining road appearance models at the cost of raising the number of possible ambiguities between regions of the input image with similar shape. A known alternative to these region-based approaches is to estimate the road's vanishing point (Rasmussen, 2004, 2008; Kong et al., 2010). This is usually done by extracting the dominant texture orienta-

tions, which are usually aligned with ruts, tire tracks, and road borders. This approach is interesting in particular when the road and background share the same appearance. The orientations-based and regions-based approaches can also be integrated into hybrid architectures (Alon et al., 2006; Song et al., 2007).

As mentioned, these models are the basis of most work on trail detection. 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 in a self-supervising manner (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 references. 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 first over-segmenting the image and then scoring a set of trail hypotheses, built by aggregating sets of generated segments, against the global shape constraint. Accurate image over-segmentation being a computationally demanding task reduces the system's ability to perform in real-time and usually requires clear edges segmenting the object from the background, which is not often the case in natural trails. This is aggravated by the considerable level of interruptions natural trails exhibit. Moreover, a global shape constraint limits the type of trails that can be detected. Finally, due to the fact that dominant orientations in natural trails seldom indicate the global orientation of the trail, the vanishing point concept, a powerful concept in the road domain, hardly applies.

Instead of adapting a road detection method to the trail detection problem, this paper proposes the exploitation of a distinguishing characteristic of trails present in natural environments: they are rather conspicuous structures. This induces us to propose the use of visual saliency models to detect them. This approach does not impose any hard constraints on the appearance or shape of both trail and background, nor it requires learning. In a parallel study, Rasmussen et al. (2009) proposed the use of local appearance contrast for trail detection, which only superficially resembles the concept of visual saliency. Visual saliency includes contrast information between trail and local

surroundings, plus contrast information between trail and overall scene. 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. In addition and in opposition to our model, Rasmussen et al. (2009) make the assumption that trails exhibit a perfect triangular shape when seen from perspective. They also assume that both left and right sides share the same appearance. Although these assumptions comply with a large set of situations, they lack support in more demanding ones.

Seeking robustness and parsimony, we follow the 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). 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, trails 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 trails 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 trail detection in natural environments. We can also conclude that our model is the first swarm-based computation of visual saliency.

5.3 Proposed Model

As a large extent of other robotics and computer vision work, the one herein proposed also benefits from biological inspiration. This inspiration is summarised in Section 5.3.1, which is followed by an overview of the proposed model in Section 5.3.2. The subsequent sections detail its key elements.

5.3.1 Biological Inspiration

The fundamental aspects considered in this work are the use of visual attention for trail detection and the use of multiple agents for its computation. Visual attention is known to be widespread in the animal kingdom (Land, 1999) and it has been exten-

sively studied in humans (Oliva and Torralba, 2007; Wolfe et al., In Press). By focussing perception, computation is saved and robustness enhanced. As a consequence, faster robot motion, lower cost, and reduced robot size are enabled. Studies on human subjects support the hypothesis that multiple covert, i.e., mental, attention processes coexist in the brain (Pylyshyn and Storm, 1988; Doran et al., 2009). This evidence is the motivation for this work to use a multi-agent approach.

Agents perform local covert visual attention loops, whereas the self-organising collective behaviour maintains global spatio-temporal coherence. Additionally, these agents being sensorimotor coordinated units can exploit the benefits of active vision (Bajcsy, 1988; Aloimonos et al., 1988; Ballard, 1991) at the information processing level. These include the ability of agents to actively select and shape their sensory input. As a result, noise-to-signal ratio can be increased, rotation and scale invariance augmented, and sensorimotor history used to induce long-range influences and improve agent behaviour (Scheier et al., 1998; Nolfi and Marocco, 2002; Beer, 2003; Fend et al., 2003; Balkenius et al., 2004; Floreano et al., 2004; Nolfi, 2005; Suzuki and Floreano, 2006; Pfeifer and Bongard, 2006; Kim and Moeller, 2006; Sporns and Lungarella, 2006; de Croon and Postma, 2007; Choe et al., 2008; Mirolli et al., 2010; Tuci et al., 2010).

The use of multiple agents allows the modeller to exploit biological knowledge obtained from similar processes that can be found in Nature, such as the *collective intelligence* (Franks, 1989) exhibited by social insects, a sort of *swarm cognition* with similarities with neuronal processes (Passino et al., 2008; Couzin, 2009; Marshall and Franks, 2009; Santana and Correia, 2010a; Trianni and Tuci, 2010; Turner, 2011; Trianni et al., 2011). The specific metaphor exploited in this work is the one of ant foraging. According to this metaphor, local agents self-organise in order to find a global solution. This is attained by having these agents interacting indirectly through pheromone-like fields, a phenomenon known as *stigmergy* (Grassé, 1959). Being loosely coupled, simple, and sensorimotor coordinated, these agents build up a robust and fast to compute system.

The option of using dynamical neural fields (Amari, 1977; Rougier and Vitay, 2006) for evidence accumulation across frames and improved focus of attention is also bio-inspired, in the sense that neural fields have a long history in the modelling of human cognition (Beer, 1995; Thelen and Smith, 1996).

5.3.2 System Overview

Fig. 5.2 illustrates the different phases involved in the proposed model's operation. At each new frame $\mathbf{I}$, two conspicuity maps, $\mathbf{C}^C \in [0,1]$ for colour and $\mathbf{C}^I \in [0,1]$ for intensity information, are computed (see Section 5.3.3). The intensity of a pixel in a given conspicuity map signals how much the pixel detaches from the background at several scales, in the scope of a given visual feature.

A set of n virtual ants (hereafter called p-ants) is deployed on each conspicuity map (see Section 5.3.4.1). These p-ants interact based on the ant-foraging metaphor for several iterations in order to build two pheromone maps, $\mathbf{P}^C \in [0,1]$ and $\mathbf{P}^I \in [0,1]$ (see Section 5.3.4.2). The behaviour of p-ants is designed to exploit a priori knowledge about typical trails approximate layout. Therefore, the activation of pheromone maps is expected to match the trail's location better than the activation of conspicuity maps, which are only sensory-driven. Thus, rather than combining both conspicuity maps to generate the final saliency map, as typically done (Itti et al., 1998; Frintrop et al., 2005), in this work $\mathbf{S}$ is obtained by combining both pheromone fields, $\mathbf{S} \leftarrow \frac{1}{2}\mathbf{P}^I + \frac{1}{2}\mathbf{P}^C$. Additionally, by allowing p-ants on a given pheromone map to also affect the other pheromone map, cross-modality influences are implicitly, i.e., through stigmergy, maintained in the system. This increases robustness by allowing p-ants to exploit multiple cues indirectly, with a residual computational overhead.

The final saliency map $\mathbf{S}$ feeds a dynamic neural field, $\mathbf{F} \in [0,1]$, which integrates pheromone (i.e., evidence) across frames and also implements both lateral excitation and long-range inhibition (see Section 5.3.5). Excitatory connections promote perceptual grouping, whereas inhibitory connections facilitate the maintenance of a coherent focus of attention across time (Rougier and Vitay, 2006). Motion compensation, between current frame $\mathbf{I}$ and previous frame $\mathbf{I}'$ is also implemented so that the dynamics of the neural field can be decoupled from the dynamics of the robot. The output of the system is given by the current state of the neural field, in which the higher the activation of a given neuron the higher its chances of being associated to a trail's pixel.

In order to allow p-ants' creation and activity to be affected by history, at the onset of each frame, both pheromone maps are initialised with a small ratio λ of the neural field after being motion compensated, $\mathbf{P}^I \leftarrow \lambda\mathbf{F}, \mathbf{P}^C \leftarrow \lambda\mathbf{F}$. This induces stability and robustness to noise and temporarily mis-behaved conspicuity maps (i.e., unable to properly discern between trail and background in the presence of distractors), as

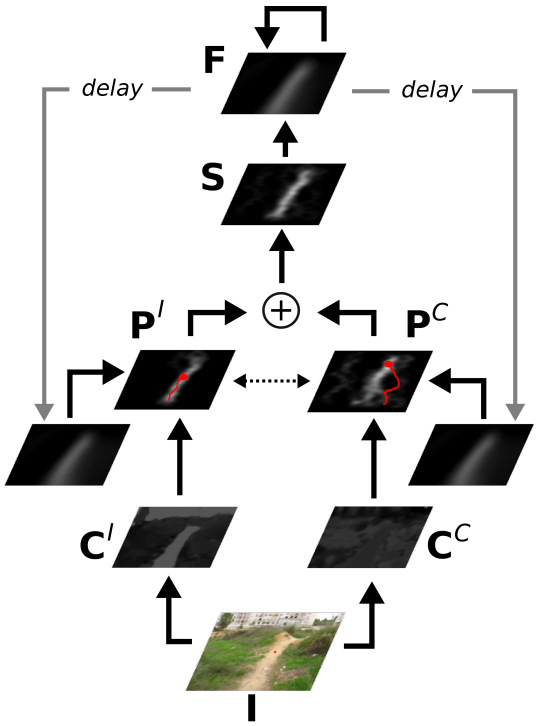


Figure 5.2: System's operation overview. The red overlays in both pheromone fields, $\mathbf{P}^C$ and $\mathbf{P}^I$, are two illustrative p-ant paths. Motion compensation aspects are not represented.

well as it enables across-frames progressive improvement.

As aforementioned, the neural field and cross-modality influences are useful to modulate the creation and behaviour of the p-ants. However, if allowed to propagate across frames, these influences may induce an undesirable neural field's activity build-up. To avoid this, two auxiliary pheromone maps, $\mathbf{P}^I_*$ and $\mathbf{P}^C_*$, are created free of influences. For instance, the auxiliary map $\mathbf{P}^I_*$ only encompasses the pheromone deposited by the p-ants associated to visual feature I. These maps are used to replace the pheromone maps, $\mathbf{P}^I \leftarrow \mathbf{P}^I_*, \mathbf{P}^C \leftarrow \mathbf{P}^C_*$, just before blending them for the purpose of creating $\mathbf{S}$. In this way, the self-influence of the neural field through the pheromone maps still exists, but it is mediated by the p-ants behaviour.

Conspicuity maps, pheromone maps, saliency map, and neural field, all share the same width w and height h. These two values are selected bearing in mind real-time performance. For a better understanding of the proposed model, Algorithm 6 outlines its pseudo-code. Details are given in the following sections.

Algorithm 6: Frame-wise model's execution pseudo-code.

Input: current frame ($\mathbf{I}$), previous frame ($\mathbf{I}'$), neural field ($\mathbf{F}$), Δt
Output: updated neural field ($\mathbf{F}$)
Data: λ and n are empirically defined constants.

use Equation 5.14 to estimate homography matrix, $\mathbf{H}$, from $\mathbf{I}$ and $\mathbf{I}'$
compensate neural field for robot motion, $\mathbf{F} \leftarrow \mathbf{H}\mathbf{F}$
compute conspicuity maps from $\mathbf{I}$, $\mathbf{C}^I$ and $\mathbf{C}^C$ [see Section 5.3.3]
initialise pheromone maps, $\mathbf{P}^I \leftarrow \lambda\mathbf{F}, \mathbf{P}^C \leftarrow \lambda\mathbf{F}$
$\mathbf{P}^I_* \leftarrow \emptyset, \mathbf{P}^C_* \leftarrow \emptyset$
// update pheromone maps
for n cycles **do** [create and execute $2n$ p-ants]
    create p-ant p_I based on $\mathbf{C}^I$ and $\mathbf{P}^I$ [see Section 5.3.4.1]
    create p-ant p_C based on $\mathbf{C}^C$ and $\mathbf{P}^C$ [see Section 5.3.4.1]
    $(\mathbf{P}^I, \mathbf{P}^C, \mathbf{P}^I_*) \leftarrow$ **execute**$(p_I, \mathbf{C}^I, \mathbf{P}^I, \mathbf{P}^C, \mathbf{P}^I_*)$ [see Algorithm 7]
    $(\mathbf{P}^C, \mathbf{P}^I, \mathbf{P}^C_*) \leftarrow$ **execute**$(p_C, \mathbf{C}^C, \mathbf{P}^C, \mathbf{P}^I, \mathbf{P}^C_*)$ [see Algorithm 7]
    remove p-ant p_I
    remove p-ant p_C
end
discard cross-modality and neural field influences, $\mathbf{P}^I \leftarrow \mathbf{P}^I_*, \mathbf{P}^C \leftarrow \mathbf{P}^C_*$
compute saliency map, $\mathbf{S} \leftarrow \frac{1}{2}\mathbf{P}^I + \frac{1}{2}\mathbf{P}^C$
use Equation 5.16 to update neural field $\mathbf{F}$ with $\mathbf{S}$
return F

5.3.3 Conspicuity Maps Computation

As mentioned, conspicuousness computation is about determining which regions of the input image detach more from the background at several scales in a given feature channel. Although in this paper only intensity and colour channels are used, additional channels (e.g., orientations, texture) could be used for improved background-trail discrimination. The following describes the biologically inspired model proposed by Itti et al. (1998) for conspicuity computation, herein properly adapted to the task at hand.

One dyadic Gaussian pyramid (Burt and Adelson, 1983), 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 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. Rather than considering the modulo of the difference, as in the original model (Itti et al., 1998), we consider both on-off and off-on centre-surround maps separately, as it has been shown to yield better results (Frintrop et al., 2005; Frintrop, 2006). Then, all centre-surround maps built from the intensity pyramid are resized to a common size and independently scaled in magnitude according to a method described in the next sections, and finally averaged together to produce the intensity conspicuity map. The same process applies to create Red-Green and Blue-Yellow conspicuity maps, each one subsequently weighted and then averaged together to produce a single colour conspicuity map. Note that all maps are 8-bit grey level images.

5.3.3.1 Typical Magnitude Scaling Functions

Magnitude scaling functions return a version of each map, obtained by a pixel-wise product of a scalar. The goal is to promote maps that have fewer conspicuous locations. As pointed out by Itti et al. (1998), this avoids that, when combining maps, salient objects appearing strongly in only a few maps are masked by noise or by less-salient objects present in a larger number of maps.

In the original model (Itti et al., 1998), the scaling factor to be applied to a given map $\mathbf{X}$ is defined by the square of the difference between map's global maximum, $M(\mathbf{X})$, and the average of all its other local maxima, $\bar{m}(\mathbf{X})$. The corresponding scaling function multiplies the magnitude of each pixel by the computed factor:

$$\mathcal{N}(\mathbf{X}) = \mathbf{X} \cdot (M(\mathbf{X}) - \bar{m}(\mathbf{X}))^2. \tag{5.1}$$

A similar approach has been proposed by Frintrop et al. (2005); Frintrop (2006). In this case, the scaling function is defined by

$$\mathcal{W}(\mathbf{X}) = \mathbf{X} / \sqrt{m(|\mathbf{X})|}, \tag{5.2}$$

where $|m(\mathbf{X})|$ is the number of the map's local maxima above a given threshold. The threshold is by default 50% of the map's global maximum (Frintrop, 2006).

Common to both methods is the use of local maxima information, which appealing it might be does not always embody the information intended to capture. Large homogeneous structures for instance, such as the sky, generally encompass only a few local maxima. In this situation, the sky would be undesirably considered highly conspicuous, despite its large foot-print in the whole image. A second aspect is that the two analysed saliency models consider that all pixels contribute equally to the saliency computation. However, excepting for extreme tilt/roll angles, the upper region of the image has little relevant information for trail detection. As a consequence, without a space-variant contribution to the final saliency map, feature maps that are only discriminative in the lower part of the image, and consequently interesting for trail detection, would not be adequately promoted.

5.3.3.2 Novel Scaling Function

In the face of the aforementioned limitations of previous scaling functions for the task of trail detection, a new one is herein proposed. Rather than considering only the map's local maxima when averaging, as it is done in $\mathcal{N}(\cdot)$ (Equation 5.1), we propose to use all pixels. Furthermore, the contribution of each pixel to the average is weighted according to its distance from the top row.

Formally, let $\mathbf{X}(c,r)$ return the grey level of the pixel in column c and row r of a given map $\mathbf{X}$. Let $w(c,r) = \sqrt{r/h}$ be the weight of pixel at position (c,r). The map's weighted average, m_w, is thus given by

$$m_w(\mathbf{X}) = \frac{\sum_{(c,r)\in\mathbf{X}} \mathbf{X}(c,r)\cdot w(c,r)}{\sum_{(c,r)\in\mathbf{X}} w(c,r)} \tag{5.3}$$

and, similarly to function $\mathcal{N}(\cdot)$ (Equation 5.1), the proposed scaling function, $\mathcal{K}(\cdot)$, takes the form

$$\mathcal{K}(\mathbf{X}) = X\cdot(M(\mathbf{X}) - m_w(\mathbf{X}))^2. \tag{5.4}$$

Prior to scaling, maps are normalised to $[0,1]$ amplitude interval, meaning that $M(\mathbf{X}) = 1$ for all cases. To reduce computational cost, the proposed system uses image operators over 8-bit images. An example of two conspicuity maps obtained with

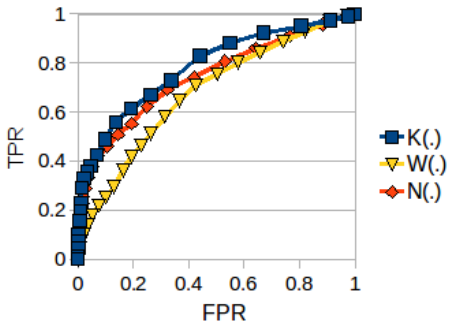


Figure 5.3: Scaling functions comparison. Each plot corresponds to the ROC curve obtained with a given scaling function.

the proposed model is depicted in Fig. 5.2.

To quantitatively assess the proposed scaling function, we generated conspicuity maps for the input images present in a test data set (Santana et al., 2010b). The two conspicuity maps produced in each image are combined to produce a saliency map, which is in turn thresholded to produce a pixel-wise trail/non-trail classification. The resulting classification is subsequently compared against hand-labelled ground-truth. This process is repeated for all images in order to obtain an average True Positive Rate (TPR) and an average False Positive Rate (FPR). Varying the threshold between minimum and maximum pixel magnitude we plot for each case a point in a TPR vs. FPR graph. The interpolation between those points produces a Receiver Operating Characteristic (ROC) curve.

Fig. 5.3 depicts the ROC curve for the proposed scaling function, $\mathcal{K}(\cdot)$ (Equation 5.4), as well as for its predecessors, $\mathcal{N}(\cdot)$ (Equation 5.1) and $\mathcal{W}(\cdot)$ (Equation 5.2). These curves show that the proposed scaling function consistently produces a better trade-off between TPR and FPR. The small difference between the ROC curves could suggest that only a small quantitative improvement was obtained with the proposed model. However, the averaging procedure used to build the curves hides the fact that none of the other two methods was able to consistently allocate higher levels of saliency to trail regions than to the background, as often as the proposed one.

Fig. 5.3 also shows that all ROC curves are considerably above the line of no-discrimination ($y = x$), meaning that saliency is correlated with trail location, an important contribution by itself. However, the correlation is still lower that the one required for high accuracy trail detection. That is, there is no single threshold on the saliency

map that clearly segments the trail for all images in the data set. Thus, a higher level analysis of the conspicuity maps is required. As will be shown, the swarm-based system described in the following section is capable of providing it.

5.3.4 Pheromone Maps Computation

This section describes how the two pheromone maps, $\mathbf{P}^I$ and $\mathbf{P}^C$, are built from the two conspicuity maps, $\mathbf{C}^I$ and $\mathbf{C}^C$. For this purpose, a given p-ant, p_m, is created and associated to a given visual feature $m \in \{I, C\}$. 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 $\mathbf{P}^m$. While moving, this p-ant deploys pheromone in each position visited in $\mathbf{P}^m$ with a magnitude ε, and a small portion of ε, υ, in $\mathbf{P}^{m'}$. After the iterations for this p-ant, a p-ant associated to the other visual feature, $p_{m'}$, is created and iterated following the same procedure. Afterwards, the two p-ants are removed from the system and the process is repeated n times, meaning that $2n$ p-ants are created and iterated (see Algorithm 6).

Allowing p-ants to affect both pheromone maps enables loosely coupled cross-modality influence, thus allowing each p-ant to exploit multiple cues indirectly, while maintaining their simplicity. As it will be shown, the deployed pheromone is a function of p-ants' sensations across their trajectories on their associated conspicuity maps. Hence, 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.

The following Section 5.3.4.1 describes the p-ant's creation process, whereas the iteration process is described in Section 5.3.4.2.

5.3.4.1 P-Ant's Creation

The chances of creating a p-ant p_m on a given location $\mathbf{o}_{p_m}$ of the conspicuity map $\mathbf{C}^m$ depends on the level of conspicuity at that location and on the level of pheromone at the same location in the corresponding pheromone map, $\mathbf{P}^m$. Hence, p-ants are progressively and probabilistically deployed where there are more chances of being a trail, under the assumptions that: (1) trails tend to be conspicuous; (2) the trail 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 p-ants deployed in the current frame builds-up mostly around the actual trail's location.

By assuming that trails often start from the bottom of the image, p-ants are deployed with a small randomly selected offset $z \in [0, 0.1 \cdot h]$ of the bottom of the conspicuity map in question, i.e., at row $r \in [h - z]$, where h is the height of the map[1]. This random small offset reduces sensitivity to any noise potentially present at the map's boundaries.

In order to determine the column where p_m is deployed, a uni-dimensional vector $\mathbf{v}^m = (v_0^m, \ldots, v_w^m)$ is first computed. The element v_k^m of $\mathbf{v}^m$ refers to the average conspicuity level of the pixels in a small window centred on column k and with a randomly selected offset from the bottom row of the map, r:

$$v_k^m = \sum_{l,j} \frac{\mathbf{C}^m(l,j)}{\delta_w \cdot \delta_h} \quad \text{with} \quad l \in [k - \delta_w/2, k + \delta_w/2], j \in [r, r - \delta_h], \tag{5.5}$$

where $\mathbf{C}^m(l,j)$ returns the conspicuity level in position (l,j), and δ_w and δ_h are the width and the height of the window, respectively. The same windowing process is applied to build a vector for the pheromone field in question, $\mathbf{u}^m = (u_0^m, \ldots, u_w^m)$. However, this time, element u_k^m corresponds to the maximum pheromone level found in the window:

$$u_k^m = \max\{\mathbf{P}^m(l,j)\}_{l,j} \quad \text{with} \quad l \in [k - \delta_w/2, k + \delta_w/2], j \in [r, r - \delta_h], \tag{5.6}$$

where $\mathbf{P}^m(l,j)$ returns the pheromone level in position (l,j). The max operator is employed to benefit those regions where the paths of p-ants overlap more often and consequently where there is a higher consensus on the trail's skeleton position. Using these two vectors in the following test, which is repeated until it succeeds, the chances of deploying a p-ant in a randomly selected column $z_2 \cdot w$ is as high as the conspicuity and pheromone levels at the deployment region:

$$z_1 < \left(\rho \cdot u_{z_2 \cdot w}^m + (1 - \rho) \cdot v_{z_2 \cdot w}^m\right), \tag{5.7}$$

where $z_1 \in [0,1]$ and $z_2 \in [0,1]$ are numbers sampled from a uniform distribution each time the test is performed and ρ is a weight factor used to trade-off the influence of both pheromone and conspicuity information. By starting with a small value, ρ_0, and by

[1] Rows are indexed in increasing order from the top to the bottom of the map.

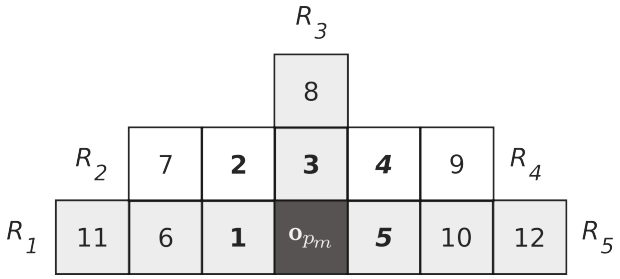


Figure 5.4: P-ants' sensory and action spaces. Regions surrounding current p-ant's position, $\mathbf{o}_{p_m}$, are segmented into a set of receptive fields, $R_1 = \{1,6,11\}, R_2 = \{2,7\}, R_3 = \{3,8\}, R_4 = \{4,9\}, R_5 = \{5,10,12\}$, whose composing pixels are numbered as in the figure. If a given action $a \in A$ is selected, then the next p-ant's position will be the closest pixel to the p-ant, represented by the pixels in bold.

linearly growing at each iteration by an amount $\Delta\rho$, ρ operates as an adaptive process, compelling the system to move from a conspicuity-driven operation (exploration) to a pheromone-driven operation (refinement/exploitation).

5.3.4.2 P-Ant's Execution

Before specifying p-ants behaviours, it is necessary to specify their 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 p-ant's current position, $R_1 \ldots R_5$ (see Fig. 5.4). For a given visual feature m and p-ant's position $\mathbf{o}_{p_m}$, $\mathbf{C}^m(R_k, \mathbf{o}_{p_m})$ and $\mathbf{P}^m(R_k, \mathbf{o}_{p_m})$ return the average conspicuity and pheromone levels of the pixels constituting receptive field R_k, respectively. Parameter $\mathbf{o}_{p_m}$ is used to transform the p-ant's centred receptive field onto the map's frame of reference. To refer directly to the pixel-wise conspicuity and pheromone levels at the p-ant's position, $\mathbf{C}^m(\mathbf{o}_{p_m})$ and $\mathbf{P}^m(\mathbf{o}_{p_m})$ are used, respectively. An action $a \in A$ moves the p-ant to one of the five neighbour pixels not behind the current p-ant's position. The action space is thus defined by the set $A = \{1,2,3,4,5\}$ (see Fig. 5.4).

At each iteration, of a maximum η, p-ant p_m executes a set of behaviours $B = \{greedy, track, centre, ahead, commit\}$, which independently vote on each possible action in A. Following a typical approach of behaviour coordination (Rosenblatt, 1995), the most voted action is the one taken by the p-ant. Table 5.1 summarises, for each behaviour, which regions in the neighbourhood of the p-ant are associated to the most preferred action.

In order to allow the system to operate with unstructured trails, these behaviours are simple and make little assumptions regarding the trail's structure. Each behaviour exploits a specific trail's shape or appearance a priori knowledge in order to partially contribute to the goal of producing a p-ant's trajectory representative of the trail skeleton. For instance, under the assumption that trails are somewhat monotonous structures, p-ants should move under the influence of some inertia. This is implemented by having the *commit* behaviour voting more strongly on the action that is most similar to the one selected in the previous iteration.

Behaviour	**Voting Preferences**
greedy	Regions of higher levels of conspicuity, under the assumption that trails are salient in the input image.
track	Regions whose average level of conspicuity is more similar to the average level of conspicuity of all the pixels visited by the p-ant, under the assumption that trails' appearance is homogeneous.
centre	Regions that maintain the p-ant equidistant to the boundaries of the trail hypothesis being pursued. That is, the p-ant will prefer regions that are closer to the centroid $\mathbf{x}_{p_m}$ of the horizontal segment in the conspicuity map where the current p-ant is. The segment is obtained by considering all pixels, represented by the set S_{p_m}, that are connected to the p-ant's current position, $\mathbf{o}_{p_m}$, through a set of pixels sharing the same row and similar, within a given margin ζ, conspicuity level of the former. See Fig. 5.5 for an illustration.
ahead	Upwards regions under the assumption that trails are often vertically elongated.
commit	Region targeted by the motor action at the previous iteration, under the assumption that trails' orientation tend to be monotonous.

Table 5.1: P-Ant behaviours for trail detection.

Formally, for a given p-ant p_m, behaviours are described as functions that return a vote in the interval $[0,1]$ for each possible action $a \in A$:

$$f_{greedy}(p_m, a) = \mathbf{C}^m(R_a, \mathbf{o}_{p_m}), \tag{5.8}$$

$$f_{track}(p_m, a) = 1 - \left| \mathbf{C}^m(R_a, \mathbf{o}_{p_m}) - \sum_{v \in V_{p_m}} \frac{v}{|V_{p_m}|} \right|, \tag{5.9}$$

$$f_{centre}(p_m, a) = \left| d_{p_m} \cdot \left(\frac{6 \cdot \mathcal{H}(-d_{p_m}) - a}{5} \right) \right|, \tag{5.10}$$

$$f_{ahead}(p_m, a) = 1 - \frac{|3 - a|}{2}, \tag{5.11}$$

$$f_{commit}(p_m, a) = 1 - \frac{\left| a'_{p_m} - a \right|}{4}, \tag{5.12}$$

where $\mathcal{H}(\cdot)$ is the Heaviside function, a'_{p_m} is the p-ant's action selected in the previous iteration (see Algorithm 6), V_{p_m} is a list whose elements are scalars representing the conspicuity level at each p-ant's previously visited position, and d_{p_m} is the normalised deviation of $\mathbf{o}_{p_m}$ to centroid $\mathbf{x}_{p_m}$, $d_{p_m} = \frac{col(\mathbf{o}_{p_m}) - col(\mathbf{x}_{p_m})}{|S_{p_m}|}$, with $col(\cdot)$ returning the column coordinate of a given map position (see third row of Table 5.1 and Fig. 5.5).

As will be shown, all these behaviours contribute for p-ants trajectories that closely represent the trail's skeleton. The absence of an explicit scoring function, which would require a model-based imposition of constraints on the trail's shape, hampers a post-ranking of all deployed p-ants to determine the "best trajectory". Moreover, not all p-ants will be deployed on the trail and so not all are able to follow the actual trail. To overcome these challenges two ingredients of the system are determinant.

The first ingredient comes in the form of positive feedback raising from the amplification of random fluctuations. With additive random fluctuations at p-ants actuation level, those that are deployed off the trail will diverge, whereas p-ants deployed on the trail will converge towards its vanishing point, thanks to the *centre* behaviour. Hence, there will be higher concentrations of pheromone on trail regions. This happens because the presence of the trail tends to be a global constraint which is only felt by the p-ants deployed on it. In a sense, the trail operates as an attractor for the self-organising system.

The second ingredient is the use of stigmergy in the form of pheromone-based interactions. By making p-ants attracted by high pheromone concentration regions, we positively reinforce the difference between diverging and converging p-ants (symmetry breaking). Hence, this second ingredient ensures that, along time, the structure imposed by the presence of the trail on the *centre* behaviour is stronger than the effects of random fluctuations. This effect is magnified by the fact that p-ants are deployed

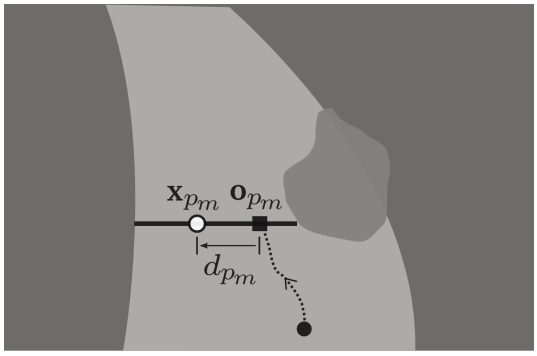


Figure 5.5: Illustrating example of key aspects of the *centre* behaviour. The dotted line represents the p-ant's motion since the first iteration until the current one. The pixels composing the thicker horizontal line define the set S. The p-ant will try to approach this line's centroid $\mathbf{x}_{p_m}$, represented by the white filled circle, which is deviated from the current agent's position, $\mathbf{o}_{p_m}$, by $|d_{p_m}|$ pixels.

according to the level of pheromone already present in the pheromone maps. Moreover, the fact that robot forward motion tends to make the neural field skew towards the bottom of the image allows regions of highest activity in further regions of the visual field more likely to invoke p-ants. The use of pheromone-based interactions has the additional advantage of overcoming the brittleness of controlling p-ants based on myopic behaviours. The local interruption of a trail, that could inhibit the *centre* behaviour from properly leading the p-ant along the trail, is overcome by having p-ants progressively building a pheromone "bridge" over the interruption thanks to *commit* and *ahead* behaviours.

In order to take these considerations into account, in each iteration a p-ant p_m selects its action a_{p_m} by maximising the following utility function, which incorporates behaviours' votes, pheromone-based interactions, and random fluctuations:

$$a_{p_m} = \arg\max_{a \in A} \left(\sum_{b \in B} \alpha_b f_b(p_m, a) + \mathbf{P}^m(R_a, \mathbf{o}_{p_m}) + \gamma q \right), \tag{5.13}$$

where α_b is a user defined weight accounting for the contribution of behaviour $b \in B$ 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. To match the randomness magnitude with the scale of the image, which is typically smaller for pixels in upper regions of the image, the weight γ starts with an initial value γ_0 and exponentially decays by a constant factor γ_τ at each iteration.

In case an immediate loop is detected, namely, the p-ant moving recurrently from one pixel to another, then the action for the current iteration is randomly selected. Finally, the p-ant's position $\mathbf{o}_{p_m}$ is updated according to the selected action. Algorithm 7 outlines the overall iteration process.

Algorithm 7: execute$(\cdot)$, p-ant's execution pseudo-code.

Input: p-ant (p), conspicuity map ($\mathbf{C}^m$), pheromone map ($\mathbf{P}^m$), pheromone map of other visual feature ($\mathbf{P}^{m'}$), pheromone map without cross-modality and neural field influences ($\mathbf{P}^m_*$)
Output: updated pheromone maps, $\mathbf{P}^m$, $\mathbf{P}^{m'}$, and $\mathbf{P}^m_*$
Data: $\eta, \varepsilon, \upsilon$ are empirically defined constants.

$a'_{p_m} \leftarrow 3$ *// default previously selected action is forward motion*
initialise list whose elements will be scalars representing the conspicuity level at each p-ant's visited position, $V_{p_m} \leftarrow \emptyset$
for η iterations **do** [execute p-ant p for η times]
    use Equation 5.13 to obtain p-ant's action, a_{p_m}, based on $\mathbf{C}^m$, $\mathbf{P}^m$, V_{p_m}, and a'_{p_m}
    append conspicuity of the new p-ant's position, $V_{p_m} \leftarrow V_{p_m} \bigcup \{\mathbf{C}^m(R_{a_{p_m}}, \mathbf{o}_{p_m})\}$
    use obtained p-ant's action, a_{p_m}, to update p-ant's position, $\mathbf{o}_{p_m}$
    update $\mathbf{P}^m$ at pixel $\mathbf{o}_{p_m}$, $\mathbf{P}^m(\mathbf{o}_{p_m}) \leftarrow \mathbf{P}^m(\mathbf{o}_{p_m}) + \varepsilon$
    update $\mathbf{P}^m_*$ at pixel $\mathbf{o}_{p_m}$, $\mathbf{P}^m_*(\mathbf{o}_{p_m}) \leftarrow \mathbf{P}^m_*(\mathbf{o}_{p_m}) + \varepsilon$
    update $\mathbf{P}^{m'}$ at pixel $\mathbf{o}_{p_m}$, $\mathbf{P}^{m'}(\mathbf{o}_{p_m}) \leftarrow \mathbf{P}^{m'}(\mathbf{o}_{p_m}) + \upsilon \cdot \varepsilon$
    store selected action, $a'_{p_m} \leftarrow a_{p_m}$
end
return $(\mathbf{P}^m, \mathbf{P}^{m'}, \mathbf{P}^m_*)$

5.3.5 Evidence Accumulation

Once p-ants' activity has ceased, the instantaneous saliency map, $\mathbf{S}$, feeds a 2-D dynamic neural field $\mathbf{F}$ (Amari, 1977; Rougier and Vitay, 2006), which is a lattice of laterally connected neurons. Its goal is to integrate evidence across time, to consider competition between multiple focus of attention, and to promote perceptual grouping.

The dynamical characteristic of the neural fields, displayed in the form of inertia, is the key element that enables information to be integrated across time. However, if not properly handled, this property causes the field to smear when the robot moves.

A way of avoiding this undesirable effect is to shift the neural field's activity according to the robot motion estimate by using asymmetrical kernels in the neurons (Zhang, 1996). However, the neural field being a representation of the environment through a projection process and considering that the robot may incur in both rotation and translation, it is more straightforward and consequently effective to affect the neural field's activity directly, i.e., to consider the neural field's state as an image to be transformed with a warping operator. In this line, the following three steps explicitly compensate the neural field for the camera motion engaged between the previous and current frames (see Algorithm 6):

1. Estimate the homography matrix $\mathbf{H}$ that describes the projective transformation between the current frame, $\mathbf{I}$, and the previous one, $\mathbf{I}'$. This step is detailed in Section 5.3.5.1.

2. Obtain a motion compensated version of the previous neural field's state by using the estimated homography matrix, $\mathbf{F} \leftarrow \mathbf{HF}$.

3. Update $\mathbf{F}$ with the pheromone map $\mathbf{S}$. This step is detailed in Section 5.3.5.2.

5.3.5.1 Homography Matrix Estimation

To estimate the projective transformation $\mathbf{H}$, a set of corner points (Tomasi and Shi, 1994) is first detected in the previous frame, $\mathbf{I}'$. These points are then tracked in the current frame, $\mathbf{I}$, with a pyramidal implementation of the Lucas-Kanade feature tracker (Bouguet, 1999). The resulting sparse optical flow is then used to estimate the projective transformation relating both frames, i.e., the 3×3 homography matrix $\mathbf{H}$, such that:

$$\mathbf{u}'_i = \mathbf{H}\mathbf{u}_i, \tag{5.14}$$

where $\mathbf{u}_i$ is a corner point found in $\mathbf{I}$ and $\mathbf{u}'_i$ its correspondence in $\mathbf{I}'$. Due to noise in the tracking process, the homography matrix is calculated as the least-squares solution that minimises the back-projection error (Bradski and Kaehler, 2008). 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-satisfaction of these two relaxed constraints is sufficient to maintain a robust operation. If a minimum of four correspondences between corner points is not found, the homography matrix is set to the identity matrix, $\mathbf{H} = \mathrm{diag}(1,1,1)$.

5.3.5.2 Neural Field Update

The neural field $\mathbf{F}$ is a 2-D lattice of $w \times h$ neurons, each one corresponding to one pixel of the saliency map. The neurons have "Mexican-hat"-shaped lateral coupling, implemented by Difference of Gaussians (DoG). This inter-neuron coupling helps in the formation of a coherent focus of attention (Rougier and Vitay, 2006). 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 $w(\mathbf{x},\mathbf{x}')$ between a neuron in position $\mathbf{x}$ and a neuron in position $\mathbf{x}'$ is given by a DoG function of the Euclidean distance between both. In addition to lateral connectivity, the neural field also has afferent interactions with pheromone field $\mathbf{S}$. The weight $d(\mathbf{x},\mathbf{y})$ of a connection between an element of $\mathbf{S}$ in position $\mathbf{y}$ and a neuron of $\mathbf{F}$ in position $\mathbf{x}$ is given by a Gaussian function of the Euclidean distance between both. This operation enlarges neurons' receptive field to reduce sensitivity to noise.

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

$$\begin{aligned}\tau\frac{\partial \mathbf{F}(\mathbf{x},t)}{\partial t} = &-\mathbf{F}(\mathbf{x},t)+\\ &\int w(\mathbf{x},\mathbf{x}')\, f\left(\mathbf{F}(\mathbf{x}',t)\right)\mathrm{d}\mathbf{x}'+\\ &\int d(\mathbf{x},\mathbf{y})\mathbf{S}(\mathbf{y},t)\mathrm{d}\mathbf{y}+\psi,\end{aligned} \tag{5.15}$$

where $f(x) = x$ in this paper, τ is a time constant and $\psi = 0$ is the neuron threshold. For numerical integration, and assuming a delay between consecutive frames, the Euler forward method is used to obtain an approximation of the neural field, which in matrix

form results in the rearranged expression

$$\mathbf{F} \leftarrow \mathbf{F} + \tau\Big(-a \cdot \mathbf{F} + b \cdot \big(DoG^{k_1,k_2}_{\sigma_1,\sigma_2} * \mathbf{F}\big) + c \cdot \big(G^{k_3}_{\sigma_3} * \mathbf{S}\big)\Big), \tag{5.16}$$

where: $*$ is the convolution operator; a, b, and c are empirically defined weights specifying the contribution of each term; and $DoG^{k_1,k_2}_{\sigma_1,\sigma_2} = G^{k_1}_{\sigma_1} - G^{k_2}_{\sigma_2}$, with G^{k}_{σ} as a Gaussian kernel of size $k \times k$ and width σ.

This matrix-based formulation of the neural field results in a synchronous update policy. That is, all neurons are updated based on the previous frame's network state. The problem with this update policy is its potential to induce undesirable activity oscillations in face of symmetries at the sensory input. However, due to robot motion, these singular configurations are unlikely to occur during a relevant amount 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. 5.6). That is, focus of attention's spatio-temporal competition is an implicit property of the system.

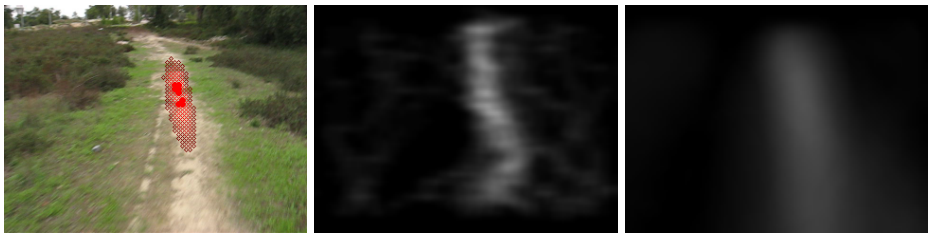
5.4 Experimental Results

5.4.1 Experimental Setup

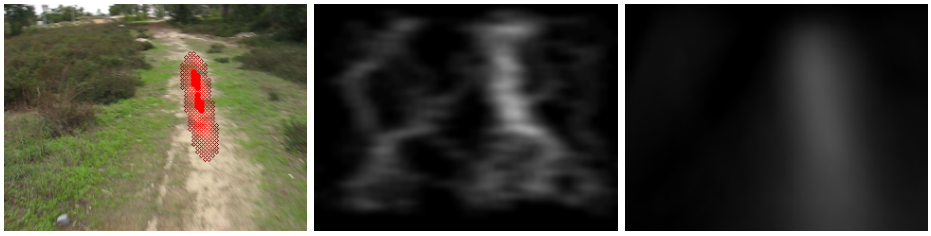
An extensive data set of 25 colour videos, encompassing a total of 12023 frames with a resolution of 640×480, has been obtained with a hand-handled camera (see Fig. 5.7). Experimental results were obtained running the model off-line. The camera was carried at an approximate height of 1.5 m, at an approximate speed of $1\,\mathrm{ms}^{-1}$. To validate the ability of the model to bring the robot back on trail, the camera was sporadically moved off trail. The considerable level of induced camera oscillations, typical in off-road robots, generated blurred and consequently noisy images[2].

The trail detector was implemented without code optimisation, and tested in a Core2 Duo 2.8 GHz, running Linux, and using OpenCV (Bradski and Kaehler, 2008) for low-level routines.

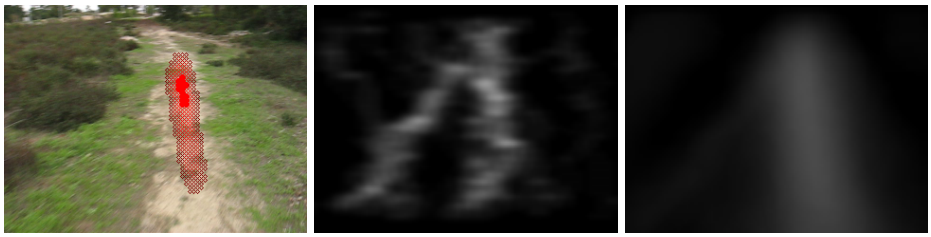
[2] Videos with the proposed model's output are available in the authors' website (Santana et al., 2010c)



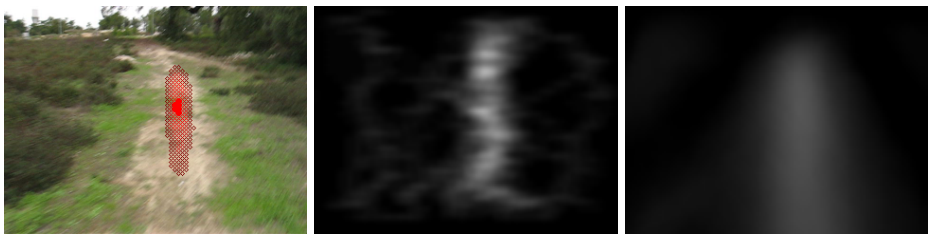
(a) $t = 190$ (b) $\mathbf{S}(190)$ (c) $\mathbf{F}(190)$



(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) $\mathbf{F}(280)$

Figure 5.6: Example of neural field competition in a situation represented by four ordered frames obtained from video 11 of the tested data set (see Fig. 5.7). Each row includes the saliency map (centre), **S**, and the neural field (right), **F**, for a given input image (left). The redness of the blobs overlaid in the input images corresponds to the activity level of the neural field above 0.85, representing the model's estimate of the trail's location. The trail is visible for several frames prior to frame 220, thus eliciting high level of activity in the neural field at frame 190. Although the transient appearance of a trail-like grass segment in the bottom-left region of the input image is felt in the pheromone field between frames 220 and 250, this distractor is actively inhibited in the neural field. The outcome is that the system's output, i.e., the red overlay on the input image, steadily segments the trail from the background.

Figure 5.7: 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 redness of the blobs overlaid in the images correspond to the activity level of the neural field above 0.85, representing the model's estimate of the trail's location.

The following describes the model's parameterisation used in the experiments. The number of p-ants per map, n, has been empirically defined to 20. A smaller number may not ensure convergence, whereas a larger one did not exhibit considerable improvement in the tested data set. The same reasoning applies to the number of iterations applied to each p-ant, η, which has been set to 50. The amount of pheromone deposited by a given p-ant on its associated pheromone map (see Section 5.3.4), ε, has been set to 0.008.

The small portion of ε deposited in the other pheromone map (see Section 5.3.4), υ, has been set to 0.3. These values should not be set too high to avoid pheromone saturation, which would inhibit the emergence of collective behaviour. The ratio of the

robot motion compensated neural field used to initialise the pheromone maps at the onset of each frame (see Section 5.3.2) has been set to $\lambda = 0.1$.

The contribution of each behaviour is $\alpha_{greedy} = 0.45$, $\alpha_{track} = 0.35$, $\alpha_{centre} = 0.10$, $\alpha_{ahead} = 0.05$, and $\alpha_{commit} = 0.05$. By making $\alpha_{greedy} > \alpha_{centre} + \alpha_{ahead} + \alpha_{commit}$, we ensure that p-ants exploit more strongly the conspicuity cue than the a priori knowledge on the expected trail's shape. This follows from the observation that trails in natural environments are highly unstructured, still, some a priori knowledge helps p-ants from being stuck or deviated by distractors in the environment. With a relatively high α_{track}, the collective influences considerably the individual to further reduce the problems associated with noise and distractors. Note however that making $\alpha_{track} < \alpha_{greedy} + \alpha_{centre} + \alpha_{ahead} + \alpha_{commit}$ is important to ensure that self-organisation occurs under exogenous influence; otherwise, a non-situated consensus among p-ants could be reached.

The margin ζ used to centre the p-ant on the segment it is currently on has been set to 0.06. The width, δ_w, and the height, δ_h, of the window used to create p-ants in Section 5.3.4.1 have been set to 10 and 20, respectively. The initial values of the random factor ρ (see Equation 5.7), ρ_0, and its increment at each iteration, $\Delta\rho$, have been set to 0.3 and 0.02, respectively. The initial values of the random factor γ (see Equation 5.13), γ_0, and the rate of its exponential decay at each iteration, γ_τ, have been set to 0.4 and 0.02, respectively.

The neural field free parameters (see Section 5.3.5.2) 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 $\tau = 0.03$. The system showed robustness to small variations around these values as long as the proportions are roughly maintained.

5.4.2 Results

This section presents quantitative results obtained with the proposed model in the presented data set. The trail is considered correctly detected if the biggest blob of neural field activity above 0.85 is fully localised within the trail's boundaries. In case of ambiguity, i.e., the co-occurrence of two similar blobs, the neural field, $\mathbf{F}$, is used to assess which blob is being reinforced, and consequently should be taken as the output. Fig. 5.7 shows key frames in each tested video, where it is possible o see that besides localising the trail, the red blob is also clearly oriented with the trail. Note however

	Classical model $\mathbf{S} \leftarrow \frac{1}{2}(\mathbf{C}^I) + \frac{1}{2}(\mathbf{C}^C)$	Proposed model $\mathbf{S} \leftarrow \frac{1}{2}(\mathbf{P}^I) + \frac{1}{2}(\mathbf{P}^C)$
Aggregate detection rate [%]	23.97 ± 27.73	91.32 ± 11.39

Table 5.2: Aggregate trail detection results in the 25 videos data set. Aggregate detection rate (mean $\pm$ standard deviation) computed as the average of the detection rates obtained per video. Refer to Table 5.4 in Appendix A for details.

that, for an automatic system it may be more useful to use directly the neural field activity level. A typical use would be to consider that the higher the activity in the neural field the higher the traversability of the terrain. The obtained results support this possibility - see the high correlation between neural field activity and trail location in Fig. 5.2, Fig. 5.6, and Fig. 5.10.

Fig. 5.8 shows a set of input images obtained from Rasmussen et al. (2009) as well as their corresponding conspicuity maps, computed with the scaling function proposed in Section 5.3.3.2. It is observable that at least one of the conspicuity maps tends to segment the trail from the background. However, the trail is not always represented by the most conspicuous segment. In these cases, the typical blend of conspicuity maps used to produce the final saliency map, from which the object of interest is directly captured (Itti et al., 1998; Frintrop et al., 2005), is prone to fail. The first hypothesis being tested in this work is that the intermediate pheromone maps are able to introduce the required added value to ensure robust operation in these situations. Table 5.2 confirms this hypothesis. That is, in the tested data set, the proposed swarm-based saliency model, $\mathbf{S} \leftarrow \frac{1}{2}(\mathbf{P}^I) + \frac{1}{2}(\mathbf{P}^C)$, predicts the trail location 3.8 times more than a classical saliency model based only on the conspicuity maps, $\mathbf{S} \leftarrow \frac{1}{2}(\mathbf{C}^I) + \frac{1}{2}(\mathbf{C}^C)$. For the sake of fair comparison, the neural field $\mathbf{F}$, which is fed by $\mathbf{S}$, is used to generate the output in both cases. To handle the probabilistic nature of the p-ants behaviours, the results refer to the average of 5 runs performed in each video.

The second hypothesis being assessed is whether the proposed model exhibits accuracy and computational efficiency enough to ensure robust trail following in off-road environments, which is confirmed by the average success rate of 91.32 % (see Table 5.2) and by the 20 Hz operation (see Table 5.3). Remarkably, the swarm-based pheromone maps computation, which is the only trail-specific operation, takes only 2 ms on average per frame. The timing reported for the neural field update also includes optical flow computation, homography estimation and neural field wrapping.

Figure 5.8: Intensity, $\mathbf{C}^I$ (middle row), and colour, $\mathbf{C}^C$ (bottom row), conspicuity maps computed with the scaling function proposed in Section 5.3.3.2, for a set of images (top row) obtained from Rasmussen et al. (2009). These figures show that the trail is usually conspicuous and segmented in at least one of the conspicuity maps.

	Neural field	Conspicuity maps computation	Pheromone maps computation	Total
Time [ms]	12	36	2	50

Table 5.3: Computation time.

These results are more stringent if the difficulty of the tested data set is taken into account. To the best of our knowledge no previous work has been tested against a data set with trails simultaneously as narrow, unstructured, and discontinuous as the ones herein considered. Moreover, differently from previous work, the model succeeds in situations where the trail is not starting from the bottom of the image (e.g., Fig. 5.1(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, composed by more than 2800 frames. Along this video, besides being often interrupted and highly unstructured, the trail also exhibits a variable width. Moreover, the terrain surrounding the trail is also 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. Failure in

5% of the experiments refer to situations where the trail is noticeable in the neural field, but with similar activation of non-trail spots. In these cases, as in other lower performance videos, ambiguity between trail and surroundings could be reduced by considering additional perceptual modalities, such as texture and depth. Nevertheless, it should be noticed that many of the observed failures are in part caused by extreme camera rotations, which hampers a proper neural field's motion compensation. Without a proper motion compensation in these transient situations, the neural field provides p-ants with misguiding information on the most likely trail location. The result is a strong competition between p-ants on and off the trail, i.e., between evidence and conspicuity information, which lasts for a few frames until symmetry is broken.

When the trail is highly conspicuous in the environment, as most often occurs, ambiguity is rarely present in the system's output. When this assumption fails, and distractors are scattered, the model is still be able to frequently perform correctly, as demonstrated by the quantitative results. This robustness owes to the synergistic operation between neural field inertia and p-ants' sensorimotor coordination capabilities, which allow an opportunistic exploitation of the trail-background prioritised segmentation present in the conspicuity maps.

In a second experiment, the proposed model was tested on an additional set of 15 images (see Fig. 5.9), employed by Rasmussen et al. (2009) to assess the ability of their model to perform without the support of 3-D data. In this second data set, our model only fails to determine the trail location in the image depicted in the bottom-right of Fig. 5.9. The cause for this failure is that without a symmetry constraint it is impossible to define as non-trail the region signalled by the system. This could be easily overcome by explicitly adding a *symmetry* behaviour to the p-ants. However, the more specialised the system gets, the less robust it is in the face of unforeseen trails. Fig. 5.10 depicts the successful output of the proposed model in a situation where the ability to operate without hard assumptions on the trail's shape or appearance is key.

The paths present in these images being well structured, uninterrupted, and large, are aligned with the assumptions made by Rasmussen et al.. That is, these paths have a triangular shape and good contrast symmetric borders. As a result, the model-based work of Rasmussen et al. successfully determines the location of the path in all images. However, being model-free and able to exploit both local and global contrast, as is the case of our model, is important to properly handle less structured paths, such as the ones considered in the first data set (see Fig. 5.7) or in situations as the one in Fig. 5.10.

Figure 5.9: Results of the proposed model on the 15 images data set obtained from (Rasmussen et al., 2009). The red blob corresponds to the proposed model's output after 50 frames running on the same image. This high number of frames aims at demonstrating that the system actually converges towards a good solution. The redness of the blobs overlaid in the images correspond to the activity level of the neural field above 0.85, representing the model's estimate of the trail location.

(a) Input image, $\mathbf{I}$ (b) Colour conspicuity map, $\mathbf{C}^C$ (c) Neural field, $\mathbf{F}$

Figure 5.10: Proposed model operating on an image of trail with non-monotonous shape and not starting from the bottom, taken from a high vantage point. The state of the neural field after 50 frames running on the same image clearly shows the ability of the model to stably determine the location of the trail.

5.5 Conclusions

This paper proposed a swarm-based visual saliency model capable of embedding a priori knowledge on the on the overall layout of the object being sought. The model

was shown to perform well and fast on the difficult task of localising unstructured trails in natural environments. In particular, the model exhibited 91 % of success rate at 20 Hz against a highly demanding and heterogeneous data set.

These results owe to large extent to the multi-agent design, which enables a robust self-organisation of visual search, perceptual grouping, and multiple hypotheses tracking. In other words, the resulting system solves a complex problem in a bottom-up way. The first outcome is a low computational footprint, which in turn promotes faster robot motion. The second outcome is that, being purely bottom-up, the system does not rely on explicit internal representations, i.e., a priori models, such as the "triangle" concept. Besides fostering robustness in unstructured environments, this is also an important asset if trail detection is expected to emerge as a product of incremental self-supervised learning. In this line, an additional advantage of being purely bottom-up is that the system's design space is restricted to the simple perception-action rules, i.e., behaviours, encapsulated in the set of homogeneous agents. Given their simplicity, it should be possible to substitute these rules by simple recurrent neural networks, thus confining the design space to a small set of weights. Hence, the proposed approach paves the way to feasible learning of complex skills.

Another key and novel aspect of the model is that by using visual saliency both local and global cues on the trail location are naturally exploited. That is, saliency maps provide contrast information not only between the trail and its surroundings, but also between the trail and the overall scene. Typically, this is the case because the appearance of the materials composing trails (e.g., soil) differ from the appearance of the materials composing its surroundings (e.g., grass) and the remainder elements present in overall scene (e.g., trees, sky). The complementarity of both local and global contrast information results in a robust handling of less structured trails.

The high success rate across the heterogeneous data set shows that the selected parameterisation is not overfit to a specific environment, highlighting its robustness. Nonetheless, an automatic p-ant's behaviour generation and parameterisation is desirable, and will be exploited in future work. Furthermore, on-line specialisation and generalisation of p-ants' behaviour can also be addressed for improved performance and robustness, respectively.

Other perceptual modalities, such as texture and depth, and even alternative conspicuity computation methods can also be considered. Additionally, we plan to test the swarm-based saliency model on other visual search tasks. Related to this, the hy-

pothesis partially analysed in this article regarding the usefulness of visual saliency as a general purpose prioritised image segmentation process should be further validated.

A limitation of the proposed model is its ability to handle bifurcations in the trail. To handle this limitation we expect to explicitly track multiple blobs of activity in the neural field. The selection of which blob to track can be hinted by feedback of the robot's action selection process. This approach has been explored by us in the context of swarm-based obstacle detection (Santana and Correia, 2010a). The parallel nature of conspicuity maps, neural field, and swarm-based saliency will be exploited in future implementations on parallel hardware, such as Graphics Processing Units (GPU).

Finally, the obtained results add to previous evidence on the usefulness of visual attention for the control of off-road robots (Santana and Correia, 2010a, 2011; Santana et al., 2011). The proposed method also contributes to the emerging swarm cognition field, which attempts to uncover the basic principles of cognition, i.e., adaptive behaviour, recurring to self-organising principles, mainly those exhibited by social insects (Santana and Correia, 2010a, 2011; Trianni and Tuci, 2010).

Acknowledgements

This work was partially supported by FCT/MCTES grant No. SFRH/BD/27305/2006. We also acknowledge useful comments and support provided by our colleague Magno Guedes. We also want to thank Vítor Matos, from University of Minho, for his input on neural field's activity shift aspects.

5.6 Appendix A: Results Detailed

Video ID	Nr. of frames	Classical model detection rate [%]	Proposed model detection rate [%]
1	278	44.60	100.00 ± 0.00
2	204	61.76	100.00 ± 0.00
3	422	4.74	85.88 ± 2.24
4	135	0.00	100.00 ± 0.00
5	2854	32.48	86.10 ± 0.87
6	186	27.96	99.89 ± 0.24
7	121	0.00	100.00 ± 0.00
8	124	0.00	100.00 ± 0.00
9	309	18.77	89.77 ± 1.46
10	147	49.66	94.15 ± 0.78
11	386	0.00	100.00 ± 0.00
12	158	0.00	68.48 ± 9.94
13	134	40.30	98.96 ± 1.25
14	676	44.23	99.05 ± 0.08
15	683	26.50	81.93 ± 1.23
16	770	4.55	76.96 ± 1.55
17	403	34.99	94.39 ± 0.92
18	335	97.01	99.04 ± 0.13
19	230	84.78	97.91 ± 0.99
20	439	6.38	55.63 ± 0.37
21	490	3.67	97.88 ± 0.47
22	230	10.87	100.00 ± 0.00
23	600	6.00	93.37 ± 0.76
24	802	0.00	85.26 ± 0.88
25	907	0.00	78.32 ± 1.01

Table 5.4: Detailed trail detection results obtained with a classical model, $\mathbf{S} \leftarrow \frac{1}{2}(\mathbf{C}^I) + \frac{1}{2}(\mathbf{C}^C)$, and with the proposed model, $\mathbf{S} \leftarrow \frac{1}{2}(\mathbf{P}^I) + \frac{1}{2}(\mathbf{P}^C)$, in the 25 videos data set. To handle the probabilistic nature of the p-ants behaviours, the results for the proposed model (mean $\pm$ standard deviation) refer to the average of 5 runs performed in each video.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis is a pioneering work on the use of visual attention and swarm cognition for the problems of robotic terrain perception and local navigation on off-road environments. The experimental validation of the proposed models bears phenomenological support to the two main hypotheses raised in this thesis: (1) parsimony, robustness, and performance of off-road robots is improved if visual attention mechanisms are employed in their control systems and (2) the synthesis of self-organising robot cognitive behaviour is facilitated if the social insects collective behaviour metaphor is used as design pattern.

The positive results obtained with visual attention for obstacle and trail detection, as well as for local navigation, confirm the first hypothesis. The positive results obtained from the extensive use of swarm models for trail detection and local navigation also bear support for the second hypothesis. These models rely heavily on self-organising properties, which endow the system with robustness in the face of unforeseen situations and allow complex solutions to emerge from the interaction of simple, and thus fast to compute, elements.

Different from typical swarm and adaptive behaviour related research, the proposed models were validated on "dirty" real life problems and assessed from an engineering performance-oriented perspective, as is typically done in robotics and computer vision related research. Although this approach is a limited theoretical effort, it permits a better analysis in context. This methodological option had the additional

benefit of generating the following by-products of the main research questions, which are important contributions themselves:

- A method for the detection of obstacles with more than 10 cm height up to a range of 20 m on uneven terrain at 5 Hz. Owing greatly to the extensive use of visual attention mechanisms, a considerable performance improvement over well-known alternative obstacle detection techniques was attained.

- A method for local navigation of stereo-based off-road robots, which owing to its self-organising visual attention mechanism only requires an average of 0.9 % of the visual input to be analysed. The system, which includes perception, visual odometry, mapping, and action selection, performs at 10 Hz, which is sufficient for typical off-road robots.

- A method for detection of trails in natural environments with the ability of exhibiting 91 % success rate at 20 Hz. Relying on self-organising visual saliency computation, the method was shown 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. This means that the model is well fit to find trails and not only to follow them.

Finally, from all these results it is possible to conclude that this thesis contributed to: (1) make field robots more parsimonious and more robust, and thus, closer to being truly useful in economically and socially important tasks; (2) boost swarm cognition as a tool for the synthesis of embodied cognitive agents.

Part of the work described in this thesis was one of the contributions to the first workshop in swarm cognition (Santana and Correia, 2009) and to a journal special issue dedicated to the topic (Santana and Correia, 2011). Moreover, this thesis describes one of the first developments on a practical application of swarm cognition and, to our best knowledge, the first in use on an embodied autonomous system.

6.2 Future Work

Motivated by the good results obtained with the proposed pioneer work on the synthesis of embodied swarm cognition, we plan to further explore the theme in both

theoretical and experimental domains. This is a recent paradigm whose potential to substitute processing previously done by others, such as neural networks, is substantial but still demanding further validation. An important research avenue refers to the evaluation of which situations the swarm cognition paradigm is advantageous over others. This is a broad question that may elicit different answers depending on the application, i.e, as an adaptive behaviour synthesis or modelling tool.

It is also important to further understand which metaphors from the social insects collective behaviour domain, in addition to ant foraging, are useful in addressing the different facets of cognition. Moreover, a design methodology still needs to be developed. This methodology should include a systematic way of: (1) decomposing an arbitrary cognitive behaviour onto a set of homogeneous agents; (2) integrating adequate sensorimotor coordination mechanisms on them; (3) selecting the social insects metaphor that best fits the problem at hand and migrating its self-organising properties to the population of agents; (4) quantifying and qualifying the resulting cognitive behaviour.

An advantage of considering swarm-based solutions is the potential to exploit computational parallel technologies. That is, the computational models being already parallel by design, match directly with a parallel architecture. As a consequence, a potential future work is the integration of all contributions into a single framework capable of running on a parallel machine, such as a GPU.

The system's overall behaviour being bottom-up emerges from the interaction of simple, homogeneous, elements. Hence, this approach may enable learning to occur in a well specified low dimensional space. In this line, the way the behaviour of each p-ant can be learned from experience will be pursued. A fitness function to be studied in the case of the local navigation problem is the inverse of the action selection's time to convergence. This would drive the learning process to generate attention policies capable of operating on a by-need basis. In the face of slower action selection mechanisms, the perceptual process cannot benefit from a feedback as frequent as the one considered in this thesis. In this case, p-ants should be able to infer, from learning, the action selection's feedback between updates. Conversely, when the time available for p-ants to operate is highly constrained, these could cast estimates of obstacles locations, also from learning, given what they have found meanwhile.

In the trail detection model p-ants should also be able to learn objects co-occurrence contextual information to guide their operation. For instance, the presence of some-

thing similar to a "tree" might suggest that trails are lower on the image. de Croon and Postma (2007) uses this type of a priori knowledge to guide an individual agent in the task of object detection, but such an approach has not been used in a swarm-based model.

In an integrated swarm solution, obstacle detection can guide trail detection p-ants towards feasible regions of the image space and vice-versa. A rough 3-D structure of the scene can also be obtained from monocular cues (Michels et al., 2005; Saxena et al., 2008), and from that, one can infer the trail's most likely location. The other way around should also be possible: the trail detector's output is another monocular cue that can be used to infer the 3-D structure of the scene.

The 3-D structure inferred from monocular cues can be used to overcome the limitations of stereo-vision in texture-less surfaces and poor lighting conditions (Saxena et al., 2008). Such an improvement is important for the stereo-based developments of this thesis, such as the saliency-based hybrid obstacle detector presented in Chapter 2. The developments carried out in this detector will be migrated to the swarm-based local navigation model presented in Chapter 4. A possibility is to move from a single to a couple of coordinated swarms, one composed of p-ants applying the small obstacle detector and another composed of p-ants applying the large obstacle detector. The computational cost of each detector must be reflected in the cardinality of the corresponding swarm. To verify whether the swarm-based approach for trail detection generalises to other perceptual tasks, a third swarm capable of determining which detected obstacles are false positives (e.g., tall and dense vegetation) can be introduced.

6.3 Dissemination

This section lists the set of publications related to the work described in this dissertation (see Section 6.3.1) and additional work carried out by the author during the PhD period (see Section 6.3.2), mostly in the context of the project AMI-02 for the Portuguese Ministry of Defence. This project aimed at the development of an autonomous robot to support humanitarian demining operations. The author was responsible for the design and development of the autonomous robot (see Fig. 1.1), which was also used for the PhD experimental work.

The work developed by the author will be continued in three different projects. One is an accepted UE-FP7-ECHORD project, whose goal is to develop an aerial-aquatic

robotic team for monitoring fluvial environments. Another is a national funded QREN ongoing project for the development of off-road surveillance robots. In a third national funded QREN project we expect to exploit swarm cognition models in the context of surveillance sensor networks.

6.3.1 Publications

Journals

- P. Santana, M. Guedes, L. Correia, and J. Barata. Stereo-based all-terrain obstacle detection using visual saliency. *Journal of Field Robotics*, 28(2):241-263, 2011.
- P. Santana and L. Correia. Swarm cognition on off-road autonomous robots. *Swarm Intelligence*, 5(1):45-72, 2011.
- P. Santana and L. Correia. A swarm cognition realization of attention, action selection and spatial memory. *Adaptive Behavior*, 18(5):428-447, 2010.

Submitted Journals

- P. Santana, N. Alves, L. Correia, and J. Barata. Finding natural trails with swarm-based visual saliency. Submitted to *Journal of Field Robotics*.

Conference Proceedings

- P. Santana, L. Correia, M. Guedes, and J. Barata. Visual attention and swarm cognition towards fast and robust off-road robots. To appear in *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE)*. IEEE Press, 2011.
- P. Santana, N. Alves, L. Correia, and J. Barata. Swarm-based visual saliency for trail detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 759-765. IEEE Press, Piscataway, 2010.
- P. Santana, M. Guedes, L. Correia, and J. Barata. A saliency-based solution for robust off-road obstacle detection. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3096-3101. IEEE Press, Piscataway, 2010.

- P. Santana, N. Alves, L. Correia, and J. Barata. A saliency-based approach to boost trail detection. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1426-1431. IEEE Press, Piscataway, 2010.
- P. Santana, M. Guedes, L. Correia, and J. Barata. Saliency-based obstacle detection and ground-plane estimation for off-road vehicles. In *Proceedings of the International Conference on Computer Vision Systems (ICVS)*, pages 275-284. Springer-Verlag, Berlin, 2009.
- P. Santana and L. Correia. Swarm-based active vision. In V. Trianni and E. Tuci, editors, In *Proceedings of the Swarm Cognition Workshop, held at the 31st Annual Meeting of the Cognitive Science Society*, 2009.
- P. Santana, P. Santos, L. Correia, and J. Barata. Cross-country obstacle detection: Space-variant resolution and outliers removal. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 1836-1841. IEEE Press, Piscataway, 2008.
- P. Santana, C. Cândido, P. Santos, L. Almeida, L. Correia, and J. Barata. The Ares robot: case study of an affordable service robot. In *Proceedings of the European Robotics Symposium (EUROS)*, pages 33-42. Springer-Verlag, Berlin, 2008.
- P. Santana and L. Correia. Improving Visual Odometry by Removing Outliers in Optic Flow. In *Proceedings of the Conference on Autonomous Robot Systems and Competitions (Robotica)*, pages 117-122. 2008.

Submitted Conference Proceedings

- P. Santana, R. Mendonça, L. Correia, and J. Barata. Swarms for robot vision: the case of adaptive visual trail detection and tracking. Submitted to European Conference on Artificial Life (ECAL). MIT Press, 2011.

6.3.2 Additional Publications

Journals

- P. Santana, J. Barata, and L. Correia. Sustainable robots for humanitarian demining. *International Journal of Advanced Robotics Systems*, 4(2):207-218, 2007.

Book Chapters

- P. Santana, L. Correia, and J. Barata. Locomotion and localisation of humanitarian demining robots. In M. Habib and Y. Baudoin (Eds.), *Using Robots in Hazardous Environments: Landmine Detection, de-Mining and Other Applications*. Woodhead Publishing Ltd., 2010.

- P. Santana, L. Correia, and J. Barata. Developments on an affordable robotic system for humanitarian demining. In M. Habib (Ed.), *Humanitarian Demining, Innovative Solutions and the Challenges of Technology*, pages 263-288. I-Tech Education and Publishing, Vienna, Austria, 2008.

Conference Proceedings

- P. Santana, C. Santos, D. Chaínho, L. Correia, and J. Barata. Predicting Affordances from Gist. In *Proceedings of the International Conference on Simulation of Adaptive Behavior (SAB)*, pages 325–334. Springer-Verlag, Berlin, 2010.

- C. Cândido, P. Santana, L. Correia, and J. Barata. Shared control of a pan-tilt camera on an all-terrain mobile robot. In *Proceedings of the IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 177-183. IEEE Press, Piscataway, 2008.

- V. Santos, P. Santana, L. Correia, and J. Barata. Teleoperation mechanisms in a multi-agent system. In *Proceedings of the IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 170-176. IEEE Press, Piscataway, 2008.

- P. Santana, M. Salgueiro, V. Santos, and J. Barata. A Knowledge-Based Component for Human-Robot Teamwork. In *Proceedings of the International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, pages 228-233. INSTICC Press, 2008.

Bibliography

H. Abelson, D. Allen, D. Coore, C. Hanson, G. Homsy, T. Knight Jr, R. Nagpal, E. Rauch, G. Sussman, and R. Weiss. Amorphous computing. *Communications of the ACM*, 43 (5):74–82, 2000.

M. Agrawal and K. Konolige. Real-time Localization in Outdoor Environments using Stereo Vision and Inexpensive GPS. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, pages 1063–1068. IEEE Computer Society Washington, DC, USA, August 2006.

J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988.

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.

G. A. Alvarez and P. Cavanagh. Independent resources for attentional tracking in the left and right visual hemifields. *Psychological Science*, 16(8):637–643, 2005.

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

J. C. Andersen, N. A. Andersen, and O. Ravn. Vision assisted laser scanner navigation for autonomous robots. In *Proceedings of the International Symposium on Experimental Robotics (ISER)*, pages 111–120, Rio de Janeiro, Brazil, July 2008.

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.

W. R. Ashby. *Design for a Brain*. Chapman and Hall, London, 1952.

A. D. Bagdanov, A. d. Bimbo, W. Nunziati, and F. Pernici. A reinforcement learning approach to active camera foveation. In *Proceedings of the 4th ACM International*

Workshop on Video Surveillance and Sensor Networks, pages 179–186. ACM, New York, 2006.

R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, 1988.

M. Bajracharya, B. Tang, A. Howard, M. Turmon, and L. Matthies. Learning long-range terrain classification for autonomous navigation. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 4018–4024. IEEE Press, Piscataway, May 2008.

C. Balkenius, A. P. Eriksson, and K. Astrom. Learning in visual attention. In *Proceedings of the International Conference on Pattern Recognition (ICPR), Workshop on Learning for Adaptable Visual Systems (LAVS 2004)*, volume 4, IEEE Computer Society, Washington, DC, 2004.

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

D. H. Ballard, M. M. Hayhoe, P. K. Pook, and R. P. N. Rao. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20:723–767, 1997.

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.

P. Batavia and S. Singh. Obstacle detection in smooth high curvature terrain. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3062–3067. IEEE Press, Piscataway, 2002.

P. H. Batavia and S. Singh. Obstacle detection using adaptive color segmentation and color stereo homography. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, volume 1, pages 705–710. IEEE Press, Piscataway, 2001.

M. A. Bedau. Artificial life: Organization, adaptation and complexity from the bottom up. *Trends in cognitive sciences*, 7(11):505–512, 2003.

R. Beer. Toward the evolution of dynamical neural networks for minimally cognitive behavior. In *Proceedings of the International Conference on Simulation of Adaptive Behavior (SAB)*, pages 421–429. A Bradford Book, 1996.

R. D. Beer. A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1-2):173–215, 1995.

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

R. Behringer and N. Muller. Autonomous road vehicle guidance from autobahnen to narrow curves. *IEEE Transactions on Robotics and Automation*, 14(5):810–815, 1998.

P. Bellutta, R. Manduchi, L. Matthies, K. Owens, and A. Rankin. Terrain perception for DEMO III. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 326–331. IEEE, October 2000.

A. Beyeler, J. Zufferey, and D. Floreano. Vision-based control of near-obstacle flight. *Autonomous Robots*, 27(3):201–219, 2009.

J. J. Biesiadecki and M. W. Maimone. The mars exploration rover surface mobility flight software: Driving ambition. In *Proceedings of the IEEE Aerospace Conference*, volume 5, pages 1–15. IEEE, March 2006.

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.

E. Bonabeau, M. Dorigo, and G. Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, Oxford, 1999.

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

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.

A. Broggi, C. Caraffi, R. I. Fedriga, and P. Grisleri. Obstacle detection with stereo vision for off-road vehicle navigation. In *Proceedings of the IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition (CVPR) - Workshops, pages 65–72. IEEE Computer Society, Washington, DC, June 2005.

R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1):139–159, 1991.

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

C. Caraffi, S. Cattani, and P. Grisleri. Off-road path and obstacle detection using decision networks and stereo vision. *IEEE Transactions on Intelligent Transportation Systems*, 8(4):607–618, 2007.

A. Castano and L. Matthies. Foliage discrimination using a rotating ladar. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–6. IEEE Press, Piscataway, September 2003.

T. Chang, T. Hong, S. Legowik, and M. Abrams. Concealment and obstacle detection for autonomous driving. In *Proceedings of the Robotics and Applications Conference*, pages 1–15, Santa Barbara, CA, October 1999.

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.

D. R. Chialvo and M. M. Millonas. How swarms build cognitive maps. In L. Steels, editor, *The Biology and Technology of Intelligent Autonomous Agents*, volume 144, pages 439–450. NATO ASI Series, 1995.

Y. Choe, H. F. Yang, and N. Misra. Motor system's role in grounding, receptive field development, and shape recognition. In *Proceedings of the 7th International Conference on Development and Learning (ICDL)*, pages 67 – 72. IEEE Computer Society, Washington, DC, 2008.

N. Chumerin and M. V. Hulle. Ground plane estimation based on dense stereo disparity. In *Proceedings of the International Conference on Neural Networks and Artificial Intelligence (ICNNAI)*, pages 209–213, Minsk, Belarus, May 2008.

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.

L. Correia. Self-organisation: a case for embodiment. In C. Gershenson and T. Lenaerts, editors, *Proceedings of the Workshop on the Evolution of Complexity, at ALife X, Bloomington, IN, USA, June 3rd*, 2006.

T. Cour and J. Shi. Recognizing objects by piecing together the segmentation puzzle. In *Proceedings of the IEEE Cnference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE Computer Society, Washington, DC, 2007.

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. Dang and C. Hoffmann. Fast object hypotheses generation using 3d position and 3d motion. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops*, pages 56–61. IEEE Computer Society, Washington, DC, June 2005.

G. de Croon and E. O. Postma. Sensory-motor coordination in object detection. In *Proceedings of the IEEE Symposium on Artificial Life (CI-ALIFE)*, pages 147–154. IEEE CIS, Los Alamitos, 2007.

B. de Haan, P. S. Morgan, and C. Rorden. Covert orienting of attention and overt eye movements activate identical brain regions. *Brain research*, 1204:102–111, 2008.

J. L. Deneubourg, S. Goss, N. Franks, and J. M. Pasteels. The blind leading the blind: Modeling chemically mediated army ant raid patterns. *Journal of insect behavior*, 2(5): 719–725, 1989.

J. Dewey. The reflex arc concept in psychology. *Psychological Review*, 3:357–370, 1896.

E. D. Dickmanns, B. Mysliwetz, and T. Christians. An integrated spatio-temporal approach to automatic visual guidance of autonomous vehicles. *IEEE Transactions on Systems, Man and Cybernetics*, 20(6):1273–1284, 1990.

M. M. Doran, J. E. Hoffman, and B. J. Scholl. The role of eye fixations in concentration and amplification effects during multiple object tracking. *Visual Cognition*, 17(4):574–597, 2009.

G. Dubbelman, W. van der Mark, J. C. van den Heuvel, and F. C. A. Groen. Obstacle detection during day and night conditions using stereo vision. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 109–116. IEEE Press, Piscataway, October 2007.

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.

M. Eimer, B. Forster, J. V. Velzen, and G. Prabhu. Covert manual response preparation triggers attentional shifts: Erp evidence for the premotor theory of attention. *Neuropsychologia*, 43(6):957–966, 2005.

M. Fend, S. Bovet, H. Yokoi, and R. Pfeifer. An active artificial whisker array for texture discrimination. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1044–1049. IEEE Press, Piscataway, 2003.

C. Fermuller and Y. Aloimonos. Vision and action. *Image and Vision Computing*, 13: 725–755, 1995.

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.

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

D. Floreano, K. Toshifumi, D. Marocco, and E. Sauser. Coevolution of active vision and feature selection. *Biological Cybernetics*, 90(3):218–228, 2004.

D. Floreano, P. Durr, and C. Mattiussi. Neuroevolution: from architectures to learning. *Evolutionary Intelligence*, 1(1):47–62, 2008.

N. R. Franks. Army ants: a collective intelligence. *American Scientist*, 77(2):138–145, 1989.

S. Frintrop. *VOCUS: a visual attention system for object detection and goal-directed search*. PhD thesis, INAI, Vol. 3899, Germany, 2006.

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.

S. Frintrop, P. Jensfelt, and H. Christensen. Simultaneous robot localization and mapping based on a visual attention system. *LNCS*, 4840:417–430, 2007.

D. Gennery. Traversability analysis and path planning for a planetary rover. *Autonomous Robots*, 6:131–146, 1999.

B. P. Gerkey, R. T. Vaughan, and A. Howard. The player/stage project: Tools for multi-robot and distributed sensor systems. In *Proceedings of the International Conference on Advanced Robotics (ICAR)*. IEEE, 2003.

W. Gerstner and W. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, Cambridge, 2002.

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.

J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Ass., Hillsdale, NJ, 1979.

S. B. Goldberg, M. W. Maimone, and Matthies. Stereo vision and rover navigation software for planetary exploration. In *Proceedings of the IEEE Aerospace Conference*, pages 2025–2036. IEEE, March 2002.

M. A. Goodale. Action without perception in human vision. *Cognitive Neuropsychology*, 25(7):891–919, 2008.

P. Grandjean and L. Matthies. Perception control for obstacle detection by a cross-country rover. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 20–27. IEEE Press, Piscataway, May 1993.

P.-P. Grassé. La reconstruction du nid et les coordinations inter-individuelles chez bellicositermes et cubitermes sp. la théorie de la stigmergie: Essai d'interprétationdu comportement des termites con- structeurs. *Insectes Sociaux*, 6:41–80, 1959.

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.

M. K. Habib. Humanitarian demining: Reality and the challenge of technology - the state of the arts. *International Journal of Advanced Robotic Systems (ARS)*, 4(2):151–172, 2007.

R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun. Learning long-range vision for autonomous off-road driving. *Journal of Field Robotics*, 26(2):120–144, 2009.

M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005.

D. Hernandez, J. Cabrera, A. Naranjo, A. Dominguez, and J. Isern. Gaze control in a multiple-task active-vision system. In *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS)*. Applied Computer Science Group, March 2007.

S. Hirose and E. Fukushima. Snakes and strings: New robotic components for rescue operations. *The International Journal of Robotics Research*, 23(4-5):341, 2004.

T. H. Hong, C. Rasmussen, T. Chang, and M. Shneier. Fusing ladar and color image information for mobile robot feature detection and tracking. In *Proceedings of the International Conference on Intelligent Autonomous Systems (IAS)*, pages 1–6, Marina Del Ray, CA, March 2002.

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.

T. Huntsberger, H. Aghazarian, A. Howard, and D. Trotz. Stereo vision–based navigation for autonomous surface vessels. *Journal of Field Robotics*, 28(1):3–18, 2011.

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 and C. Koch. Computational modelling of visual attention. *Nature Reviews, Neuroscience*, 2:1–10, 2001.

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.

V. John, E. Trucco, and S. Ivekovic. Markerless human articulated tracking using hierarchical particle swarm optimisation. *Image Vision Computing*, 28(11):1530–1547, 2010.

D. Johnson, D. Naffin, J. Puhalla, J. Sanchez, and C. Wellington. Development and implementation of a team of robotic tractors for autonomous peat moss harvesting. *Journal of Field Robotics*, 26(6-7):549–571, 2009.

A. Kelly and A. Stentz. Rough terrain autonomous mobility—part 2: An active vision, predictive control approach. *Autonomous Robots*, 5(2):163–198, 1998.

D. Kim and R. Moeller. Passive sensing and active sensing of a biomimetic whisker. In *Proceedings of the International Conference on the Simulation and Synthesis of Living Systems (ALife X)*, pages 282–288. The MIT Press, Cambridge, MA, 2006.

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

J. Z. Kolter, Y. Kim, and A. Y. Ng. Stereo vision and terrain modeling for quadruped robots. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 1557–1564. IEEE Press, Piscataway, May 2009.

H. Kong, J. Audibert, and J. Ponce. General road detection from a single image. *IEEE Transactions on Image Processing*, 19(8):2211–2220, 2010.

K. Konolige. Small vision systems: Hardware and implementation. In *Proceedings of the International Symposium on Robotics Research (ISRR)*, pages 111–116. Springer, London, 1997.

K. Konolige and D. Beymer. Sri small vision system users manual. Technical report, SRI International, Menlo Park, CA, May 2007.

K. Konolige, M. Agrawal, M. R. Blas, R. C. Bolles, B. Gerkey, J. Solà, and A. Sundaresan. Mapping, navigation, and learning for off-road traversal. *Journal of Field Robotics*, 26 (1):88–113, 2009.

C. Kwok and D. Fox. Reinforcement learning for sensing strategies. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 4, pages 3158–3163 vol.4. IEEE Press, Piscataway, September-October 2004.

R. Labayrade, D. Aubert, and J. P. Tarel. Real time obstacle detection in stereovision on non flat road geometry through v-disparity representation. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 646–651. IEEE, June 2002.

S. Lacroix, A. Mallet, D. Bonnafous, G. Bauzil, S. Fleury, M. Herrb, and R. Chatila. Autonomous rover navigation on unknown terrains: Functions and integration. *International Journal of Robotics Research*, 21(10-11):917–942, 2002.

J. F. Lalonde, N. Vandapel, D. F. Huber, and M. Hebert. Natural terrain classification using three-dimensional ladar data for ground robot mobility. *Journal of field robotics*, 23(10):839–861, 2006.

M. Land. Motion and vision: why animals move their eyes. *Journal of Computational Physiology A*, 185:341–352, 1999.

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.

K. Low, G. Podnar, S. Stancliff, J. Dolan, , and A. Elfes. Robot boats as a mobile aquatic sensor network. In *Proceedings of ESSA Workshop*, 2009.

W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, November 2006.

R. Manduchi, A. Castano, A. Talukder, and L. Matthies. Obstacle detection and terrain classification for autonomous off-road navigation. *Autonomous Robots*, 18(1):81–102, 2005.

D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co., Inc. New York, NY, 1982.

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

J. A. R. Marshall, R. Bogacz, A. Dornhaus, R. Planqué, T. Kovacs, and N. R. Franks. On optimal decision-making in brains and social insect colonies. *Journal of the Royal Society Interface*, 6(40):1065–1074, 2009.

S. Martel and M. Mohammadi. Using a swarm of self-propelled natural microrobots in the form of flagellated bacteria to perform complex micro-assembly tasks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 500–505. IEEE Press, Piscataway, 2010.

L. Matthies. *Dynamic Stereo Vision*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1989.

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.

D. Meger, P. E. Forssen, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, D. G. Lowe, and B. Dow. Curious george: An attentive semantic robot. *Robotics and Autonomous Systems*, 56(6):503–511, 2008.

L. Merino, F. Caballero, J. Martínez-de Dios, J. Ferruz, and A. Ollero. A cooperative perception system for multiple UAVs: Application to automatic detection of forest fires. *Journal of Field Robotics*, 23(3-4):165–184, 2006. ISSN 1556-4967.

J. Michels, A. Saxena, and A. Ng. High speed obstacle avoidance using monocular vision and reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 593–600. ACM, 2005.

A. D. Milner and M. A. Goodale. *The Visual Brain in Action*. Oxford University Press, Oxford, 1995.

M. Minsky. *The society of mind*. Simon & Schuster, New York, 1988.

M. Mirolli, T. Ferrauto, and S. Nolfi. Categorisation through evidence accumulation in an active vision system. *Connection Science*, 22:331–354, December 2010.

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.

S. Moorehead, R. Simmons, D. Apostolopoulos, and W. L. Whittaker. Autonomous navigation field results of a planetary analog robot in antarctica. In *Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation*, pages 237–242, Noordwijk, the Netherlands, June 1999.

J. Moren, A. Ude, A. Koene, and G. Cheng. Biologically based top-down attention modulation for humanoid interactions. *International Journal of Humanoid Robotics*, 5 (1):3–24, 2008.

M. Mossio and D. Taraborelli. Action-dependent perceptual invariants: From ecological to sensorimotor approaches. *Consciousness and Cognition*, 17(4):1324–1340, 2008.

F. Mufti, R. Mahony, and J. Heinzmann. Saptio-temporal ransac for robust estimation of ground plane in video range images for automotive applications. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 1142–1148. IEEE, October 2008.

R. Murphy and S. Stover. Rescue robots for mudslides: A descriptive study of the 2005 La Conchita mudslide response. *Journal of Field Robotics*, 25(1-2):3–16, 2008.

B. Nabbe and M. Hebert. Where and when to look: how to extend the myopic planning horizon. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 920 – 927. IEEE Press, Piscataway, 2003.

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.

P. Newman and K. Ho. Slam-loop closing with visually salient features. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 635–642. IEEE Press, Piscataway, April 2005.

S. Nolfi. *Handbook of Categorization in Cognitive Science*, chapter Categories Formation in Self-Organizing Embodied Agents, pages 869–889. Elsevier, Oxford, 2005.

S. Nolfi and D. Marocco. Active perception: a sensorimotor account of object categorization. In *Proceedings of the 7th International Conference on Simulation of Adaptive Behavior (SAB)*, pages 266–271. MIT Press, August 2002.

A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007.

F. Orabona, G. Metta, and G. Sandini. Object-based visual attention: a model for a behaving robot. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops*, pages 89–96. IEEE Computer Society, Washington, DC, June 2005.

J. K. O'Regan and A. Noe. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24:939–1031, 2001.

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.

K. Patel, W. Macklem, S. Thrun, and M. Montemerlo. Active sensing for high-speed offroad driving. In *Proceedings of the IEEE International Conference Robotics and Automation (ICRA)*, pages 3162–3168. IEEE Press, Piscataway, 2005.

R. Pfeifer and J. C. Bongard. *How the Body Shapes the Way We Think - A New View of Intelligence*. The MIT Press, Cambridge, MA, 2006.

R. Pfeifer and C. Scheier. *Understanding intelligence*. The MIT Press, Cambridge, MA, 1999.

C. Plagemann, S. Mischke, S. Prentice, K. Kersting, N. Roy, and W. Burgard. Learning predictive terrain models for legged robot locomotion. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3545–3552. IEEE Press, Piscataway, September 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.

J. Poppinga, A. Birk, and K. Pathak. Hough based terrain classification for realtime detection of drivable ground. *Journal of Field Robotics*, 25(1-2):67–88, 2008.

Z. W. Pylyshyn and R. W. Storm. Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, 3(3):179, 1988.

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. Grouping dominant orientations for ill-structured road following. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1. IEEE Computer Society, Washington, DC, 2004.

C. Rasmussen. Roadcompass: following rural roads with vision+ ladar using vanishing point tracking. *Autonomous Robots*, 25(3):205–229, 2008.

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.

R. Ratcliff and P. L. Smith. A comparison of sequential sampling models for twochoice reaction time. *Psychological Review*, 111:333–367, 2004.

G. Rizzolatti, L. Riggio, I. Dascola, and C. Umiltá. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1A):31–40, 1987.

J. K. Rosenblatt. DAMN: a distributed architecture for mobile navigation. In *Proceedings of the AAAI Spring Symposium on Lessons Learned from Implemented Software Architectures for Physical Agents*, Stanford, CA, 1995.

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.

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 *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 962–967. IEEE Press, Piscataway, 2008.

R. Rusu, A. Sundaresan, B. Morisset, K. Hauser, M. Agrawal, J. Latombe, and M. Beetz. Leaving Flatland: Efficient real-time three-dimensional perception and motion planning. *Journal of Field Robotics*, 26(10):841–862, 2009.

P. Santana and L. Correia. Behaviour cooperation by negation for mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Biomimetics, (ROBIO)*, pages 982–987. IEEE Press, Piscataway, 2006.

P. Santana and L. Correia. Improving visual odometry by removing outliers in optic flow. In *Proceedings of the 8th Conference on Autonomous Robot Systems and Competitions (Robotica)*, pages 117–122, 2008.

P. Santana and L. Correia. Swarm-based active vision. In V. Trianni and E. Tuci, editors, *Proceedings of the Swarm Cognition Workshop, held at the 31st Annual Meeting of the Cognitive Science Society*, Amsterdam, 2009.

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

P. Santana and L. Correia. Swarm cognition experimental results supporting videos. `http://www.uninova.pt/~pfs/videos_AB2010/videos.html`, 2010b.

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

P. Santana, J. Barata, and L. Correia. Sustainable robots for humanitarian demining. *International Journal of Advanced Robotics Systems*, 4(2):207–218, June 2007.

P. Santana, C. Cândido, P. Santos, L. Almeida, L. Correia, and J. Barata. The Ares robot: case study of an affordable service robot. In *Proceedings of the European Robotics Symposium (EUROS)*, pages 33–42. Springer-Verlag, Berlin, Germany, 2008a.

P. Santana, P. Santos, L. Correia, and J. Barata. Cross-country obstacle detection: Space-variant resolution and outliers removal. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 1836–1841. IEEE Press, Piscataway, September 2008b.

P. Santana, M. Guedes, L. Correia, and J. Barata. Saliency-based obstacle detection and ground-plane estimation for off-road vehicles. In *Proceedings of the International Conference on Computer Vision Systems (ICVS)*, pages 275–284. Springer-Verlag, Berlin, Germany, 2009.

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, N. Alves, L. Correia, and J. Barata. Trail detection experimental results supporting videos. `http://www.uninova.pt/~pfs/traildetection.html`, 2010c.

P. Santana, M. Guedes, L. Correia, and J. Barata. A saliency-based solution for robust off-road obstacle detection. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3096–3101. IEEE Press, Piscataway, 2010d.

P. Santana, M. Guedes, L. Correia, and J. Barata. Stereo-based all-terrain obstacle detection using visual saliency. *Journal of Field Robotics*, 28(2):241–263, 2011.

A. Saxena, S. Chung, and A. Ng. 3-d depth reconstruction from a single still image. *International Journal of Computer Vision*, 76(1):53–69, 2008.

H. Schafer, M. Proetzsch, and K. Berns. Stereo-vision-based obstacle avoidance in rough outdoor terrain. In *Proceedings of the International Symposium on Motor Control and Robotics (ISMCR)*, pages 1–9, Brussels, Belgium, November 2005.

C. Scheier, R. Pfeifer, and Y. Kuniyoshi. Embedded neural networks: exploiting constraints. *Neural Networks*, (11):1551–1596, 1998.

H. Seraji. Traversability index: A new concept for planetary rovers. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2006–2013. IEEE Press, Piscataway, 1999.

H. Seraji. Safety measures for terrain classification and safest site selection. *Autonomous Robots*, 21(3):211–225, 2006.

S. Singh, R. Simmons, T. Smith, A. Stentz, V. Verma, A. Yahja, and K. Schwehr. Recent progress in local and global traversability for planetary rovers. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, volume 2, pages 1194–1200. IEEE Press, Piscataway, 2000.

A. Slocum, D. Downey, and R. Beer. Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In *Proceedings*

of the International Conference on Simulation of Adaptive Behavior (SAB), pages 430–439. The MIT Press, Cambridge, MA, 2000.

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.

N. Soquet, D. Aubert, and N. Hautiere. Road segmentation supervised by an extended v-disparity algorithm for autonomous navigation. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 160–165. IEEE, 2007.

O. Sporns and M. Lungarella. Evolving coordinated behavior by maximizing information structure. In *Proceedings of ALife X*, pages 3–7. The MIT Press, Cambridge, MA, 2006.

N. Sprague, D. Ballard, and A. Robinson. Modeling embodied visual behaviors. *ACM Transactions on Applied Perception*, 4(2):Article 11, 2007.

S. Squyres, A. Knoll, R. Arvidson, J. Ashley, J. Bell III, W. Calvin, P. Christensen, B. Clark, B. Cohen, P. de Souza Jr, et al. Exploration of Victoria crater by the Mars rover Opportunity. *Science*, 324(5930):1058, 2009.

G. Sukhatme, A. Dhariwal, B. Zhang, C. Oberg, B. Stauffer, and D. Caron. Design and development of a wireless robotic networked aquatic microbial observing system. *Environmental Engineering Science*, 24(2):205–215, 2007.

R. Sukthankar, D. Pomerleau, and C. Thorpe. Panacea: An active sensor controller for the alvinn autonomous driving system. In *Proceedings of International Symposium on Robotics Research (ISRR)*. Springer, London, 1993.

M. Suzuki and D. Floreano. Evolutionary active vision toward three dimensional landmark-navigation. In *Proceedings of the 9th International Conference on the Simulation of Adaptive Behavior (SAB)*, pages 263–273. The MIT Press, Cambridge, MA, 2006.

A. Talukder, R. Manduchi, A. Rankin, and L. Matthies. Fast and reliable obstacle detection and segmentation for cross-country navigation. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, volume 2, pages 610–618. IEEE, 2002.

C. Tessier, M. Berducat, R. Chapuis, F. Chausse, and A. Cemagref. A new landmark and sensor selection method for vehicle localization and guidance. In *Proceedings of the 2007 IEEE Intelligent Vehicles Symposium*, pages 123–129. IEEE Press, Piscataway, 2007.

E. Thelen and L. B. Smith. *A dynamic systems approach to the development of cognition and action*. The MIT Press, Cambridge, MA, 1996.

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, W. Burgard, and D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, Cambridge, MA, 2005.

S. Thrun, M. Montemerlo, and A. Aron. Probabilistic terrain analysis for high-speed desert driving. In *Proceedings of the Robotics Science and Systems Conference*, pages 1–7, Philadelphia, PA, August 2006a.

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, 2006b.

P. Tokekar, D. Bhadauria, A. Studenski, and V. Isler. A robotic system for monitoring carp in Minnesota lakes. *Journal of Field Robotics*, 27(6):779–789, 2010.

C. Tomasi and J. Shi. Good features to track. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600. IEEE Computer Society, Washington, DC, 1994.

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 and E. Tuci, editors. *Annual Meet. of the Cognitive Science Society (CogSci), Workshop on Swarm Cognition*, Amsterdam, 2009.

V. Trianni and E. Tuci. Swarm cognition and artificial life. In *Proceedings of the European Conference on Artificial Life (ECAL)*, volume LNCS/LNAI 5777, 5778. Springer-Verlag, Berlin, Germany, 2010.

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.

E. Tuci, G. Massera, and S. Nolfi. Active categorical perception of object shapes in a simulated anthropomorphic robotic arm. *IEEE Transactions on Evolutionary Computation*, 14(6):885 –899, 2010.

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.

C. Urmson, C. Ragusa, D. Ray, J. Anhalt, D. Bartz, T. Galatali, A. Gutierrez, J. Johnston, S. Harbaugh, H. Kato, W. Messner, N. Miller, K. Peterson, B. Smith, J. Snider, S. Spiker, J. Ziglar, W. Whittaker, M. Clark, P. Koon, A. Mosher, and J. Struble. A robust approach to high-speed navigation for unrehearsed desert terrain. *Journal of Field Robotics*, 23(8):467–508, 2006.

W. van der Mark, J. Heuvel, and F. Groen. Stereo based obstacle detection with uncertainty in rough terrain. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 1005–1012. IEEE Press, Piscataway, 2007.

V. van Veen, M. K. Krug, and C. S. Carter. The neural and computational basis of controlled speed-accuracy tradeoff during task performance. *Journal of Cognitive Neuroscience*, 20(11):1952–1965, 2008.

F. J. Varela, E. Thompson, and E. Rosch. *The embodied mind*. The MIT Press/Bradford Books, Cambridge, MA, 1991.

P. Vernaza, B. Taskar, and D. D. Lee. Online, self-supervised terrain classification via discriminatively trained submodular markov random fields. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2750–2757. IEEE Press, Piscataway, 2008.

S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal. Overt visual attention for a humanoid robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2332–2337. IEEE Press, Piscataway, October 2001.

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

C. Wellington and A. Stentz. Online adaptive rough-terrain navigation in vegetation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 96–101. IEEE Press, Piscataway, April 2004.

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

J. M. Wolfe, M. L.-H. Võ, K. K. Evans, and M. R. Greene. Visual search in scenes involves selective and nonselective pathways. *Trends in cognitive sciences*, doi:10.1016/j.tics.2010.12.001, In Press.

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

C. Ye. Navigating a mobile robot by a traversability field histogram. *IEEE Transactions on Systems Man and Cybernetics-Part B-Cybernetics*, 37(2):361–372, 2007.

Y. Yu, G. K. I. Mann, and R. G. Gosine. A task-driven object-based attention model for robots. In *Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1751–1756. IEEE, 2007.

K. Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112, 1996.

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.