ALIZ-E

Adaptive Strategies for Sustainable Long-Term Social Interaction
EU FP7 project, Grant Agreement no. 248116 Seventh Framework Programme (FP7)

Objective ICT 2009.2.1: Cognitive Systems and Robotics

Deliverable D3.1
User model and robot roles: User, social environment, affect and interaction models for mediating the interactive robot behaviours in accordance to its roles

Deliverable submission: Month 12
12th March 2011

Contents

1 Executive summary

This report presents the results of WP3 for the first year of the ALIZ-e project. The overall objective of WP3 is to facilitate adaptation of interaction, between the humans and robots, to the human preferences and social environment. We see interaction as an inherent aspect of long-term interaction: How does the robot adapt to the user, retain the human's interest, and contribute to obtaining personal goals? In Year 1,WP3 focused on defining characteristics of the user, task and context, and robot strategies for adapting to these factors, in one-to-one child-robot interaction.

In this report, we present the results of the first year.

2 Role of adaptive user and task modelling in ALIZ-e

ALIZ-e develops a social robot for long-term interaction with diabetic children. WP3 contributes to this overall objective by providing input to other WPs about when and how to change behaviour, most relevantly WP2 "non-verbal behaviour" and WP4 "verbal behaviour". The user and task model will also make use of knowledge acquired during experiments and tests from other WPs. The general approach is to start with "realistic scenarios". WP5 provides the input for this. For these scenarios, the corresponding scientific goals for adaptive user and task modelling are indentified and a selection is made for empirical investigation (i.e., the research questions or claims on robot's "core adaptation functions"). Both the realistic scenarios and scientific goals provide input for the user model's content and shape the roles of the robot (educator, motivator, and buddy). Subsequently, the claims are tested in an abstracted version of the scenario (focusing on specific aspects of the realistic scenario). These controlled (small) experiments with a prototype lead to the final step: incorporating the experiment's results and implications into the realistic scenarios (WP5). This general approach is done in a yearly cycle.

3 Contribution to the ALIZ-e scenarios and prototypes

In the first year, WP3 contributes to the scenarios and (autonomous and wizard of oz) prototypes by providing elementary user models for the Simon game and the dance game, which are specified in WP5. At the moment, we started with a third scenario (the quiz) and in enabling the child to switch between scenarios (the child will not always get what he/she wants; e.g. when a goal has not reached yet the robot acknowledges that it knows the child wants to switch but that first the goal has to be reached). The different scenarios have a focus on different roles:

- The Simon game has a focus on buddy (mostly for fun).
- The dance scenario has a focus on the motivator role (moving is healthy), but also incorporates aspects of the educator role (learning dance steps) and buddy (it is also fun to dance).
- The quiz scenario has its focus on the educator role (learning the child about healthy living), but also incorporates the motivator role (the child should be encouraged to play the quiz for a longer period of time).

The knowledge gathered in the domain analysis for user and task modelling also provides inputs for the further development of scenarios and prototypes.

4 Tasks, objectives and results

4.1 Planned work

During the first year, WP3 investigated and supported the development of a user model and robot roles for diabetic children (MS3.1). The work was planned to contribute to three tasks: "define

models of user, social environment, affect and interaction" (T3.1), "Long-term behaviour support" (T3.2), "Personalize robot behaviours" (T3.3).

The concrete objectives were: to define a user model supporting social intelligent robot interaction based on domain analysis, study how adaptive behaviour can contribute to long-term interaction and personalize the robots behaviour. The first year adaptation was planned to include some personal characteristics of the user.

4.2 Actual work performed

WP3 is divided into four main tasks. Task 3.1 only runs in the first year, task 3.2 in the first two years, task 3.3 in 4 years and task 3.4 runs from the second year until the end of the project (M54). Actually, all these tasks were started in the first year (also task 3.4).

T3.1 Define models of user, social environment, affect and interaction

For this task the goal was to make an overview of information that should to be incorporated in the user model. General information that the user model should contain is information of the user group, the social environment, the situated affects and interaction tasks. Based on the domain analysis of WP5, we conducted two series of interviews: one with care providers and one with children with diabetes and their parents. To elicit self-management determinants, we interviewed six Dutch diabetes care givers (i.e., children's diabetes nurses, a paediatrician, psychologist and dietician), who are designated to support children aged 8-12. These interviews were complemented with a literature review. In addition, nine children, aged 8-12, were interviewed. We asked about important life domains (e.g., school, sports, hobbies, family, friends and activities in the weekend), perception of their diabetes management and the effect of the diabetes self-management on quality of life aspects (i.e., mental wellbeing, physical wellbeing, and social wellbeing). Results of the interviews show the following (for details see appendix [Groot, 2010] (thesis not included, but will be made available on ALIZ-e website) [Hoondert et al., 2010, Hoondert and Blanson Henkemans, 2011]). First, the studied children group does not experience significant problems with diabetes management and their glucose levels are relatively well regulated, mostly due to the dominant involvement of parents and guardians. Second, during puberty, the child's need for self-control increases and he or she may rebel against his or her parents, at the cost of diabetes self-management and leading to unhealthy blood glucose levels. Lastly, an important factor in the child's decision to perform self-management is that, although they make the best of their life, managing diabetes goes at the cost of their quality of life (i.e., mental and social wellbeing).

Furthermore, field tests in Italy have been performed to get a better understanding about how children perceive the robot [Nalin et al., 2011]. This understanding is extended by all experiments performed with children in the project. Based on the interviews and field tests, the roles defined in ([Looije et al., 2010, Kanda et al., 2004]), and the scenarios defined in WP5, a working document has been created in which the goals, and success factors (metrics) per scenario are made explicit. Furthermore, robot behaviours with a beneficial role for the success factors (based on literature) are summed up which lead to an overview of user characteristics that should be incorporated into the user model.

T3.2 Long-term behaviour support

In this task the focus is on studying how the robot's behaviour can contribute to long-term goal achievement and the child's experiences in the hospital. The first year focused on two aspects: 1) conducting experiments in which the child worked multiple times with a robot, and 2) measuring the motivation of children during different interactions and increasing the motivation by using techniques from literature. Regarding the first point, an experiment was performed that looked into the effect of situatedness of emotions (in the future different roles have different context) and the effect of experience with emotions over time (learning). Both aspects play a role in the personalization. For the experiment, we developed emotional postures of the NAO, which were subsequently tested on recognition by children (14 children between 8-9 years old) during

two sessions and with sentences that placed the emotion into context or without. In addition, children's recognition of NAO-emotions was compared to their recognition of iCat-emotions. The annexes contain a paper [Cohen et al., 2011] reporting this experiment. Regarding the second point, measuring child's experiences and supporting motivation, all experiments of this WP contained measures for experience and motivation. When comparing two robots the child was asked which one was more fun, but also more objective measures were used, such as emotion recognition and a free choice period. A free choice period is a period where children are free to choose the activity they want to do, amongst which is playing with the robot.

T3.3 Personalize robot behaviours

In this task, the empathic, persuasive and engaging behaviours of the robot are implemented and tailored to the user (personalization) and task demands (adaptation). In the first year the focus was on automatic personalization based on cognitive characteristics (introversion/extraversion, memory), physical characteristics (capability to dance/balance) and performance (mathematical problem solving, longer/more difficult sequences with simon game or dancing). Three experiments were performed and three are planned. The first experiment involved adaptation to personality (extraversion/introversion). In this experiment, 14 children in the age of 8-12 played the simon game with two robots. One exhibited extravert behaviour and the other introvert. Interaction and movements were adapted. The goal of the experiment was threefold: Firstly, do children have a preference for a certain robot, secondly, does this preference correlate with their own personality and, thirdly, is it possible to automatically derive personality from the behaviour of the child with help from questionnaires? The master thesis on this is not appended to this deliverable, but will be available from the ALIZ-e website [Robben, 2011]. In the second experiment, 30 children in the age of 8-12 interact three times (with three days in between each interaction) with the robot playing the simon game. During the simon game the children have to solve mathematical problems, both the difficulty of the simon game and the difficulty of the mathematical problems increase based on the performance of the child. With one half of the group, at each interaction, the difficult level starts at the beginning level. With the other half of the group the difficulty level is set at the level reached during the last interaction. The results are analysed at this moment. The third experiment was a pilot for the dance scenario. Two children (age 7 and 11) danced a short choreography with the robot. The dance included four movements: ranging from simple arm movements (single layer motion) to a combination of arms and legs (layered motion), and two balancing motions (lifting one leg to the side). In the annexes the report of this experiment is appended [Ros et al., 2011]. Planned experiments involve adapting a dance training to the child, adapting a quiz to the child and make it possible for the child to initiate a scenario switch (e.g. from quiz to dance). The aspects to base the personalization on were derived from T3.1. Furthermore, theoretical and empirical knowledge on how to measure these aspects and how to adapt to these aspects were incorporated in the different experiments.

T3.4 Context-aware human-robot interaction

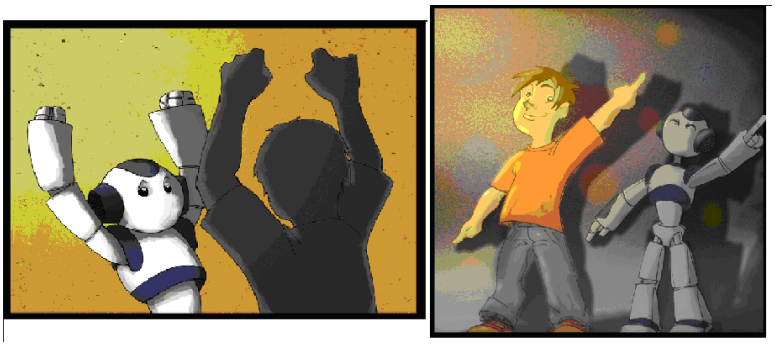


Figure 1: NAO in two scenarios, Simon game and dancing

The robot needs to be aware of its and the child's environment, and how this environment affects the child's interaction needs and preferences. For example, when a physician is in the same room as the child and the robot has learned the child has a question about its disease, the robot can remind the child about this. Furthermore, the interaction can differ when the child is alone in comparison to when there are more people in the room. Hence, it is important NAO to direct its attention to the person who is talking and to decide where the sound is coming from. In the first year, VUB has looked into the sound source localisation. A localisation algorithm is typically based on the difference in time delay on arrival (TDOA) of the audio signal at the different microphones in the array. We propose TDOA to be estimated based on computationally efficient Generalized Cross-Correlation (GCC) [Knapp and Carter, 1976]. The TDOA estimates can then be used to determine an estimate for the location of a sound source, both in 2D and in 3D. The specific NAO's microphone configuration (which is forming a sensor quadruple consisting of two orthogonal sensor pairs with a common centre) can be used to simplify these calculations. An evaluation of NAO's inbuilt localisation and the proposed approach is discussed in [Brouckxon et al., 2011].

The Voice Activity Detector (VAD) is another important element for facilitating the child-robot interaction. Our studies led us to believe that most VAD algorithms described in the literature are greatly affected by the type of background noise. Motivated by the fact that NAO will interact within an environment of unknown background noise conditions, we have examined the performance of different VAD algorithms for different background noise types that exhibit certain characteristics. In addition, a robust energy based voice activity detection algorithm has been proposed. The proposed energy based VAD algorithm can be configured for different noise conditions. Moreover, we showed that it outperforms conventional spectrum based VADs under certain noise conditions and certain configurations [Dekens and Verhelst, 2011].

Another aspect which is important for context-awareness is detecting and tracking objects. For visual attentional selection of objects of interest, we propose a novel biologically-inspired region-based focus-of-attention mechanism simulating the middle stages of visual human attention. It contains means for early identification and segmentation of perceptual object (called proto-regions or proto-objects), visual saliency computation, and object-based attention [Geerink and Sahli, 2011].

A very special object in the surroundings of the robot will be the face of the child. The child should not only be detected and tracked, but also recognized. In T3.4 a first version of a face recognition module has been built [Enescu and Sahli, 2011]. The proposed recognition module is seen as a pipeline consisting of face detection, face feature tracking, image processing and face recognition. Coping with adverse conditions, such as illumination changes and in-plane rotations, for an improved recognition accuracy is accomplished by the first three stages. In designing the final recognition stage, we were guided by two principles: i) real-time operation with minimal data storage, and ii) online learning and adaptivity. The novelties in our module consists of: i) building a simple and fast recognition system using face templates and sum of absolute differences as a similarity measure (instead of less performing Euclidean distances); ii) eyes alignment for meaningful face-to-templates comparisons via eye tracking and image warping; iii) providing a simple strategy for online learning the face templates (as opposed to batch learning) and for their

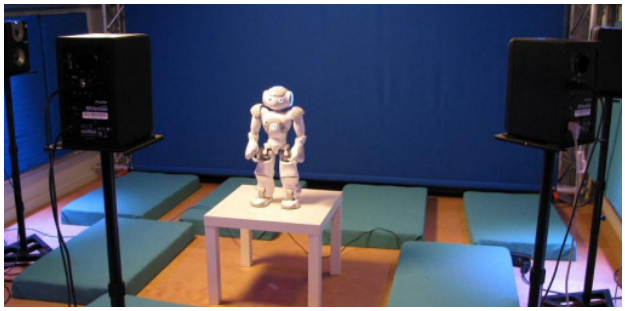


Figure 2: NAO in the sound-localisation experiment setup

subsequent update.

4.3 Relation to user-centric design

In WP3 we followed a user-centric design approach as follows from the situated Cognitive Engineering framework we apply. In T3.1 experts and children with diabetes were consulted, because both groups will be using the Nao. Some experiments with users in the same age as our diabetic children focus group were performed. In T3.2 and T3.3 experiments with children in the same age as the end users were performed. Also in pilot experiments the opinion of the children influenced the experimental setup in T3.1-T3.3. T3.4 has not used children yet, but it looks at relevant surroundings in both environment as people.

4.4 Relation to the state-of-the-art

Below we briefly discuss how the obtained results of the individual tasks relate to the current state of the art (the papers in the appendix provide more detailed references to the state-of-the-art).

T3.1 Define models of user, social environment, affect and interaction Our approach to user modelling is based on ongoing research on situated Cognitive Engineering (sCE) [Neerincx and Lindenberg, 2008]. sCE is advancing the state of the art because it is a theoretical and empirical approach that focuses both on design and evaluation, in contrast to approaches that only look at design [Bartneck and Forlizzi, 2005] or evaluation [Weiss et al., 2009]. In addition, sCE incorporates a human factors and technological perspective, making a complete integrated development process possible. This process consists of five components: 1) an integrated analysis of operational, human factors and technological drivers or constraints, 2) a requirements baseline specified by claims, use cases and core functions, 3) a built prototype, 4) evaluation and 5) refinement of 1 through 3 based on the results of the evaluation. In WP3, the first and second component are handled in T3.1 (see also the annexes about the interviews with the diabetic care team and with children with diabetes and their parents). T3.2-T3.4 provide implementations that are tested and evaluated and provide input back to the documents prepared in T3.1. In the User model document in the annex a short introduction of sCE is provided. User models are mostly used in teacher-student contexts. The students have to learn something and the teacher adapts to the knowledge the user already has (e.g. by using the zone of proximal development theory (ZPD) [Vygotsky, 1978]). We on the other hand look, next to these aspects, at aspects of the user that are less straight forward. Examples are personality and balance, for both it holds that it is not either 0 or 1. There are for instance introvert children and extrovert children but the scale ranges from 0 to 1 and is almost never completely 0 or 1, furthermore the level of introversion can change over interaction moments. The same holds from balanced, there will be few children completely imbalanced, but probably quite some that are not perfectly balanced. The challenges in this are to define the scale and be able to both recognize and adapt to this automatically.

T3.2 Long-term behaviour support Long-term behaviour support is essential for children with a disease that influences their whole lifestyle. In the healthcare domain there are examples of virtual agents for long-term behaviour support [Bickmore and Picard, 2005] and also in the domain of robotics [Kidd and Breazeal, 2008]. Both Kidd and Bickmore notice in their research that the interaction with the device decreases when the novelty effect wears of. Both try to reduce this effect by making the character less predictable and adaptive to the user. In contrast to their research, we focus on children and long term behaviour support. A similarity is that motivation is both important for adults and children. To look at the effect of multiple interactions and context, an experiment with these aspects was performed (see annex [Cohen et al., 2011]). We looked explicitly at the intrinsic motivation in two experiments. The rsults of one experiment are analyzed and show a ceiling effect, as children like the different conditions to a maximum [Robben, 2011]. The results from the second study are currently analyzed.

T3.3 Personalize robot behaviours Personalization and adaptation are important factors for long-term behaviour support [Bickmore and Picard, 2005, Kidd and Breazeal, 2008, Meneu et al., 2009]. This is not only evident from the healthcare domain, but also from other domains [De Bra et al., 2010, Li et al., 2007, Vasilyeva et al., 2008]. Our work will be based on the HAMMER architecture [Demiris, 2007] and the BDI based GOAL architecture [Hindriks, 2009] to learn the user models and subsequently act accordingly. In the first year, we explored these architectures for their capabilities to address the complexity of the user model and to support intentional behaviour. Several elements of this model and behaviour are tested with children in the same age group as the diabetic children target group; a challenge is to develop a model that can be applied to a suite of games or scenarios in a transparant, consistent and coherent way (current user models, most often apply to one game or application).

T3.4 Context-aware human-robot interaction In literature, a number of algorithms have been developed to automatically localise a single dominant speaker in the environment by using the microphone array signals. More complex location estimation techniques like MUSIC [Di Claudio and Parisi, 2001, Dmochowski et al., 2007] and ESPRIT [Roy et al., 1986] also exist and are based on high-resolution spectral estimation. These techniques can be used to simultaneously locate multiple sound sources, but require a significantly higher computational load and in some cases lead to lower temporal and spatial resolution of the location estimates. Therefore, we propose TDOA to be estimated based on computationally efficient and robust Generalized Cross-Correlation (GCC).

The proposed energy based Voice Activity Detector uses the feature of smoothed energy contained in the frequency region of interest and is configurable to different background noise types. An elaborate evaluation of the different configurations of the energy based VAD algorithm and a comparison with state-of-the-art can be found in [Dekens and Verhelst, 2011].

The object localization system uses a novel biologically-inspired region-based focus-of-attention mechanism simulating the middle stages of visual human attention. It uses object-based attention instead of the point-based attention used by most state-of-the-art systems, a detailed comparison with state-of-the-art can be found in [Geerink and Sahli, 2011].

The face recognition system is beyond the state-of-the-art due to its real-time behavior and ability to update the learned face database over time. The system uses state-of-the-art methods [Sung and Poggio, 1998, Viola and Jones, 2004, Wu et al., 2008] for face detection to initiate the face tracking and recognition process. Tracking is accomplished using the Active Shape Model [Cootes et al., 1995], which provides the position of the eyes. Based on this information, the image is warped so that the face becomes vertical, to enhance recognition accuracy. For face recognition, after a number of tests with recent state-of-the-art methods, we have decided to actually extend an earlier approach, namely the template-based method [Brunelli and Poggio, 2002]. The advantages of this method include accuracy, real-time operation and face learning easiness. We have then devised a simple and effective strategy for online learning and update of the face templates (stored faces). Finally, by storing dissimilar templates, our new method allows recognising faces with out-of-plane rotations and non-rigid deformations such as faces experiencing emotion and speaking.

References

[Bartneck and Forlizzi, 2005] Bartneck, C. and Forlizzi, J. (2005). A design-centred framework for social human-robot interaction. In *13th IEEE International Workshop on Robot and Human Interactive Communication, 2004. ROMAN 2004.*, pages 591–594. IEEE.

[Bickmore and Picard, 2005] Bickmore, T. and Picard, R. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

[Brouckxon et al., 2011] Brouckxon, H., Athanasopoulos, G., and Verhelst, W. (2011). Microphone array technology for robots: Sound source localisation. Technical report, Vrije Universiteit Brussel, dept. ETRO.

[Brunelli and Poggio, 2002] Brunelli, R. and Poggio, T. (2002). Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052.

[Cohen et al., 2011] Cohen, I., Looije, R., and Neerincx, M. (2011). Child's recognition of emotions in robot's face and body. In *Proceedings HRI 2011.*

[Cootes et al., 1995] Cootes, T., Taylor, C., Cooper, D., Graham, J., et al. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

[De Bra et al., 2010] De Bra, P., Kobsa, A., and Chin, D., editors (2010). *User Modeling, Adaptation, and Personalization*. Springer.

[Dekens and Verhelst, 2011] Dekens, T. and Verhelst, W. (2011). On the noise robustness of voice activity detection algorithms. To be submitted to InterSpeech.

[Demiris, 2007] Demiris, Y. (2007). Prediction of intent in robotics and multi-agent systems. *Cognitive Processing*, 8(3):151–158.

[Di Claudio and Parisi, 2001] Di Claudio, E. D. and Parisi, R. (2001). *"Multi-Source Localization Strategies", Chapter 9 in "Microphone Arrays: Signal Processing Techniques and Applications"*. Springer.

[Dmochowski et al., 2007] Dmochowski, J. P., Benesty, J., and Affes, S. (2007). Broadband music: opportunities and challenges for multiple source localization. In *Proc. IEEE WASPAA, pp. 18–21.*

[e project team, 2010] e project team, A. (2010). Scenarios with corresponding user and robot models, working document. Technical report, ALIZ-e.

[Enescu and Sahli, 2011] Enescu, V. and Sahli, H. (2011). Face detection, tracking and recognition with online learning.

[Geerink and Sahli, 2011] Geerink, T. and Sahli, H. (2011). Attentional selection of objects of interest.

[Groot, 2010] Groot, F. (2010). Medical, social and personal factors that influence self-management in children with diabetes: Current situation in the netherlands and implications for a social robot. Master's thesis, University of Amsterdam.

[Hindriks, 2009] Hindriks, K. V. (2009). *Multi-Agent Programming*, chapter Programming Rational Agents in GOAL, pages 119–157. Springer US.

[Hoondert and Blanson Henkemans, 2011] Hoondert, V. and Blanson Henkemans, O. (2011). The influence of personal and social determinants and disease management on the quality of life of dutch children with type 1 diabetes. In *NVDO Young researchers workshop.*

[Hoondert et al., 2010] Hoondert, V., Blanson Henkemans, O., Looije, R., Alpay, L., Janssen, J., and Neerincx, M. (2010). Alize-e: Social robots supporting children with diabetes. Technical report, TNO.

[Kanda et al., 2004] Kanda, T., Ishiguro, H., Imai, M., and Ono, T. (2004). Development and evaluation of interactive humanoid robots. *Proceedings of the IEEE*, 92(11):1839–1850.

[Kidd and Breazeal, 2008] Kidd, C. and Breazeal, C. (2008). Robots at home: Understanding long-term human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. IROS 2008.*, pages 3230–3235. IEEE.

[Knapp and Carter, 1976] Knapp, C. H. and Carter, G. C. (Aug. 1976). The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Process*, 24:320–327.

[Li et al., 2007] Li, L., Yang, Z., Wang, B., and Kitsuregawa, M. (2007). Dynamic adaptation strategies for long-term and short-term user profile to personalize search. *Advances in Data and Web Management*, pages 228–240.

[Looije et al., 2010] Looije, R., Neerincx, M., and Cnossen, F. (2010). Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies*, 68(6):386–397.

[Meneu et al., 2009] Meneu, T., Traver, V., Fernandez, C., Serafin, R., Dominguez, D., and Guillen, S. (2009). Life assistance protocols (lap)–a model for the personalization of lifestyle support for health and wellbeing. In *International Conference on eHealth, Telemedicine, and Social Medicine, 2009. eTELEMED'09.*, pages 140–144. IEEE.

[Nalin et al., 2011] Nalin, M., Bergamini, L., Giusti, A., Baroni, I., and Sanna, A. (2011). Children's perception of a robotic companion in a mildly constrained setting: How children within age 8–11 perceive a robotic companion. In *Proceedings HRI 2011*.

[Neerincx and Lindenberg, 2008] Neerincx, M. and Lindenberg, J. (2008). *Naturalistic Decision Making and Macrocognition*, chapter Situated cognitive engineering for complex task environments., pages 373–390. Aldershot, UK: Ashgate Publishing Limited.

[Robben, 2011] Robben, S. (2011). Facilitate bonding between a child and a social robot: Exploring the possibility of a robot adaptive to personality. Master's thesis, University of Nijmegen.

[Ros et al., 2011] Ros, R., Demiris, Y., Baroni, I., and Nalin, M. (2011). Adapting robot behavior to user's capabilities: a dance instruction study. In *Proceedings HRI 2011*.

[Roy et al., 1986] Roy, R., Paulraj, A., and Kailath, T. (Oct. 1986). Esprit - a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-34:1340–1342.

[Sung and Poggio, 1998] Sung, K.-K. and Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51.

[Vasilyeva et al., 2008] Vasilyeva, E., Pechenizkiy, M., and De Bra, P. (2008). Adaptation of elaborated feedback in e-learning. *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 235–244.

[Viola and Jones, 2004] Viola, P. and Jones, M. (2004). Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154.

[Vygotsky, 1978] Vygotsky, L. (1978). *Mind and society: The development of higher psychological processes*. Cambridge , MA; University Press.

[Weiss et al., 2009] Weiss, A., Bernhaupt, R., Lankes, M., and Tscheligi, M. (2009). The usus evaluation framework for human-robot interaction. In *AISB2009: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction*.

[Wu et al., 2008] Wu, J., Brubaker, S. C., Mullin, M. D., and Rehg, J. M. (2008). Fast asymmetric learning for cascade face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):369–382.

5 Annexes

5.1 Groot, Femke (2010), Medical, social and personal factors that influence self-management in children with diabetes: Current situation in the Netherlands and implications for a social robot

Bibliography Groot, Femke (2010), Medical, social and personal factors that influence self-management in children with diabetes: Current situation in the Netherlands and implications for a social robot, Masterthesis, University of Amsterdam. [Groot, 2010].

Abstract **Background** - Diabetes type 1 is a chronic disease that will increase exponentially among children in the near future. Professional healthcare is essential for these children to prevent serious complications later in life. Unfortunately, due to the ageing of the population and a decline in the labor market for health personnel, health professionals will be confronted with an increasing load. To contribute to this challenge a social robot that is capable of supporting these children as a coach, teacher and companion will be developed and evaluated within the project ALIZ-E. In the underlying study, that is part of the ALIZ-E project, factors that influence diabetes management of Dutch children are examined and implications for the social robot are being discussed.
Method- Several members of diabetes healthcare teams throughout the Netherlands have been interviewed about different aspects of their experience with 8 to 10 year old children with diabetes. Conclusions drawn from these interviews are complemented with a variety of brochures, web based information, research articles, and books. By this, the result of the study is based on both science and practice.
Results - Findings suggest that Dutch children in the age cohort between 8 to 10 years do not show significant problems with coping and their diabetes management in general. A probable explanation for this is that their parents are still largely responsible for the execution of the different diabetes activities. However, to prevent problems with responsibility and acceptance during puberty, diabetes management has to be handed over to the child in a stepwise fashion and adjusted to the development of the child before he enters puberty.
Conclusion - Medical, social and personal factors that promote or slow down the process of acquiring all the needed knowledge and skills for successful diabetes management presumably have most impact on the child's attitude towards diabetes self-management and future well-being. Factors that promote progress are for instance regular visits to the diabetes healthcare team, sufficient diabetic education, cooperation between the parents, high agreeableness, high conscientiousness and low neuroticism as characteristics of both the child and the primal caregiver of the child and the use of coping strategies that are appropriate for the age and development of the child. Therefore, it is recommended that the different robot roles should support these factors and do not interfere with them. Further research should also include the experiences of children and parents.

Relation to WP This work directly contributes to Tasks T3.1.

5.2 Hoondert, Vera et al. (2010), ALIZE-E Social Robots Supporting Children with Diabetes

Bibliography Hoondert, Vera et al. (2010),ALIZE-E Social Robots Supporting Children with Diabetes, report. TNO. [Hoondert et al., 2010].

Abstract Social robots can offer long-term support to children with diabetes. It requires attending to diabetes management determinants. Earlier studies entail professionals' perspective; this study focuses on children's perspective.

Relation to WP This work directly contributes to Task T3.1.

5.3 Hoondert, Vera and Blanson Henkemans, Olivier. (2011), The influence of personal and social determinants and disease management on the quality of life of Dutch children with type 1

Bibliography Hoondert, Vera and Blanson Henkemans, Olivier. (2011), The influence of personal and social determinants and disease management on the quality of life of Dutch children with type 1 diabetes[Hoondert and Blanson Henkemans, 2011].

Abstract **Background** In the Netherlands approximately 15,000 children suffer with diabetes, 98% of them are having type 1 diabetes. It is important that they regulate their glycaemic blood level effectively, otherwise it can result in various short and long-term complications. To regulate their glucose level, the children have to learn about their disease, develop self-management skills and control their carbohydrate intake level and activity level. Still, these children are developing emotionally and cognitively and self-management of the disease is complex. Moreover, the children have to find a balance between the disease management and their quality of life. The ALIZ-E project looks at opportunities for social robots that support these children with the complexity of their diabetes self-management. As starting point, it is important to have more understanding about how Dutch children cope with their disease, their perception on their diabetes self-management and which influence diabetes has on their quality of life. Previous research focus on determinants from a care givers perspective. In this study we focus on the children perspective. **Methods** To gain insight in children perception of their diabetes, we interviewed children between the ages 8-12 who have type 1 diabetes for more than six months. In this semi-structured interview we asked the children about important life domains, their perception on their diabetes self-management and how this influences their experienced quality of life. Alongside the interview with the child, we asked the parents for completing a questionnaire. **Results** The first results show that children in our study accept diabetes as a part of their life, but do not want to be considered as being diabetic too much by friends. The children strongly appreciate the support of teachers, parents and siblings, but there are differences in the need for support from classmates. Their diabetes management greatly depends on the support of their parents and school. Diabetes can also negatively affect the school performance of the children. Although children make the best of their life, managing diabetes goes at the cost of their mental and social wellbeing. Finally, children like having friends with diabetes, due to common ground. **Conclusions** These results have the following implications for the social robot. First, the self-management support should take place on the child's request. Second, the robot needs to attend on the social environment of the child, the robot needs to be aware of the diabetes on the social and mental wellbeing of the child. Finally, the robot can play a role in balancing diabetes management and school work and the social robot can act as a friend, who also has diabetes. With these findings we can work further on the development of social robots that contribute to the self-management children with diabetes, while maintaining a good quality of life.

Relation to WP This work directly contributes to Task T3.1.

5.4 Nalin, Marco et al. (2011), Children's perception of a Robotic Companion in a mildly constrained setting: How children within age 8 - 11 perceive a robotic companion

Bibliography Nalin, Marco et al. (2010),Children's perception of a Robotic Companion in a mildly constrained setting: How children within age 8 - 11 perceive a robotic companion, HRI 2011 [Nalin et al., 2011].

Abstract This paper presents the results of a study, conducted by Scientific Institute San Raffaele in Milan, involving 35 children in between 8 and 11 years old. The purpose of this study was to assess the children's perception of a robotic companion. The interaction was organized in small groups (3-4 children per session), for a quite short duration (15min), and was structured in form of game, where the children had to discover how to activate all the robot's capabilities (four in total, one of which including physical contact with the robot). The robot was controlled through a Wizard of Oz interface, thanks to which an operator was able to activate the specific behaviors. The study demonstrated that all the kids accept favorably the presence of the robot, and that they are willing to spend more time with it. Furthermore the study indicated that children have the tendency to humanize the robot, assigning it functions, behaviors and emotions that are typical of human beings. Another interesting result is that all the children claimed (through a proper questionnaire) that the robot could be able to support them, in case they were feeling down or worried about something.

Relation to WP This work directly contributes to Task T3.1.

5.5 ALIZ-e project team (2010), Scenarios with Corresponding User and Robot Models

Bibliography ALIZ-e project team (2010), Scenarios with Corresponding User and Robot Models, working document [e project team, 2010].

Abstract The ALIZ-E project aims at a prolonged or recurrent child-robot interaction that has a positive effect on child's well-being. Three robot roles have been identified, and multiple scenarios are drawn up in which the interaction between children and robot take place. For a good execution of these scenarios and adequate adaptive robot behaviors, it is important to define the underlying user and robot models that drive the personalized interaction (and can be parameterized). This document gives an overview of the different elements and for each scenario what these elements entail.

The scenario elements are the following:

1. Goal: What goals are aspired during the scenario?
2. Success factors: What outcomes (specific and measurable) indicate if a goal is realized or not?
3. Robot behavior: What robot behavior contributes to these factors?
4. Scenario specific user model: What does the robot need to know about the user to perform this behavior, specific for this scenario?
5. Generic user model: What does the robot need to know about the user to perform appropriate behavior, in general?
6. Robot model: What classes of characteristics must the robot have and which memory elements to express prolonged personalized behavior?

In the first year, ALIZ-E mainly focuses on the Simon, Dance and Quiz games. The annex shows a first version of the resulting user model aspects, the working document can be made available on the ALIZ-e website.

Relation to WP This work directly contributes to Task T3.1.

5.6 Cohen, Iris, Looije, Rosemarijn and Neerincx, Mark (2011), Child's Recognition of Emotions in Robot's Face and Body

Bibliography Cohen, Iris, Looije, Rosemarijn and Neerincx, Mark (2011), Child's Recognition of Emotions in Robot's Face and Body, HRI 2011 [Cohen et al., 2011].

Abstract Social robots can comfort and support children who have to cope with chronic diseases. In previous studies, a "facial robot", the iCat, proved to show well-recognized emotional expressions that are important in social interactions. The question is if a mobile robot without a face, the Nao, can express emotions with its body. First, dynamic body postures were created and validated that express fear, happiness, anger, sadness and surprise. Then, fourteen children had to recognize emotions, expressed by both robots. Recognition rates were relatively high (between 68the recognition was better for the iCat (95Nao (68recognitions. In a second session, the emotions were significantly better recognized than during the first session for both robots. In sum, we succeeded to design Nao emotions, which were well recognized and learned, and can be important ingredients of the social dialogs with children.

Relation to WP This work directly contributes to Task 3.2.

5.7 Robben, Saskia (2011), Facilitate Bonding Between a Child and a Social Robot: Exploring the Possibility of a Robot Adaptive to Personality

Bibliography Robben, Saskia (2011), Facilitate Bonding Between a Child and a Social Robot: Exploring the Possibility of a Robot Adaptive to Personality. Masterthesis, University of Nijmegen [Robben, 2011].

Abstract In this thesis we investigate whether it is worth to let a social robot adapt itself to personality (extroversion) to aid bonding between a robot and a child. After a short exploratory study, we designed an experiment where 14 children aged 10 played a mimicking game for 10 minutes with two different robots. One robot moved slightly slower and gave nurturing feedback (the 'introvert' robot) and the other moved slightly faster and gave challenging feedback (the 'extrovert' robot). Our first question was whether the children noted the differences, and the second was if the children had a personality related preference. To answer these questions we assessed personality (extroversion) of the children with a subset of the Big Five Questionnaire for Children (BFQ-C), furthermore we adopted both subjective and objective metrics to analyze the interaction regarding fun, bonding and trust. Further questions that needed to be addressed are whether the robot is able to infer extroversion from the behavior of the child and secondly whether the robot is able to learn meaningful adaptation rules. The main conclusion is that the children very much liked playing the game with both of the robots. There was no difference found between the two robots, it is possible that they are not distinguishable enough, and the experiment lasted not long enough to pass the novelty effect and reveal the subtle differences. There was no personality related preference. However the score on extroversion was densely distributed which raises some questions on personality assessment with children of this age in general. Our conclusion is that our robot was already perceived as a social being to a great extend, regardless of personality adaptation. To further facilitate bonding there are a lot of other aspects which need attention first, for example the gender of robot or child or the development of optimal scenarios and dialogues. We have to make giant leaps in child-robot interaction elsewhere before returning to personality.

Relation to WP This work directly contributes to Task 3.3.

5.8 Ros, Raquel et al. (2011), Adapting Robot Behavior to User's Capabilities: a Dance Instruction Study

Bibliography Ros, Raquel et al. (2011), Adapting Robot Behavior to User's Capabilities: a Dance Instruction Study, HRI 2011 [Ros et al., 2011].

Abstract The ALIZ-E project's goal is to design a robot companion able to maintain affective interactions with young users over a period of time. One of these interactions consists in teaching a dance to hospitalized children according to their capa- bilities. We propose a methodology for adapting both, the movements used in the dance based on the user's cognitive and physical capabilities through a set of metrics, and the robot's interaction based on the user's personality traits.

Relation to WP This work directly contributes to Task 3.3.

5.9 Brouckxon, H. and Athanasopoulos, G. and Verhelst, W (2011), Microphone Array Technology for Robots: Sound Source Localisation

Bibliography Brouckxon, H. and Athanasopoulos, G. and Verhelst, W (2011), Microphone Array Technology for Robots: Sound Source Localisation, Technical report, Vrije Universiteit Brussel [Brouckxon et al., 2011].

Abstract A short overview on what has been done on the subject of sound localisation in the first year of the ALIZ-e project.

Relation to WP This work directly contributes to Task T3.4.

5.10 Dekens, T and Verhelst, W. (2011 / unpublished), On the Noise Robustness of Voice Activity Detection Algorithms

Bibliography Dekens, T and Verhelst, W. (2011 / unpublished), On the Noise Robustness of Voice Activity Detection Algorithms [Dekens and Verhelst, 2011].

Abstract In this paper, we show that the performance of voice activity detection algorithms (VAD) can be highly dependent on the type of background noise and we introduce a new VAD algorithm that is based on relative energy measurements in different frequency bands. The obtained experimental results are compared to the results obtained with two other spectrumbased VADs and it is concluded that a VAD, configured to use around 3 frequency bands can cope best with a large variety of background sounds.

Relation to WP This work directly contributes to Task T3.4.

5.11 Geerink, T. and Sahli, H. (2011 / unpublished),Attentional Selection of Objects of Interest

Bibliography Geerink, T. and Sahli, H. (2011 / unpublished),Attentional Selection of Objects of Interest [Geerink and Sahli, 2011].

Abstract object-based attention requires considering the following aspects:

- Early identification and segmentation of perceptual objects. These early identified objects are called protoregions or proto-objects.
- The relationship between object-based and space-based attention
- Grouping/segmentation and object-based attention. A grouping is a hierarchical structure of objects and space. A grouping may be a point, an object, a region, or a structured grouping.
- Visual saliency and visual attention. The salience of a grouping measures how much this grouping contrasts with its surroundings and depends on various factors, such as feature properties, perceptual grouping, dissimilarity between the target and its neighborhood.

Following the above criteria, we propose a novel biologically inspired region-based focus of attention mechanism simulating the middle stages of attention, with specific algorithmic details. Moreover, the behavioral responses of the model and human attention using eyetracker are compared and their correlation is measured.

Relation to WP This work directly contributes to Task T3.4.

5.12 Enescu, V. and Sahli, H. (2011 / unpublished), Face detection, tracking and recognition with online learning

Bibliography Enescu, V. and Sahli, H. (2011 / unpublished), Face detection, tracking and recognition with online learning [Enescu and Sahli, 2011].

Abstract This paper describes our work on designing a face recognition module for the humanoid robot Nao, within the EU Aliz-e project. We regard this task as a pipeline consisting of face detection, face features tracking, image processing and, finally, face recognition. The first three stages are meant to increase resistance of recognition against adverse conditions such as illumination changes and in-plane rotations. As to the last stage, our goals are i) to achieve real-time operation with minimal data storage, and ii) to learn new faces or update the already stored ones in an online manner.

Relation to WP This work directly contributes to Task T3.4.

TNO | Innovation for Life

Vera Hoondert
Olivier Blanson Henkemans
Rosemarijn Looije
Laurence Alpay
Joris Janssen
Mark Neerincx
TNO Prevention & Care, Netherlands

SEVENTH FRAMEWORK PROGRAMME

vrije Universiteit

ALIZE-E Social Robots Supporting Children with Diabetes

Social robots can offer long-term support to children with diabetes. It requires attending to diabetes management determinants. Earlier studies entail professionals' perspective; this study focuses on children's perspective.

CHILDREN WITH DIABETES

Diabetes is one of the fastest growing chronic illnesses amongst adults, but also amongst children. In the Netherlands, approximately 15,000 children suffer from diabetes. To regulate their glucose level, these children learn about their illness and develop self-management skills (i.e., inject insulin) corresponding to their carbohydrate intake and activity level. Children between the age 8 and 12 learn to cope with their diabetes after diagnosis and subsequently prepare themselves for puberty. This is a period in which they experience a strong cognitive, emotional and physical development.

ALIZ-E

Personal computer assistants, such as social robots, offering support as a companion, educator and motivator can contribute to self-management. Moreover, to realize long-term interaction, they can be programmed towards socially intelligent behavior, including assessing and responding to personal determinants for diabetes self-management. The European 7th framework program ALIZ-E is looking at social robots for long-term interaction with children with diabetes and their environment.
For this study, we are focusing on child-specific diabetes self-management determinants and looking at how social robots can attend to them.

CHILD-SPECIFIC DIABETES SELF-MANAGEMENT DETERMINANTS

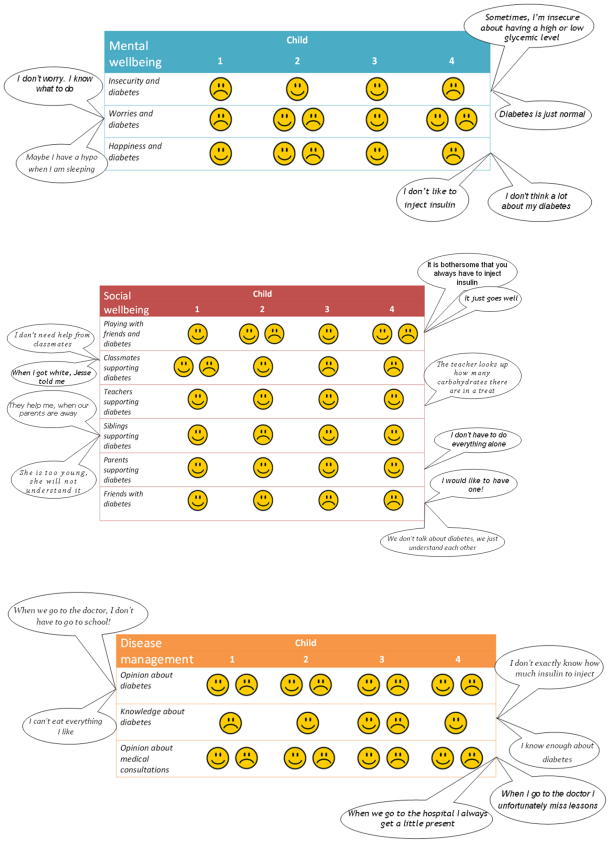
Previous research on self-management determinants, conducted from a health care professionals' perspective, shows the following:

- The children group does not experience significant problems with diabetes management and their glucose levels are relatively well regulated;
- The parents and guardians' support of diabetes self-management (through cohesion and communication) forms an important determinant for successfull self-management;
- Parent involvement is important due to the complexity of self-management for children, who are still developing cognitively and emotionally;
- During puberty, the child's need for self-control increases and he or she may rebel against his or her parents, at the cost of diabetes self-management and leading to unhealthy blood glucose levels.

CHILDREN'S PERSPECTIVE ON DIABETES SELF-MANAGEMENT

Currently, we study the children's perspective on their diabetes and self-management in relation to quality of life. We conducted semi-structured interviews with children with diabetes ($N=4$). We asked about important life domains (e.g., school, sports, hobbies, family, friends and activities in the weekend), perception of their diabetes management and the effect of the diabetes self-management on quality of life aspects (i.e., mental wellbeing, physical wellbeing, and social wellbeing). From these interviews we elicited important shared determinants:

- Mental wellbeing,
- Social wellbeing and
- Disease management.



DISCUSSION

Results show that mental and social wellbeing and disease management are important topics for children with diabetes. They have both positive and negative experiences with or views on these topics. Overall:

- Children accept diabetes as part of their life, but do not want to be considered different from friends;
- Children like having friends with diabetes, due to a common ground;
- There are differences in the need for support from classmates;
- Children appreciate the support form parents, teachers, and siblings;
- Diabetes management depends on the support of parents and school;
- Although children make the best of their life, managing diabetes goes at the cost of their mental and social wellbeing;
- Diabetes (e.g., symptoms and hospital visits) negatively affect school performance.


Kids with social robot

IMPLICATIONS

These results have the following implications for the interaction between children and social robots, supporting diabetes self-management:

- As children do not want diabetes to define them, diabetes management support should take place preferably on the child's request;
- The social robot needs to attend to the social environment of the child;
- The social robot needs be aware of the effect of diabetes on mental and social wellbeing;
- The social robot can play a role in balancing diabetes management and school work.
- The social robot can act as a friend, who also has diabetes.

CONTACT:
Dr. Olivier Blanson Henkemans
TNO Quality of Life • P.O. Box 2215 • 2301 CE Leiden • T: 088 866 61 86
E : olivier.blansonhenkemans@tno.nl

Kvl-L.10-12.1234Ep

Children's perception of a Robotic Companion in a mildly constrained setting

How children within age 8–11 perceive a robotic companion

Marco Nalin
Fondazione Centro San Raffaele del Monte Tabor
via Olgettina 60, 20132
Milan, Italy
nalin.marco@hsr.it

Linda Bergamini
Studio di Psicologia Nacsis
via Uberti 26, 20129
Milan, Italy
bergamini@nacsis.it

Alessio Giusti
Fondazione Centro San Raffaele del Monte Tabor
via Olgettina 60, 20132
Milan, Italy
giusti.alessio@hsr.it

Ilaria Baroni
Fondazione Centro San Raffaele del Monte Tabor
via Olgettina 60, 20132
Milan, Italy
baroni.ilaria@hsr.it

Alberto Sanna
Fondazione Centro San Raffaele del Monte Tabor
via Olgettina 60, 20132
Milan, Italy
sanna.alberto@hsr.it

ABSTRACT

This paper presents the results of a study, conducted by Scientific Institute San Raffaele in Milan, involving 35 children in between 8 and 11 years old. The purpose of this study was to assess the children's perception of a robotic companion. The interaction was organized in small groups (3-4 children per session), for a quite short duration (15min), and was structured in form of game, where the children had to discover how to activate all the robot's "capabilities" (four in total, one of which including physical contact with the robot). The robot was controlled through a Wizard of Oz interface, thanks to which an operator was able to activate the specific behaviors. The study demonstrated that all the kids accept favorably the presence of the robot, and that they are willing to spend more time with it. Furthermore the study indicated that children have the tendency to humanize the robot, assigning it functions, behaviors and emotions that are typical of human beings. Another interesting result is that all the children claimed (through a proper questionnaire) that the robot could be able to support them, in case they were feeling down or worried about something.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Psychology; H.1.2 [**Models and Principles**]: User/Machine Systems—*human factor, software psychology*

General Terms

Experimentation, Human Factors

Keywords

Robotic companion, children's perception, affective interaction

1. INTRODUCTION

Scientific Institute San Raffaele (HSR), in Milan, is studying how *long term robotic companions* can become an important support tool for easing the experience of the hospitalization of young patients. San Raffaele is participating in the Aliz-e project[1], which long term aim is to implement believable, long-term, affective child-robot interaction, in order to support the child's well-being and facilitate therapeutic activities in the hospital setting. Our initial target group ranged from 8–12 years old children with metabolic disorders (i.e., diabetes and obesity). The project will be implemented over a period of four and a half years in an incremental, iterative fashion. The first goal is to implement behaviors for establishing an initial relationship between robot and child through mutual engagement in simple games and interactions. This relationship will provide the substrate upon which more complex interactions will be based for subsequent goals. The first year of the project thus focuses on 1) establishing a bond between the child and the robot, 2) gaining the child's trust, and 3) introducing primary concepts about metabolic disorders and healthy lifestyles.

On this line, the first step to achieve the project results is to determine the acceptance of the robot by the children. Some preliminary experiments were conducted using the robot Nao, bought from Aldebaran Robotics[2], and the conclusions were that it appears that the robot has a promising capacity to engage children from a wide age range (preliminary experiments involved children within age 4–11), and that it is perceived as an non-threatening source of entertainment (i.e., as something between a peer and a toy). Lin et al. [6] studied the possible application of the robot as support to children's education and observed that most children

[1]http://www.aliz-e.org
[2]http://www.aldebaran-robotics.com/

have a good impression and positive attitude toward seeing the robot in classroom. These observations lend support to the notion that such a robotic platform could be effective in communicating healthcare information to child users in entertaining and memorable ways.
Once verified children's acceptance of the robotic companion, the second step was determining how children were actually perceiving this companion. As described by Khan et al. [5], we are in the case that the robot cannot of course become a human (ontologically speaking), but people (psychologically speaking) believe the robot is a human, and act accordingly. This paper presents the results of preliminary experiments, carried out in two classes of a primary school in Italy, to verify this perception. In the next section we will describe the methodology used to obtain measurable results, while the following sections will be dedicated respectively to the presentation of the results and their discussion.

2. GOALS

The purpose of this study was to assess the children's perception of a robotic companion. Particular interest was focused on three main aspects: 1) perception of the robot characteristics (like age, sex, etc.), 2) perception of the "robot being" (and how close this was to "human being", in experiencing emotions and feelings, but also in carrying out autonomous life and activities), and 3) perception of their relationship with the robot.

3. METHODOLOGY

Previous experiments at San Raffaele Hospital, to assess the children's acceptance of the robotic companion, were strongly biased on the image of the robot that the experimenters presented to the children. While on one hand this was useful for the sole purpose of making the robotic companion believable, on the other hand it had the drawback that in those preliminary experiments it was very hard to really understand the children's perception of the companion. What is very common in HRI is to recur to design patterns (for example see Khan et al. [4]) to study a particular interaction, but these kind of study strongly bias the robot perception in the child. As opposite, this study was organized to understand how in an ideally unconstrained (in reality: mildly constrained) setting the child would perceive the robotic companion. Experiments involved two classes of an Italian primary school, including in total 35 children. These experiments were designed to capture as much as possible the children's expectations and perceptions of the robotic companion, without influencing their opinion beforehand. For the experiment, three Nao robots purchased from Aldebaran Robotics were given. They came in three different colors (red, green and blue), and names with the letters of the Greek alphabet (beta, gamma and delta). These names were chosen to avoid a strong sexual characterization of the robots (like Jack, Mindy, etc.), but providing a simil-name different from numbers or letters (e.g., R2D2, C-3PO, etc.), which could have provided instead a strong characterization of the robot as an object.

The experiment consisted of four phases:

1. **Presentation:** The robot introduced itself to the full classroom, providing only basic information about its physical appearance and the sensors equipments. After this brief presentation (approximately 5min), the classroom was divided into small groups of 4-5 children.

2. **Small group interactions:** In the second phase, each group was free to discover what the robot could do and 15 minutes were given to each small group to experiment. Children were then told that the robot was able to do "several" actions, and that they had to find out what they had to do to activate these actions. Furthermore they were instructed not to touch the robot for the first 10 minutes of the interaction. Each robot had four behaviors installed that an operator was able to select through a Wizard of Oz interface. The four behaviors were 1) dancing (activated by showing a dance, playing a music, singing, etc.), 2) laughing (activated by telling jokes, or showing funny things), 3) walking and following a child (activated by vocal commands ordering to follow them, the robot was following and looking at the closest face), and 4) repeat with the hands and the head the movement that the children were making (activated by moving the hands and the head of the robot). The first three interactions were done without touching the robot, while in the last one they were free to touch it. The interaction was led by a psychologist, who guided the children in making hypothesis relating to the robot capabilities and then in testing out these hypothesis on the robot. The groups that completed the interaction were led to another room, where they were asked to make a drawing of the robot. This was just a mean of separating them from the other children who had yet to experiment with the robot, so as to avoid any word of mouth or exchange of opinions.

3. **Questionnaire:** The third phase was plenary with the full classroom again. Children were requested to fill a questionnaire that will be presented in the next section. The questionnaire was designed to collect from the children their perceptions and an evaluation of their relationship with the robot.

4. **Question and answer session:** The final phase of the activity consisted in answering the children's questions about the robot. This phase was led by both the psychologist and the engineers who created the behaviors of the robot. This phase was quite useful, because from the questions it was possible to gather important information on the children's perception of the robot.

4. RESULTS

In this section the answers gave by the children to the questionnaire will be illustrated.

4.1 Perception of robot characteristics

A group of question was related to the robot characteristics, to understand what the children think about its appearance, background, and abilities. When asked about what the robot could do, all the kids said that the robot was able to do at least one of the actions given as multiple choice: most of them ticked that it was able to play sports, read and write, travel, paint and draw. Some children also added actions by their own, like walking, telling tales, teaching,

running, dancing and jumping. Furthermore they agreed on its capability to do something more than just simple games or quizzes, listing among them, teaching (again), helping children, amusing, playing sport (again) and making happy those who need it.
As far as age is concerned, 40% of the children said that the robot was between 1 and 5 years old, followed by who thought that it was between 5 and 10 years old (25.7%), and those thinking that it was younger that one year old (22.9%). The interesting thing is even though they knew that the robot had batteries and everybody realized that quite clearly, still 20% of the children answered that it didn't have batteries.
To the question on the sex of the robot, most of the kids answered "neither male nor female" (48.6%), while 34.3% of the children ticked "male" and the rest (22.9%) selected "female".

4.2 Humanization of the robot

The second group of questions was meant to understand how close do the children feel the robot's character with respect to themselves or, more in general, to human beings. Most of the results indicated that children perceived the robot as very similar to humans: 88.6% of them thought that the robot had feelings, 60% ticked that it worked like a human (some of them, in the "other" option, specified "it works like a human-computer", "it works like a human with batteries") and 88.6% said that it seemed to be human (see figures 1, 2, 3 for details).
All the children (100%) answered that the robot could have helped them if they were feeling down or worried about something.

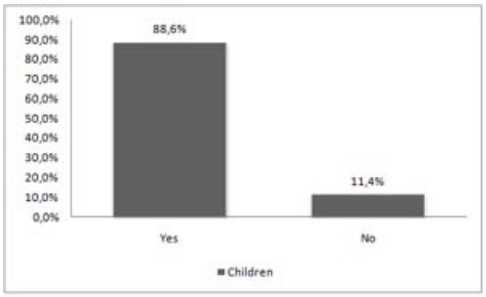


Figure 1: Does the robot have feelings?

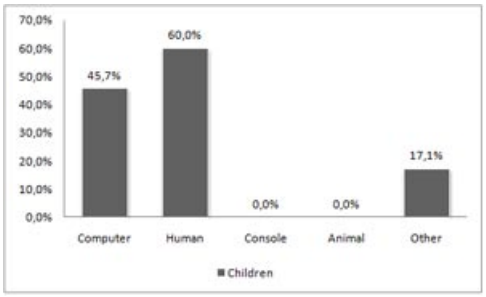


Figure 2: How does the robot work like?

4.3 Relational aspects of the robot

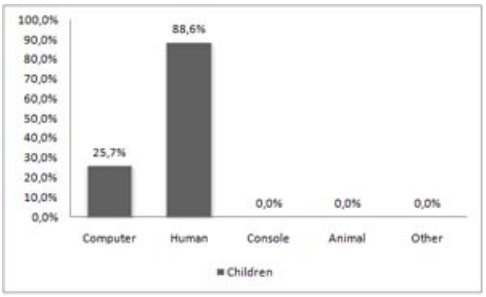


Figure 3: What does the robot seem to be?

The last group of questions was meant to understand both the perceived relational aspects of the robot and the relationship established with the children in this short study. The whole group of children, when asked in the questionnaire, claimed that they were curious and happy during the interactions with the robot, nobody selected negative options (e.g., scared, dislike, etc.), and everyone expressed the willingness to spend more time with it.
85.7% of the children, when directly asked about what the robot could be for them, with the possibility to select among "teacher", "coach" and "companion", selected this last role (followed by 22.9% selected teacher and 11.4% coach). These possible roles are those identified by the Aliz-e project as the most likely to work in the hospital context to provide support in the different moments of the hospitalization experience, and they were extracted from the previous research of R. Looije [7].
As far as the robot's personal relationship is concern, all of them said that it had friends, and all of them but one wrote that the robot has also a family. To the question "what would you like to do with the robot?" (multiple choices were allowed), 88.6% of the kids answered play, 65.7% talk, 60.0% play sport and 34.3% study.
All the children except one expressed the desire to be friend of the robot.

5. DISCUSSION

The results of the presented study clearly indicated that the robot is able to create believable relationships with the children. From the observation of the interaction, it resulted that beside being something new, unknown and weird, the robot doesn't raise any negative reactions such as refusal or fear.
From the physical appearance point of view, the Nao robot was very suitable for this purpose: it is a humanoid robot, with a lot of degrees of freedom and fluidity in the movements, still being a cartoon-like character, and not a realistic human character. As hypothesized by Mori [9], and later confirmed by Bartneck et al. [2], too much realism in the robot appearance reduces the person's familiarity with the robot, and could potentially lead to disgust and repulsion with the so-called "zombie effect". Nao's appearance is very close to the cartoons characters the children are used to.
Furthermore the study demonstrated a strong tendency of the children to humanize the robot, i.e., to attribute to the robot characteristics, behaviors and feelings that are typical of human beings. As described in the results section, 88.6% of the sample indicated that the robot has its own feelings. This can be strongly biased by the fact that the robot was

able (i.e., programmed) to laugh. Indeed most of the children, during the discussion of the questionnaire, adduced this fact as the justification for the given answer. Simple emotional expressions like this are developed also in the humans since the early stage of development, as described by R. Vasudevi [10], and they are indication of being in tune with other persons. The robot was provided with this capability to simulate the same mechanism to generate a reaction in the child.

On the other hand, this ability to express emotions through a simple laugh, both with the voice (i.e., verbal parameters and para-verbal parameters, like intonation, speed pauses, speed, etc.) and with the body (i.e., non-verbal parameters, as the robot was also moving the head and the arms), proven to be an effective expedient to give this impression to the users, and can be used to establish a first bond. These concepts were derived from the NLP theory by Dilts et al. [3] and Bandler and Grinder [1].

Other arguments to the facts that the robot was seen more as a "human" being than an object was demonstrated by the social relationships attributed to the robot (97.1% said that the robot has a family and 100% said that it has friends). Furthermore, most of them perceived that the robot was functioning like a human being (60.0%) and was behaving/looking like a human being (88.6%).

The conclusion of the humanization is supported also by the questions that the children posed to the engineers in the last phase of the activity, e.g., "how long does it live?", "how old is it?", etc. (note that in the Italian language there is no distinction between "it" and "he", thus questions sounded like "how long does he live?", etc.), as well as by the attachment demonstrated by the children (e.g., they wanted to give the robot a gift like a drawing and a toy car, and they were asking when they would see it again). Finally the study pointed out that there is a high chance that this humanization and anthropomorphism of the robot behaviors and expressions can be used in a different context (like a hospital), to provide support and stimulus for the children to build a believable and trustful relationship between robots and children. This hypothesis is supported also by the children's answers, when directly asked if they think that the robot could be able to relief them in case they are sad or worried, 100% of the sample group answered positively. Deriving from this last data the fact that children somehow trust the robot, an interesting hypothesis to be investigated in the future will be to verify if the robot can also enforce coping mechanisms in the child (as claimed by Marchetti et al. [8].), to be used to better react and participate to the treatment process.

6. CONCLUSIONS

This paper presents the results of a study conducted in a school with over 35 children between 8 and 11 years old. The study was meant to understand how children perceive a robotic companion and what do they believe that it can do, without biasing them with proper presentations in advance. A first result was that all of them accepted favorably the presence of the robot, and, more important, they would like to spend more time with it to get to know it better. The second main result is that children humanize the robot, assigning it functions, behaviors and emotions that are typical of human beings. Older children (11 years old) are more concrete (e.g., the robot can work, play, etc.), while younger children (8 years old) are still very open to magic and fantastic possibilities (e.g., the robot can fly in the space, etc.), but all of them attribute it human characteristics. The third and last result of the study is that the children believed that the robot could be able to support them in a hard moment (100% of the sample answered positively). The long term vision of this study is the creation at San Raffaele Hospital in Milan, of a robotic companion that is able to support children during the hard experience of hospitalization. In this case, probably the ability of the robot to express feelings through verbal and non-verbal expressions, and the children's ability to perceive these feelings, will be a strong point for creating a bond between the robot and the child, and leverage on this bond can be used to provide to young patients a proper motivational support.

7. ACKNOWLEDGMENTS

This work is supported by the EU Integrated Project ALIZ-E (FP7-ICT-248116).

8. REFERENCES

[1] G. J. Bandler, R. *The Structure of Magic II: A Book About Communication and Change*. Science and Behavior Books, Palo Alto, CA, 1975.

[2] K. T. I. H. H. N. Bartneck, C. Is the uncanny valley an uncanny cliff? In *16th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2007*, pages 368–373, Jeju, Korea, 2007. IEEE.

[3] G. J. B. R. B. L. D. J. Dilts, R. *Neuro-Linguistic Programming: Volume I: The Study of the Structure of Subjective Experience*. Meta Publications, Cupertino, CA, 1980.

[4] F. N. G. K. T. I. H. R. J. H. S. R. L. Kahn, P. H. Jr. Design patterns for sociality in human-robot interaction. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pages 97–104. New York, NY: ACM Press, 2008.

[5] J. I. H. F. B. K. T. Kahn, P. H. What is a human? toward psychological benchmarks in the field of human-robot interaction. In *Proceedings of the 15th International Workshop on Robot and Human Interactive Communication(RO-MAN âĂŹ06)*, pages 364–371. Psycataway, NJ: Institute of Electrical and Electronics Engineers (IEEE), 2006.

[6] Y.-C. Lin, T.-C. Liu, M. Chang, and S.-P. Yeh. Exploring children's perceptions of the robots. In M. Chang, R. Kuo, Kinshuk, G.-D. Chen, and M. Hirose, editors, *Learning by Playing. Game-based Education System Design and Development*, volume 5670 of *Lecture Notes in Computer Science*, pages 512–517. Springer Berlin / Heidelberg, 2009.

[7] N. M. A. d. L. V. Looije, R. Children's responses and opinion on three bots that motivate, educate and play. *Journal of Physical Agents*, 2(2), 2008.

[8] D. T. E. P. S. Marchetti, A. *Fiducia e coping nelle relazioni interpersonali*. Carocci Faber, Rome, Italy, 2008.

[9] M. Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[10] R. Vasudevi. *How Infants Know Minds*. Harvard University Press, Harvard, MA, 2010.

Scenarios with Corresponding User and Robot Models, working document (Y1 results)

ALIZ-e Goals:

- Long term interaction
  - Bonding
  - Fun
  - Challenging
  - Knowledge transfer

User model aspects that seem to be relevant after experience with the three first year scenarios:

- General user info
  - Name
  - Age
  - School/class level
  - Medical info
  - Hobbies
  - Personal interests (e.g., sport, school subjects and television programs)
  - Gender
  - Family
  - Word use
  - Conversation content
  - Preferences
  - Amount/distribution of verbal/nonverbal behavior (metric needed): is the user generally "talkative", gestures a lot, uses predominantly speech/gesture
  - Language style: e.g., how does the user address the robot, formal/informal style, imperative/descriptive style, ego-/other-centric style, use of expressive vocalizations (Note: We need interaction data to determine which aspects play an important role.)
  - ASR/language comprehension error rates
  - Acceptance of robot / Attitude to robot
- State of the user (bored, excited/ positive negative)
- Attention (shared attention, eye contact)

- Dancing and Simon game
  - Strength
  - Recall /Performance

- Dancing specific
  - Balance
  - Flexibility
  - Layered movement (number of body parts moving in different ways)
  - Synchronization
  - Strength
  - Rhythm

- Simon game specific:
  - Sequence orders

- Quiz specific
  - Knowledge level

Child's Recognition of Emotions in Robot's Face and Body

Iris Cohen
TNO/University Utrecht
i.cohen1@hotmail.com

Rosemarijn Looije
TNO Human Factors
rosemarijn.looije@tno.nl

Mark A. Neerincx
TNO/Delft University of Technology
mark.neerincx@tno.nl

ABSTRACT
Social robots can comfort and support children who have to cope with chronic diseases. In previous studies, a "facial robot", the iCat, proved to show well-recognized emotional expressions that are important in social interactions. The question is if a mobile robot without a face, the Nao, can express emotions with its body. First, dynamic body postures were created and validated that express fear, happiness, anger, sadness and surprise. Then, fourteen children had to recognize emotions, expressed by both robots. Recognition rates were relatively high (between 68% and 99% accuracy). Only for the emotion "sad", the recognition was better for the iCat (95%) compared to the Nao (68%). Providing context increased the number of correct recognitions. In a second session, the emotions were significantly better recognized than during the first session for both robots. In sum, we succeeded to design Nao emotions, which were well recognized and learned, and can be important ingredients of the social dialogs with children.

ACM Classification Keywords:
I.2.9 **[Artificial Intelligence]**: Robotics, H.1.2 **[Models and Principals]:** User/Machine Systems - *Software psychology*

General Terms:
Human Factors, Experimentation

Authors Keywords:
Social robots, persuasive technology, emotions, children, affective body posture, facial expression.

1. INTRODUCTION
With computers and robots stepping out of their industrial environment and into the human society, they can be of surprising help in the healthcare. "Non-interactive" robots can provide support for surgical, rehabilitation and medication delivery purposes, whereas "highly interactive" robots can provide more cognitive, affective and social support [6]. Our research focuses on the latter, social robots, and their potential to comfort or support children who have to cope with a chronic disease.

The iCat is a robotic research platform with movable eyes, eyelids, eyebrows and lips. It has the ability to show facial expressions and thus show emotions. The six basic emotions were programmed into the iCat and validated by Kessens et al. [5]. The Nao robot is a humanoid robot that doesn't have moveable facial features like the iCat does, but has the ability to alter its body posture and thus show affective postures. The Nao also has the ability to show different colors in its eyes and this can help to strengthen the emotions being expressed. The colors that will be used are those investigated by Kaya and Epps [4]. This article describes two experiments. In the first one, a set of dynamic emotional postures for the Nao will be created and validated. In the second the question is if the emotions are recognized more accurate with the iCat (facial expression) or with the Nao (body posture). An additional question is if the emotion recognition is better when the emotional expression is presented in a corresponding context for both robots. A third question concerns repeated exposures. Does it get easier to recognize emotions in a second time contact, than after the first session for both robots?

HRI'11, March 6–9, 2011, Lausanne, Switzerland.
ACM 978-1-4503-0561-7/11/03.

2. EXPERIMENT 1
First, emotional postures for the Nao needed to be created. Four emotional postures, anger, fear, happiness and sadness were based on research by Bianchi-Berthouze and Kleinsmith [2] and one posture, surprise was based on research by Coulsen [3]. For anger, fear and happiness, two postures were created for the Nao, for sad, three postures were created and for surprise one posture was created.

2.1 Methods
A signal-detection task was used to validate the postures. Every emotion had a trial, where the participants had to indicate whether they saw a certain emotion (signal), or a different emotion (noise).

2.2 Results
The hit-rates and false-alarm rates were then calculated (see table 1). D' was calculated by the Z-score of the hit-rate minus the Z-score of the false alarm-rate and shows the discriminability between the trial emotion (signal) and different emotions (noise).

Table 1
Hits, false alarms, hit rate, false alarm rate and d'.

	H	FA	M	h	f	d'	A'
Angry1	12	1	3	0.75	0.01	2.92	
Angry2	14	1	2	0.87	0.01	**3.39**	
Fear1	14	3	2	0.87	0.04	**2.93**	
Fear2	11	3	4	0.69	0.04	2.27	
Happy1	14	0	2	0.87	0		**0.97**
Happy2	12	0	4	0.72	0		0.94
Sad1	11	1	5	0.69	0.01	2.73	
Sad2	7	1	1	0.88	0.01	**3.46**	
Sad3	3	1	4	0.38	0.01	1.99	
Surprise	22	3	10	0.69	0.05	**2.16**	

A higher d', means a higher discriminability between the trial emotion and a noise emotion. The posture with the highest d' was chosen as the posture for that emotion. For anger and

sadness, the second posture was chosen and for fear and happiness the first posture was chosen. The surprise posture was altered to make a better recognizable emotion with the help of vocal feedback from the participants.

2.3 Conclusion

With these results the final postures were chosen and for surprised the final posture was altered. The postures are shown in figure 2 and can be seen at the following link: http://mmi.tudelft.nl/SocioCognitiveRobotics/index.php/SocioCognitiveRobotics

3. EXPERIMENT 2

With the robots both having validated emotional expressions (see figure 1 and 2), the actual experiment was conducted.

Figure 1. A scared, happy, angry, sad and surprised iCat

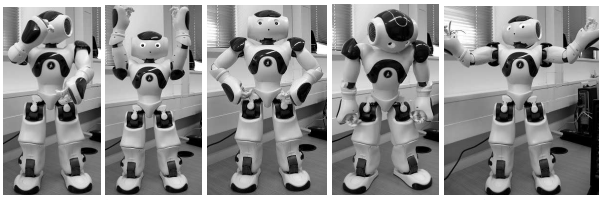


Figure 2. A scared, happy, angry, sad and surprised Nao

3.1 Methods

Participants

Fourteen children between 8 and 9 years, were recruited from an elementary school in the Netherlands.

Procedure

The children came to the experiment room and interacted twice with both robots. A 'context' interaction and a 'no-context' interaction. The orders of the interactions were counterbalanced to reduce order effects. In the context interactions, a story was told by the computer. The robots showed emotions appropriate to the story. In the 'no-context' interaction, the robots showed the emotions without a story. On a questionnaire the children could fill in what emotion they thought the robot had shown.

3.2 RESULTS

Correct recognition rates for the emotions are calculated. For the iCat, fear had a recognition of 88.39%, happy had a rate of 73.21%, angry had 99.11%, sad 94.64% and for surprised, 69.64% was correctly recognized. For the Nao, the percentages of correct recognitions for fear, happy, angry, sad and surprised were 87.5%, 89.28%, 96.43%, 67.86% and 68.75% respectively. Then, an ANOVA (figure 3) showed that there was no overall significant difference of recognition accuracy between the iCat and Nao emotions $F(1, 1118)=1.24$, $p=0.27$. Emotions expressed in context were significantly better recognized than emotions expressed without a context $F(1, 1118) = 29.79$, $p = .00$. And last, in the second session, emotions were better recognized than in the first session ($F(1, 1118) = 18.76$, $p = .00$).

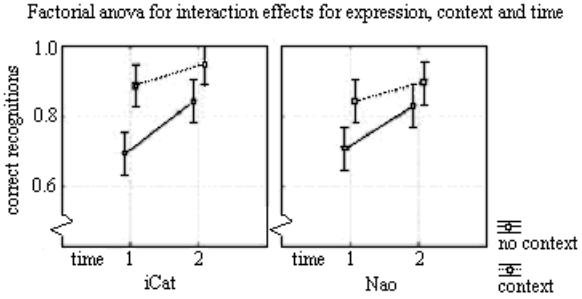


Figure 3. Factorial ANOVA with three variables: robot, session and context.

3.3 Conclusion

The main research questions focused on the difference in emotion recognition for the two robots, the effects of context and learning effects.

Only for the sad emotion, the iCat's expression was significantly better recognized than the Nao's posture, but when looking at the entire set of emotions, no difference was found between facial and body posture expression.

It was expected that the emotions will be better recognized when they are being expressed in the context condition [1]. The conducted ANOVA shows that this is indeed the case.

The last and final hypothesis stated that in the second session the correct recognition rates would be higher than in the first session. This was based on Nelson's review [7] which conclude that the skill for emotion recognition increases with multiple experiences. This hypotheses was supported by the results.

4. REFERENCES

[1] Barrett, L.F., Lindquist, K. A., & Gendron, M. 2008. Language as context for the perception of emotion. *Trends cognitive science. 11* (8).

[2] Bianchi-Berthouze, N., & Kleinsmith, A. (2003). A categorical approach to affective gesture recognition. *Connection science. 15* (4).

[3] Coulsen, M. (2004). Attributing emotion to static body postures: recognition accuracy, confusion, and viewpoint dependence. *Journal of Nonverbal Behavior, 28*(2).

[4] Kaya, N., & Epps, H. (2004). Relationship between color and emotion: a study of college students. *College student journal. 38.* 3

[5] Kessens, J.M., Neerincx, M.A., Looije, R., Kroes, M., & Bloothooft, G. (2009). Facial and vocal emotion expression of a personal computer assistant to engage, educate and motivate children. *IEEE* 2009.

[6] Nejat, G., Sun, Y., & Nies, M., (2009). Assistive robots in health care settings. *Home health care management & practice. 21*(3).

[7] Nelson, C.A. (2001). The development and neural bases of face recognition. *Infant and child development. 10*, 3-8.

Adapting Robot Behavior to User's Capabilities: a Dance Instruction Study*

Raquel Ros, Yiannis Demiris
Imperial College
London, UK
{rrosespi, y.demiris}@imperial.ac.uk

Ilaria Baroni, Marco Nalin
Fondazione Centro San Raffaele del Monte Tabor
Milan, Italy
{baroni.ilaria,nalin.marco}@hsr.it

ABSTRACT

The ALIZ-E[1] project's goal is to design a robot companion able to maintain affective interactions with young users over a period of time. One of these interactions consists in teaching a dance to hospitalized children according to their capabilities. We propose a methodology for adapting both, the movements used in the dance based on the user's cognitive and physical capabilities through a set of metrics, and the robot's interaction based on the user's personality traits.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: [User/Machine Systems]; I.2.9 [**Artificial Intelligence**]: Robotics—*Commercial robots and applications*

General Terms

Design, Measurement, Experimentation, Human factors

1. INTRODUCTION

When designing a robot that will interact with users while performing a task we must bear in mind two main aspects: (*i*) how will the robot perform the task, i.e. the engineering point of view (which actions is the robot going to perform?, how and when is it going to perform them?); and (*ii*) how will the robot interact with the user and how this interaction will modify the performance of the task, i.e. the social point of view (when should the robot start the interaction? what type of feedback to provide? when to stop the interaction?).

The goal of the ALIZ-E project is to design a robot companion able to maintain believable multimodal affective interactions with young users over an extended period of time. The robot will be tested at San Raffaele Hospital, providing support for hospitalized children with diabetes and obesity.

In the first stage, the robot is supposed to establish a bond with the child through verbal and non-verbal interaction. One of these interactions consist in having dance sessions with the hospitalized children according to his/her capabilities. Creative dance appears to be a valuable instrument for stimulating or enhancing physical and social/emotional growth in people [1]. On the one hand, it motivates exercise through a fun activity for children, and on the other hand, it also helps the development of social skills such as cooperation, self-confidence and communication. To this end, we have interviewed dance teachers to both, learn about dance teaching techniques (i.e., the engineering aspect of the task) and strategies to adapt the dance sessions based on the interaction with the user (i.e., the social aspect of the task). In both cases, the user's profile is taking into account as we describe in next sections.

This paper mainly focuses in the design of the dancing task based on the user's model, although we also propose the influence of social queues in the task performance. We conclude the work describing a first experience where the Nao robot teaches a dance to two children.

2. ADAPTING THE TASK

A common and simple technique to teach a choreography dance consists of a two-steps methodology. First the teacher shows the movements used in the choreography independently, focusing and evaluating whether they are correctly perform in technical terms (i.e., position and movement of the different body parts) – the *exploration* stage. Once these have been learned in a satisfactory way, the teacher proceeds to teach the choreography using the movements shown before – the *dancing* stage. The sequence is gradually enlarged including new movements as the sequence is memorized and correctly performed. If at a given point a movement is not correctly performed the teacher reviews that single motion and then continues with the rest of the sequence.

Two main questions have to be addressed in order to adapt the performance of the task to the user's cognitive and physical capabilities: (*i*) movement selection, i.e., which movements to include in the choreography and (*ii*) performance evaluation, i.e., how well the user has performed. In both cases we need to define metrics that will allow the robot to determine those movements that the user is able to perform (or would be able after practice) based on their complexity and to evaluate the user's performance within his/her own limitations.

Based on the literature on fundamental motor skills development [2, 1, 3] and after discussion with professional dance teachers[2] we propose the following measures:

- *balance*: how much effort is needed to balance the body to not to fall from a given position (e.g. two feet, one

*This research was supported by FP7 grant agreement no. 248116 ALIZ-E project.

[1]http://www.aliz-e.org

HRI'11, March 6–9, 2011, Lausanne, Switzerland.
ACM 978-1-4503-0561-7/11/03.

[2]Language of Dance Centre (London, UK) and Scuola di danza Ida Petrullo (Costamasnaga, Italy).

foot). Jumps correspond to a specific type of balance where there is no support during a short period of time.

- *flexibility*: how forced are the body joints when performing the motion (e.g. lean forward, somersault).
- *layered movement*: number of body parts moving in different ways (e.g. moving the arms and the legs at the same time would correspond to two layers). This measure can also be seen as self-coordination, where the coordination of the different body parts its more complex as more parts are implied.
- *synchronization*: synchronizing the self-movements with other partners (robot, other users).
- *strength*: how much strength from a body part is required to perform the motion (e.g. elevating the leg requires a strength in the thigh).
- *speed*: how fast the movement is performed. Some movements may be easier when being performed fast (lifting a leg), while others, when being performed slow (moving different body parts at the same time).

While the measures described above characterize movements individually, the following measures can determine the complexity of the overall choreography:

- *recall*: number of different movements the dance is composed of. The more movements it is composed of, the harder it will be to memorize.
- *rhythm*: following the rhythm is more complex than just performing the movements at any pace.

3. ADAPTING THE INTERACTION

Following the general ALIZ-E project methodology, affective aspects including personality and emotions will be considered for achieving an adaptive robot behavior in order to persuade and guide the user when performing the task. Based on the current state of the art (e.g. [4, 5]) we plan to adapt the following key-points in the teaching methodology:

- *movement selection*: the complexity of the movements varies from "easy" to perform to very "challenging" within the user's capabilities. Based on the user personality traits the robot may start the activity with simpler movements or with more challenging ones.
- *performance evaluation*: we define parameterized evaluation metrics that will allow to obtain outcomes ranging from "flexible", where a minimum effort of reproducing the motion will produce a positive evaluation, to "strict", where only correct motions, from a technical point of view, will be considered a success.
- *feedback*: different types of feedback should be provided to the user to assist the learning stage of the movements and the recalling phase when memorizing the dance sequences. Based on his/her cognitive capabilities verbal descriptions of the movements can be given in a technical way (e.g. raise both arms extending your arms to the maximum); in a pictorial fashion (e.g. stretch your arms as if you were trying to touch the sky); comparing movements with well known characters (e.g. fly like Superman); or even providing visual images (in this case additional resources as cards with pictures or a screen can be used).

Moreover, variations during the dance session are essential based on the current emotional state of the user. Thus, besides taking into account the general user's personality traits, in the project we also consider his/her current emotional state to fine tune the way the session will take place according to the aforementioned aspects, as well as the duration of the session (probably shorter sessions are more appropriate for non-motivated users).

4. CONCLUSION

A first experience with two children (7 and 11-years-old, both girls) has been performed with the Nao robot. The aim was to observe a first interaction of the dancing robot with them. The choreography has been designed by a dance teacher, and a short version of it (45 sec.) was tested with the children. It included four main movements: ranging from simple arm movements (single layer motion) to combination of arms and legs (layered motion), and two balancing motions (lifting one leg to the side). The dance included music, although for this first essay following the rhythm was not required. Both children enjoyed the session and did not get bored (even if the robot crashed at some point having to initialize it again).

We have confirmed the need of designing adaptable evaluation metrics, not only to base them on each person's capabilities, but also to consider differences on the robot's motions and the user's ones (e.g. closing the legs sliding is easy for the robot, but it seemed hard for the children to reproduce it –probably because the shoes were not appropriate as well–). Regarding the feedback provided to the user, the robot should dance along with the child and not only observe how he/she is doing after showing the motion for the first time. Children seem to mirror the robot's actions, and if it stops, the children stopped as well. Thus, giving continuous feedback (specially visual) will improve the child's comprehension about the action to perform, and will also increase his/her self-confidence for dancing.

5. REFERENCES

[1] I. E. V. Rossberg-Gempton, J. Dickinson, and G. Poole, "Creative dance: Potentiality for enhancing social functioning in frail seniors and young children," *The Arts in Psychotherapy*, vol. 26, pp. 313–327, 1999.

[2] T. I. Hilgenkamp, R. van Wijck, and H. M. Evenhuis, "Physical fitness in older people with id-concept and measuring instruments: A review," *Research in Develop. Disabilities*, vol. 31, pp. 1027–1038, 2010.

[3] B. Hands, "Changes in motor skill and fitness measures among children with high and low motor competence: A five-year longitudinal study," *Journal of Science and Medicine in Sport*, vol. 11, pp. 155–162, 2008.

[4] R. Looije, M. A. Neerincx, and F. Cnossen, "Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors," *Int. J. Hum.-Comput. Stud.*, vol. 68, pp. 386–397, 2010.

[5] A. Tapus and M. J. Matarić, "Socially assistive robots: The link between personality, empathy, physiological signals, and task performance," in *AAAI*, 2007.

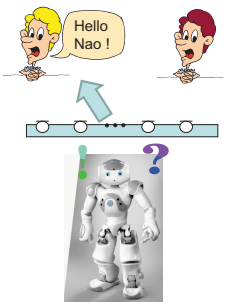
Aliz-E Workshop Speech - Audio

Microphone Array Technology for Robots: Sound Source Localisation

H. Brouckxon, G. Athanasopoulos and W. Verhelst
Vrije Universiteit Brussel, dept. ETRO-DSSP

Vrije Universiteit Brussel

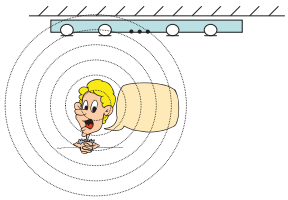
Sound Source Localisation – Why?



As part of an acoustic attention system, sound source localisation can be very useful:

- Who is talking?
- Where does that sound come from?

Microphone Arrays – Spatial Sampling of the Wavefield



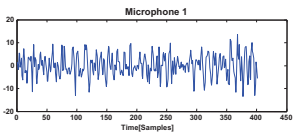
Signal, picked up by microphone i at location $\vec{r}_i$:

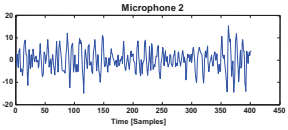
$$x_i(t) = A_i(\vec{r}_s - \vec{r}_i).s\left(t - \Delta_t(\vec{r}_s - \vec{r}_i)\right)$$

By measuring the time difference on arrival (TDOA) of the sound at different microphones, a microphone array can localise sound sources

$$TDOA(i,j) = \Delta_t(\vec{r}_s - \vec{r}_j) - \Delta_t(\vec{r}_s - \vec{r}_i)$$

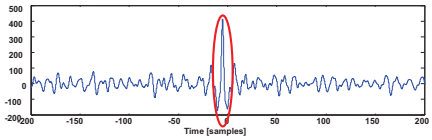
TDOA Measurement - GCC





Generalised Cross Correlation:

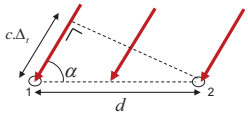
$$R_{xy}^{(g)}(l) = \frac{1}{N}\sum_{k=0}^{N-1}\psi_{xy}(k)X(k)Y^*(k)e^{j2\pi kl/N}$$



Time-domain maxima in the GCC indicate possible TDOA values

From TDOA to Location (FarField Quadrupole)

- Based on TDOA, a microphone pair (1,2) allows determination of the incidence angle relative to the pair's axis:

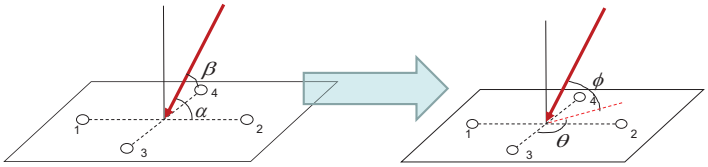


$$\alpha = \arccos\left(\frac{c.\Delta_t}{d}\right)$$

$\Delta_t = TDOA$

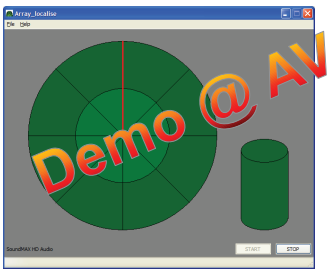
$c = \text{speed of sound}$

- Combine two orthogonal pairs (1,2) and (3,4) to obtain spherical coordinates:

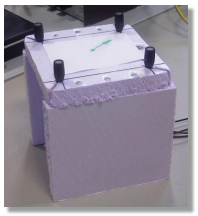


Development Setup and Software

- Square-shaped microphone array (10x10 cm)
- >4-channel Audio Interface
- Offline Matlab Implementation
- Real-time pc implementation (C++ and control Matlab)

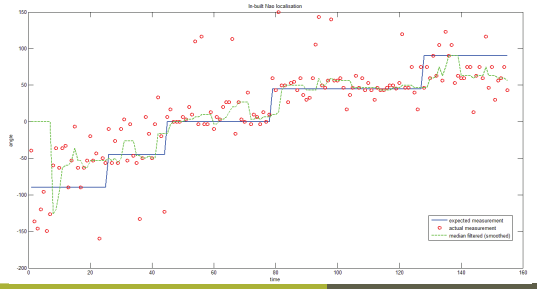


Demo @ AV-Lab Studio



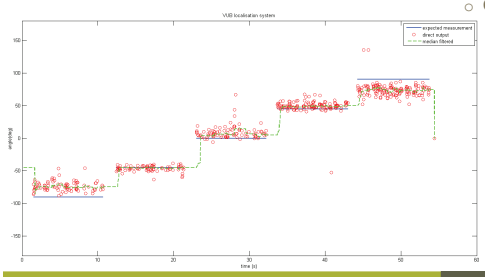
Nao inbuilt localisation

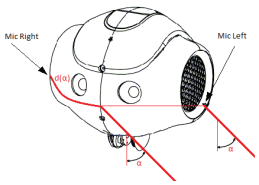
- Localisation is triggered only when sound is detected (front microphone)
- Evaluated with the default Nao's Sound Detector parameters
- Measured with speech signal of 60dB at 1 meter
- 5 discrete sound source locations were used



The Aliz-e approach for localisation

- Based on ETRO-VUB implementation
- Evaluated under the same conditions as Nao's inbuilt localisation
- Future steps:
  - TDOA calibration for NAOs head
  - Noise reduction for improving accuracy
  - GCC shadowing





Localisation & Noise Suppression

Set up

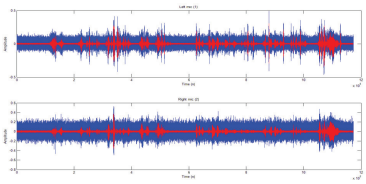
- 2 loudspeakers located at 1.5m distance from Nao:
  - Left loudspeaker at -45° angle
  - Right loudspeaker at +45° angle

Signals

- Left loudspeaker: speech of 60dB (pink noise equivalent)
- Right loudspeaker: pink noise of 65dB

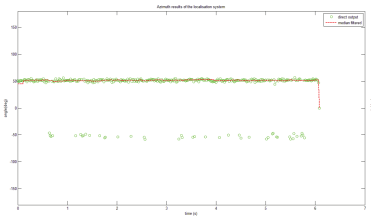
Noise Suppression

- Noise suppression algorithm: Ephraim-Malah

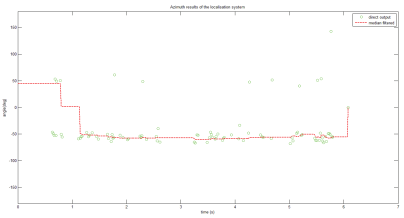


Use Case: Noise Suppression for localisation

- Original (non suppressed) signals:



- Processed (suppressed noise) signals:

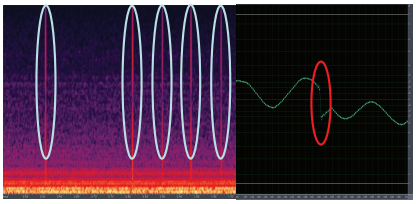


- Noise Suppression effect:
  - Most of the noise related measurements are removed
  - Speech SNR increases → more speech related measurements
- Why noise suppression:
  - Reduction of Nao's noise
  - Reduction of environmental noise: e.g. air-co, fan, computers, etc.

A brief look at Nao's audio

- 4 channel audio with DC offset
- Optimal SNR measured in controlled acoustic conditions
  - ch1 (L): 38.86 dB
  - ch2 (R): 36.7 dB
  - ch3 (F): 39.09 dB
  - ch4 (B): 22.78 dB
- **Test case**: for 60dB speech at 1 meter the SNR is
  - ch1 (L): ~10.5 dB
  - ch2 (R): ~7.5 dB
  - ch3 (F): ~8.5 dB
- SNR improves by at least 4 dB when removing the head cover

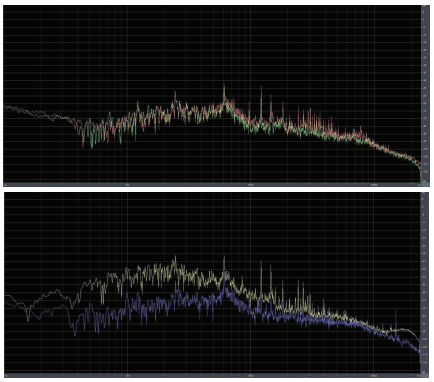
- Lets take a look at the spectrogram of the back microphone



- High frequency spikes are present (with no evident periodicity)
- These spikes are due to "discontinuities" in Nao's audio data (they appear in pairs for mic 1-2 & 3-4)

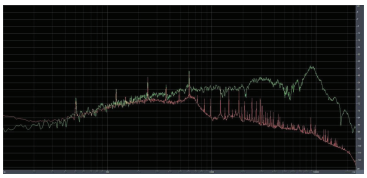
A closer look at noise characteristics

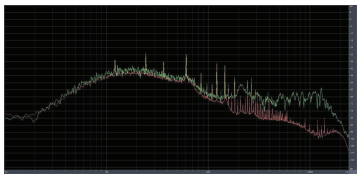
- Noise spectrum for microphones 1(L) and 2(R)
- Noise spectrum for microphones 3(F) and 4(B)
- Observations:
  - High wide band noise, especially for mic 2 and 4 which are the closest to the fan
  - Fan noise fundamental freq. is 656Hz (with 1st and 2nd harmonics strongly present)
  - Plastic cover acts like a sound box (with no cover the fan noise fundamental freq. is 609Hz)



Nao's microphones

Microphone's frequency response vs. noise level for channel 1 (L) and 4 (B), measured with MLS of 73 dB at 1m



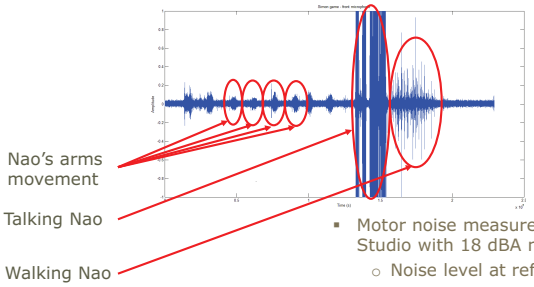


Open questions:

- Microphones location and quality
- High noise level masking the microphones' characteristics (especially below ~700Hz)

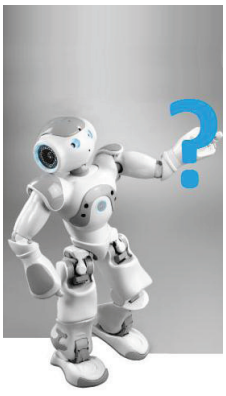
Towards a moving and talking Aliz-e

Simulation of Simon scenario with audio capturing:



- Motor noise measurements at AV-Lab Studio with 18 dBA noise floor:
  - Noise level at reference point: 41 dBA
  - Right arm move: 56 dBA
  - Left arm move: 51 dBA
  - Motor noise measured up to 58 dBA

Questions and discussion time



On the Noise Robustness of Voice Activity Detection Algorithms

Tomas Dekens and Werner Verhelst

Interdisciplinary Institute for Broadband Technology – IBBT, Vrije Universiteit Brussel, department ETRO-DSSP, Belgium

tdekens@etro.vub.ac.be, wverhels@etro.vub.ac.be

Abstract

In this paper, we show that the performance of voice activity detection algorithms (VAD) can be highly dependent on the type of background noise and we introduce a new VAD algorithm that is based on relative energy measurements in different frequency bands. The obtained experimental results are compared to the results obtained with two other spectrum-based VADs and it is concluded that a VAD, configured to use around 3 frequency bands can cope best with a large variety of background sounds.

Index Terms: voice activity detection, audio classification, noise robustness

1. Introduction

A voice activity detector is a sound classification unit that expects a noisy speech signal as input and then decides which parts of this signal contain speech and which parts don't. This information could be of use in several speech processing applications. E.g., in speech coding or automatic speech recognition, where its goal is to only retain frames that contain speech. In noise suppression it is important to know which parts of the signal contain no speech at all, as these parts can be used to estimate the noise characteristics, which are needed if the noise has to be filtered out of the signal.

Our previous studies involving VAD led us to believe that most VAD algorithms described in the literature are greatly affected by the type of background noise they have to deal with. This should not come as a surprise, as a great deal of these algorithms relies on the dissimilarity between the noise sounds and speech to make the VAD decision. This conjecture is the motivation for the experiments we describe in this paper. We will examine some noise types that exhibit certain characteristics, which hamper the VAD process each in its own way. We propose an energy based VAD algorithm that we configure in 3 different ways, such as to determine the influence of the considered noise types on the different VAD mechanisms.

In section 2 of this paper we describe the energy based VAD algorithm and explain the different configurations we used. In section 3 we elaborate on the experiments that were conducted and we show some test results. Finally, in section 4 some conclusions are drawn.

2. Energy based VAD

We developed an energy based voice activity detector, that we named eVAD. The feature used in the eVAD algorithm is the smoothed energy, contained in the frequency region of interest. Let $Y(m,k)$ be the short-time Fourier transform (STFT) of the input signal $y(t)$, with m the frame number and k the frequency index. The smoothed energy is then calculated as:

$$E(m,\hat{k}) = mean\left\{ \frac{2}{Nfft} \sum_{k=k_1}^{k_2} |Y'(m+j,k)|^2 \right\}_{j=-N}^{j=+N} \qquad (1)$$

$$0 \le k_1, k_2 \le Nfft/2$$

Where $Y'(m,k)$ is the same as $Y(m,k)$, except when k corresponds to DC ($k=0$) or half the sampling frequency ($k=Nfft/2$), then $Y'(m,k)$ is $Y(m,k)/\sqrt{2}$ to ensure that the energy at these frequency bins is only counted once. One can see the parameter N in (1) can control the extent to which the feature is smoothed. The index $\hat{k}$ represents the frequency range $[k_1, k_2]$ that is used to calculate the energy E.

During an initialization phase the first frames of the input signal are used to calculate the noise energy in the considered frequency region using (1); the mean of these noise frame energies gives us the initial estimation of the smoothed noise energy $E_{Noise}(m_1,\hat{k})$. Note that in our experiments we ensured that each signal does indeed start with a noise-only part. Next, the smoothed energy is calculated for each signal input frame m. This energy is then divided by E_{Noise} and the logarithm is taken:

$$Eratio(m,\hat{k}) = 10\log_{10}\frac{E(m,\hat{k})}{E_{Noise}(m,\hat{k})} \qquad (2)$$

This ratio is an indication of the difference the signal and the expected background noise exhibit in terms of energy content in the considered frequency region, at the current time instant m. A large difference indicates that, besides noise, other signals are present and could be a speech activity cue. In the next paragraphs we will describe how we use the energy ratios (2) corresponding to certain frequency ranges $\hat{k}$ in different configurations of eVAD.

2.1. Configuration 1: One band

The first configuration of eVAD uses one frequency band $\hat{k}$ to calculate the *Eratio* according to (2). If it is known that speech will only cover a fraction of the full signal frequency range or if the noise energy can be expected to be small compared to the signal energy in a certain frequency band, this frequency region can be selected. However, in our experiments we will assume the noise characteristics are unknown and the full frequency band, i.e., $k_1 = 0, k_2 = Nfft/2$, will be used.

The calculated energy ratio *Eratio* is compared to a certain threshold. When the ratio is smaller than this threshold the frame is considered to contain only noise and the noise energy is updated:

$$E_{Noise}(m+1,\hat{k}) = \alpha E_{Noise}(m,\hat{k}) + (1-\alpha)E(m,\hat{k}) \qquad (3)$$

$$0 \le \alpha \le 1$$

If the ratio is larger than the threshold, speech is detected and the current noise energy estimate will be kept:

$$E_{Noise}(m+1,\hat{k}) = E_{Noise}(m,\hat{k}) \tag{4}$$

According to the expected SNR of the input signals, an appropriate threshold value can be selected. The algorithm, however, also gives the possibility to work with a self-adaptive threshold. For this an on-line SNR estimation is performed. This SNR estimation requires a speech (actually speech + noise) energy estimate $E_{Signal}(m,\hat{k})$. If speech is detected in a certain frame, this estimate is updated in a similar fashion as the noise energy estimate:

$$E_{Signal}(m+1,\hat{k}) = \alpha E_{Signal}(m,\hat{k}) + (1-\alpha)E(m,\hat{k}) \tag{5}$$

If a frame is classified as a noise only frame, the estimate is left unchanged. A certain percentage of the logarithm of the ratio

$$E_{Signal}(m,\hat{k}) \big/ E_{Noise}(m,\hat{k}) \tag{6}$$

is then used as a threshold value.

Low energy phonemes such as /s/ or /t/ are hard to detect with an energy based VAD. If these phonemes occur in the middle of a speech fragment this doesn't pose a problem. The surrounding phonemes will ensure the energy feature curve won't drop to too low values. One can also select a minimum duration that a detected pause should last before it is actually classified as a pause. This can eliminate pauses that would be detected due to a short unwanted drop in the energy curve. On the other hand, very often a speech fragment starts or ends with such a low energy phoneme and this can not be dealt with by selecting a minimum pause length. That is why a detected speech portion is extended in time (front and back) by a certain amount, i.e. the regions before and after a detected speech fragment will be classified as speech regions as well. Because it is rather uncertain whether these regions are speech or not, they are not used to update the estimated noise or signal energies. To cope with short bursts of high energy like e.g. clicks, a minimum speech length can also be selected. If a speech region is detected that is shorter than this minimum length, it is classified as noise, but it is not used to update the noise energy.

Besides the relative energy ratio (2) it is also useful to use an absolute power measure. This can make the VAD deaf to signals whose power is below a certain value. So if

$$\frac{1}{Nwin}\sum_{n=0}^{Nwin} \left| win(n)y(n+Sm) \right|^2 < \delta \tag{7}$$

frame *m* will be classified as a noise fame, where $Nwin$ is the length of the window used and *S* is the frame shift. This configuration of eVAD is henceforth referred to as eVAD(1band).

2.2. Configuration 2: All frequency bins

Most spectrum based VAD algorithms compare the properties of the signal to those of the estimated background noise at the frequency bin level, possibly followed by some averaging, leading to one value that can be compared to a threshold. An example of such a VAD is the Long Term Spectral Divergence (LTSD) [2]. Also, statistical VADs such as the Statistical Likelihood Ratio [3] and the Inverse Normalized Noise Likelihood Ratio (INNL) [4] estimate probability density functions at the frequency bin level.

We also configured eVAD to compare energies at every frequency bin, which leads to the following expressions:

$$E(m,k) = mean\left\{ \frac{2}{Nfft} \left| Y'(m+j,k) \right|^2 \right\}_{j=-N}^{j=+N} \tag{8}$$

$$Eratio(m,k) = 10\log_{10}\frac{E(m,k)}{E_{Noise}(m,k)} \tag{9}$$

$$Eratio_{mean}(m) = \frac{1}{k_2 - k_1 + 1}\sum_{k=k_1}^{k_2} Eratio(m,k) \tag{10}$$

In our experiments we will again use the full band, $k_1 = 0, k_2 = Nfft/2$. The energy ratio of (10) will then be compared to a threshold to make the speech/no speech decision.

If the current frame is classified as a noise frame, the different noise energies $E_{Noise}(m,k)$, $k = k_1,...,k_2$ are updated using (3). In the other case the current estimates are left unchanged. (7) can be used to avoid speech detections if the absolute power of a frame is too low. We will call this configuration eVAD(all bins).

Note that this configuration would lead exactly to the LTSD VAD if the maximum would have been used instead of the mean in (8) and if the logarithm would be taken after computing the mean in (10).

2.3. Configuration 3: W bands

The approach taken in the third configuration lays somewhere between those taken in the first two configurations. Here the useful frequency region is subdivided into *W* frequency bands $\hat{k}_1$, $\hat{k}_2$, $\hat{k}_3$,…, $\hat{k}_W$. We discriminate two ways to employ the *W* corresponding energy ratios $Eratio(m,\hat{k})$ that will be calculated as indications of speech activity.

2.3.1. L detections

The first method of speech detection compares each of the *W* $Eratio(m,\hat{k})$ ratios to a certain threshold. If a number of ratios *L* or more are higher than their corresponding threshold, the frame is classified as a frame that contains speech. If the number of energy ratios that exceed their threshold value is lower than *L*, the frame is considered to contain only noise and the *W* different noise energies $E_{Noise}(m,\hat{k})$, $\hat{k} = \hat{k}_1, \hat{k}_2,..., \hat{k}_W$ are updated using (3). If self-adaptive thresholds are desired, (6) can be used to calculate in each frame the threshold used in each frequency band. We will refer to this method as eVAD(W bands/L detections).

2.3.2. Mean

The second method entails calculating the mean of the W different energy ratios:

$$Eratio_{mean} = \frac{1}{W}\sum_{\hat{k}=\hat{k}_1}^{\hat{k}_W} Eratio(m,\hat{k}) \tag{8}$$

This value is compared to one threshold. Again, if the frame is classified as a noise-only frame, the *W* different noise energies $E_{Noise}(m,\hat{k})$, $\hat{k} = \hat{k}_1, \hat{k}_2,..., \hat{k}_W$ are updated using (3). If the frame is expected to contain speech, the current estimates are kept (4). We call this method eVAD(Wbands/mean).

In both methods (7) will still be used to overrule speech detections if the absolute power of a frame is below a certain value.

3. Experiments

In order to evaluate the different configurations of the eVAD algorithm, we used the algorithm to detect pauses in the Dutch subset of a test database [1]. During the recording of this database four different (Dutch) speakers uttered 8 different sentences, which resulted in 24 sound files. The mean duration of the utterances is 37 seconds. On average, 74.59% of such an utterance contains speech; the other 25.41% consists of pauses. The sampling frequency of the sound files is 16kHz. The speech files were manually labeled in order to have a reference for the speech-pause detection. Only pauses of 200ms or longer were labeled as a pause, since it is not our intention to detect e.g. inter-word pauses [1].

Besides the eVAD algorithms, the LSTD algorithm [2] and the INNL algorithm [4] were also evaluated in the experiments. During the tests, different kinds of noises were added to the clean speech at different SNRs. Subsequently, the three VAD algorithms were applied on the 24 utterances using 32ms long speech frames with 50% overlap and different threshold values. For each threshold value the speech detection probability (SDP) was calculated by dividing the number of frames that were correctly classified as speech by the total number of speech frames, and the false alarm probability (FAP) was calculated as the number of pause frames that were classified as speech frames divided by the total number of pause frames. Note that the SDP and FAP are exactly the same as, respectively, the sensitivity and one minus the specificity of the binary classification.

The smoothing parameter N in (1) was for every configuration set to 6. This leads to a low-pass smoothing, with a rise and fall time of about 160ms. The weight parameter α in (3) was set to 0.95. Since we want to get an impression of how well the features used in the different algorithms and configurations are a representation of speech presence or absence, we used no extension of detected speech fragments, nor any absolute power or minimum pause or speech duration constraints.

In configuration 1 and 2 we used the full frequency band, meaning that in the case of configuration 2, we work with $Nfft/2+1$ different bands. In configuration 3, three frequency bands are utilized: (0-500) Hz, (500-2000) Hz and (2000-8000) Hz. In the case where each of the three energy ratios is compared to a threshold, the same threshold was used for the three different bands. This was done in order not to overcomplicate the SDP-FAP relation and as a result make it possible to compare the results to those obtained with the other methods. Note, however, that the best performance can probably be seen when the threshold values are not identical. The amount of ratios L that need to be higher than their threshold was set to 1 (leading to the name eVAD(3bands/1band) seen in Figures 1 and 2).

Some results of the experiments can be seen in Figures 1 and 2. Each point on the curves displayed in these figures shows the FAP and SDP values that were obtained in the case of a certain noise type and with a certain threshold value.

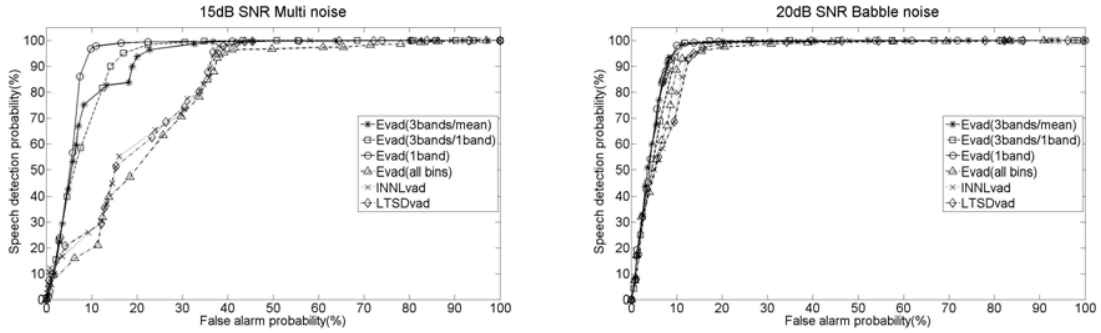


Figure 1: *Speech detection probability vs False alarm probability. Left: 15dB SNR multi-noise, right: 20dB SNR babble noise*

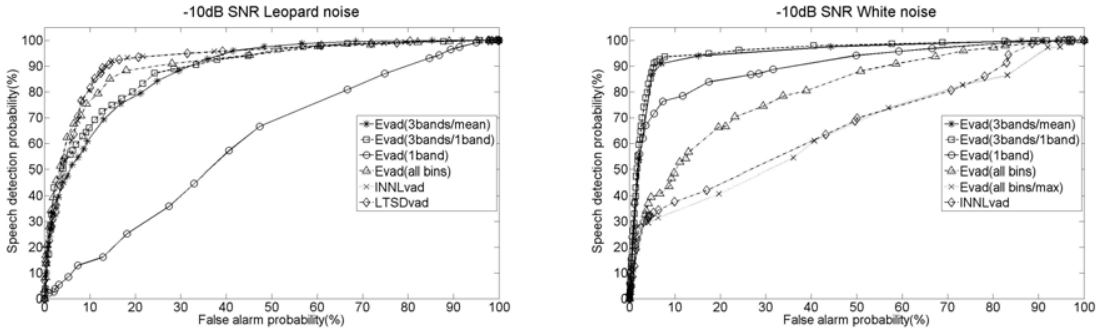


Figure 2: *Speech detection probability vs False alarm probability. Left:-10dB SNR tank noise, right: -10dB SNR white noise*

An obvious noise property that can upset most VAD mechanisms is non-stationarity. For this reason we used a non-stationary noise type we called multi-noise. This is nothing more than noise whose characteristics change drastically every 5 seconds and was produced by repeatedly concatenating 5 second fragments from 5 different normalized noise files (pink noise, jet fighter noise, tank noise, factory noise, car noise). It is clear that VADs that compare the spectrum of the signal to that of the background noise will suffer most from sudden changes in the background noise spectrum. The background noise energy however stays more or less constant, which implies that a VAD that analyzes the global energy should still be able to attain good classification scores, as long as the overall SNR remains high enough (we kept the SNR stable at 15dB in our experiments). This hypothesis is proven by what can be seen in the left hand panel of Figure 1, where the results are displayed for multi-noise at a 15dB SNR. The eVAD(1 band) attains a high SDP at low FAP values. LTSD, INNL and eVAD(all bins) are three methods that rely strongly on changes in the spectral curve; that is why for these methods a high FAP needs to be tolerated if high SDP values are desired. The two eVAD(3 bands) methods possess a spectral sensitivity that lays between those of the other two families of methods. When noise with changing characteristics is present, the energy of the noise changes in every band but not to the same extent as the energy at the frequency bin level. As expected from all this, the SDP-FAP curves of the eVAD(3 bands) methods indicate the second best performances.

Most VAD algorithms do not struggle when stationary noise types are encountered, as long as the SNR stays within an acceptable range. When the SNR becomes too low, problems can arise. To investigate this, we added two high power stationary noise types to the clean speech signals: one where the noise energy is spread over the whole frequency range (white noise), and one where the noise energy content stays confined to a relatively narrow band (tank noise). The right hand panel of Figure 2 shows how high power tank noise (SNR = -10dB) affects the VAD algorithms. In this case the spectrum based VADs (LTSD, INNL, eVAD(all bins)) do a better job in distinguishing speech from noise. This can be explained as follows: tank noise only covers the lower end part of the spectrum (0-1500Hz). The rest of the spectrum is unaffected by this type of noise. This means that when speech starts, overall the spectrum still changes to a large extent. The energy however doesn't, because of the high amount of energy inserted in the low part of the spectrum by the added noise. Again, the two eVAD(3 bands) methods seem to lead to the second best results. In this case, however, these results are certainly not as awful as the worst results that are obtained by eVAD(1 band). Although the global energy of the signal does not change much when speech starts, the energy in one of the bands used in eVAD(3 bands) (the highest one in this case) is altered to a large extent when speech emerges in the signal, leading to correct speech detections.

The left hand panel of figure 2 shows the SDP-FAP curves that are obtained when high power white noise (SNR = -10dB) is added to the speech. It is obvious that looking at the global energy is not a good idea, as the change in energy when speech starts will almost not be noticeable. In this case however, looking at the global spectral shape is not a good idea either. The spectrum of the signal will be more or less flat at all times; when speech starts its energy at higher frequencies is so low compared to that of the noise that no spectral changes can be seen. In the very low frequency part of the spectrum (~0-500 Hz) on the other hand some peaks can be seen that emerge from the flat spectrum. This is not enough to influence the global spectral shape features used in these algorithms much, hence their mediocre performances. This is, on the other hand, a sufficient change for the eVAD(3 bands) methods, since the energy content modification in the lowest band is relatively high. That is why we see much better results when these algorithms are employed.

The last noise type we consider is babble noise. Since the spectrum of babble noise is somewhat comparable to that of speech, the spectrum is altered to a lesser degree when speech starts than if a noise type with the same power but a completely non-speech like spectrum were present. This affects all spectrum based VADs. INNL, LTSD and eVAD(all bins) however suffer more from the non-stationarity of the noise. eVAD(3bands) and eVAD(1 band) are slightly less disturbed. This is reflected in the curves that are displayed in the right hand panel of figure 1, which are obtained when babble noise is added at an SNR of 20dB. When the SNR decreases the accuracies of all VAD algorithms worsen rapidly, but the same order in terms of performance holds.

Of note should be that selecting a more appropriate band instead of the full available band when using eVAD(1 band), LTSD, INNL or eVAD(all bins) could lead to much better results (e.g., by not using the low frequency part in eVAD(1 band) when tank noise is present, or by using exactly only the low frequency part in eVAD(1 band), LTSD, INNL or eVAD(all bins) when high level white noise is present). However, this demands that the noise characteristics are known beforehand, which generally speaking is not the case.

To summarize, methods such as LTSD, INNL or eVAD(all bins) are very sensitive to global spectral changes and are useful when high energy noise is present but highly concentrated around certain frequencies; they fail when the noise characteristics change over time. Looking at the global energy (eVAD(1 band)) is a well suited technique when the (low energy) background noise has variable spectral properties, but inappropriate for use with frequency concentrated noise. eVAD(3 bands) certainly is the best option when a flat spectrum noise such as white noise is encountered. Also in the other cases we investigated, using 3 energy bands leads to acceptable results.

4. Conclusions

In this paper we described an energy based voice activity detection algorithm. We showed that certain configurations of this VAD can outperform conventional spectrum based VADs under certain noise conditions. A configuration where 3 frequency bands are used generally leads to the most satisfactory results when the noise characteristics are unknown.

5. Acknowledgements

Parts of the research reported on in this paper were performed in the context of the EU-FP6 project SAFIR (IST-507427), EU-FP7 project ALIZ-E (ICT-248116) and SBO project AMASS++ that was supported by the Flemish government agency for Innovation by Science and Technology – IWT.

6. References

[1] M. Demol, W. Verhelst, and P. Verhoeve, "A Study of Speech Pauses for Multilingual Time-Scaling Applications", in proc. ISCA-ITRW Multiling 2006, Stellenbosch, South Africa, April 9-11, 2006.

[2] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, A. Rubio, "Efficient voice activity detection algorithms using long-term speech information", Speech Communication 42 (2004), pp. 271-287.

[3] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” IEEE Signal Process. Lett., vol. 6, no. 1, pp. 1–3, Jan. 1999.
[4] Dekens T., Demol M, Verhelst W. and Beaugendre F, (2007), "Voice Activity Detection based on Inverse Normalized Noise Likelihood Estimation," proceedings of the XIII-th Convention of Electrical Engineering, CIE 2007, Santa Clara, Cuba.

Attentional Selection of Objects of Interest

Thomas Geerinck, Hichem Sahli

Vrije Universiteit Brussel (VUB), Dept. Electronics & Informatics (ETRO)
Pleinlaan, 2, B-1050 Brussels, Belgium
Tel.: +32-2-6291300 Fax: +32-2-6292883 {tgeerinc,hshali}@etro.vub.ac.be

Abstract

Key words:

1. Introduction

It is well known that the human visual system employs an attention mechanism, due to limited processing resources, to selectively process important information that is currently relevant to visual behaviors or visual tasks [56, 133, 61]. This mechanism deals efficiently with the balance between computing resources, time cost and fulfilling different visual tasks in normal, cluttered and dynamic environments.

In computer vision, attentional mechanisms are mainly used to reduce the amount of data for complex computations. They employ a method of determining important and salient units of attention, and select them sequentially for more complex computations. The latter are either pre-attentively applied to the complete input data in parallel, or attentively and serially to the selected unit of attention only. Complex accurate computations are usually done attentively, while simple computations are assigned to the pre-attentive part. The main task of pre-attentive computations is to provide the necessary information for controlling visual attention, e.g., by localizing promising candidates for further processing.

When trying to model visual attention in a machine vision context, several related issues arise. We present the most common ones [133]:

- Bottom-up vs. top-down information and mutual interaction. The bottom-up approach uses basic features, conjunctions of features or even learned features as saliency information to guide visual attention. Attention can also be controlled by top-down or goal-driven information relevant to current visual behaviors. The deployment of attention is then determined by an interaction between bottom-up and top-down attentional priming or setting.
- Early vs. late selection. To what extent is visual processing serial or parallel, and what interplay exists between these factors? It is suggested that in a pre-attentive stage primary visual features are detected and extracted automatically in parallel. A second stage of attention processes shifts the focus of attention serially to scan subsets of the incoming information available from the previous stage.
- Shifts in focus of attention. There exist two traditional assumptions in literature to account for this shift in units of attention. The space-based attention theory claims that attention is allocated to a region of space, like a spotlight [34, 110, 137]. Object-based attention theory argues that attention is actually directed to an object or a group of objects to process any properties of selected objects rather than regions of space [32, 30, 126].

To date, there have been a number of attentional models for psychophysics or for machine vision that use the hypothesis of the ”spotlight” or ”zoom-lens” analogy for visual attention. Most of them are derived from Treisman’s Feature Integration Theory (FIT) [137]. The dominant line of thought following this metaphor, describes attention as being space-based. Attention is believed to be focused on one single contiguous region in space at each time.

Recently, this view has been challenged from the so-called object-based theories of attention [8, 148], which claim that instead of a simple region of space, the partitioning into objects is taken into account and determines the distribution of attention. Since traditional models have only concentrated on mechanisms of visual attention based on selectivity by spatial locations, they inherently lack mechanisms to account for object-based visual selection, and hence are not perfectly suited to work in real-world natural scenes. Space-based attention models may fail to work in environments that are cluttered or where objects overlap or share some common properties.

In contrast to the traditional theory of space-based attention, object-based attention suggests that visual attention can directly select discrete objects in a spatiotemporal context, without excluding continuous spatial locations within the visual field. Unlike space-based theories, spatial locations that do not contain any object are not considered in attentional selection. The converging evidence for object-based attention has been reviewed in [126]. The primary difference between object-based and space-based theories is the nature of the underlying unit of attentional selection [134]:

1. attention may need to work in discontinuous spatial regions or locations at the same time
2. attention may need to select an object composed of different visual features but from the same region of space
3. attention may need to select objects, locations, and/or visual features as well as their groupings for some structured objects.

There has been a rapidly increasing interest in object-based attention (c.f. [61, 151, 134, 150, 40, 39]) but research into useful systematic theories is still a very open research area, especially practical models of object-based attention for real-world applications.

In this article, we argue that a computational model for object-based attention requires considering the following aspects:

- Early identification and segmentation of perceptual objects. These early identified objects are called *proto-regions* or *proto-objects*.
- The relationship between object-based and space-based attention
- Grouping/segmentation and object-based attention. A grouping is a hierarchical structure of objects and space. A grouping may be a point, an object, a region, or a structured grouping.
- Visual saliency and visual attention. The salience of a grouping measures how much this grouping contrasts with its surroundings and depends on various factors, such as feature properties, perceptual grouping, dissimilarity between the target and its neighborhood.

Following the above criteria, we propose a novel biologically inspired region-based focus of attention mechanism simulating the middle stages of attention, with specific algorithmic details. From the viewpoint of modeling object-based visual attention, our approach, for proto-region detection, uses an innovative saliency driven perceptual grouping process, extending the pixel-based saliency map to salient groups/objects. Proto-objects are defined as blobs of uniform color in the image. As object-based image segmentation is beyond current computer vision techniques, the proposed method segments an image into regions (proto-regions), which are then merged using a perceptual organization approach. At the same time, an attention region (AR) is created based on the saliency map and salient regions from the image. A hierarchical perceptual grouping is used to select the salient regions, which are then clustered into the proto-object, named Object Of Interest (OOI), using new region merging criteria. Unlike other algorithms, the proposed method allows multiple OOIs to be segmented according to the saliency map.

The main contributions of the proposed model can be summarized in the following 3 aspects. Firstly, region is chosen as the perceptive unit, which makes the method more effective in terms of perception. Secondly, compared with traditional attention models our model provides saliency maps with meaningful region information, by eliminating misleading high-contrast edges. Finally using both *global effect* and *contextual difference* the proposed focus of attention shifts in unit of perceptual objects rather than spatial regions.

To our knowledge, several models of object-oriented visual attention have been proposed in the literature (e.g. [134, 118]), however, none of them has been completely implemented.

This article is organized as follows. In section 2 the predominant attention models considering space as the unit of attentional selection are briefly reviewed. Section 3 discusses the relation between objects and attention. In section 4, our model of object-based attention is presented and in section 5 it is validated thoroughly with experimental results. Finally, section 6 gives some conclusions and proposes evaluation methods for the developed approaches.

2. Space-based Models of Visual Attention

A wide variety of visual attention models, simulating human perception, exists in the field of psychology. In this section, we give an overview of visual attention models considering space as the unit of attentional selection.

2.1. Treisman's Feature Integration Theory [137]

The *Feature Integration Theory (FIT)*, introduced in 1980 [137], is considered as the seminal work for computational visual attention. The theory evolved towards current research findings. Figure 1 depicts the main ideas of the FIT scheme. The reader is referred to [139] for more details.

In [137], it is stated that "different visual features are registered automatically and in parallel across the visual field, while objects are identified separately and only thereafter at a later stage, which requires focused attention". Information from the resulting *feature maps* - topographical maps that highlight saliency according to the respective feature - is collected in a *master map of location*. This map specifies *where* (in the image) the entities (points, regions, objects) are situated, but not *what* they are. Scanning serially through this map directs the focus of attention towards selected scene entities and provides data useful for higher perception tasks. Information about the target entities is gathered into so called *object files* [139].

2.2. Wolfe's Guided Search [17]

Another very important work, in the field of visual attention, is the *Guided Search Model (GSM)* of Wolfe [17, 155, 153, 154]. Figure 2 depicts the model architecture. It shares many concepts with the FIT, moreover, it gives details allowing computational implementations. Like FIT, it models several feature maps. Unlike FIT it does not follow the idea that there are separate maps for each *feature type* (red, green, ...), it defines only one map for each *feature dimension*, and within each map different feature types are represented. However, Wolfe mentions

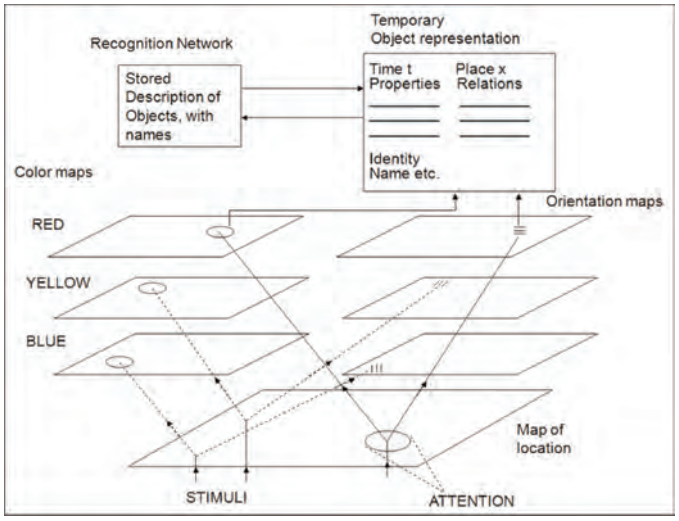


Figure 1: Model of *Feature Integration Theory (FIT)* [138]. Features such as color and orientation are coded automatically, pre-attentively and in parallel. Each feature dimension consists of several *feature maps* (red, yellow, blue for color). The saliency values of the feature are coded in the *master map of locations*. When attention is focused on one location in this map, it allows retrieval of the features that are currently active at that location and creates a temporary representation of the object in an *object file*.

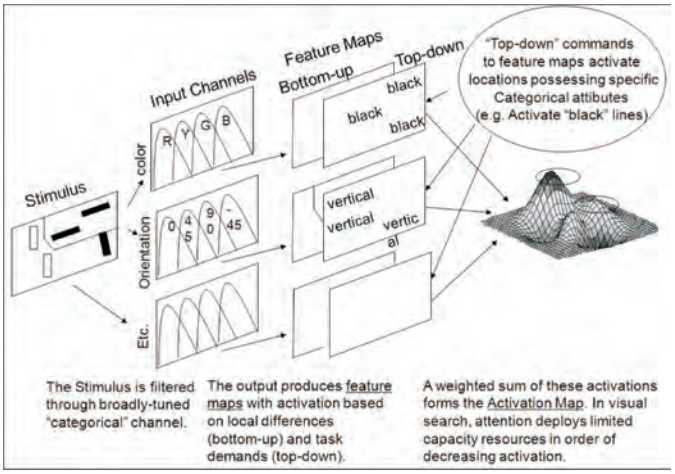


Figure 2: The *Guided Search model* of Wolfe [155]. One map for each feature dimension codes the properties of a scene concerning several feature types. Additionally to these bottom-up maps, top-down maps highlight the regions with task-specific attributes. A weighted sum of these activations forms the *activation map*.

that there is evidence for differences between features. For example, there may be multiple color maps but only one orientation map [99]. The features considered in the implementation are color and orientation.

Comparable to the *master map of location* in FIT, there is an *activation map* in GSM in which the feature maps are fused. But in contrast to at least the early versions of FIT, in GSM the attentive part profits from the results from the pre-attentive one. The fusion of the feature maps is done by summing them.

Additionally to this bottom-up behavior, the model also considers the influence of top-down information. To realize this, for each feature there is not only a bottom-up but also a top-down map. The latter map selects the feature type which distinguishes the target best from its distractors. This is not necessarily the feature with the highest activation for the target. Only one feature type is chosen.

2.3. Additional Psychophysical Models

Besides the FIT and the GSM models, there is a wide variety of psychophysical models on visual attention. The often used metaphor of attention is a *spotlight* coming from the *zoom lens model* [34]. In this model, the scene is investigated by a spotlight with varying size. Many attention models fall into the category of connectionist models, referring to models based on neural networks. They are composed of a large number of processing units connected by inhibitory and excitatory links. Examples are the *dynamic routing circuit* [101], SeLective Attention Model (SLAM) [109], SEarch via Recursive Rejection (SERR) [52], and Selective Attention for Identification Model (SAIM) [45].

A formal mathematical model is presented in [77]: the CODE Theory of Visual Attention (CTVA). It integrates the COntour DEtector (CODE) theory for perceptual grouping. The theory is based on a *race model* of selection. In these models, a scene is processed in parallel and the element that first finishes processing is selected (the winner of the race). That means, a target is processed faster than the distractors in a scene. Newer work concerning CTVA can be found, for example, in [14].

2.4. Koch & Ullman [67]

The first approach for a computational architecture of visual attention was introduced by Koch and Ullman [67] (see Figure 3). It served as a foundation for later implementations and for many current computational models of visual attention. The idea is that several features are computed in parallel and their *conspicuities* are collected in a *saliency map*. A *Winner-Take-All (WTA)* network determines the most salient location in this map, which is routed to a *central representation*, where more complex processing might take place.

The model is based on the FIT of Treisman [137]. The feature maps, that represent in parallel different features, as well as the central map of attention (Treisman's *master map of location*) are adopted.

An important contribution of Koch and Ullman's work is the WTA network - a neural network that determines the most salient location in a topographical map - and a detailed description of its implementation. The WTA network shows how the selection of a maximum in neural networks is performed, by single units that are only locally connected. This approach is strongly biologically motivated and shows how such a mechanism might be realized in the human brain. However, from the implementation point of view, WTA brings a computational overload to the system.

The most salient location, selected by WTA, is then routed into a central representation which at any instant contains only the properties of a single location in the visual scene. The idea

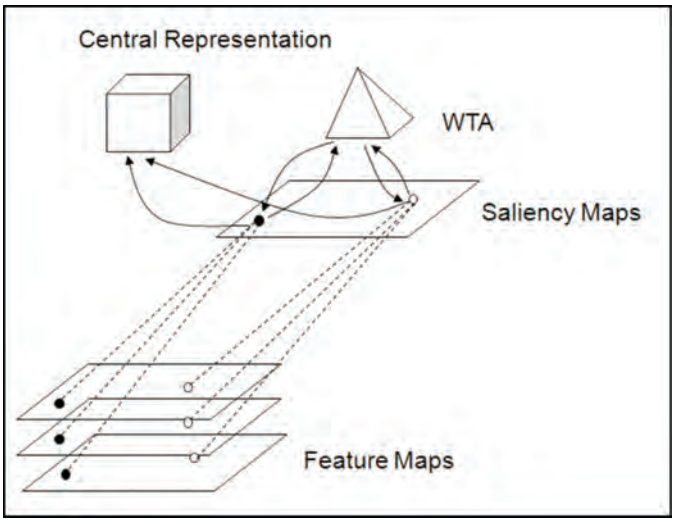


Figure 3: The Koch-Ullman model [67]. Different features are computed in parallel and their *conspicuities* are represented in several *feature maps*. A central *saliency map* combines the saliencies of the features and a *winner take all network (WTA)* determines the most salient location. This region is routed to the *central representation* where complex processing takes place.

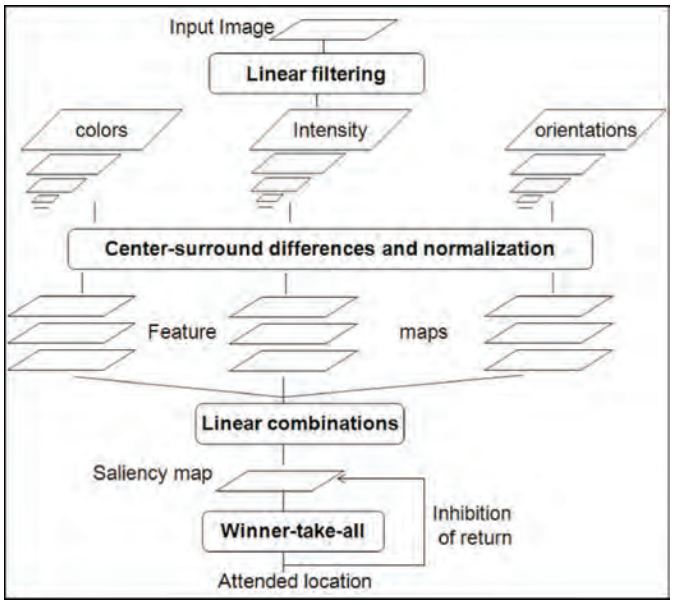


Figure 4: Model of the *Neuromorphic Vision Toolkit (NVT)* by Itti et al. [59] From an input image, three features are computed: color, intensity, and orientation. For each feature, an *image pyramid* is built to enable computations on different scales. *Center-surround mechanisms* determine the conspicuities concerning the features which are collected in a *central saliency map*. A *winner take all network* determines the most salient location in this map which yields the focus of attention. *Inhibition of return* inhibits this region in the saliency map and enables the computation of the next focus.

is that more complex vision processes are restricted to selected information. Due to this routing, the approach is also referred to as *selective routing*. Finally, a mechanism is suggested for inhibiting the selected region causing an automatic shift towards the next most conspicuous location (*inhibition of return (IOR)*).

2.5. Milanese [92]

The implementation of the visual attention system by Milanese [92, 93] is based on the Koch and Ullman model [67] and uses filter operations for the computation of the feature maps. Hence, it is one of the first *filter-based models*. These models are especially well-suited to be applied to real-world scenes since the filter operations provide useful tools for the efficient detection of scene properties like contrasts or edges' orientations.

As features, Milanese considers two color opponencies - *red-green* and *blue-yellow* -, 16 different orientations, local curvature and, intensity. To compute the feature-specific saliency, he proposes a *conspicuity operator*, referred to as *center-surround mechanism* or *center-surround difference*, which compares the local values of the feature maps to their surround. The resulting contrasts are collected in the so called *conspicuity maps*, a term that was since then frequently used to denote feature-dependent saliency.

The conspicuity maps are integrated into the saliency map by a relaxation process that identifies a small number of locations of interest, highlighted on the saliency map. A process determining the order in which to select the locations from this map is not proposed.

In [93], Milanese includes top-down information from an object recognition system. The idea is that object recognition is applied to a small number of regions of interest that are provided by the bottom-up attention system. The results of the object recognition are displayed in a top-down map which highlights the regions of recognized objects. This top-down map competes with the conspicuity maps for saliency, resulting in a saliency map combining top-down and bottom-up cues. The effect is that known objects appear more salient than unknown ones. The top-down information only influences the conspicuity maps (feature dimensions) and not the feature maps (feature types). Therefore, it is not possible to strengthen properties like "red" or "vertical". Furthermore, the system depends strongly on the object recognition system. It is not able to learn the features of an object independently.

2.6. Itti et al. [59]

One of the currently most used attention systems is the *Neuromorphic Vision Toolkit (NVT)*, a derivative of the Koch-Ullman model [67], that is steadily kept up to date [59, 57, 90, 58, 96]. The good documentation and the availability of the source code [1] makes the model a very popular basis for many research groups. Figure 4, shows the basic structure of the model. The ideas of the feature maps, the saliency map, the WTA and the IOR mechanisms were adopted from the Koch-Ullman Model; the approaches of using linear filters for the feature computation, of determining the contrasts by center-surround differences, as well as the conspicuity maps were adopted from Milanese [92]. The main contributions of this work are detailed elaborations on the realization of theoretical concepts, a concrete implementation of the system and the application to artificial and real-world scenes. The authors describe in detail how the feature maps for intensity, orientation,

and color are computed: all computations are performed on *image pyramids*, a common technique in computer vision that enables the estimation of features at different scales. Additionally, they propose a number of different techniques for combining different feature maps, including a weighting function promoting maps with fewer peaks and suppressing those with many ones; and a non-linear procedure introduced in [58].

The system contains several details that were chosen for efficiency reasons or because they represent a straight-forward solution to complex requirements. This approach may lead to some problems and inaccurate results in several cases. For example, the center-surround mechanism is realized by the subtraction of different scales of the image pyramid, a method that is fast but not very precise. Then, the conspicuity of the feature intensity is collected in a single intensity map, although neurobiological findings show that there are cells for both on-off and off-on contrasts [104] and psychological work suggests considering separate detectors for darker and lighter contrasts [139]. The same is true for the computation of the color-opponency maps: one red-green and one blue-yellow map are computed instead of considering red-green as well as green-red and blue-yellow as well as yellow-blue contrasts separately. Furthermore, the chosen color space RGB represents colors differently to human perception, which seems not appropriate for a system simulating human behavior.

Some of these detailed drawbacks were pointed out by Draper and Lionelle [29] who showed that the NVT lacks robustness according to 2D similarity transformation like translations, rotations, and reflections. They pointed out that these drawbacks result from weaknesses in implementation rather than from the design of the model itself. To overcome these drawbacks, they introduced an improved version of the system, called Selective Attention as a Front End (SAFE), which shows several differences and is more stable with respect to geometric transformations. It may be noted, that although these invariances are important for an object recognition task - the task Draper had in mind - they are not obviously required and maybe not even wanted for a system that aims at simulating human perception since usually human eye movements are not invariant to these transformations, too.

To evaluate the quality of the NVT, a comparison with human behavior was performed in [107]. The authors compared how the saliency computed by the system matched with human fixations on the same scenes and found a significant coherence which was highest for the initial fixation. They also found that the coherence was dependent on the kind of scene: for fractal images it was higher than for natural scenes. This was explained by the influence of top-down cues in the human processing of natural scenes, an aspect left out in the NVT.

Miau et al. [89], [90] investigated the combination of the NVT with object recognition, considering the simple biologically plausible object recognition system HMAX, from MIT [121], and the recognition with support vector machines. Walther et al. [149] continued these investigations, also in combination with the HMAX object recognition model. In [152], they combine the system with the well-known recognition approach of Lowe [78] and show how the detection results are

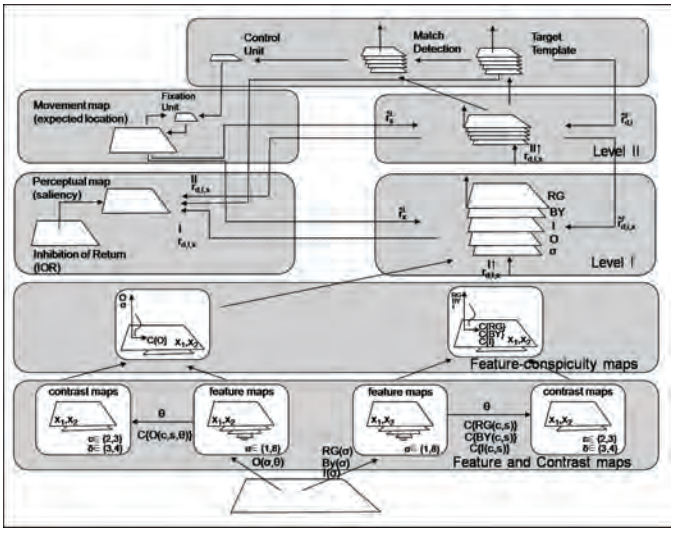


Figure 5: The attention system of Hamker [43]. From the input image, several feature and contrast maps are computed and fused into feature-conspicuity maps and finally into the perceptual map. Additionally, target information influences the processing. Match detection units determine whether a salient region in the perceptual map is a candidate for an eye movement.

improved by concentrating on regions of interest.

A test platform for the attention system - the robot platform *BeoBot* - was presented in [23], [54], [55]. It was shown how the processing can be distributed among different CPUs enabling a fast, parallel computation.

2.7. Hamker [43]

The attention system of Hamker [43, 42] aims mainly at modeling the visual attention mechanism of the human brain. Its objective is more on explaining human visual perception and gaining insight into its functioning than on providing a computational implementation. The model is based on current computer models [67], [59]. Hamker's model, shown in Figure 5, shares several aspects with the architecture of Itti et al. [59]: contrasts are computed for several features - intensity, orientation, red-green, blue-yellow and additionally spatial resolution - and combines them in feature conspicuity maps. The conspicuities of these maps are combined in a *perceptual map* that corresponds to the common saliency map.

In addition to this bottom-up behavior, the system belongs to the few existing ones that consider top-down influences. It is able to learn a target, that means it remembers the feature values of a presented stimulus. This stimulus is usually presented on a black background; hence the system concentrates on the target's features but is not able to consider the background of the scene. This means a waste of important information since it is not possible to favor features that distinguish a target well from its background. When searching for a red, vertical bar among red, horizontal bars, the color red is not relevant; in this case it would be useful to concentrate on orientation. To achieve a stable and robust system behavior, it would be necessary to learn the features of a target from several training images.

Hamker distinguishes between *covert* and *overt* shifts of attention, the latter corresponding to eye movements. The covert

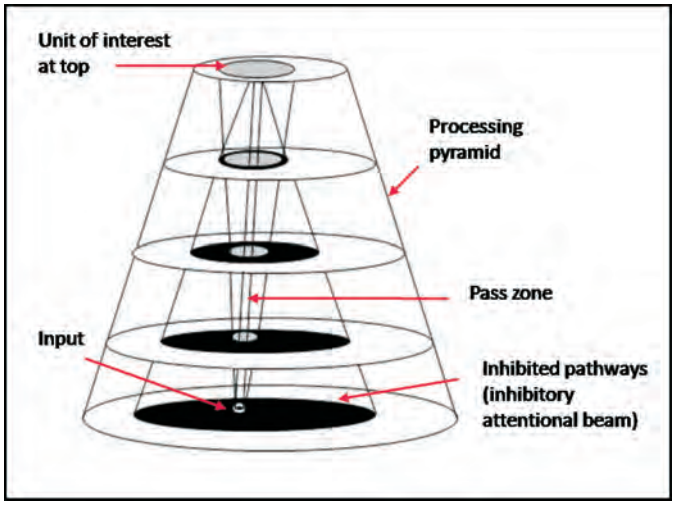


Figure 6: The *inhibitory attentional beam* of Tsotsos et al. [140] The selection process requires two traversals of the pyramid: first, the input traverses the pyramid in a feedforward manner. Second, the hierarchy of WTA processes is activated in a top-down manner to localize the strongest item in each layer while pruning parts of the pyramid that do not contribute to the most salient item.

focus of attention is directed to the most salient region in the perceptual map. Whether this region is also a candidate for an eye movement is determined by so called *match detection units* that compare the encoded pattern with the target template. If these patterns are similar, an eye movement is initiated towards this region and the target is said to be detected. The match detection units are an interesting approach in this system. However, it may be noted that this is a very rough kind of object recognition which is only based on few simple features and does not consider spatial configuration of features, and lacks rotation invariance.

2.8. Additional Attention Systems

Beside the mentioned attention models, there is a wide variety of models in the literature. Many differ only in minor changes from the above described approaches, for example, they consider additional features. Among them we can refer to the work of Backer [5], who presents a model of attention with two selection stages. The first stage resembles standard architectures like [67], but the result is not a single focus but a small number (usually four) of salient locations. In the second selection stage, one of these locations is selected and yields a single focus of attention. The model explains some of the more unregarded experimental data on multiple object tracking and object-based inhibition of return.

Beside the mentioned models that are based on feature computations with linear filters, there is another important class of attention models: the *connectionist models*. These models process the input data mainly with neural networks. Usually, these models claim to be more biologically plausible than the filter models. Since this approach differs strongly from the approach presented in this thesis, they are mentioned only briefly.

One of the most known models in the field of *connectionist models* is the *Selective Tuning Model* of visual attention by Tsotsos et al. [140], [141] (Figure 6). It consists of a pyramidal architecture with an *inhibitory beam*. The beam is rooted at the selected item at the top of the hierarchy and has a *pass zone* and an *inhibit zone*. The pass zone is the pathway that is selected for further processing; in the inhibit zone, all locations are inhibited that do not belong to the selected item. It is also possible to include target-specific top-down cues into the processing. This is done by either inhibiting all regions with features different from the target features or regions of a specified location. Additional excitation of target features as proposed by [95] is not considered. The model has been implemented for several features, for example luminance, orientation, or color opponency [140], and currently in a sophisticated approach also for motion, considering even the direction of movements [141]. Note that in each version only one feature dimension is processed; the binding of several feature dimensions has not yet been considered but is subject for future work, as per Tsotsos.

An unusual adaptation of Tsotsos's model is provided in [116]: the distributed control of the attention system is performed by game theory concepts. The nodes of the pyramid are subject to trading on a market, the features are the goods, rare goods are expensive (the features are salient), and the outcome of the trading represents the saliency.

Another model based on neural networks is the *FeatureGate* Model described in [16]. Beside bottom-up cues it also considers top-down cues by comparing pixel values with the values of a target object; but since the operations only work on single pixels and so are highly sensitive to noise, it seems to be not applicable to real-world scenes.

3. Objects and Attention

3.1. Space-based vs. Object-based Visual Attention

Visual Attention is a complex and extensive process, a set of processes that is difficult to define precisely. Some of the central aspects of our everyday notion of attention are reviewed by Pashler (1998) [126]. Intuitively, attention seems to be an extra processing capacity which can both intentionally and automatically select - and be effortfully sustained on - particular stimuli or activities. As such, we can roughly define visual attention as the mechanism that allocates limited visual resources for processing selected aspects of the retinal image more fully than non-selected aspects [104]. By using this intelligent visual selection, the visual system can flexibly explore the contents and layout of a complex visual field [82].

In the vast psychophysics literature concerning visual attention [126], the nature of the underlying units of attentional selection engenders two groups of theories. Traditional models characterize attention in spatial terms, as a spotlight (or a 'zoom lens') which can move in the visual field, focusing processing resources on whatever falls within that spatial region - be it an object, a group of objects, part of one object and part of another, or even nothing at all, as described in the previous section. Recent models of attention, in contrast, suggest that (in some cases) the underlying units of selection are discrete visual objects, and that the limits imposed by attention may then

concern the number of objects which can be simultaneously attended.

In the following we present the most influential evidence for spatial selection, evidence for object-based attention, extracted from experimental paradigms. This includes selective looking, divided attention, attentional cuing, and multi-element tracking. With this short review, we emphasize on the major themes in the study of objects and attention, without exhaustively discussing empirical details. The reader can refer to the review papers of Driver and Baylis [30], Kanwisher and Driver [64], and Scholl [126].

Evidence for spatial selection

The contrast between objects and locations is the main motivation driving the study of object-based attention. Does attention always select spatial areas of the visual field, or may attention sometimes directly select discrete objects? The canonical evidence for spatial selection, which gave rise to the dominant 'spotlight' and 'zoom lens' models for spatial attention, comes from spatial cuing studies. Posner, Snyder, and Davison [110], for instance, showed that a partially valid cue to the location where a target would appear, speeded the response to that target, and slowed responses when the cue was invalid and the target appeared elsewhere.

These types of results suggested that attention was being deployed as a spatial gradient, centered on a particular location and becoming less effective as the distance from that location increased.

Early suggestions from 'selective looking'

Some of the earliest evidence for object-based selection, came from the work of Neisser [97, 98]. Subjects, given a 'selective looking' task (two spatially superimposed movies), failed to notice unexpected events which happened in the unattended scene. As such, this early work, though subject of methodological flaws, provides evidence that attention does not simply consist of a single unitary region of spatial selection.

'Same-object advantages' in divided attention

In studies of divided attention [32, 8, 7], it was concluded that observers were less accurate at reporting two properties from separate objects, but were able to judge two properties of a single object without extra cost. This has been termed a 'same-object advantage'.

Space-based theories cannot easily account for such results, since spatial location does not vary with the number of perceived objects. However, the interpretation of these divided attention tasks is still controversial. It has been argued that the results of these divided attention studies are due to the fact that automatic attentional spread must fill a greater area with two objects than with one [26]. The details of this interpretation still implicate object-based attention, but the mechanism responsible is seen to be automatic spread of attention.

Multiple Object Tracking

The object-based nature of attentional selection is also apparent in dynamic situations, in which object tokens must be maintained over time (multiple object tracking (MOT)). Experiments from Pylyshyn and Storm [115], and Sears and Pylyshyn [127] state that the observed tracking performance cannot be accounted for by a single spotlight of attention which cyclically visits each item in turn. Also, attention has been found to speed response times to attended objects, and this advantage appears to be target-specific in MOT. Third, it is indicated that attention is split between the targets rather than being spread among them.

Object-attention concerns objecthood and object-based selection in a spatiotemporal context. Unlike space-based theories, spatial locations that do not contain any object are not considered in attentional selection. The object-based hypothesis is based on the assumptions that perceptual (but pre-conceptual) organization of a visual scene into discrete units occurs before attention is allocated, and that attention then selects or enhances visual stimuli as organized into objects rather than undifferentiated regions of visual space [118]. The contrast with the spotlight metaphoric model is clear. Since, on the spotlight model, everything in the spotlight is assumed to be processed in parallel, features from two nearby or overlapping objects should be attended as easily as a single object, whereas on the object-based model this would not be the case. In literature (see [61] for a review), there is a well-established body of evidence in support of the idea that dividing attention between objects results in less efficient processing than attending to a single object.

It should be noted that spotlight and object-based attention theories are not contradictory but rather complementary [68, 77]. Nevertheless, the object-based theory accounts for many phenomena better than the spotlight model does. Form the above discussions on space-based and object-based attention, it seems clear, that these two notions should not be treated as mutually exclusive. Attention may well be object-based in some contexts, location-based in others, or even both at the same time. The 'units' of attention could vary depending on the experimental paradigm, the nature of the stimuli, or even the intentions of the observer. The relation between space-, feature-, and object-based attention is not yet clear. Available evidence suggests that different, but interacting, systems may be involved (e.g. [68]).

3.2. Attention and Perceptual Grouping

Several research works have emphasized that scenes are organized into perceptual groups defined by the Gestalt principles of similarity (common attribute), continuity (form a completed shape), proximity (close to one another), common fate (move together), etc. [104]. In this section, we consider how the attended objects, serving as units of attention, relate to other units, including perceptual groups, parts, and visual surfaces.

Attention and Perceptual Groups

Driver and Baylis [30] (also [33]) combined perceptual grouping work with attention demonstrations, and replicated

some evidence for object-based selection, when Gestalt groups are used as stimuli instead of single objects. Such evidence suggests that 'object-based' attention and 'group-based' attention may reflect the operation of the same underlying attentional circuits.

Attending to Parts

Just as multiple objects can be perceptually grouped together, so can individual visual objects be composed of multiple parts. In the study of attention, recent research has demonstrated 'same-part advantages' (section 3.1) for complex objects composed of hierarchical part arrangements. These studies suggest that it may be worthwhile in future work to bring the literatures on attention and perceptual part structure into closer contact [130, 147].

Attending to surfaces

The previous paragraphs considered both multi-object units such as groups, and intra-object units such as parts. Visual surfaces constitute another level of representation which can encompass both of these categories: complex objects can consist of multiple surfaces, while multiple objects can be arrayed along a single surface.

From their experiments, He and Nakayama [44] indicate that attention can efficiently select individual surfaces. In another experiment, using a cuing study (similar to that of [33]), they demonstrate that in some cases attention must spread along surfaces. He and Nakayama conclude that the visual system can direct selective attention efficiently to any well-formed, perceptually distinguishable surface. In this context 'well-formedness' must be seen as local co-planarity and collinearity of surface edges.

As the previous three paragraphs have emphasized, there may be a hierarchy of units of attention, ranging from intra-object surfaces and parts to multi-object surfaces and perceptual groups. It remains an open question whether attention to each of these levels reflects the operation of the same or distinct attentional circuits.

3.3. Visual Object

From the above discussion, it can be said with a fair degree of certainty, that under certain experimental conditions, the allocation of attention depends on spatial properties of visual stimuli, and that under other experimental conditions, factors of perceptual organization play a more dominant role in distributing attention [63]. At the heart of the concerns discussed below is the problem of what exactly is meant by 'visual object'. To understand this concept, we consult two approaches in the literature: the object taxonomy of Jarmasz [61]; and the coherence theory of Rensink [118]. We also consider how the visual system organizes visual stimuli into the objects used by attention and how to simulate this behavior. The simulation is achieved either by Gestalt principles, or by other low-level mechanisms that are independent of higher-level conceptual knowledge. Subsequently, we wonder if these mechanisms are sufficient or if an observer's background knowledge and current mental states are also required at this early level.

Object taxonomy of Jarmasz [61]

Jarmasz [61] proposes a four-way taxonomy of objects that can play a role in vision:

- *c-objects*: physical objects, or what philosophers call concrete particulars
- *p-objects*: mental representation of visual objects or objects of phenomenal experience
- *v-objects*: virtual objects; 2D devices that are perceived as c-objects
- *a-objects*: attentional objects; intentional objects involved in attentional selection

The object-based attention thesis can now be restated as: "the attentional system selects *a-objects* in order to create *p-objects*, which are supposed to allow a person to know about and act upon the *c-objects* and *v-objects* that gave rise to the *a-objects*." The 'objects' of object-based attention are thus a-objects.

What, then, are *a-objects*? The standard answer given by several researchers is that *a-objects* are perceptual groupings whose formation is governed by the Gestalt principles of perceptual organization [8, 68]. Until the advent of cognitive psychology, the Gestalt principles constituted the only available theory of perceptual organization, and were thus integrated into cognitive psychology by Neisser [97]. The choice of Gestalt groupings for *a-objects*, the objects that are selected by attention, was thus a natural one for object-based attention.

Coherence theory of Rensink [118]

On the other hand, the attention theory of Rensink [118, 119], based on a study of change-blindness phenomena, suggests that attention may endow structures with a coherence lasting only as long as attention is directed to it. These thoughts are formulated in Rensink's coherence theory of attention, stating:

- Prior to focused attention, low-level '*proto-objects*' are continually formed rapidly and in parallel across the visual field. These *proto-objects* can be fairly complex, but have limited coherence in space and time. Consequently, they are volatile, being replaced when any new stimulus appears at their retinal location.
- Focused attention acts as a metaphorical hand that grasps a small number of *proto-objects* from this constantly regenerating flux. While held, these form a stable object, with a much higher degree of coherence over space and time. Because of temporal continuity, any new stimulus at that location is treated as the change of an existing structure rather than the appearance of a new one.
- After focused attention is released, the object loses its coherence and dissolves back into its constituent *proto-objects*. There is little or no "after-effect" of having been attended.

According to the coherence theory, a change in a stimulus can be seen only if it is given focused attention at the time the change occurs. Since only a small number of items can be attended at any time [108, 115], most items in a scene will not have a stable representation. Thus, if attention cannot be automatically directed to the change, the changing item is unlikely to be attended, and change-blindness will likely follow. Moreover, unattended objects have limited spatiotemporal coherence. From visual search experiments, described in [119], proof is provided for the limited spatial coherence of *proto-objects*, relatively complex assemblies (by rapid grouping/segmentation) of fragments that correspond to localized structures in the world. [117] indicates that *proto-objects* are the lowest level structures directly accessible to attention, with much of their underlying detail being accessed only by deliberate effort. As such, *proto-objects* serve as the highest outputs of low-level vision, but also the lowest level operands upon which higher level attentional processes can act.

The proof for limited temporal coherence of *proto-objects* comes largely from studies on visual integration and change-blindness. Early level structures are either overwritten by subsequent stimuli or else fade away within a few hundred milliseconds, making them inherently volatile [120, 119]. Given that unattended structures have only limited spatial and temporal coherence, it follows that focused attention must provide the coherence that knits proto-objects into larger-scale objects and allows them to retain their continuity over time.

Discussion

Based on the presented evidence on object-based attention, we can state correctly that focused attention is intimately involved with the perception of objects (*c-objects* according to Jarmasz' taxonomy). Essential properties of an object include the requirement that it be discrete, be differentiated from its background, and have a coherent unity across space and time. Attention makes use of surface representations generated by early visual perceptual processes. These surface representations serve as units of object-based attention. Following Jaramsz' nomination these surface representations are called attentional objects or *a-objects*. *A-objects* are believed to be perceptual groupings whose formation is governed by perceptual organization processes. Being the only available theory of perceptual organization, the Gestalt principles are put forward as theoretical answer to the formation of *a-objects* as perceptual groupings. Once *a-objects* are formed, attention selects them, and creates a mental representation of the visual objects, the so called *p-objects*. Rensink, on the other hand, claims that the so called *proto-objects* are the units of attention. *Proto-objects* are fairly complex, rapidly formed surface representations generated by early visual perception. Attention selects a number of *proto-objects*, with limited coherence, and forms a (mental representation of a) stable object, with high coherence over space and time.

Comparing Jarmasz' view on objects as unit of attention with Rensink's thoughts, we conclude that both *a-objects* and *proto-objects* refer to the perceptual entities (groups) formed by early visual perception. From this point on, if we refer to objects as the unit of attention, we will use the term *proto-object*. Reasoning about the simulation mechanism behind the formation of the *proto-objects*, brings us, following Jarmasz, to the Gestalt principles of perceptual organization. However, other low-level mechanisms, independent of higher-level conceptual knowledge, can also be considered to simulate the rapid, pre-attentive formation of *proto-objects*. We also emphasize the distinction between perceptual groups and objects, which could themselves be comprised of many perceptual groups.

To my opinion, the formulation by Jarmasz [61] describes best the relationship between object-based and space-based attention. Attention is not a reflexive mechanism based on any particular property or set of visual stimuli. Instead of being space-based or object-based, attention is space- and object-mediated. Spatial and object features are concepts used by the visual system to deploy attention.

3.4. Segmentation, Perceptual Grouping, and Attention

Without segmentation and grouping, object-based attention may lose its selection units. In general, segmentation processes - the processes that bundle parts of the visual field together as units - probably exist at all levels of visual processing. Some of these processes are early, using 'quick and dirty' heuristics to identify likely units for further processing. This results in a visual field which has been segmented into *proto-objects*, which are thought to be volatile in the sense that they are constantly regenerated [118]. In this scheme, the *proto-objects* serve as the potential units of attention. Once a *proto-object* is actually attended, additional object-based processes come into play. In Rensink's coherence theory [118], deploying attention to a *proto-object* gives rise to a more coherent representation of that object. It seems likely, however, that this attentional processing could in some cases override the earlier parsing characterized by the *proto-objects*. For instance, the additional attentional processing on a set of *proto-objects* may result in a higher-level representation of that portion of the visual field as a pair of intertwined objects, or as only a part of a more global object or group of objects. In general, since such processes can occur at multiple levels, 'segmentation' cannot be considered as synonymous with object-based attention. In conclusion we state that the units of some rapid and rough segmentation processes (*proto-objects*) may serve as the focus of attention, while the units of other perceptual grouping processes may be in part the result of (proto-)object-based attention.

Indeed, Mack et al. [83] and Rock et al. [122] have presented results that suggest that perceptual organization does not occur without attention [61]. Attention and visual perception are mutually dependent, likely interactive and concurrent processes. As such, perceptual grouping/organization is deeply intertwined with object-based attention and hierarchical selectivity (or multiple selective levels) by features, objects, or their hierarchically structured groupings.

Duncan [32] states, "The study of visual attention and perceptual organization must proceed together". However, one of the remaining questions is, when, where, and how the properties of an object, or elements of a grouping become a perceptive

object or a grouping? Another question is, how the mutual impact between perceptual grouping and attention is evaluated or measured?

Trying to answer these types of questions, the interactions between the different processes need to be modeled. Rensink [119] posits that attentional interaction with lower-level structures is taking place via a nexus, a single structure containing a summary description of the attended object, for example, its size, overall shape, and dominant color. When a *proto-object* is attended, a link is established between it and the nexus, enabling a two-way transmission of information between these structures. Information going up the link allows the nexus to obtain descriptions of selected properties from the attended *proto-object*. Information going down the link can in turn provide stability to the volatile *proto-object*, allowing it to be maintained or to rapidly regenerate. The nexus and its *proto-objects* form a local hierarchy, with two levels of description (object- and part-level).

In conclusion, attention influences, and is influenced by, perceptual organization in a way that favors information that is relevant to the actions and intentions of an agent.

3.5. Modeling Perceptual and Relevance-based Influence on Attention

The literature ([61]) argues that attention is best understood as cognitively mediated, and not a mere reflexive response to putative salient properties of visual stimuli. The central claim of this account is that, in order to do what attention is generally said to do (filter information, enhance processing, integrate visual features into unified percepts), the visual system uses basic visual features (color, shape, location, surfaces, motion, depth information, and so on) as tools in order to direct attention according to an observer's background knowledge, goals, intentions, and particular task demands. The deployment of attention itself then organizes and transforms these basic visual stimuli into a meaningful parsing of ones visual environment, through the effects noted above.

Visual saliency can attract visual attention if the current top-down attentional setting is not fully loaded (or in other words, the current attention can be gained without top-down control) [158]. In this regard, we wonder, what is the visual salience of a feature, an object, or a grouping? And what is the neural substrate to execute the saliency computation and judgement? How does visual saliency drive visual attention? The most important requirement to model visual attention in practice is how visual salience of a perceptual unit (whether a perceptual object or grouping) can be quantitatively measured, so that the saliency mapping of a visual field truly reflects its competitive situation for visual attention.

As concluded above, perceptual organization and attention are deeply intertwined favoring information relevant to the task at hand. Therefore, perceptual- (bottom-up) and relevance-based (top-down) influences on attention need to be modeled. Insight into modeling the interaction between bottom-up and top-down influences on attention, is acquired by studying existing models of object-based attention: the triadic architecture of Rensink [118], and the conative model of attention of Jarmasz [62].

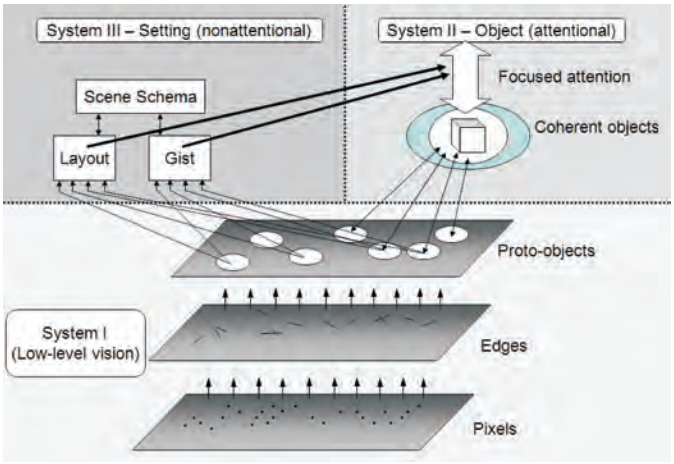


Figure 7: The *triadic architecture* of Rensink [118] suggests that visual perception is carried out via the interaction of three different systems: in the low level system, early level processes produce volatile *proto-objects* rapidly and in parallel. In system II, focused attention grabs these objects and in system III, setting information guides the attention to various parts of the scene.

The triadic architecture *of Rensink [118]*

The triadic architecture [118, 119] is depicted in Figure 7, and consists of three subsystems. First, the low-level vision system, which produces *proto-objects* rapidly and in parallel. The *proto-objects* result from linear and non-linear processing of the input scene and are "quick and dirty" representations of objects or object parts that are limited in space and time. Second, a limited capacity attentional system forms these structures into stable object representations. Finally, a non-attentional system provides setting information, for example, on the *gist* - the abstract meaning of a scene, e.g., beach scene, city scene, etc. - and on the *layout* - the spatial arrangement of the objects in a scene. This information influences the selection of the attentional system, for example, by restricting the search for a person on the sand region of a beach scene and ignoring the sky region. Whereas the two first modules resemble the traditional approaches of pre-attentive and attentive processing stages, the third part of the model provides some relevance-based information about the scene at hand and extends existing models in this way. As such, this model integrates low-level, rapid, rough perceptual organization; intertwined attention and perceptual organization; and top-down relevance-based influences on the attentional selection task. However, in this model, these three modules are modeled as rather largely independent systems. Also, the detailed interaction between perceptual organization and attentional selection is not fully described.

Conative model of attention by Jarmasz [62]

A different model, accounting better for the integration of perceptual and relevance-based influences on attention is described in [62]. It is argued that in this model, attention uses visual objects (the products of early perceptual organization)

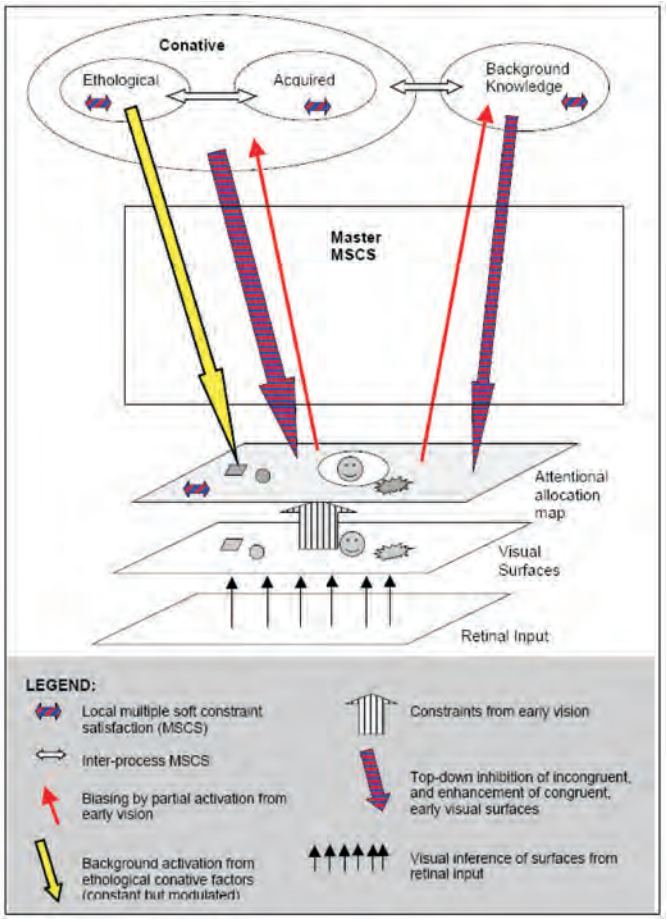


Figure 8: Functional architecture of a conative model of attention, as proposed by Jarmasz [62].

as tools to direct and guide action, and that in doing so, attention shapes its tools. As action depends heavily on an agent's goals, motivations, and needs; themes that have traditionally been grouped under the heading of conation in psychology. The model is called a conative model of attention. As such, visual attention is assumed to interface perception and conation. A three-part architecture for attention is proposed that integrates conative, conceptual, and stimulus-driven factors in a single attentional system. Two of the constituents are the two types of determinants of attentional allocation: the stimulus-driven products of early visual perception on the one hand, and the higher-order factors, such as memory and conation, on the other. The third constituent is the mechanism that combines the two types of determinants of attentional allocation into some structure that can direct attention, corresponding to the unit of object-based attention: the *proto-object*.

Figure 8 illustrates the different types of determinants in the architecture [62]. The stimulus-driven determinants provide the *proto-objects* of the visual field inferred from various low-level visual cues. The organism driven determinants provide information that facilitates perceptual organization and direct attention to elements or groups of elements relevant with an organism's intentions and needs.

In conclusion, early perceptual organization provides a rough initial parsing of a visual scene into elements that the visual perception system uses to organize attention into an attentional structure, namely, a hierarchical organization of *proto-objects*.

In the absence of organism-driven factors, attention will be guided exclusively by the stimulus-driven cues, from which the first *proto-objects* are inferred. The stronger the cues of a particular *proto-object*, the higher the *proto-object* is placed in the attentional hierarchy. When organism-driven factors are strong, stimuli most relevant to an organism's needs and goals will receive most attention. When organism-driven and stimulus-driven constraints on attention are compatible, visual attention picks out stimuli that are relevant to an organism's intentions and actions in an effortless fashion. When the congruence between organism-driven and stimulus-driven constraints is reduced, the deployment of attention is more effortful and less efficient.

What remains as open issues at this point are the mechanisms, both neural and computational, which underlie the interrelation of organism-driven and stimulus-driven determinants of attentional deployment. These are outside the scope of this article.

3.6. Conclusion

Given the space-based attentional models reviewed in section 2, we can conclude that the space-based models of attention made good contributions to the implementation of location-based visual selection. However, the presented study on object-based attention identifies several limitations of space-based visual selection models, and defines primordial requirements for modeling attention to overcome these shortcomings. Research into object-based attention has received increasing interest, but research into useful systematic theories is still open research, especially computational models of object-based attention for real-world applications. We summarize the issues involved in developing biologically plausible object-based attention models as:

1. A recent study [36] shows that object-based and space-based attention share common neural mechanisms in the brain. Object-based and space-based attention are not exclusive but operate at multiple selection levels in the visual system depending on visual tasks. They achieve the coherent selection by objects, features, locations, and their groupings. Grouping is not a simple equivalent of segmentation, but a key means to integrate both object-based and space-based attention together.
2. Object-based attention holds that the underlying unit of attentional selection is an object or a grouping of objects, features, locations, called *proto-object*. These perceptual *proto-objects* are identified and segmented early.
3. Segmentation (perceptual organization) and attention are mutually constrained and influenced [31]. Without segmentation and perceptual grouping, attention may lose its selection units.

4. Attention is controlled by bottom-up and top-down influences. This interaction, especially the top-down influence on attention, biases competition for attentional selection towards objects which are relevant to the current behavior.
5. In order to simulate human-like visual (re-)exploration behavior, an object's visual saliency should be evaluated in a spatiotemporal context. This way, visual saliency of an object varies with multiple resolutions and over time.
6. Grouping-based competition for attention should be performed by integration of object features, location features, and their distribution over the visual field. Grouping-based competition considers object-based hierarchical selectivity, involving object-based selection between objects and within an object, as required for real-world scenes [133]. Attention can work at multiple processing levels to execute selectivity by features, objects, locations, or their groupings.
7. Saliency mapping, grouping-based competition, and inhibition of return mechanism for control of attention must operate in a spatiotemporal context, for achieving human-like visual behavior in machine vision.

One issue concerning visual attention modeling has been omitted in this work, namely, the discussion about covert attentional selection and overt foveal eye movements. Visual attention covertly shifts in the visual field to select interesting objects when the fovea is fixated. Visual attention can perform visual selection without eye movements but eye movements require visual attention to function so as to assist attention to scrutinize the potential objects of visual selection in the periphery of the field of view [48]. Therefore, the shifts of attentional selection are clearly distinct from eye movements. Recently more and more active vision systems attempt to employ attentional mechanisms to help eye movements for their goal locating (e.g. [135, 5]), but research into integrating both of the shifts of (covert) attentional selection and eye movements in one system is lacking. To be complete, a biologically plausible vision system should consider foveal sensing together with visual attention but importantly make a clear distinction between them.

In conclusion, modeling and implementing object-based visual attention must engender a framework satisfying the enlisted issues and requirements. To our knowledge, only Sun [133] proposed a hierarchical object-based attention framework. This framework aims at integrating object-based and space-based attention, and employing grouping-based competition for attentional selection to achieve object-based hierarchical selectivity of visual (covert) attention and attention-guided overt saccadic eye movements. The concept of grouping is defined as the underlying unit of attention selection and is used to link object-based and space-based attention together so as to obtain hierarchical selectivity within visual attention. A grouping is defined as a hierarchically structured unit, and can be a point, a feature, an object, a group of objects or features, or a region.

[133] provides answers and implementation details to the above requirements 1, 4, and 6, and also includes considerations on how to integrate visual (covert) attention and attention-guided overt foveal eye movements. However, early segmentation into *proto-objects*, intertwined perceptual organization and attention, focus of attention and inhibition of return mechanisms in a spatiotemporal context are not described with implementation details.

Complementary to [133], we propose in the following section our approach to object-based attention modeling, based on the foundational discussions presented in the current chapter. Our approach will formulate explicit answers, with implementation details to requirements 2, 3, 5, 6, and 7. We will only consider covert attentional selection, in a bottom-up approach, omitting (overt) saccadic eye movements and top-down influences on attention.

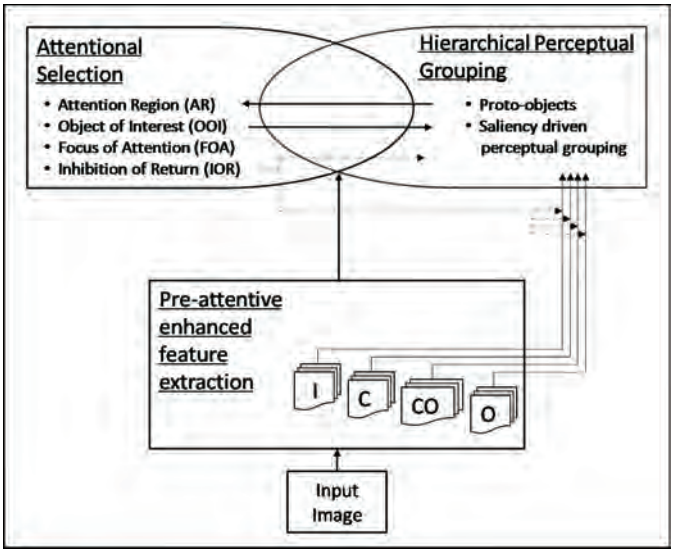


Figure 9: Functional architecture of the proposed bottom-up model for object-based visual attention.

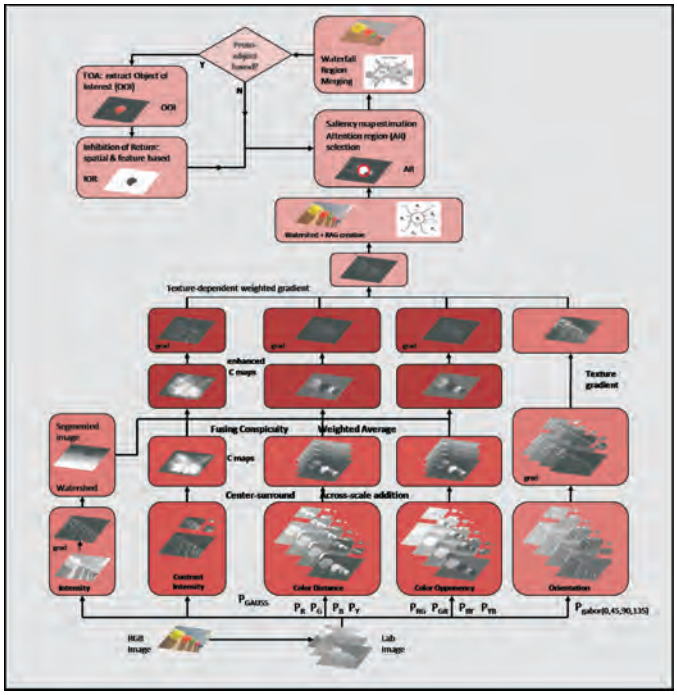


Figure 10: Structural framework of the proposed bottom-up model for object-based visual attention. (P = pyramid)

4. Modeling Selective Attention and Perceptual Grouping to Salient Objects

4.1. Visual Attention Model - General Overview

Based on the models of Jarmasz [61, 63, 62] and Rensink [118, 119] and the discussion of section 3, we propose a new biologically inspired computational model for object-directed attention. In our approach, we extend attention to the middle stages and integrate the selection process with the perceptual grouping process. Integration is achieved through an innovative saliency driven perceptual grouping strategy, extending the traditional pixel-based saliency map to salient *proto-objects*.

Figure 9 gives a general overview of the proposed scheme. It is composed of:

- A Pre-attentive bottom-up enhanced feature extraction module. To achieve salient region localization, pixel-based feature extraction methods are enhanced with region information from rapid, rough image segmentation results, eliminating misleading high-contrast edges.

- An Attentional Selection and A Hierarchical Perceptual Grouping Module, which form an integrated process and perform the selection of the most salient *proto-objects* and a perceptual grouping for forming perceptual meaningful objects.

The proposed object-oriented attentional selection model is in conformity with the conclusions drawn in the previous section 3 and possesses as such the following key-features:

- Proto-objects are generated and utilized as units of visual attentional selection. The incorporation of region-based information occurs from the pre-attentive filtering stage on, with the goal of enhancing extracted feature information.

- The visual region saliency is defined and estimated. It measures how different a proto-object is from its surround. The more salient a proto-object is, the more advantaged it is to compete with other proto-objects for visual attention.

- Intertwined attentional selection and perceptual grouping processes are obtained by saliency driven perceptual grouping.

- Integrated competition for attention between groupings follows from the fact that, at any given moment, enhanced responses to one grouping will decrease responses to other competitors. When a grouping gains the dominance of selective attention, all other relevant processing in the visual system to this grouping and all sub-groupings belonging to this grouping share the same dominance.

- The relationship between object-based attention and space-based attention is comprised in the definition of the visual region saliency. The saliency of a proto-object is biased by a measure of the spatial distribution of visual region saliency over the field of view. This measure is the compositional balance indicator (CBI), defined in section 4.6, and reflects the spatial probability distribution of saliency over the whole scene.

- An inhibition-of-return (IOR) mechanism is incorporated in the model, allowing multiple objects of interest (OOIs) to be selected consecutively, hereby generating the scan path. Two types of inhibition mechanisms are considered, namely spatial based and feature based.

Figure 10 depicts the structure of the proposed bottom-up model for visual attention. In conformity with [39], three different *feature dimensions* (in the following simply called features) are computed: intensity, color, and orientation. For each feature dimension, the pixel-based responses are computed on different scales and for different *feature types*, e.g. red, green, blue and yellow for the color feature. First, we compute an *image pyramid*. For each level in the pyramid, the feature type responses form the *scale maps*. These represent the saliencies at different scales for all feature types. Then, the scale maps are fused, across scales, into *feature maps* representing different feature types. Subsequently, the feature maps are combined, across feature types, into *conspicuity maps*, one for each feature, thereby strengthening important aspects and ignoring others.

Taking into consideration the critical view on pyramidal segmentation algorithms described in [12], a hybrid approach combining low-level saliency and region information is used to produce *enhanced conspicuity maps* (*enhanced C maps*). The conspicuity maps are enhanced with region information (from image segmentation), by averaging the conspicuity values in each region. This approach has also been followed by [74] to produce region enhanced saliency using a single feature intensity. The output of this phase is a "multi-spectral" image combining all the enhanced conspicuity maps.

The obtained "multi-spectral" image is then segmented using the watershed transform. To apply the watershed, the gradient of the "multi-spectral" image is obtained by combining, using the approach of [100], the gradients of the texture (from the orientation responses) and the gradients of the enhanced conspicuity maps. This approach allows obtaining a final gradient capturing all perceptual edges in the input RGB image. The method is general, in the sense that it makes limited assumptions about scene content. Both textured and non-textured areas are accommodated, and the region size is irrelevant. The processing is adapted to local properties of the image, allowing suppressing the intensity gradient in textured areas but leaving it unmodified in smooth regions.

From the segmentation results, the region-based saliency map is obtained. Subsequently, the most salient region is selected as *attention region* (*AR*).

Salient region extraction, from the saliency map, provides a good starting point for semantic-sensitive content representation. However, perceived salient region extraction from images is still an unsolved problem. One reason is that low-level features are often not sufficient to classify some regions unambiguously without the incorporation of high-level and human perceptual information into the classification process. Another reason for the problems is perception subjectivity. Different people can differ in their perception of high-level concepts, thus a closely related problem is that the uncertainty or ambiguity of classification for some regions cannot be resolved completely based on measurements methods.

In order to alleviate for the above constraints, we propose a new method for salient region extraction where the segmented image is analyzed by a number of perceptual attributes based on the *mise-en-scene* principles [24], which are used by the film-makers to guide our attention across the screen, shaping our sense of the space that is represented and emphasizing certain parts of it. The used perceptual attributes, to select a salient attention region, are its *contrast* with its surroundings, its *Orientation Conspicuity*, its *Compactness*, and its *Compositional Balance* [24] which can be interpreted as the extent to which the areas of image space have equally distributed masses and points of interest.

In conformity with the discussion in section 3, the attention region corresponds to the *proto-object* of Rensink [118] or the *a-object* of Jarmasz [63] at the earliest stage, corresponding to the fairly complex, rapidly formed perceptual entities (groups) created by early visual perception. At the early stage of simulated visual perception, before the perceptual grouping processes are involved, we refer to the attention region as a *proto-region*, rather than a *proto-object*. As stated by Rensink [119], attentional interaction with lower-level structures is taking place via a nexus, a single structure containing a summary description of the attended proto-region. As such, *proto-regions* serve as the highest outputs of low-level vision, but also the lowest level operands upon which higher level attentional processes can act.

Starting from the attention region (AR), as most salient *proto-region*, a saliency driven perceptual grouping process of segmented regions is proposed to obtain perceptually meaningful regions, the proto-object. The perceptual organization process groups segmented proto-objects and shapes the attention region.

Indeed, a meaningful image segmentation groups the pixels into disjoint regions that consist of uniform components. Facing the absence of contextual knowledge, the only alternative which can enrich our knowledge concerning the significance of our segmented groups is the creation of a hierarchy guided by the knowledge which emerges from the superficial and deep image structure [144]. Our main goal here, is to create a hierarchy among the gradient watersheds which preserves the topology of the initial watershed lines and extracts homogeneous objects of a larger scale. The *waterfall* algorithm [11, 85] is used here for producing a nested hierarchy of partitions.

In our implementation of the waterfall, the saliency measure of a boundary is based on a collection of energy functions used to characterize desired single-segment properties and pair-wise segment properties. The single segment properties include *area*, *convexity*, *compactness* and *color variances* within the segment. The pair-wise properties include *color mean differences* between two segments and *edge strength*.

Using these energy functions as region merging criteria, the saliency driven perceptual grouping process results in the formation of *Object Of Interest* (*OOI*). The proposed method allows multiple OOIs to be segmented according to the saliency map, by incorporating an inhibition of return (IOR) mechanism, which resets the selected OOI. Iteratively applying

the process of OOI determination and IOR mechanism, creates a scan path, a chronological list of attended objects of interest. As clarified above, the AR corresponds to an early formed proto-object, termed proto-region. Focused attention grasps the attention region, being the most salient proto-region. The perceptual organization process, in concordance with focused attention, forms and shapes the proto-segments, hereby shaping the AR into a coherent and stable proto-object, termed Object of Interest, in conformity with Rensink's coherence theory of visual attention [118]. The OOI corresponds to a reshaped hierarchical proto-object, comprised of several lower-level hierarchically organized proto-objects. The subtle differentiation in nomenclature reflects what is called *grouping* in literature [133]. As defined in [133], a grouping is a hierarchically structured unit, and can be a point, a feature, an object, a group of objects or features, or a region.

The following sections describe the detailed implementation of each step of the attention model. The general algorithm is given in the Table 1, and the detailed implementation in the following sections.

4.2. Feature Maps Estimation

According to the standard attention model of Koch and Ullman [67], human perception uses three features, namely, intensity, color and orientation. Many other features have been also incorporated in the attentional selection process, e.g. size, curvature, and motion [138].

The distinct feature channels are acquired by using linear filters tuned to specific stimulus dimensions, such as luminance, red, green, blue, and yellow hues, or various local orientations. The number and response properties of the linear filters implemented in the model have been chosen according to what is known of their neuronal equivalents in the early stages of visual processing in primates. In addition, such decomposition is performed at a number of spatial scales, to allow the model to represent smaller and larger objects in separate subdivisions of these channels.

It is important to note that the luminance efficiency function for the human eye, i.e. the eye's response to light or constant luminance, peaks at a wavelength of about 550nm, corresponding to yellow-green light [37]. This should be taken into account when trying to more closely reproduce the response properties of luminance selective neurons.

Similar to the luminance case, the color representation is approximate. It does not account for the fact that the three types of color-sensitive cone photoreceptors in the human retina have their peak sensitivities at wavelengths of light which are not necessarily matched to the primary colors used in digitized images. (Although these three types of cones are often referred to as "red", "green", and "blue" types, with peak sensitivities at 580nm, 545nm, and 440nm, they actually maximally respond to orange, yellow and blue hues, respectively; also sensitivity to blue is almost ten times smaller than to red and green [37].

- input:	RGB image
- step 1:	feature maps estimation
output:	I'_{on}, I'_{off}, $O'_{0^o}, O'_{45^o}, O'_{90^o}, O'_{135^o}$, C'_R, C'_G, C'_B, C'_Y $CO'_{R^+G^-}, CO'_{G^+R^-}, CO'_{B^+Y^-}, CO'_{Y^+B^-}$
- step 2:	enhanced conspicuity maps
output:	$I_e(x,y)$ $O_e(x,y)$ $C_e(x,y)$ $CO_e(x,y)$
- step 3:	multispectral gradient estimation
output:	$GS(x,y)$
- step 4:	watershed segmentation, small region merging
output:	RAG: $G = (P^0, E^0)$ $P^0 = (R_1^0, ..., R_n^0)$, set of regions E^0, set of arcs between regions R_i^0, R_j^0
- step 5:	region saliency estimation and AR selection
	if no AR: stop
	else: goto step 6
- step 6:	waterfall region merging, perceptual organization OOI determination
output:	RAG^k: $G = (P^k, E^k)$ OOI
- step 7:	focus of attention and inhibition of return
	- extract (inhibit) focused OOI and update RAG
	- re-estimate saliency map
	goto step 5

Table 1: The general algorithm of the proposed attention model. In step 5, the stopping criteria is expressed: 'if no attention region (AR) is selected, stop the algorithm'.

Concerning the orientation, the filters approximate the receptive field sensitivity profile (impulse response) of orientation-selective neurons in primary visual cortex [53].

The features intensity, color and orientation are used at the basis of the model. A multi-scale approach is adopted to enable the detection of important edges, and hence salient regions, of different sizes. Instead of rescaling the filters, resulting in extremely time-consuming computations, the images are rescaled.

In the literature, either an empirically fixed image scale is selected [81], or a fixed number of image scales is defined for computation of the feature responses [59, 57]. Determining automatically a suitable scale for each image is a difficult task. In [74], Liu et al. proposed a method which calculates a scale-invariant saliency map based on pixels/blocks. The underlying idea of this multi-scale method is to calculate the image feature contrast at an image scale matching the feature scale. In other words, features will stand out at an image scale matching to their feature scales. For example, the large scale features will be highlighted at a coarse scale and the small-scale features will be highlighted at a fine scale. The number of levels, n_l, is calculated from the original image size (w, h) as $\log_2(min(w, h)/10)$.

In the following sections we summarize the perceptual features we used in our system, and adopt the same notation as in [39].

The Intensity Feature

The intensity feature emphasizes regions with strong intensity contrasts in the input image. The intensity feature map is created by *center-surround mechanisms*. These mechanisms are inspired by the ganglion cells in the visual receptive fields of the human visual system, which respond to intensity contrasts between a center region and its surround. The cells are divided into two types: on-center cells responding excitatorily to light at the center and inhibitorily to light at the surround, whereas off-center cells respond inhibitorily to light at the center and excitatorily to light at the surround [104].

First, the RGB color input image is converted to gray-scale. From the gray-scale image, a *Gaussian image pyramid* $\mathcal{P}_{Gauss}$ is computed by iteratively applying a 3x3 Gaussian filter to the image resulting in a smoothed image, followed by sub-sampling the image. This strategy results in an image pyramid with n_l different scales s_0 to $s_{n_{l-1}}$ (Figure 11). The following computations are all performed on scales s_1 to $s_{n_{l-1}}$, to enable robustness to noise [59].

As in [39], in the center-surround mechanism, the center c is given by a pixel in one of the scales s_1 to $s_{n_{l-1}}$. The averages of the pixels surrounding the center c are calculated for two different surround radii (3 and 7 pixels). Two surrounding σ's are defined. The two center-surround differences are determined as the difference between the c value and the respective surround average σ. According to the human system, we determine two feature types for intensity: the on-center difference responding strongly to bright regions on a dark background, and the

(a) (b)

Figure 11: (a) Input image; (b) the derived Gaussian image pyramid.

off-center difference responding strongly to dark regions on a bright background [53]. This yields intensity scale maps $I''_{i,s,\sigma}$ with $i \in \{(on), (off)\}$, $s \in \{s_1, ..., s_{n_{l-1}}\}$, $\sigma \in \{3, 7\}$ (Figure 12). Unlike the human system, where the surrounding region of the ganglion cells is circular, we consider a rectangular surround, for simplicity reasons.

The maps for each center-surround variation are summed up by *across-scale addition*: first, all maps are resized to scale s_1 (resizing scale s_i to scale s_{i-1} is done by duplicating each pixel), then the maps are added up pixel by pixel. This yields the intensity feature maps I':

$$I'_i = \bigoplus_{s,\sigma} I''_{i,s,\sigma} \tag{1}$$

with $i \in \{(on), (off)\}$, $s \in \{s_1, ..., s_{n_{l-1}}\}$, $\sigma \in \{3, 7\}$, and $\bigoplus$ denoting the across-scale addition. The two intensity feature maps are shown in Figure 12 on the right.

Figure 12: Left: the intensity scale maps $I''_{i,s,\sigma}$. First row: the *on-maps*. Second row: the *off-maps*. Right: the two intensity feature maps $I'_{(on)}$ and $I'_{(off)}$.

Compared to the approach of [39] presented above, [74], based on [81], proposes a slightly different method to calculate the contrast intensity feature. The contrast pyramid is obtained by calculating the contrast map at each scale. The contrast intensity value $CI_{i,s}(x, y)$ at scale s is defined as the weighted sum of the differences in intensity I_i ($i \in \{(on), (off)\}$) between the pixel (x, y) at scale s and each other pixel in its neighborhood:

$$CI_{i,s}(x, y) = \sum_{\mathbf{q} \in \Theta} w_s(x, y) d(I_{i,s}(x, y), I_{i,s}(\mathbf{q}))$$

with $w_s(x, y) = 1 - r_s(x, y)/r_{s,M}$, Θ being the neighborhood of pixel (x, y) at scale s, d the distance between the two values. $r_s(x, y)$ is the distance from (x, y) to the center of the image, and $r_{s,M}$ is the maximal distance to the image center. The weighting factor $w_s(x, y)$ is used to account for the heuristics that the center of an image is usually more visually salient.

The resulting contrast intensity feature map is estimated from the contrast intensity pyramid by across scale addition of the contrast maps at all the scales. This yields the intensity feature maps I':

$$I'_i = \bigoplus_s CI_{i,s} \tag{2}$$

with $i \in \{(on), (off)\}$, $s \in \{s_1, ..., s_{n_{l-1}}\}$. The two intensity feature maps, obtained with this approach are shown in Figure 13.

Figure 13: The two intensity feature maps $I'_{(on)}$ and $I'_{(off)}$ resulting from the across scale addition of the corresponding scale maps.

We will adopt the approach of [74, 81], because this method accounts for the heuristics that the center of the image is usually more salient, as opposed to the method of [39].

The Orientation Feature

The orientations are computed by convolution with Gabor filters [53] detecting bar-like features according to a specified orientation. Also steerable filters or a wavelet based approach [100] can be adopted.

A 2D Gabor function is an oriented complex sinusoidal grating modulated by a 2D Gaussian function. A 2D Gabor function $g(x, y)$ [160]:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp[-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}) + 2\pi j u_0 x]$$

In the frequency domain, the Gabor filter bank [124] is defined as:

$$G_{ij} = G(\omega_r - \omega_{r_i^o}, \omega_{\varphi_j^o}) \tag{3}$$

where (r, φ) are polar coordinates, $\omega_{r_i^o}$ is the logarithm of the center frequency at scale $i \in [1, M_G]$, $\omega_{\varphi_j^o}$ is the j^{th} orientation ($j \in [1, N_G]$)and $G_{\omega_r\omega_\varphi}$ is defined as:

$$G(\omega_r, \omega_\varphi) = exp[\frac{-\omega_r^2}{2\sigma_{r_i}^2}] exp[\frac{-\omega_\varphi^2}{2\sigma_{\varphi_j}^2}] \tag{4}$$

where σ_{r_i} and σ_{φ_j} are the parameters of the Gaussian. The N_G orientations are taken equidistant (Eq. 5) and the scales are obtained by dividing the frequency range $(\omega_{max} - \omega_{min})$ into M_G octaves (Eq. 6).

$$\begin{aligned} \sigma_{\varphi_j} &= \frac{\pi}{2N_G} \\ \omega^o_{\varphi_j} &= 2(j-1)\sigma_{\varphi_j} \end{aligned} \tag{5}$$

$$\begin{aligned} \sigma_{r_i} &= 2^{i-1}\sigma \\ \omega^o_{r_i} &= \omega_{min}(1 + 3(2^{i-1} - 1))\sigma \end{aligned} \tag{6}$$

where $\sigma = \frac{\omega_{max} - \omega_{min}}{2(2^{M_G} - 1)}$ which yields $2\sigma, 4\sigma, ..., 2^{M_G}\sigma$ octaves. Note that the maximum frequency cannot be larger than the Nyquist frequency and the DC-component of the image is removed before filtering.

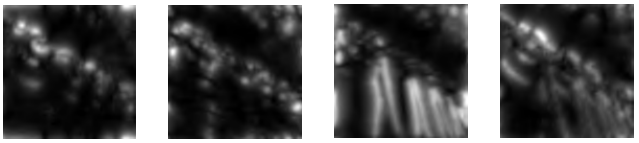
The gray-level image, $I(x, y)$, is passed through a bank of Gabor filters, and a set of filtered images $O''_{\theta,s}(x, y)$ are obtained. We consider four orientations, namely $\theta \in \{0^o, 45^o, 90^o, 135^o\}$, and three scales, namely $s \in \{0, 1, 2\}$. $O''_{\theta,s}(x, y)$ is the modulus of the convolution of the input image $I(x, y)$ with Gabor filter $g_{\theta,s}(x, y)$:

$$O''_{\theta,s}(x,y) = |I(x,y) * g_{\theta,s}(x,y)| \tag{7}$$

Following [39], the orientations scale maps $O''_{\theta,s}$ are summed up by across-scale addition for each orientation, yielding four orientation feature maps O'_θ of scale s_1, one for each orientation:

$$O'_\theta = \bigoplus_s O''_{\theta,s} \tag{8}$$

with $\theta \in \{0°, 45°, 90°, 135°\}$, and $s \in \{s_1, ..., s_{n_{l-1}}\}$. The orientation feature maps are depicted in Figure 14.



(a) $O'_{0°}$ (b) $O'_{45°}$ (c) $O'_{90°}$ (d) $O'_{135°}$

Figure 14: The four orientation feature maps.

In contrast to the NVT [59], [39] does not use the center-surround technique explicitly for computing the orientation maps. The oriented center-surround difference that is determined by cells in the human cortex is already determined implicitly by the Gabor filters.

The Color Features

To compute the color feature maps, the RGB color image is converted into the LAB color space. The CIE LAB color space is one of the most popular uniform color spaces [38]. In uniform color spaces, the distance between colors corresponds quite good to the difference between two colors as perceived by a human observer. This fact is important trying to simulate the human perception mechanism.

LAB is a complementary color system, i.e. it is based on the three elementary color pairs: red-green (+a,-a), yellow-blue (+b,-b) and black-white L$\in$ $[0, 100]$. Luminance is already considered in the intensity feature, so we ignore this channel here. In the human visual cortex, color is perceived by a color double-opponent system with the color opponent cells red-green, green-red, blue-yellow, and yellow-blue. Red-green cells are excited by red light and inhibited by green, and so on. The representation of red and green as well as of blue and yellow in separate maps enables color-specific pop-outs.

From the LAB image, an LAB image pyramid, $\mathcal{P}_{LAB}$, is generated by applying a Gaussian filter. From the pyramid $\mathcal{P}_{LAB}$, four color pyramids $\mathcal{P}_R$, $\mathcal{P}_G$, $\mathcal{P}_B$, and $\mathcal{P}_Y$ are generated for the distinct colors red, green, blue, and yellow. We note that $\mathcal{P}_{LAB}$ is a multi-channel pyramid, where each pixel of a scale level in the pyramid corresponds to a vector (p_l, p_a, p_b), whereas $\mathcal{P}_R$, $\mathcal{P}_G$, $\mathcal{P}_B$, and $\mathcal{P}_Y$ are single channel images, where each pixel of a layer of these pyramids is a scalar.

The maps of these color pyramids show to which degree a color is represented in an image, i.e., the maps in the pyramid $\mathcal{P}_R$ show how "red" the image regions are: the brightest values are at red regions and the darkest values at green regions (since green has the largest distance to red in the color space). The pixel value $\mathcal{P}_{R,s}(x, y)$ in map s of the "red" pyramid $\mathcal{P}_R$ is obtained by estimating the distance between the color of the corresponding pixel $\mathcal{P}_{LAB}(x, y) = (p_a, p_b)$ and the prototype red color $r = (r_a, r_b) = (255, 127)$ for a maximal value of 255 in the color space. This yields:

$$\mathcal{P}_{R,s}(x, y) = \sqrt{(p_a - r_a)^2 + (p_b - r_b)^2} \tag{9}$$

Figure 15 depicts one level of the color pyramids. The color contrast is computed by the on-center differences:

$$C''_{\gamma,s,\sigma} = center - surround(s, \sigma) \tag{10}$$

with $\gamma \in \{red, green, blue, yellow\}$, $s \in \{s_1, ..., s_{n_{l-1}}\}$, and $\sigma \in \{3, 7\}$. The off-center-on-surround difference is not needed, because these values are represented in the opponent color pyramid. The maps of each color are rescaled to the scale s_1 and summed up into four color feature maps C'_γ:

$$C'_\gamma = \bigoplus_{s,\sigma} C''_{\gamma,s,\sigma} \tag{11}$$

with $\gamma \in \{red, green, blue, yellow\}$, $s \in \{s_1, ..., s_{n_{l-1}}\}$, and $\sigma \in \{3, 7\}$. Figure 15, shows the obtained color feature maps.

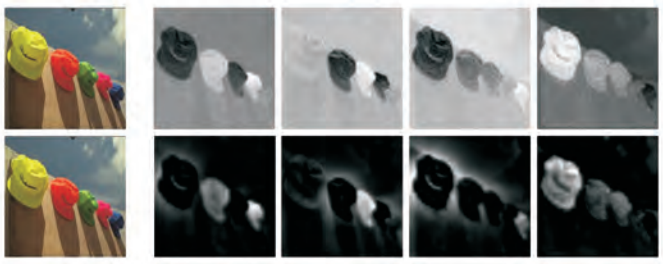


Figure 15: Left: the input image. Right: first row: P_R, P_G, P_B, P_Y maps at scale s_2. Second row: the color feature maps after applying the center-surround difference.

Color Opponency Features

The red, green, blue channels of each image are separated and the yellow channel is constructed as the arithmetic mean of the red and green channels. Successively these four channels are combined to generate four color opponent channels, similar to those of the retina. Each channel, indicated as R^+G^-,G^+R^-,B^+Y^-, and Y^+B^-, has a *center-surround receptive field* (RF) with spectrally opponent color responses. That is, for example, a red input in the center of a particular RF increases the response of the channel R^+G^-, while a green one in the surrounding will decrease its response. The spatial response of the RF is expressed by a difference-of-Gaussians (DoG) over the two subregions of the RF, 'center' and 'surround'. This operation, considering for example the R^+G^- channel, is expressed by:

$$R^+G^-(x, y) = \alpha R * g_c - \beta G * g_s \tag{12}$$

The two Gaussians g_c and g_s, are not balanced: the ratio β/α is chosen to 1.5. The unbalanced ratio preserves the achromatic information: that is, the response of the channels to a uniform gray area is zero. The ratio σ_s/σ_c, the standard deviations of the two Gaussians, is chosen equal to 3 [**?** 28].

Figure 16 depicts one level of the color opponency pyramids. The color contrast is computed through the on-center differences and subsequently across-scale addition as in the previous section. This is shown in Figure 16, bottom.

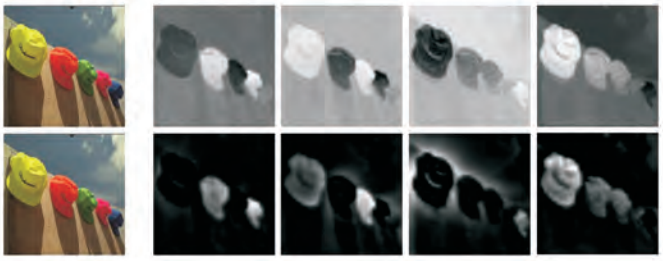


Figure 16: Left: the input image. Right: first row: $P_{R^+G^-}$, $P_{G^+R^-}$, $P_{B^+Y^-}$, and $P_{Y^+B^-}$ at scale s_2. Second row: the color opponency feature maps after applying the center-surround difference.

4.3. Enhanced Conspicuity Maps

The next step is the generation of the *conspicuity maps*. Conspicuity, a term introduced by Milanese [92], denotes the saliency of a feature dimension. This is done by summing the normalized maps into conspicuity maps for features: intensity (I), orientation (O), color (C), and color opponency (CO).

$$\begin{array}{ll} I = \sum_i \mathcal{W}(I'_i); & i \in \{on, off\} \\ O = \sum_\theta \mathcal{W}(O'_\theta); & \theta \in \{0^o, 45^o, 90^o, 135^o\} \\ C = \sum_c \mathcal{W}(C'_c); & c \in \{red, green, blue, yellow\} \\ CO = \sum_c \mathcal{W}(CO'_c); & c \in \{R^+G^-, G^+R^-, B^+Y^-, Y^+B^-\} \end{array} \tag{13}$$

In order to determine the most important maps and raise their influence, we should apply a normalization operator. Itti [59] uses for the map X the following normalization $N(X) = X \cdot (M - \overline{m})$, with M being the global maxima and $\overline{m}$ the average of the local maxima. It emphasizes maps with one strong peak and suppresses those which contain many almost equivalent peaks. The main problem with this approach, as also mentioned in [58], is that if there are two equally high maxima, the difference yields zero, implying that this map is ignored completely, while humans would consider both maxima as salient.

Another approach for map normalization is an iterative nonlinear normalization procedure, based on local competition between neighboring salient locations. While this non straight forward approach yields acceptable results, it is complex and time consuming.

In this work, we propose:

$$\mathcal{W}(X) = X/\sqrt{m} \tag{14}$$

Where m is the number of local maxima in a pre-specified range from the global maximum [39]. This ranging threshold is determined by analysis of the distribution of the maximum values, by choosing for example the median of the maximum.

The effect of the weight $\mathcal{W}$ is shown in Figure 17: the map I'_{on} with a single peak is weighted higher than the map I'_{off} with multiple peaks. This enables the white dot pop-out in the conspicuity map.

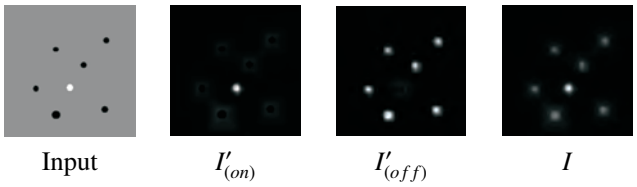


Figure 17: The effect of the uniqueness weight function $\mathcal{W}$. $I'_{(on)}$ has a higher influence than $I'_{(off)}$ and the white dot pops out in the conspicuity map.

The conspicuity maps of the image of Figure 11 are illustrated in Figure 18.

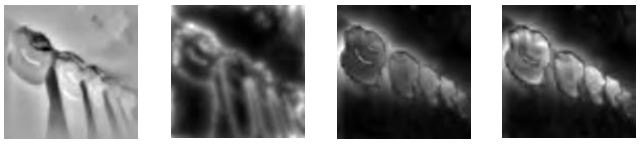


Figure 18: The conspicuity maps for intensity (a), orientation (b), color (c), and color opponency (d).

Image conspicuity, denoting the saliency of a feature dimension, enhances contrast, e.g. edges. However, it fails to provide enough information for localization of salient objects. The low-level features, as described in previous sections, do not necessarily map well to salient objects. For instance, high-contrast edges between regions usually stand out, which will mislead applications into identifying the salient object.

Recently several researchers have addressed the region based saliency/conspicuity analysis. In [72] salient region extraction is performed using first k-means clustering to segment the image into homogenous regions, and next use k-means clustering again to classify the regions into salient and non-salient groups based on the observation that salient regions usually are in the image center. In [51] a salient region is extracted by applying convex-hull to the salient points. In [50], regions are identified by estimating subspaces in the 2D space of polar transformed image features.

Other research combines low-level saliency with high-level information to achieve attentive region/object localization. Feature conspicuity is combined with high-level information such as faces and text to find regions of interest [80], [19], [73].

In this work, we propose a hybrid approach combining low-level saliency and region information into enhanced conspicuity maps (enhanced C maps). To achieve salient region localization, the C maps of equation 13 are enhanced with region information from image segmentation (see 4.5, section "Image Segmentation"), by averaging the conspicuity values in each region, yielding enhanced conspicuity maps denoted as I_e, O_e, C_e, and CO_e. On top of providing conspicuous regions, the method eliminates misleading high-contrast edges. The same approach was followed in [74, 129, 81].

The output of this phase is a "multi-spectral" image combining all the enhanced conspicuity maps, as illustrated in Figure 19. Note that our approach only enhances the conspicuity maps for the intensity, color and color opponency feature. In a subsequent phase, the derived feature information (enhanced C maps) will be combined with texture information, which incorporates orientation conspicuity.

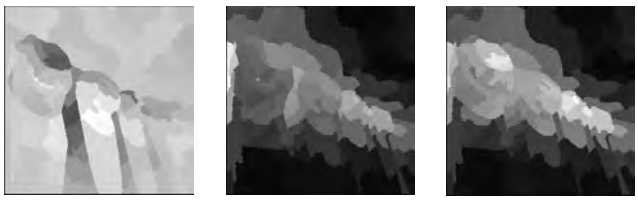


Figure 19: The enhanced conspicuity maps for intensity (a), color (b), and color opponency (c).

4.4. "Multi-spectral" Gradient Estimation

Traditionally the gradient image is computed as $|\nabla I(x, y)|$, where I is the gray scale intensity. This captures the perception that region boundaries are likely where the intensity gradient is large. However, this formulation ignores the fact that humans are able to discriminate not just between regions of homogeneous intensity, but also between those of homogeneous texture.

The gradient of the "multi-spectral" image is obtained by combining the gradients of the texture (from the orientation responses) and the gradients of the enhanced conspicuity maps, using the approach of [100], based on [47]. This approach allows obtaining a final gradient that captures all perceptual edges in the input RGB image. The method is general, in the sense that it makes limited assumptions about scene content. Both textured and non-textured areas are accommodated, and the region size is irrelevant. The processing is adapted to local properties of the image. allowing suppressing the intensity gradient in textured areas but leaving it unmodified in smooth regions.

The proposed method integrates a measure of spatial variation in texture with the traditional gradient computation of the enhanced conspicuity maps, and consists of a number of conceptual stages, as illustrated in Figure 20:

1. Compute a texture representation that characterizes a local area surrounding each pixel.
2. Post process the texture features to make them suitable for meaningful gradient extraction.
3. Generate gradient images for each of the texture features, as well as for considered contrast features.
4. Normalize/weight the contribution of each gradient image.
5. Combine the various gradient images to form the single valued gradient surface.

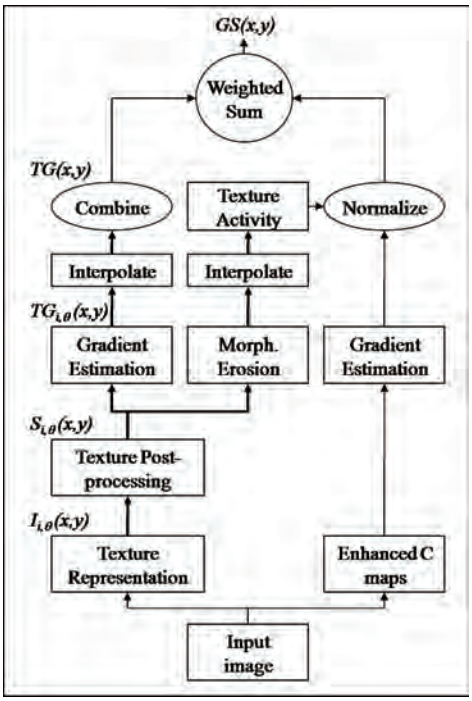


Figure 20: Block diagram of combined gradient extraction incorporating texture information. Thick connecting arrows indicate multiple images (e.g. filter subband respones).

Texture Representation

Texture decomposition is a common way to describe texture in image processing. Gabor filters have gained credence as a perceptually meaningful texture representation because of evidence from psychophysical experiments. Such studies have indicated that the human visual system decomposes the visual field into perceptual channels that are bands of spatial frequency. These channels are evenly spaced in angle with octave bandwidths in the radial direction. Hence, the texture content is usually represented as a vector valued image, in which each decomposition band describes the energy at a given frequency and orientation.

Various Gabor filter based approaches have been explored for texture segmentation. They are generally grouped into two categories: (1) filter-bank approaches, where the Gabor filters are selected from predetermined subbands, wavelets, or other decompositions partitioning the frequency plane; and, (2) filter-design approaches, where the Gabor filters are designed and tuned for a specific texture segmentation task.

In [100] for example, complex wavelets are adopted alternatively to Gabor functions for texture analysis. While retaining the useful properties of scale and orientation sensitivity and approximate shift invariance, they offer a computationally attractive alternative.

In correspondence with the orientation feature estimation described above, we apply a bank of Gabor filters $g_{\theta,s}$. A set of maps $O''_{\theta,s}$ are obtained through Equation 7, with $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ and $s \in \{0, 1, 2\}$. The magnitudes of the responses encode the energy content of the texture feature, as illustrated in Figure 21.

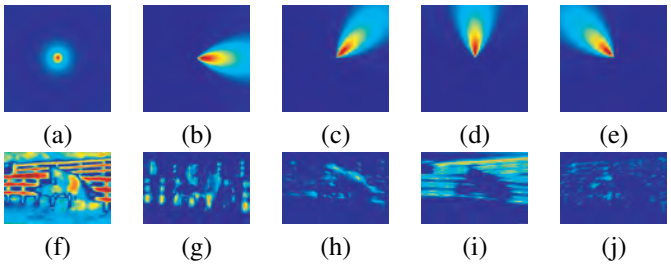


Figure 21: Gabor filter bank: (a) low-pass component and (f) its filter response; (b)-(e) Gabor filters at the different orientations and (g)-(j) the corresponding responses.

Texture Post-processing

In their raw form, the responses of the Gabor filter bank are not useful for the task at hand. A naive approach would be to simply calculate the magnitude of the gradient of the response's image for each subband, as noted in [47]. The problem is that the Gabor filter function not only responds to extended patches of texture with a particular scale and orientation, but also to lines and step edges at that orientation across several scales. Figure 22 illustrates this problem, showing the response of the first level subband oriented at 45^o. If a standard linear derivative-approximating filter is applied to this image, the Gabor filter response gives rise to a double edge in the gradient magnitude (Figure 22(b)). Watershed transformation of this result would result in a spurious narrow region along the boundary.

The solution is to median filter the texture subband magnitudes before the application of the gradient operator (see Figure 22(c) and (d)). Median filtering is well known as a nonlinear edge-preserving smoothing or noise removal technique. In this case, the "noise" in question is any Gabor filter response with a small spatial extent, indicating a local edge rather than an extended area of texture. Thus, the size of the median filter is related to the extent of the filter bank impulse response at that level. In [100], the response to step edges increases slowly in area with subsequent levels of the filter bank and so the support of the median filter was defined to be $(7 + 2n)$, where n is the current level of the Gabor transform. The constant term reflects the size of the filters, while the increase of two samples with each level counteracts

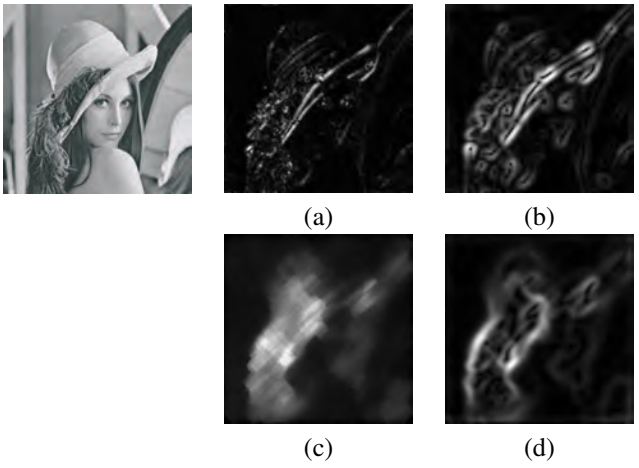


Figure 22: Gabor filter response subband at level one, 45^o (each image has been independently scaled for display purposes). Median filtering removes the response to edges and smooths within textured regions. (a) Raw subband. (b) Gradient of (a). (c) Median filtered subband. (d) Gradient of (c).

the spreading effect of the iterated filtering. Smaller median filters leave behind artifacts of the edges, while larger ones over-smooth, removing texture detail as well as edge responses.

A separable median filter is implemented, as in [100] but, unlike [47], its orientation is adapted to the orientation selected by each subband. Thus, both scale and orientation adaptation is employed in the median filtering. Noting that this processing is not "separable" in the sense of linear two-dimensional (2-D) filters, the order of application of the one-dimensional (1-D) filters is chosen with care.

$$S_{i,\theta}(x,y) = MedFilt_{\theta}(MedFilt_{(\theta+\frac{\pi}{2})}(I_{i,\theta}(x,y))) \tag{15}$$

The first median filter neighborhood extends in a line normal to the subband orientation. By choosing the direction orthogonal to that of the artifact, the extent of the median filter may be minimized. This is desirable, as it avoids over-smoothing and retains as much of the fine structure as possible, while at the same time minimizing computation. A second pass is then made with a neighborhood at right angles to the first (i.e., parallel to the subband orientation) to reduce noise.

Texture Gradient

Let $S_{i,\theta}$ be a median filtered texture subband. The gradient magnitude [15], $TG_{i,\theta}$ of each subband is given by:

$$TG_{i,\theta} = \sqrt{(S_{i,\theta} * G'_x)^2 + (S_{i,\theta} * G'_y)^2} \tag{16}$$

where G'_x and G'_y are the Gaussian partial derivative filters in the x and y directions, and $*$ denotes the convolution.

Since the texture edges will appear at different scales with different resolutions, it is desirable to adapt the support of the derivative filter depending on the level of the transform. As in the case of the median filtering, we take advantage of the subsampling of the filter bank, allowing the use of a single gradient filter size over all the texture subbands. This results in correct multiscale gradient estimation.

Finally, a texture-gradient $TG(x,y)$ is then given by:

$$TG(x,y) = \sum_{i,\theta} w_{i,\theta}\widehat{TG}_{i,\theta}(x,y) \tag{17}$$

with

$$\widehat{TG}_{i,\theta}(x,y) = \frac{TG_{i,\theta}(x,y)}{\max_{x,y}(TG_{i,\theta}(x,y))}$$

and

$$w_{i,\theta} = \frac{N_i}{\sum_{x,y}\widehat{TG}_{i,\theta}(x,y)^2}$$

N_i being the number of pixels of the subband. The effect of the normalization can be understood as the product of two separate scalings: first the coefficients are scaled into the range $[0,1]$, then the result is normalized by its energy. This punishes functions with values all about the same level, while amplifying those with a small number of relatively large peaks.

Gradient Combination

The final single-valued gradient surface is computed as the combination of the texture gradient and modulated feature gradients

$$GS(x,y) = \frac{1}{3*Activity(x,y)}\left(\frac{|\nabla I_e(x,y)|}{\omega_{I_e}} + \frac{|\nabla C_e(x,y)|}{\omega_{C_e}} + \frac{|\nabla CO_e(x,y)|}{\omega_{CO_e}}\right) + \frac{TG(x,y)}{\omega_T} \tag{18}$$

The weight ω_T is just the median value of the texture gradient, while ω_{I_e}, ω_{C_e}, ω_{CO_e} are defined to be four times the median feature gradient. Normalizing each component by its median aligns the noise floor of each function. The additional factor of four reflects the fact that the feature gradient has sharp peaks, while the texture gradient is smoother (and, therefore, the latter must be amplified, to avoid being dominated by the former).

Texture gradient and feature gradients (gradients of the enhanced conspicuity maps) are combined to obtain a final gradient capturing all perceptual edges in the image, by modulating the feature gradients by a measure of texture activity, $Activity(x,y)$. We aim at suppressing the feature gradients in textured areas but leave it unmodified in smooth regions. When the texture gradient is then added, the combined result will be dominated by feature gradients in smooth regions and texture gradient in textured regions, as required. The texture activity measure is defined as:

$$Activity(x,y) = e^{\mathcal{R}_{half}\left(\frac{E_{text}(x,y)}{\alpha} - \beta\right)} \tag{19}$$

with $\mathcal{R}_{half}()$ is just the half-wave rectification to suppress negative exponents:

$$\mathcal{R}_{half}(\zeta) = \begin{cases} 0 & : \quad \zeta < 0 \\ \zeta & : \quad \zeta \geq 0 \end{cases}$$

and

$$E_{ext} = \sum_{i,\theta} \epsilon_B\left(S_{i,\theta}(x,y)\right)$$

with ϵ_B is the morphological erosion.

The activity measure is unity wherever the texture response is below a threshold. When the response exceeds the threshold, the exponential function ensures the activation rises rapidly. The values of α and β are determined empirically [100]. Currently, $\alpha = 2$ and $\beta = 7$. The texture energy, is computed from the up-sampled subband features. However, the texture features respond in a slightly larger area than desired, due to the spatial integration involved. To remedy this, morphological erosion is applied, in order to cause the texture response to recede slightly from region edges.

Figure 23 illustrates the obtained image gradient, combining the different conspicuity maps and the texture.

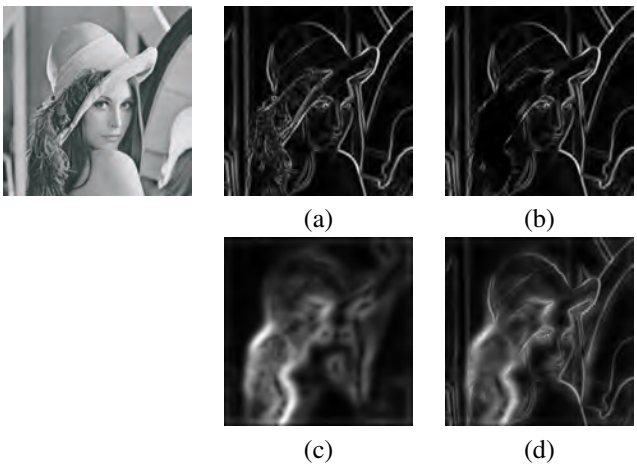


Figure 23: Combination of gradients (images independently scaled for display purposes). (a) Intensity Gradient. (b) Texture modulated gradient: corrected to suppress edges within textured regions. (c) Texture gradient. (d) Combined gradient.

An alternative approach for texture segmentation is proposed in [160]. The presented operation aims at smoothing the image and reducing noise, and highlighting the pixels on the boundaries of two textures, hereby, producing an equivalent edge magnitude map, which emphasizes the texture boundaries.

To conclude the current section, we present results obtained by adopting the proposed approach for texture gradient and feature gradient combination of our enhanced feature conspicuity maps, incorporating region information. As shown in Figures 24 and 25 the approach is able to identify homogenous textured regions. The algorithm has been tested on a dataset of wildlife footage. The texture features are orientation specific, leading to separation of similar patterns in different rotations. However, the effect of oversegmentation is still present, however less dominant than with image segmentation based on traditional gradient estimation. We can conclude that the gradient estimation procedure, as proposed suits perfectly the envisaged subsequent steps.

4.5. Image Segmentation

Extracting regions from an image is a well-studied field [27, 111, 144, 112, 113, 145, 131][]. It aims at parti-

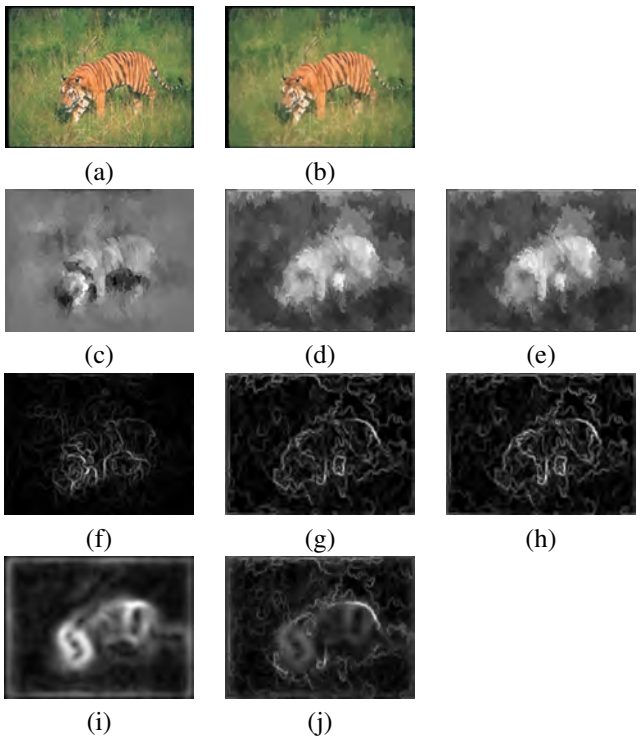


Figure 24: Gradient estimation results on wildlife image. (a) Original image. (b) Mosaic image after watershed segmentation. (c) Enhanced intensity conspicuity map. (d) Enhanced color conspicuity map. (e) Enhanced color opponency conspicuity map. (f) Gradient of enhanced intensity conspicuity map. (g) Gradient of enhanced color conspicuity map. (h) Gradient of enhanced color opponency conspicuity map. (i) Texture gradient obtained with [100]. (j) Final combined gradient.

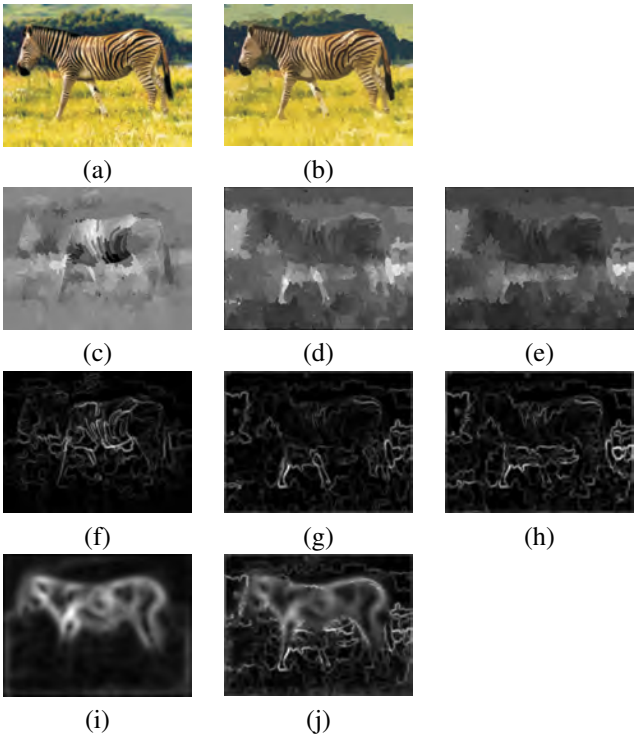


Figure 25: Gradient estimation results on wildlife image. (a) Original image. (b) Mosaic image after fast and rough segmentation. (c) Enhanced intensity conspicuity map. (d) Enhanced color conspicuity map. (e) Enhanced color opponency conspicuity map. (f) Gradient of enhanced intensity conspicuity map. (g) Gradient of enhanced color conspicuity map. (h) Gradient of enhanced color opponency conspicuity map. (i) Texture gradient obtained with [100]. (j) Final combined gradient.

tioning the image into meaningful regions, used in subsequent image analysis steps. Essentially, segmentation transforms the image representation from a pixel-based system to a region-based system, allowing higher-level descriptors in the image model. The difficulty of the segmentation process is that it is an ill-posed problem, since what is a meaningful region is often context dependent, or even unknown as in the case of natural scene images. In spite of many considerable attempts, finding a general method that can produce meaningful segments from low-level image information for a large variety of natural images remains a difficult task. Commonly, the idea is to identify connected pixels, which possess similar feature vectors or are part of a spatial neighborhood that exhibits some kind of coherence in the feature space. The coherence property is often referred to as homogeneity or uniformity in the field of image analysis. In recent years, the notion of scale has been added to account for the fact that physical objects only persist within a certain scale.

Watershed

The watershed transform [87] is a morphological segmentation tool often applied on the gradient magnitude of an image in order to guide the watershed lines to follow the crest lines and the real boundaries of the regions. The watershed transform can be seen as a region-based segmentation approach of scalar images. It segments a scalar-valued image into different regions by interpreting the image as a topographic surface.

The watershed transform can be explained in an intuitive way by a flooding procedure of the image which has been considered as a topographic relief. Figure 26 provides an illustration. Imagine that we have a surface which is flooded through its pierced regional minima. As the water progressively floods the basins, we erect a barrage at the point where water from different adjacent minima meets. When the whole surface will be flooded, the set of all created barrages will be the watersheds. When simulating this process for image segmentation, two approaches may be used: either one first finds basins, then watersheds by taking a set complement; or one computes a complete partition of the image into basins, and subsequently finds the watersheds by boundary detection [123]. Note that methods that follow this principle are denoted immersion type watersheds.

Moreover, the idea of the watershed transformation has been extended to include different flooding scenarios. Initially, the watershed algorithms used solely the height of the topographical surface. There exist watershed transformation that use constraint flooding, flooding based on size or volume that can be associated with topographical surface [143, 88, 132]. However, these methods are beyond the scope of this work.

For a thorough review of the watershed transformation and its algorithms, the interested reader is encouraged to read the review paper by Roerdink et al. [123].

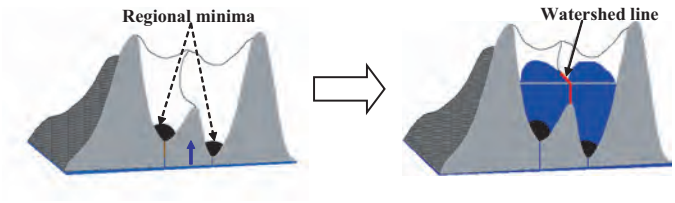


Figure 26: Watershed transformation.

The advantageous characteristics of applying the watershed transform are : (i) The fact that watersheds form closed curves, providing a full partitioning of the image domain, thus it is a pure region-based segmentation which does not require any closing or connection of the edges, (ii) Watersheds of the gradient magnitude can play the role of a multiple point detector, thus treating any case of multiple region coincidence [94], (iii) there is a 1-1 relationship between the minima and the catchment basins. Therefore, we can represent a whole region by its minimum.

Applying the watershed transformation leads to a strong over-segmentation. Two main approaches can be identified coping with this problem. The first approach, denoted the marker-controlled segmentation [87], requires a preprocessing step which identifies the relevant local minima, i.e. the markers, from which the watershed flood will start. However, the determination of the markers is not an easy task. Moreover, the need of a criterion for defining the markers makes generalization difficult. The second approach can be seen as a post-processing step. Once all the catchment basins are extracted, a procedure will merge those catchment basins that belong to almost homogenous regions.

Both approaches can be applied in an hierarchical framework with levels yielding different degrees of partitioning of the image structures. Well-know methods are the waterfall algorithm [11], flat-zones[125][25], dynamics of minima [41].

Small Region Merging

The initial segmentation is over-segmented: homogeneous visual objects are segmented into large regions, while complex visual objects are partitioned into a number of small regions. Obviously, a region belonging to a complex visual object should no be merged to a region belonging to a homogeneous visual object. To eliminate redundant small regions, a small-region merging procedure is applied. The need for merging segments of small size is straightforward. The fact that the more advanced region merging algorithms adopt distribution based merging costs, introduces the requirement that a segment must be of reasonable size to estimate the representative distributions.

This general idea is that segments with a size smaller than a certain threshold are merged with the neighboring segment that exhibits the smallest merging cost. The latter is estimated using the segment's spectral characteristics. The sequence in which segments are merged is given by this cost: the segment couple which has the smallest cost and for which at least one of the segments has a size lower than the size threshold is merged first, until all segments have been removed. Hereafter, the size threshold may be increased.

Let r_i be a region considered for merging, and $\xi(r_i)$ the set of neighboring segments of r_i by 4-connectivity. First, valid merging candidates $\xi_M(r_i)$ have to be selected from $\xi(r_i)$. Let $S(r_i)$, $\mu(r_i)$, and $\sigma(r_i)$ denote the area, spectral statistical mean (in Lab color space: $\mu(r_i) = (\mu_L(r_i), \mu_a(r_i), \mu_b(r_i))$), and spectral variance $\sigma(r_i) = (\sigma_L(r_i), \sigma_a(r_i), \sigma_b(r_i))$) of r_i respectively. A number of approaches exist.

For example, in [106], a region r_j is a valid merging candidate, if the following condition is satisfied:

$$\sigma(r_i) < \sigma(r_j), \quad S(r_i) < S(r_j), \quad |\mu(r_i) - \mu(r_j)| \leq DIFF \tag{20}$$

where $DIFF$ is a threshold value to prevent dissimilar regions from being merged, and the used region merging algorithm is as follows [106]:

Step 1:	-	$DIFF = 0$, $VAR = 0$
Step 2:	-	Find a region r_i such that $\sigma(r_i) \leq VAR$, $S(r_i) \leq S_{min}$, and $\xi_M(r_i) \neq \phi$
	-	If there is no such region, goto step 3
	-	Otherwise, find a merging pair (r_i, r_{j^*}), $r_{j^*} \in \xi_M(r_i)$ that provides the smallest value of variance after merging, and then merge them
	-	Repeat step 2
Step 3:	-	Increase VAR by one
	-	If $VAR > VAR_{MAX}$ goto step 4
	-	Otherwise, goto step 2
Step 4:	-	- Increase $DIFF$ by one
	-	If $DIFF > DIFF_{MAX}$: terminated
	-	Otherwise, goto step 2

We adopt a slightly different approach. First, a stack of region area thresholds T_S is determined: $T_S = [T_{S,0}, T_{S,1}, ..., T_{S,n}]$. Practically, T_S is of the form [2,4,8,16,32,64,128,256,...]. Next, a minimum region size threshold $T_{S_{min}}$ is set. In practice, $T_{S_{min}}$ equals 10 or 20 depending on the size of the image. This additional threshold indicates the minimum required region area to calculate representative region statistics. In a first phase of our approach, we simply would like to merge all segments smaller than $T_{S_{min}}$, following the area threshold stack, with the most corresponding neighbor region in terms of merging cost. The merging cost between region r_i and a valid merging candidate r_j is determined by the spectral mean distance:

$$d(\mu(r_i), \mu(r_j)) = \sqrt{(\Delta\mu_L(r_i, r_j))^2 + (\Delta\mu_a(r_i, r_j))^2 + (\Delta\mu_b(r_i, r_j))^2} \tag{21}$$

In the following phase, we introduce an extra constraint on the merging process, by estimating a color normalization factor T_{cd}, similar to [21]. The color threshold T_{cd} is estimated as

$T_{cd} = \mu_{cd} - \sigma_v$, with μ_{cd} the mean of the color differences D_i's (calculated as in Eq. 21), and $\sigma_v = \sqrt{1/n \sum_{i=1}^{n} (D_i - \mu_d)^2}$ the standard deviation of the $n = \frac{k(k-1)}{2}$ color differences between the k remaining segments. The merging constraint, for region r_i and valid candidate region for merging r_j is then expressed as $d(\mu(r_i), \mu(r_j)) \leq T_{cd}$. The proposed small region merging algorithm can be summarized as:

Step 1:	-	$l = 0$
	-	current size threshold $T_{S,l}$
Step 2:	-	If $T_{S,l} > T_{S_{min}}$, estimate T_{cd}
	-	Otherwise, proceed to step 3
Step 3:	-	Find all regions $\{r_i\}$ such that $S(r_i) \leq T_{S,l}$
	-	If there is no such region, goto step 4
	-	Otherwise, if $T_{S,l} \leq T_{S_{min}}$: find $\forall r_i \in \{r_i\}$ a merging pair (r_i, r_{j^*}), $r_{j^*} \in \xi_M(r_i)$ that provides the smallest merging cost, and perform the merging
	-	if $T_{S,l} > T_{S_{min}}$: find $\forall r_i \in \{r_i\}$ a merging pair (r_i, r_{j^*}), $r_{j^*} \in \xi_M(r_i)$ that provides the smallest merging cost and satisfies the color difference constraint $d(\mu(r_i), \mu(r_{j^*})) \leq T_{cd}$
Step 4:	-	Take next region area threshold in stack of region area thresholds T_S: $l = l + 1$
	-	If $T_{S,l} > T_{S,n}$ it is terminated

Region Adjacency Graph

A mosaic image, produced by applying watershed transformation, can be represented compactly in a graph [87, 146, 91]. A region adjacency graph (RAG) represents the segments and their relationships. It is an undirected graph where segments are represented as nodes, while their common borders with other segments are the arcs of the graph. The elements of the graph can be associated with weights. Commonly the idea is to associate a weight to the arc that gives an indication of the similarity or dissimilarity of the segments it separates. In the case of a directed graph the weight of an arc also depends on the nodes. The latter is useful when the dissimilarity measure includes a segment's local neighborhood. These weights may evolve as one gets a more reliable estimation of the contours. As such, representing the region and their relationships using graphs allows the application of graph-based segmentation algorithms. Figure 27 illustrates the RAG associated to a partition.

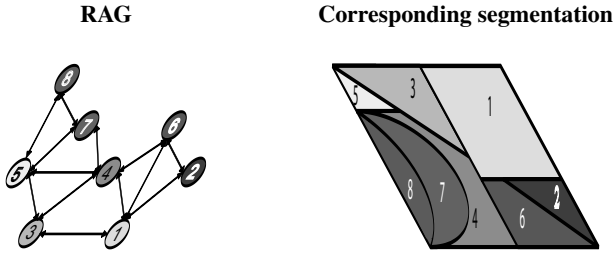


Figure 27: Region Adjacency Graph

Formally we define the RAG as follows:

$$\mathcal{G} = (\mathcal{P}; \mathcal{E}) \tag{22}$$

where $\mathcal{P} = \{r_1, r_2, ..., r_m\}$ represent the set of segments (regions, vertices or nodes) and $\mathcal{E}$ denotes the relationships between them (common boundaries, edges or arcs). Given the number of nodes (m) in a graph, we know that for an undirected graph the number of arcs falls within $[m - 1; \frac{m(m-1)}{2}]$. In our implementation each node is represented by a feature vector corresponding to the average values of the conspicuity maps.

4.6. Region Saliency Estimation and AR selection - proto-region

State-of-art

To our knowledge, salient region determination has not been researched extensively. In the literature, a straightforward approach to region saliency value is averaging the pixel-based saliency map over the pixels of a region [129, 74]. The most acknowledgeable works with respect to region saliency are described in [102, 105], where simple region features are computed and combined into a perceptual metric, representing a region's perceptual relevance or importance. The low-level factors influencing the perceptual metric are: contrast, size, shape, and color. High-level factors include location, and foreground/background distinction.

Contrast Region contrast is a very strong factor; regions with high contrast with their neighboring regions attract our attention, and therefore they might belong to regions of perceptual importance. Let r_i be a region and $\xi(r_i)$ the set of neighbors. The contrast of r_i, estimated using the intensity image, is given by:

$$Contrast(r_i) = \frac{1}{N} \sum_{j \in \xi(r_i)} \alpha_j |\mu_i - \mu_j|$$

$$\alpha_j = Length(\partial r_i \cap \partial r_j) / Perimeter(r_i)$$

where μ_j is the mean of the region r_j and α_j weights the contribution of each neighboring region to the contrast measure. That is, the more contact between regions the more it should contribute to the contrast measure. In this context, ∂r_i refers to the contour of the region r_i [22].

Size Large regions are more likely to attract our attention than small ones. The size measure is computed as

$$Size(r_i) = max\{Area(r_i)/Area_{max}, 1\}$$

where $Area_{max}$ is set to 10% of the total area to prevent excessive weighting to very large regions.

Shape It has been argued that long and thin regions are visual attractors, but also that our perception tends to favor compact regions [128]. Compact regions are valued more important by the shape factor:

$$Shape(r_i) = Area(r_i)/Perimeter(r_i)^2$$

Foreground/Background Typically, objects in the foreground attract our attention. To determine if a region is part of the background we measure the number of pixels of the region border that belong to the image border. In this way, the foreground/background measure is computed as:

$$FB(r_i) = 1 - min\{\frac{Length(\partial r_i \cap \partial\Omega)}{0.5Perimeter(\Omega)}, 1\}$$

where Ω refers to the image and $\partial\Omega$ to the image border.

Location Typically, viewers focus at the center of the image. To compute this factor, the number of pixels of the region which are within the 25% center of the image are counted: $Center(r_i)$. Regions that are entirely in the center of the image obtain the maximum weight:

$$Location(r_i) = Center(r_i)/Area(r_i)$$

Combining all these factors results in an importance map [102, 105]. After normalizing each of the factors to the range $[0, 1]$ the importance map is computed as the sum of their squared values, hereby assigning higher scores to regions with high scores in some factors.

In [75], regions are also chosen as perceptive unit of visual attention analysis. Each region is represented by its average features, RGB color, and the following measures are used for the saliency measure:

Contrast The contrast between two regions r_i, r_j is defined as:

$$FD_{i,j} = (1 - exp(-d_{i,j}/2\sigma^2)) \times 255$$

with $d_{i,j}$ the Euclidian distance between the mean colors.

Area Factor The area factor is simply the ratio of the area of the region to the image size:

$$A(r_i) = \frac{area(r_i)}{AreaOfImage}$$

Global Effect This property is represented as a distance factor between regions:

$$\Theta_2(SD_{i,j}) = 1 - exp(-SD_{i,j}^2/\sigma^2)$$

with $SD_{i,j}$ the relative spatial distance between the regions i and j, normalized to $[0, 1]$.

Contextual difference Different from the pixel based or block based method, in which a unit's contrast is determined equally by the neighboring units, the contextual difference between regions has a coefficient proportional to the adjacency degree between them. An adjacency factor is used to describe this mechanism, calculated as:

$$\Theta_3(E_{i,j}) = 1 + \frac{Length(\partial r_i \cap \partial r_j)}{Length(\partial r_i)}$$

Central effect When watching an image, observers have a general tendency to stare at the central locations, thus the central effect has been defined as:

$$\Theta_0(P_i) = 1 - exp(-\frac{P_i^2}{2\sigma^2})$$

where P_i is the relative distance of the region away from the center of the image, normalized to $[0, 1]$, and σ determines the saliency of marginal regions.

Finally, in [75], the region saliency has been defined as:

$$S(r_i) = \sum_{j \in \xi(r_i)} \Big(\Theta_0(P_i) FD_{i,j} A(r_j) \Theta_2(SD_{i,j}) \Theta_3(E_{i,j})\Big)$$

Proposed Region Saliency

We propose a different method for perceived salient region extraction from images. The segmented image is analyzed by a number of perceptual attributes based on the *mise-en-scene* principles [24]. Mise-en-scene techniques are used by the filmmakers to guide our attention across the screen, shaping our sense of the space. The used perceptual attributes are the *contrast* of a segment with its surroundings, its *Orientation Conspicuity*, its *Compactness*, and its *Compositional Balance* [24] which can be interpreted as the extent to which the areas of image space have equally distributed salient masses and points of interest (Equations 24 to 28).

Formally, the saliency of a region r_i at the current hierarchical level h is given by:

$$\mathcal{S}(r_i) = \frac{\mathcal{S}_r(r_i)}{CBI(r_i)} = \frac{CSR(r_i)OC(r_i)SI(r_i)}{CBI(r_i)} \tag{23}$$

where,

Contrast $CSR(r_i)$ is the *normalized mean color contrast* of a region with respect to the surrounding regions, defined as

$$CSR(r_i) = \sum_{r_j \in \xi(r_i)} \alpha_{i,j}(\sqrt{(\Delta\mu_L(r_i, r_j))^2 + (\Delta\mu_a(r_i, r_j))^2 + (\Delta\mu_b(r_i, r_j))^2} - T_d) \tag{24}$$

with $\alpha_{i,j}$ the ratio of the length of the common boundary of r_i and r_j, over the perimeter of r_i ($\alpha_{i,j} = \frac{Length(\partial r_i \cap \partial r_j)}{Perimeter(r_i)}$). The normalization factor T_d is estimated as

$$T_d = \mu_d - \sigma_v \tag{25}$$

with μ_d the mean of the color differences D_c's, and $\sigma_v = \sqrt{1/n \sum_{c=1}^{n} (D_c - \mu_d)^2}$ the standard deviation of the $n = \frac{k(k-1)}{2}$ color differences between the k generated segments at the current hierarchical level [20]. Indeed, regions, which have a high contrast with their surroundings are likely to be of greater visual importance and attract more our attention. For instance, bright colors set against a more subdued background are likely to draw the eye.

Orientation Conspicuity $OC(r_i)$ is the *orientation conspicuity* defined as the mean output value of the steerable filter (4 orientations, 3 scales) over the region r_i:

$$OC(r_i) = \frac{\sum_{p \in r_i} \hat{O}_p}{Area(r_i)} \tag{26}$$

with $\hat{O}_p$ being the normalized orientation map (at pixel p). Indeed, the orientation map is an important recognition cue, here, it is also employed to describe region orientation information importance, and calculated as defined in [39] (see section 4.2).

Compactness $SI(r_i)$ expresses the compactness of the region:

$$SI(r_i) = \frac{perimeter(r_i)}{Area(r_i)} \tag{27}$$

. With this parameter, we try to find a trade-off between articulated regions and more compact regions of different sizes.

Compositional Balance Indicator Let $gc(r_i)$ be the center of gravity of region r_i; $gc(r)$ the gravitational center of all regions in the image with respect to their saliency value and size, defined as $gc(r) = \frac{\sum^{regions} S_r(r_i) Area(r_i) gc(r_i)}{\sum^{regions} S_r(r_i) Area(r_i)}$. Then, CBI is defined as:

$$CBI(r_i) = \begin{cases} \|gc(r_i) - gc(r)\|; \\ \qquad gc(r) \in r_i \\ \|CSR(r_i)\| + \|CSR(r')\| + \|gc(r_i) - gc(r)\|; \\ \qquad otherwise \end{cases} \tag{28}$$

with, r' the region whose gravitational center is the nearest neighbor of the symmetrical point of $gc(r_i)$ with respect to the midline of the image, this as a measure of overall content of the image.

If the salient region is located near $gc(r)$, we know that the larger CSR and the nearer distance between its gravitational center and the $gc(r)$ region in the image is, the smaller CBI of the region is, meaning the higher the possibility that it will be a salient portion of the image frame. For the second case, the higher CBI (high $CSR(r_i)$ and high $CSR(r')$) shows that the image frame may balance two or more elements encouraging our eye moving between these regions. If $CSR(r_i)$ is high and $CSR(r')$ is low, then CBI will be lower, resulting in a higher saliency compared to the previous described situation, where both $CSR(r_i)$ and $CSR(r')$ are high.

The saliency $\mathcal{S}(r_i)$ is guided by the overall content of the image, represented inherently by the CBI factor, which depends on the saliency of the regions ($gc(r)$). While most existing work, as described above, focus on the detection of salient regions, the proposed approach for extraction of prominent/salient regions is distinctive, since, according to the mise-en-scene principles, the perceptual features are extracted for homogeneous regions, rather than the low-level features.

The region with the highest saliency value $\mathcal{S}(r_i)$ is selected as attention region (AR). If multiple regions exist with high saliency, a limited set of high salient regions can be selected, instead of only considering one AR, in correspondence with the human ability to focus attention on a limited number of objects (typically four to five maximally).

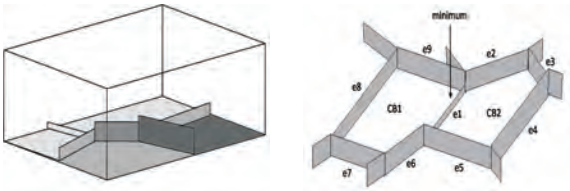
4.7. Perceptual Organization-based Proto-object Detection

A data-driven system for segmenting scenes into objects and their components is presented. This segmentation system generates hierarchies of partitions that correspond to structural elements such as boundaries of objects. The technique is based on perceptual organization, implemented as a mechanism for exploiting geometrical and appearance regularities in the shapes of objects in the image. Image segments are recursively grouped on regions being abstract descriptors encoding structural information. Here, perceptual organization is used to group primitive segments based on Gestalt-like criteria involving some basic properties of proximity, similarity, closure (compactness), and convexity.

The pivotal idea of the proposed model, of object-based visual attention is, in conformity with psychophysical findings, *grouping based salience computation*, and *integrated competition for focus of attention*. Perceptual grouping of segmented regions is envisaged, starting from the salient segment, to obtain perceptually meaningful segments. The salience of a grouping measures how different this grouping contrasts with its surroundings. After the selection of the AR, a perceptual grouping process is applied using the waterfall algorithm ([11, 85]) based on a saliency attached to each edge of the RAG. As such, a nested hierarchy of partitions is obtained, which preserves the topology of the initial watershed transformation on the combined gradient image (see section 4.5). The same idea of grouping, considering a hierarchical structure of objects and space has been proposed in [133].

The *waterfall* algorithm [11, 85] is used here for producing a nested hierarchy of partitions, $\mathcal{P}^h = \{r_1^h, r_2^h, ..., r_{m_h}^h\}; h = 1, \cdots n$, which preserves the inclusion relationship $\mathcal{P}^h \supseteq \mathcal{P}^{h-1}$, implying that each atom of the set $\mathcal{P}^h$ is a disjoint union of atoms from the set $\mathcal{P}^{h-1}$. For successively creating hierarchical partitions, the waterfall algorithm removes from the current partition (hierarchical level) all the boundaries completely surrounded by higher boundaries (see Figure 28). Thus, the saliency of a boundary is measured with respect to its neighborhood. This process can be seen as a watershed applied not to the pixels of an image but to the frontiers of a partition. The iteration of the waterfall algorithm ends with a partition of only one segment.

The original waterfall algorithm [11, 85] produces contrast-based hierarchies by removing edges according to their height, representing the distance between segments (e.g. the minimum pass point of the gradient along the frontier). The graph



(a) (b)

Figure 28: (a) Partition with valuated frontiers and (b) example of region boundary: as the value of edge e1 is smaller than the values of its neighboring edges (e2 to e9), it will be removed by the waterfall algorithm.

implementation of the waterfall algorithm allows to easily produce other hierarchies, by changing the edge valuation method. For example, edges can be valuated with volume extinction values [85, 142], a trade-off between the contrast and the size of a segment.

In our approach, the segment saliency mappings $\mathcal{S}(r_i)$ are dynamically varied according to competition conditions among the groupings at different hierarchical levels of the waterfall. In our implementation of the waterfall, the saliency measure of a boundary is based on a collection of energy functions used to characterize desired single-segment properties and pair-wise segment properties. The single segment properties include segment area, segment convexity, segment compactness and color variances within the segment. The pair-wise properties include color mean differences between two segments and edge strength along the shared boundary. The saliency of the boundary between two neighboring segments r_i and r_j [79]:

$$E(\tilde{r} = r_i \cup r_j | r_i, r_j) = E(\tilde{r}) + E(r_i, r_j) \tag{29}$$

where $E(\tilde{r} = r_i \cup r_j | r_i, r_j)$ is the cost of merging the segments r_i and r_j, $E(\tilde{r})$ is the merged segment property (saliency) and $E(r_i, r_j)$ the pair-wise property, respectively defined as follows.

$$\begin{aligned} E(\tilde{r}) &= E_{area}(\tilde{r}) \tfrac{1}{E_{hom}(\tilde{r})} \textstyle\sum_c E_{var_c}(\tilde{r}) \\ &\quad (1 + |E_{conv}(\tilde{r})|)^{sign(E_{conv}(\tilde{r}))} \\ &\quad (1 + |E_{comp}(\tilde{r})|)^{sign(E_{comp}(\tilde{r}))} \end{aligned} \tag{30}$$

$$E(r_i, r_j) = E_{edge}(r_i, r_j) E_{CMDif}(r_i, r_j) \tag{31}$$

The terms in equations 30 and 31 are defined as:

Area

$$E_{area}(r_i) = 0.002 \frac{NM}{Area(r_i)}$$

NM being the image size. Larger area is always preferred. The normalization factor 0.002 means a segment that has a 0.2% area of the whole image will have an energy of 1.0. A zero-area segment has an infinite area energy value and a whole image segment has an area energy value of 0.002. A penalty function is used to prevent to large segments. Typically, segments are penalized when there area is greater than ±25% of the whole image area.

Convexity

$$E_{conv}(r_i) = \frac{Area(ConvexHull(r_i))}{Area(r_i)} - 1.25$$

represents the segment convexity energy. We assume that segments with convexity larger than 1.25 are not preferred, and those with convexity energy smaller than 1.25 are desired. Therefore, the offset for the convexity energy is set to −1.25.

Compactness

$$E_{comp}(r_i) = \frac{Perimeter(r_i)^2}{4\pi Area(r_i)} - 1.25$$

represents the region compactness energy. The compactness energy $\frac{Perimeter(r_i)^2}{4\pi Area(r_i)}$ is always greater than or equal to 1 (1 for a circle, $4/\pi$ for a square). We assume that segments with compactness larger than 1.25 are not preferred. Again the offset for compactness energy is set to −1.25.

Homogeneity

$$E_{hom}(r_i) = 1 - \sigma(r_i) V(r_i)$$

represents the segment's intensity (I) homogeneity. Homogeneity is largely related to the local information extracted from an image and reflects how uniform a region is. In this formulation, we extend to region, the pixel-based homogeneity defined in [20], as the product of the standard deviation, σ, and the discontinuity of intensity I, V.

$$E_{hom}(r_i) = 1 - \sigma(r_i) V(r_i)$$

with

$$\sigma(r_i) = \frac{1}{|r_i|} \Big(\sum_{(x,y) \in r_i} [\mathcal{I}(x,y) - \mu_{\mathcal{I}}(r_i)]^2 \Big)^{1/2}$$

and

$$V(r_i) = \frac{1}{|r_i|} \sum_{(x,y) \in r_i} |\nabla \mathcal{I}(x,y)| \mu_{\mathcal{I}}(r_i)$$

The homogeneity energy of each segment has a range from 0 to 1. The more uniform the segment is, the larger the homogeneity value.

Dynamics of Contour

$$E_{edge}(r_i, r_j) = \frac{1}{(\partial r_i \cap \partial r_j)} \sum_{p \in (\partial r_i \cap \partial r_j)} [GS(p)] - max(M(r_i), M(r_j))$$

reflecting the dynamics of contours; with $(\partial r_i \cap \partial r_j)$ the set of border pixels between r_i and r_j, $GS(p)$ the combined gradient map, and $M(r_k)$ the minima of segment r_k.

Color Variance

$$E_{var_c}(r_i) = \frac{1}{15} \sigma_c(r_i)$$

represents the color homogeneity of a segment, with $\sigma_c(r_i)$ the standard deviation of the color $c \in \{L, a, b\}$ within segment r_i. The normalization factor for color variances is derived from statistical analysis of the color variance on image data base [79].

Color Contrast

$$E_{CMDif}(r_i, r_j) = \sqrt{(\Delta\mu_L(r_i, r_j))^2 + (\Delta\mu_a(r_i, r_j))^2 + (\Delta\mu_b(r_i, r_j))^2} - T_d$$

with T_d as defined in Equation 25.

Using these energy functions as region merging criteria, the saliency driven perceptual grouping process results in the formation of a proto-object or Object Of Interest (OOI).

Figure 29 illustrates our results compared to well-known state-of-the-art methods, including: NVT [59] (Figure 29 (a)), Walther [150] (Figure 29 (b)), and VOCUS [39] (Figure 29 (c)). Figure 29 (e) presents the first extracted OOI, which evolved during the region merging procedure starting form the AR (Figure 29 (d)). In NVT [59] a circular fixed-sized focus is drawn around the selected salient point to form the focused region. In the method of Walther [149, 150], a watershed-based algorithm is applied on the feature map with highest activation, to form the salient region around the most salient point. In VOCUS [39], the most salient region (MSR) is determined by first selecting the most salient point (pixel) in the pixel-based saliency map and, starting from this pixel, apply a seeded region growing approach. This method recursively finds all neighbors with similar values within a certain range (e.g. a difference of 25% from the maximum value is allowed).

The results indicate that, in comparison with the state-of-the-art methods, the proposed approach is able to extract the runner's body as perceptual object, as opposed to a region of space containing (part of) an object of interest. The extracted perceptual object has a well delineated shape, in conformity with human perception. The delineated object represents as such a useful front-end result for contour and shape descriptor determination in higher-level computer vision tasks (e.g. object recognition [65]).

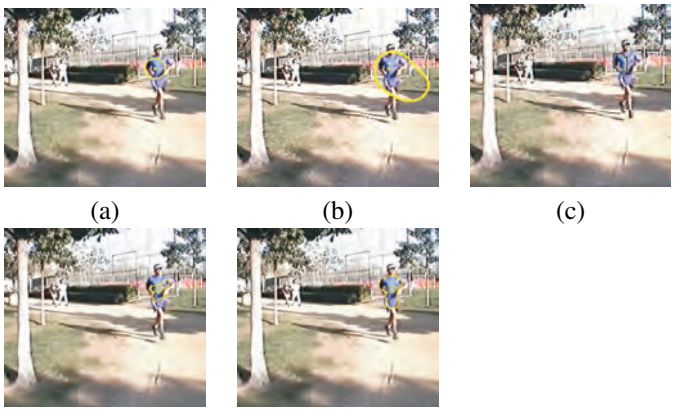


Figure 29: (a) Salient region extracted by NVT [59]. (b) Salient region extracted by the method of Walther [150]. (c) Salient region extracted by VOCUS [39]. (d) Attention region (AR) extraction with proposed approach. (e) Object of interest (OOI), which evolved starting from AR, extraction with proposed approach.

The shape delineation of extracted perceptual objects is further illustrated in Figure 30. Objects of interest are extracted from images obtained from the Berkeley Segmentation Dataset [86, 10]. The Berkeley Segmentation Dataset provides the ground truth segmentation. This dataset contains 12,000 manual segmentations of 1,000 images by 30 human subjects. Each image has been segmented by at least 5 subjects, so the ground truth is defined by a set of human segmentations. Comparison between the extracted OOI's with the proposed model and the ground truth human segmentation confirms that the proposed model is able to provide well delineated perceptual objects.

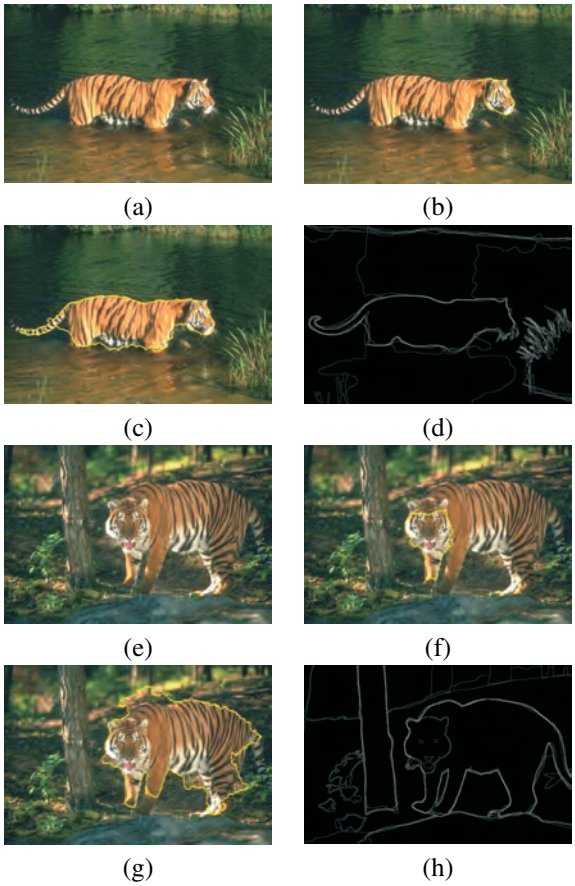


Figure 30: Segmentation results vs. human delineated objects from the Berkeley Segmentation Dataset [86, 10]. (a-e) Original images ((a) Training Image 108073, (e) Test Image 108005). (b-f) Attention region (AR). (c-g) OOI. (d-h) manually delineated shape of perceptual objects.

4.8. Iterative "Object Popping out" Detection & "Inhibition of Return"

Once an OOI is focused on, another computational problem is posed: how can we prevent attention from permanently focusing onto the same OOI? In [57], an efficient computational strategy is proposed, consisting of suppressing the currently attended location in the saliency map. Hence, the winner-take-all network naturally converges towards the next most salient location. Repeating this process generates attentional scan paths. Such inhibitory tagging of recently attended locations has been widely observed in human psychophysics

as a phenomenon called *Inhibition of Return* (IOR) [66]. A typical psychophysical experiment to evaluate IOR consists of performing speeded local pattern discriminations at various locations in the visual field; when a discrimination is performed at a location to which the observer has been previously cued, reaction times are slightly, but significantly, higher than at locations not previously visited [66]. These results indicated that visual processing at recently attended locations might be slower, possibly owing to some inhibitory tagging at attended locations.

Computationally, IOR implements a short-term memory of the previously visited locations and allows the attentional selection mechanism to focus instead on new locations. The simplest implementation of IOR, suppressing the attended location, only represents a coarse approximation of biological IOR, which has been shown to be object bound. What is also important, considering IOR, is the frame of reference in which IOR is expressed. In a dynamic scene, the IOR mechanism should take into account tracking and following of moving objects, eye movements, gaze, overt attention, and camera (body) displacement [9, 136]. This frame of reference problem should be accounted for in designing computational models of attention. For example, a tracking of the most salient region may be required or even a tracking of not only the first but the first n foci of attention, so the behavior should be adapted to the task. The reader is referred to [5] for a detailed discussion.

In the proposed model of object-based visual attention, we are only concerned with covert attention, that is, shifts of the focus of attention in the absence of eye movements. Although simple in principle, IOR is computationally a very important component of attention. Indead, IOR allows us - or a model - to rapidly shift the attentional focus over different items (OOIs) with decreasing saliency, rather than being bound to attend only to the OOI of maximal saliency at any given time. The role of IOR in active vision and overt attention poses challenges that are not addressed in this work.

Although pop-out detection and IOR are two different processes, they are very much interdependent on each other. The IOR greatly influences the process of pop-out by dictating what not to attend in the consequent attention cycle. In general, two types of inhibitions are considered, top-down and bottom-up. The top-down influence is regarded as an external stimulus from outside of the core attention mechanism. This inhibition factor may come from long term knowledge, recent experiences, and current needs. The other type of inhibition occurs within the attention mechanism to avoid repeatedly focusing on the same object. Top-down inhibition is not included in our work.

It has been established by experiments in psychophysics that inhibition takes place in terms of both location and object features [3]. Evidence is provided for inhibition in the immediate vicinity of the attended location and a U-shaped function has been reported which strongly suppresses the immediate surroundings of the attended location and gradually fades to no suppression after a limited diameter. In general, IOR has single influence directly on the resulting saliency map, meaning the saliency map will not change apart from the inhibited focused region ([59, 39]).

In contrast, in this work, we consider two types of inhibition mechanisms, namely spatial based and feature based. Introducing IOR at lower early levels of the computation (feature estimation), as suggested in [46], implies the evolution of the region saliency values of all segments following Equation 32. The spatial inhibition factor ζ_s inhibits the OOI, and prohibits this object to be refocussed on. The feature based inhibition factor ζ_f inhibits the influence of the extracted OOI on its neighboring segments. As such, the feature based inhibition factor influences a segment's saliency value and edge energy.

$$\mathcal{S}'(r_i) = \zeta_s \frac{CSR(r_i, \zeta_f) OC(r_i) SI(r_i)}{CBI(r_i)} \tag{32}$$

$$\zeta_s = \begin{cases} 0; r_i \in OOI \\ 1; otherwise \end{cases} \tag{33}$$

Following the definition of $CSR(r_i)$ in Equation 24, $CSR(r_i, \zeta_f)$ is formulated as:

$$CSR(r_i, \zeta_f) = \sum_{r_j \in \xi(r_i)} \zeta_f \alpha_{i,j} (\sqrt{(\Delta\mu_L(r_i, r_j))^2 + (\Delta\mu_a(r_i, r_j))^2 + (\Delta\mu_b(r_i, r_j))^2} - T_d) \tag{34}$$

$$\zeta_f = \begin{cases} 0; r_j \in \{OOI \cap \xi(r_i)\} \\ 1; otherwise \end{cases} \tag{35}$$

ζ_f excludes the OOI from CSR calculation. From an implementation point of view, this is achieved by simply excluding the set of segments forming the OOI from the RAG, and accordingly updating the RAG.

The human browsing behavior can be approximately modeled by two mutually exclusive statuses: the fixation status (e.g., exploiting an interesting region) and the shifting status (e.g., covertly scrolling to the next region). The fixation status corresponds to the static viewing of an attention object, and the shifting status can be simulated by covertly traveling between different attention objects. The shifting path is the shortest path between the centers of two fixation areas (i.e. objects of interest). Iteration of these two states composes the whole simulation of the shifting process, forming a scan path, a chronological list of attended OOIs. The saccade status can be described as a shifting process from the most informative region to the second one, then the third and so on.

The scan path starts with the pop-out of the most informative object of interest and is formed by subsequently focussing on OOIs in the image. In [157], *minimal perceptible time (MPT)* is introduced as a threshold for the fixation duration when focusing on an OOI. If an attention object does not stay on the

screen longer than MPT, it may not be perceptible enough to let users catch the information. Fixation durations are variable, typically ranging from 100 ms to 500 ms. The MPT of an OOI is proportional to its region saliency value.

Subsequently, the attended OOI is inhibited. The inhibition of return mechanism works as a short term memory, stores attended OOIs in a focus of attention map (FOA map), and maintains the attended area of extracted OOIs in an inhibition map. Taking into account information from the inhibition map, the segment features are updated. Hence, the region saliency map is updated as well as the edge energy values.

Since humans are known to make fixations on nearby objects, the saliency map (saliency value of each region) is weighted by the proximity to the current fixation, defined as

$$w_{saccade,i} = \frac{1}{\sqrt{(gc(OOI)-gc(r_i))^2}}; \quad r_i \neq OOI \tag{36}$$

with $gc(OOI)$ the gravitational center of the selected Object of Interest; $gc(r_i)$ the gravitational center of all remaining segment. The weighted saliency map is then calculated as

$$\mathcal{S}^w(r_i) = w_{saccade,i}\mathcal{S}(r_i) \tag{37}$$

with $\mathcal{S}(r_i)$ as defined in Equation 23, or 32. The scan path determination process continues with a new AR: the newly highest salient region excluding the already detected OOI's. The scan path is complete when (a part of) the background is selected, or when there are no more meaningful regions left.

5. Experiments and Evaluation

5.1. Illustration of Scan Path Determination

Before more extensively evaluating the proposed model, we start with a first experiment on the previously used hats test image of Figure 11, to completely illustrate the process of determining the scan path, by subsequently focussing our attention on Objects of Interest. In this test image a set of hats are well defined as objects of attention.

In Figure 31 the results of the scan path determination are presented. Figures 31 (a) and (b) depict the original image and the segmented image, respectively. Figure 31 (c) illustrates the AR, in the current case a part of the yellow hat. The rest of the figures illustrate the scan path determination process. For each Object of Interest that has gained the focus of attention, the following resulting maps are presented:

- Mosaic Image, corresponding to the iterative perceptual organization merging (hierarchical level).
- Saliency Map, displaying the saliency value (normalized between [0-255]) of each region at the moment when the scan path is updated.
- Focus of Attention (FOA) Map, representing the scan path. The current Object of Interest is added to the focus of attention map, and highlighted by means of white edges and the index in the scan sequence of Objects of Interest.
- Inhibition of Return (IOR) Map. Whenever an Object of Interest is added to the scan path, the IOR Map is updated after extraction of the OOI. More concretely, the first OOI extracted in step 1 of the scan path development, is inhibited during the subsequent scan path augmentation, and stays inhibited until the scan path is complete.

Figure 32 presents the obtained scan paths on natural traffic scene images, often used for illustrating focus of attention [2].

5.2. Performance Evaluation

Computational bottom-up model performance can not be readily assessed and there is no real consensus on any assessment method. The qualitative and quantitative evaluation of a pure bottom-up system of visual attention is difficult because of the lack of ground truth data. We do not know what the "correct" focus of attention is on a natural scene. For humans, the focus of attention will vary depending on certain motivations, ideas, etc, on top of the existing bottom-up contrast information.

Nevertheless, some objective methods have been proposed. They depend both on the type of assessed images (synthetic or natural) and the knowledge of the ground truth coming from eye tracking experiments. The data from eye tracking experiments is then transformed into a fixation density map, a similar format as the saliency map, and as such quantitative comparison is possible. In the literature, two ways to conduct

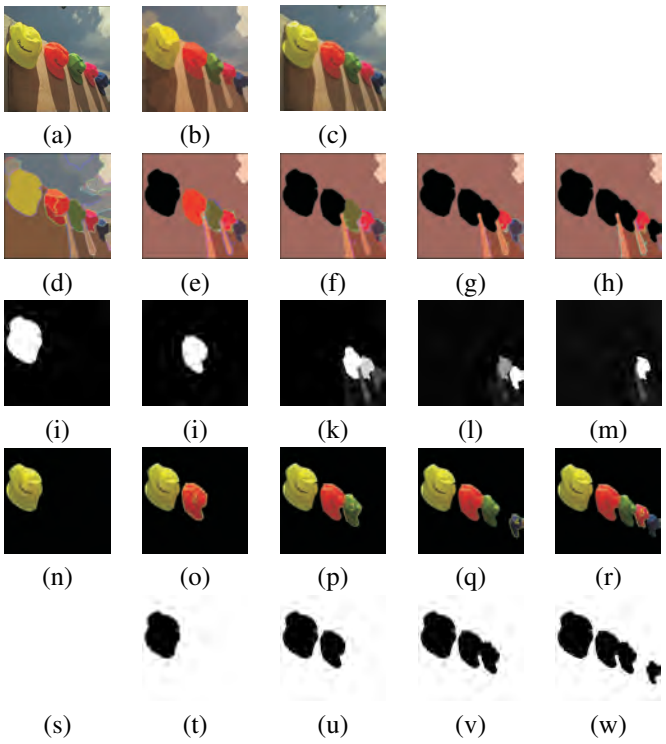


Figure 31: Hats image scan path determination. (a) Original hats image. (b) Starting mosaic image. (c) AR selection. For each Object of Interest (OOI) that has gained the focus of attention, the corresponding Mosaic Image (d-h), Saliency Map (i-m), Focus of Attention (FOA) Map (n-r), and Inhibition of Return (IOR) Map (s-w) are presented.

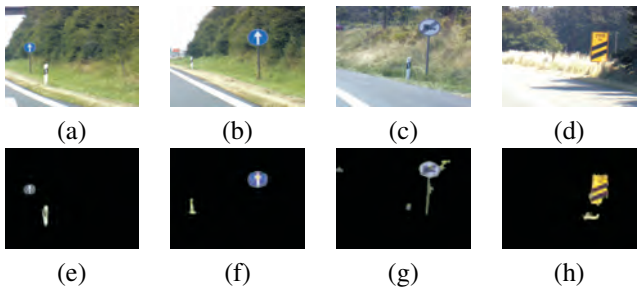


Figure 32: Scan path determination on natural traffic scene images. (a-d) Original images. (e-h) Extracted scan path.

the objective assessment are proposed [107, 156, 13, 114, 84]. The first one consists in comparing the first fixations of the scan paths whereas the second compares two fixation density maps.

Qualitative evaluation provides an insight into the effectiveness of the proposed model. In conformity with [4], [39] we enlist several evaluation methods:

- Comparison with other attention systems.
- Repeatability of the results under image transformations.
- Comparison with human perception.

In the following sections, the listed evaluation methods will be discussed from both qualitative and quantitative point of view. The performance of the proposed model is merely evaluated qualitatively. In addition, we give indications on how to further evaluate quantitatively the proposed model, making extensively use of data coming from eye tracking experiments.

Comparison with Other Attention Systems

Comparison with other systems does not enable to extensively evaluate the quality of the proposed attention system. The most difficult aspect in such a comparison is that in most examples it is not possible to say which results are better, since no ground truth is available from eye tracking experiments on the same data. The objective of the qualitative comparison with other attention systems is to analyze the differences between the models. We tend to show the added value of our object-directed visual attention model, aiming at the extraction of well delineated, meaningful perceptual objects, as opposed to regions of interest in the image.

The comparison with state of the art focus of attention models is performed in two steps. First, we compare our proposed approach for object of interest scan path determination with the famous Neuromorphic Vision Toolkit (NVT) [59], as well as the bottom-up model VOCUS [39], from which the feature extraction phase of our model is inspired. Second, we compare our presented approach with the publicly available Saliency Toolbox [150]. This system was chosen for comparison since it is well known, and it produces comparable results in terms of scan path determination of proto-objects. To take fair conditions for the experiments, we consider two experimental data sets. In the first, images of natural traffic scenes were collected from Itti's data (downloaded from [2]). For the second comparison, an image set from the Saliency Toolbox [150] is selected.

In [150] a biologically plausible model for generating and attending to proto-object regions is described, as a preliminary stage for the ultimate goal of object recognition. The attention system is based on the well known Neuromorphic Vision Toolkit (NVT) [59] implementation of the saliency map-based model of bottom-up attention by Koch and Ullman [67], which models selective attention to salient locations in a given image. This model was extended with a process of inferring the extent of a proto-object at the attended location from the

feature conspicuity maps that are used to compute the saliency map. For a selected salient location in the saliency map, the feature map with the highest contribution to the saliency of this attended location is set. On this map, the approximate extent of the proto-object at that particular location is determined as a contiguous region of high activity in that feature map by means of a flooding algorithm with adaptive thresholding.

In Figures 33 and 34 we present the results of the first test set. The results of the second test set are depicted in Figure 35. Parameters for the experiment, illustrated in Figure 35, using the Saliency Toolbox are: equal weights for all features (color, intensities, 4 orientations), lowest surround level 3, highest surround level 5, smallest c-s delta 3, largest c-s delta 4, saliency map level 3, iterative normalization with 3 iterations.

The scan path determined with the considered methods contains more or less the same salient locations/objects. However, NVT and VOCUS (Figures 33 and 34) focus on locations rather than objects. The scan path determined with the Saliency Toolbox (Figure 35) selects the most salient regions iteratively, omitting consideration of the real object borders. In contrast, the proposed approach detects well the relevant areas, similar to NVT and VOCUS and Saliency Toolbox, respectively, and additionally, extracts well delineated Objects of Interest. As such, with the proposed model contour and shape descriptors can be easily estimated, without the need for extra processing stages. This characteristic, inherent to our object-oriented model of visual attention, provides a clear advantage compared to the other models of visual attention.

Image Transformations

Evaluation of a system with respect to the repeatability of the results under geometric image transformations like rotation, translation, scale, variation of illumination, and 3D viewpoint, are interesting and desired qualities for robots acting in a dynamic environment, or for attentional systems that serve the purpose of front ends for object recognition, as shown by Draper and Lionelle ([29]).

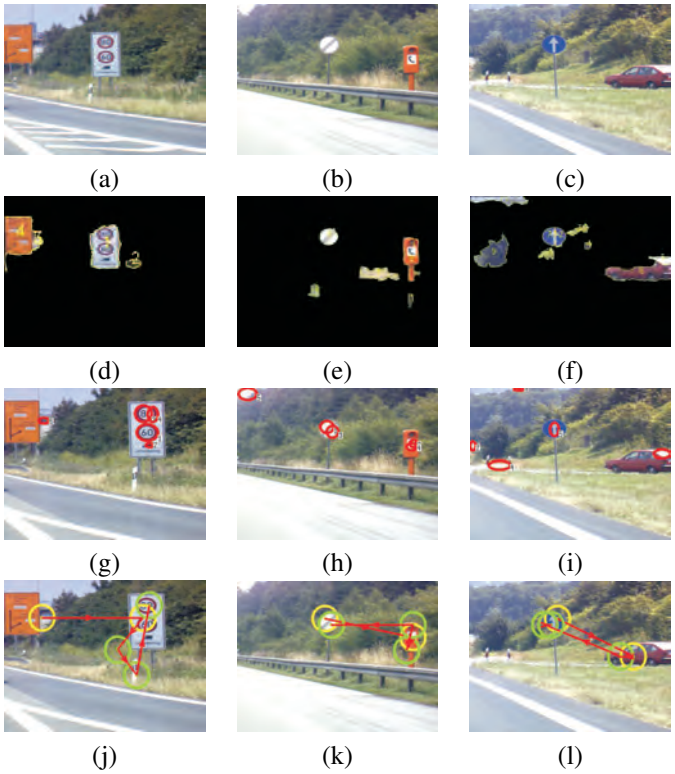


Figure 33: Comparison of the proposed scan path determination with VOCUS [39] and NVT [59] on an image test set of iLab([2]). (a-c), the original image. (d-f), the scan path composed of the extracted OOIs with the proposed approach. (g-i), the scan path of VOCUS as red ellipses. (j-l), the scan path of NVT as yellow (and green) circles.

However, even in human perception, there is no invariance concerning scan-paths: there is the tendency of humans to scan a scene in reading direction, i.e., in Europe from upper left to lower right. This yields different scan paths for rotated images. Furthermore, humans tend to prefer the center of a scene, yielding variances on translated images. In [70] the bias towards the central part of images is illustrated by showing the spatial distribution and the density of human fixations, captured by tracking eye movements. The tendency of observers to stare at the central locations of the screen is not reduced with the viewing time. It can be shown that observers continue to focus on these areas rather than to scan the whole picture. There are at least two plausible explanations. The nonuniform distribution of photoreceptors is a biological candidate. However, it seems more logical to tackle this question by introducing a top-down or higher-level explanation as proposed in [107]. The great majority of visually important information is traditionally located in the central part of the picture frame. Consequently, observers

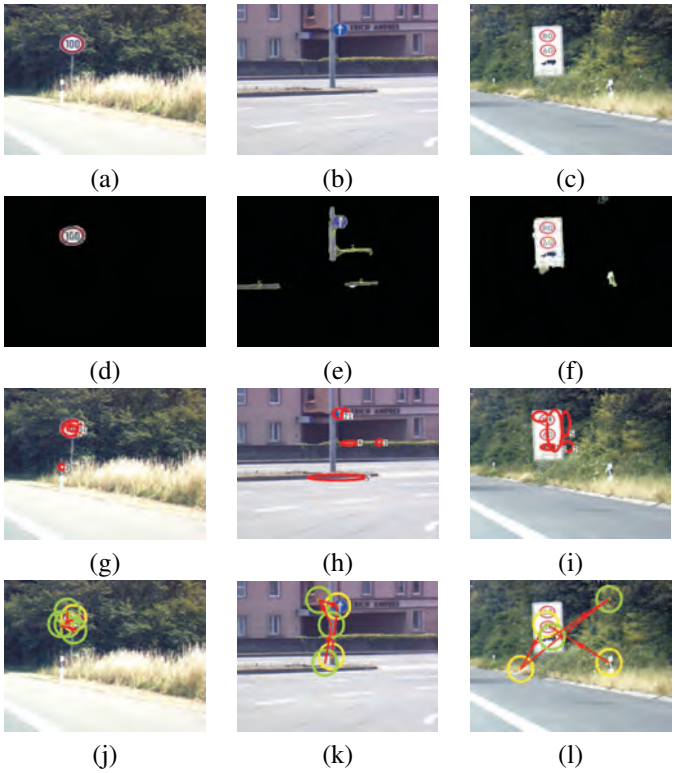


Figure 34: Comparison of the proposed scan path determination with VOCUS [39] and NVT [59] on an image test set of iLab([2]). (a-c), the original image. (d-f), the scan path composed of the extracted OOIs with the proposed approach. (g-i), the scan path of VOCUS as red ellipses. (j-l), the scan path of NVT as yellow (and green) circles.

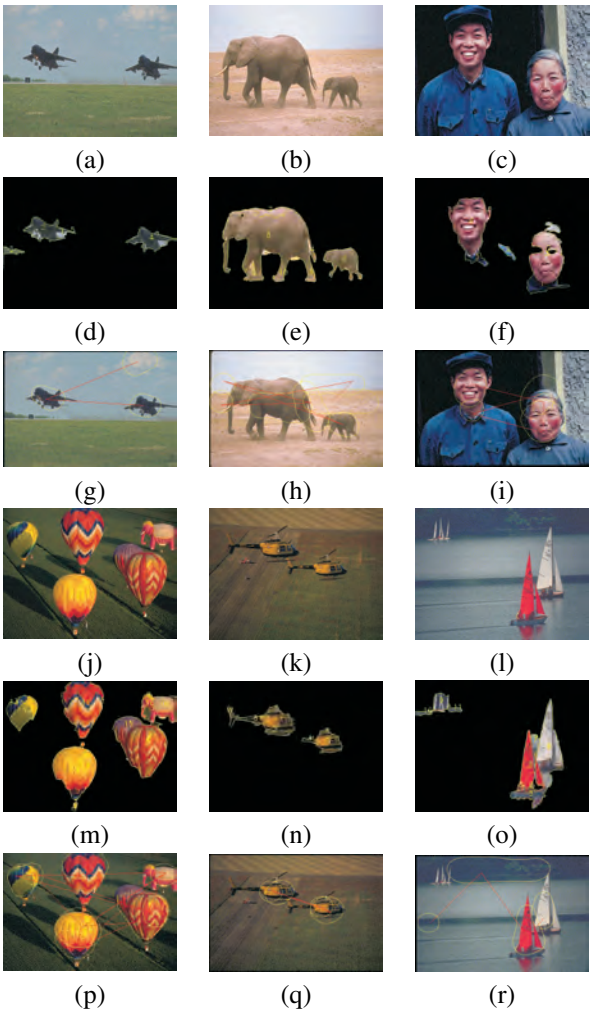


Figure 35: Comparison of the proposed scan path determination with the Saliency Toolbox [150]. (a-c, j-l), the original image. (d-f, m-o), the scan path composed of the extracted OOIs with the proposed approach. (g-i, p-r), the scan path obtained with the Saliency Toolbox.

unconsciously tend to select central locations in order to catch the potentially most important information.

Even more difficulties arise under illumination variations as well as changing of scale and 3D viewpoint. Different illuminations may highlight other parts of a scene; zooming a scene draws attention to larger object or groups of objects. Therefore, we conclude that for our biologically motivated object-oriented attentional selection system, this evaluation method will not add qualitative value.

Comparison with Human Perception

The comparison with human perception means comparing the selected and focused objects of interest (OOIs) with human eye movements and fixations. While this seems an easy and straightforward task, the contrary is certainly true. Usually, there is no general scan path of human eye movements on a scene. In [84], it is shown on examples of complex natural scenes, that there is no evidence for repetitive scan paths. Most scenes contain many objects competing for saliency and it is not clear in which order they are focused. Each individual human focuses on different parts of a scene according to top-down influences such as: preferences, emotions, and motivations. This makes the objective and quantitative comparison with human perception very difficult.

In this section we will qualitatively assess the proposed model by comparison with human perception. In [39], two approaches are suggested for the comparison with human perception. The first method is named "subjective analysis", referring to the subjective decision on whether or not the focused objects of interest make sense. This method is intuitive, and easy to perform, but also subjective and scientifically not sound. Despite these drawbacks, the method gives a good first impression on how the system performs.

The second method is to compare the system behavior with general human observation behavior on psychological data. This method is more objective and scientifically sound. In the following, we show several experiments concerning the two approaches.

Subjective Analysis

Usually, in subjective analysis some real-world images are provided to the system and the evaluation of whether the results are sensible is left to the user. Although an objective evaluation is hardly possible with this method for most real-world scenes, there are special scenes in which it is intuitively clear which regions are most salient. Such scenes are for example those which contain objects that were explicitly designed to attract attention. Many of such objects are found in traffic scenes: traffic signs,traffic lights, brake lights, and signaling lights are all designed with strong colors, strong intensities, or flashing lights. Also other security relevant objects are designed salient, e.g., fire extinguishers, emergency exit signs, and police sirens. More examples are found in sports: the balls in many games are designed with colors that distinguish them well from their background. In Figure 36 a limited set of images and corresponding objects of interest are presented. Only the first selected object of interest is highlighted and corresponds in each image to an intuitively salient (part of) object, explicitly designed to attract attention.

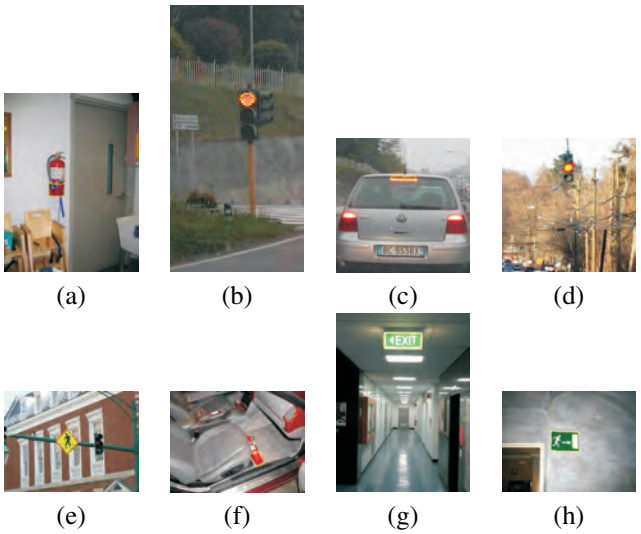


Figure 36: Foci of attention (FOAs) depicted as (yellow or red) circumvented regions (OOIs) on objects that were explicitly designed to attract attention, e.g. traffic situations, security relevant areas.

Comparison with Psychophysical Data

Although a general scan path on complex scenes for multiple observers is not completely congruent, there is a general viewing behavior on simple synthetic scenes, mainly pop-out scenes in which one item differs in one feature from all the other regions in the scene. Figure 37 presents some of these more artificial pop-out images, often used in psychological experiments on visual search [97, 155**?**]. In psychophysical experiments, the scene for a visual search task is usually an artificial composition of several items with different features. The *pop-out* item is that particular item that differs in terms of one or more of the features with respect to the other items. These other items are called *distractors*. These particular images have been processed with the proposed model of visual attention, and as shown in Figure 37 the pop-out objects are selected as objects of interest, well delineated in terms of shape.

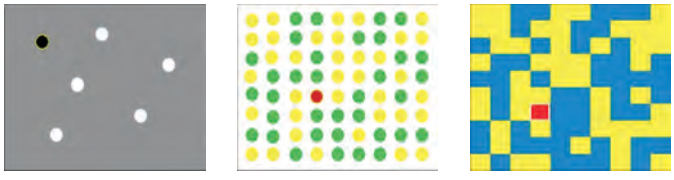


Figure 37: FOAs depicted as circumvented regions (OOIs) on some pop-out data.

One of the advantages of working with these artificial images, is the possibility to gradually change feature values and set sizes. In [39], an interesting test is executed on pop-out

scenes. Not only the uniqueness of a feature, but also the strength of the feature value has an influence on the pop-out effect: in human vision as well, a feature has to have a certain strength to pop out of a scene. An oriented bar has to differ by certain amount of degrees from its distractors. In [39] these limits have been investigated in two different cases. We will not reproduce these tests.

5.3. Indications for Quantitative Evaluation

The proposed model of visual attention is purely data-driven and no prior knowledge about the scene is given. Inspired by psychological theories (section 3), and in spite of the qualitative and subjective analysis of the previous section, we may conclude that the theoretical grounding of the saliency-based model has not been assessed extensively in quantitative terms. In order to do so, we need to examine the plausibility of the model of object-based visual attention by comparing its performance with human behavior. Performance is estimated by comparing the system behavior with human observation behavior obtained from viewing experiments.

The viewing behavior of an observer is measured by tracking the eye movements (saccades) and fixation points with an eye tracker. The availability of eye tracking technologies allow for investigations towards the relation between visual attention (covert attention) and eye movements (foveated, overt attention). As suggested in [49, 48, 114], visual attention guides eye movements (saccades) in order to place the fovea on the interesting parts of the scene. The foveated part of the scene can then be processed in more detail. In other words, saccades to a location in space or on an object of interest are preceded by a shift of visual attention to that location.

The claimed objectivity in this type of experiments is not completely conclusive. Problems range from the unnatural situation that participants to the experiment are subjected to unreported afflictions of their eyes, to time pressure while calibrating, lacking expertise to the bias in selecting data, constraints like monitor size, and a limited measuring capability of an actually complex system [71]. Still eye movement registration with an eye tracker is a brilliant and insightful technique and a large portion of the fore mentioned problems can be soothed out by simple measures like experimenter training and good equipment training.

In this section, we propose a methodology to quantitatively evaluating further the proposed model of visual attention. An experiment involves several main stages to compare human performance and model prediction. These stages include: selection of data for viewing experiment, process the selected stimuli with the proposed model of visual attention, present selected stimuli to human observers and track the eye movements, compare the behavioral responses of the model and human. As such, the plausibility of the proposed computational model can be quantitatively assessed by comparing its performance with human behavior. The basic idea is to compare the map of attention, the saliency map, produced by the computational model with a fixation density map derived from eye movement experiments. Practically, a computational saliency map is computed for a given color input image. The same color image is presented to human subjects whose eye movements are recorded, providing information about the spatial location of the sequentially foveated image parts and the duration of each fixation.

Data selection

The selection of visual stimuli as input data for viewing experiments is very important. It is stated that no general scan path on complex scenes exists, however, on simple scenes, a general viewing behavior exists. Although there is no general scan path on natural images, there are accumulation points at which humans look more often than at others. Important in this context is to measure the dissimilarity between between the participating observers' fixation behavior. On simple and artificial images, this dissimilarity is expected to be much lower, than on complex natural images.

Human fixation density map computation

Typically, during an experiment, each image is presented in random order to all observers for a limited time period (e.g. 5 seconds in [103]) and the instruction given to observers is "just look at the image".

The fixation behavior of participating human observers is measured by tracking and recording eye movements using an eye tracker. Eye tracker data typically consists of the gaze directions and their durations. Basically, the position of the pupil (and cornea) is measured at a certain frequency (250-2000Hz) and is transformed into a corresponding gaze position on the screen through calibration. The human eye has properties that often are removed in the interpretation of the measurements such as flutter, inertia, etc. Comprehensive models that combine these properties and their modeling with stages of perception are rare. From the gaze directions and their durations different derivatives might be computed in order to compare human observation behavior with the responses of the proposed model of visual attention.

In [103], under the assumption that attention is guided by the saliency of the different scene parts, the recorded fixation location serve to plot a human saliency distribution map or human attention map. The human attention map is derived as an integral of single impulses located at the positions of successive fixation points. Practically, each fixated location gives rise to a grey-scale patch whose activity is normally (Gaussian) distributed. The width (σ) of the Gaussian patch should approximate the size of the fovea. A parameter α is introduced that tunes the contribution of the fixation duration to the Gaussian amplitude. If $\alpha = 0$, the amplitude is the same for all fixations regardless of their duration. However, if $\alpha = 1$, the amplitude is proportional to the fixation duration. The human attention map $\mathcal{M}_h(x, y)$ is computed as:

$$\mathcal{M}_h(x,y) = \sum_{(k,l,t)\in F_i}(\alpha.t + (1-\alpha)).exp(-\frac{(k-x)^2+(l-y)^2}{\sigma^2}) \quad (38)$$

where $F_i(k, l, t)$ are the set of measured eye fixations ((k, l) are spatial coordinates and t is the duration of the fixation), σ is

the standard deviation of the Gaussian patch, and α tunes the contribution of fixation duration to the saliency.

In [70], the human fixation density map is computed from the collected data. For a particular picture and for each observer, the samples corresponding to saccades are filtered out. A data point is removed if the number of data included in a squared window is below a given threshold. The size of the window and the threshold are functions of the viewing distance, the accuracy of the eye tracker (0.25 degrees of visual angle) and the resolution of the display (800×600 pixels). In practice, the size of the window and the threshold are, respectively, 9×9 (corresponding to 0.25 degrees of visual angle) and 5 (corresponding to the number of data required in the previous defined window). All fixation patterns for a given picture are added together providing a spatial distribution of human fixation. The resulting map is then smoothed using a two-dimensional Gaussian filter. Its standard deviation is determined according to the accuracy of the eye tracking apparatus. The result is a fixation density map [156] which represents the observers regions of interest (RoI).

Prediction of OOI's with proposed model

The proposed model of visual attention, as most computer vision models, is inspired by psychological hypotheses. Experimental cognitive psychology often relies on hypothetical stage models so that separate experiments can be designed for each conceptual stage. The combination of studies in order to build encompassing models has to deal with incompatibilities which are mainly due to the conceptualization of the human brain. Computer vision models often abstract further away from this picture by merely applying useful techniques which are rather distantly related to biological or functional reality.

Successive runs of the model on the same static input images result in the same responses. On the other hand, data obtained from monitoring observation experiment participants will vary during multiple observations of the same image. Repeated application of stimuli to different participants gives an optimal statistical control of these error sources. The sources of statistical error with human participants are numerous and most likely non-linear. Learning may occur, participants may be distracted, slightly change position, etc. In psychology, all these sources of variation are not modeled, but instead statistically controlled. Therefore, the analogy between presenting stimuli to human participants, and feeding these stimuli to feature extraction stages of visual attention models is a double abstraction.

Objective comparison between model and human behavior

In the final phase of the experiment, the behavioral responses of the model and human are compared and a measure for correlation is obtained. Several issues remain concerning the comparison. First of all, in order to compare the human fixation density map, and the predicted Objects of Interests, the format of the map must be equal. Therefore, the extracted OOI's must be transformed in one predicted saliency map. This might be achieved for example by taking the center of an extracted Object of Interest as predicted fixation point. Also, both the human fixation density map and the predicted saliency map must take into account the order of fixation points. Third, the fixation points or locations in the image must be related to our proposed enhanced feature extraction, which incorporates region information.

Despite these important points of discussion, and under the assumption these issues are solved in a plausible manner, we overview some quantitative evaluation methods [103, 69, 70]. Two objective metrics might be used for the quantitative assessment: the linear correlation coefficient and the Kullback-Leibler divergence. Using these objective metrics, further comparison with other state of the art models of visual attention might be conducted.

The Linear Correlation Coefficient. The linear correlation coefficient ρ, given by equation 39, is widely used to compare two images for applications such as image registration, object recognition, and disparity measurement. The linear correlation coefficient measures the strength of a linear relationship between two variables. It has some interesting advantages. The first one is its capacity to compare two variables by providing a single scalar value. The correlation coefficient has a value between -1 and +1. When the correlation is close to +/-1, there is an almost perfectly linear relationship between the two variables. Let $\mathcal{M}_h(x,y)$ and $\mathcal{M}_c(x,y)$ be the human and the computational maps respectively. The correlation coefficient of the two maps is computed according the following equation:

$$\begin{aligned}\rho &= \frac{cov(\mathcal{M}_h(x,y),\mathcal{M}_c(x,y))}{\sigma_{\mathcal{M}_h(x,y)}.\sigma_{\mathcal{M}_c(x,y)}} \\ &= \frac{\sum_{(x,y)}[(\mathcal{M}_h(x,y)-\mu_h).(\mathcal{M}_c(x,y)-\mu_c)]}{\sqrt{\sum_{(x,y)}(\mathcal{M}_h(x,y)-\mu_h)^2.\sum_{(x,y)}(\mathcal{M}_c(x,y)-\mu_c)^2}}\end{aligned} \tag{39}$$

where $cov(\mathcal{M}_h(x,y),\mathcal{M}_c(x,y))$ is the covariance value between $\mathcal{M}_h(x,y)$ and $\mathcal{M}_c(x,y)$. μ_h and μ_c are the mean values, and σ_h and σ_c are the standard deviations of the two maps $\mathcal{M}_h(x,y)$ and $\mathcal{M}_c(x,y)$ respectively.

The Kullback-Leibler Divergence. The Kullback-Leibler divergence is used to compute the degree of dissimilarity between two probability density functions. Two probability density functions are deduced from the human saliency maps and the predicted saliency maps. The Kullback-Leibler divergence, noted KL, is given by:

$$KL(\mathcal{M}_h(x,y)|\mathcal{M}_c(x,y)) = \sum_{(x,y)} \mathcal{M}_c(x,y).Log(\frac{\mathcal{M}_c(x,y)}{\mathcal{M}_h(x,y)}). \tag{40}$$

When the two probability densities are strictly equal, the KL value is zero.

Another way to objectively evaluate the performances of the model consists in computing the average dissimilarity over all the observers. This could be obtained by computing the Kullback-Leibler divergence between the probability density

function for one observer and the probability density function obtained for all participants. This computation is iterated over the set of observers. The average of the Kullback-Leibler values, called KL_{avg} is given in equation 41. The behavior of an average observer can then be identified: a high KL_{avg} value means that the visual strategy of all observers is different. In others words, the dispersion inter-observers is high. A weak value means that the visual strategy of all observers is similar. The minimum value is zero and will be obtained only if all observers stare at the same locations during the same amount of time.

$$KL_{avg} = \frac{1}{N} \sum_{i} KL(\mathcal{M}_{h,i}(x,y)|\mathcal{M}_h(x,y)). \quad (41)$$

with $\mathcal{M}_h(x,y)$ the global probability density function from the data for all participants, $\mathcal{M}_{h,i}(x,y)$ the probability density function for the i observer, N the number of observers.

It is also interesting to compare the divergence value, noted $KL(\mathcal{M}_c(x,y)|\mathcal{M}_h(x,y))$ with the KL_{avg} value. $KL(\mathcal{M}_c(x,y)|\mathcal{M}_h(x,y))$ is computed from the predicted probability density and the global probability density. Three cases can be considered:

- $KL(\mathcal{M}_c(x,y)|\mathcal{M}_h(x,y)) \approx KL_{avg}$ When the two values are similar, there is a good pairing between the predicted density functions and the set of density functions obtained for each observer.
- $KL(\mathcal{M}_c(x,y)|\mathcal{M}_h(x,y)) < KL_{avg}$ When the value associated to the prediction is smaller than the KL_{avg} value, the most important part of the predicted density is well paired with the set of density functions obtained for each observer. In others words, the most conspicuous areas of the picture are well predicted. The predicted saliency map is almost fully included in the saliency map produced by the observers.
- $KL(\mathcal{M}_c(x,y)|\mathcal{M}_h(x,y)) > KL_{avg}$ When the value is greater than the KL_{avg} value, there is a weak pairing between the predicted density and the set of density functions obtained for each observer. Differences stem from the spatial locations of the most important areas in the two density functions. There are major dissimilarities between the two sets.

5.3.1. Experiment

This section reports some experimental results that illustrate the different steps of our proposed empirical quantitative validation of the model for attentional selection of objects of interest. The basic idea is to compare predicted computational and human fixation maps obtained from the same color images according to objective criteria. The human fixation map is gained from an eye tracking and recording experiment with 20 subjects between 20 and 35 years, with equal number of females and males. All of them have normal or corrected-to-normal visual abilities and normal color vision. The images were presented to the subjects with a resolution of 800×600 pixels on a 19" monitor, placed at a distance of 70cm from the subject. The view-time was 5 seconds, under the free-viewing task. The human fixation map is computed using Equation 38 with following parameter values:

- $\sigma = 37.0$
- $\alpha = 0.5$. That means that the importance of all fixations is influenced by their duration.

The predicted map is calculated from the Focus of Attention Map which represents the predicted scan-path of Objects of Interest, following the approach described in section 4. Each Object of Interest is characterized by its center of mass, and its contour, a sequence of contour points. Using Equation 38 (with $\sigma = 37.0$, and $\alpha = 0.0$), the predicted fixation is obtained in a comparable format with respect to the human fixation map.

Figure 38 depicts the selected original image. Figure 39 illustrates the recorded human fixations of "viewer 1", "viewer 7", and "all viewers", respectively. The size of a fixation circle is relative to the fixation duration. Subsequently, Figure 40 presents the human fixation map.

Figure 38: Empirical quantitative validation of the approach for attentional selection of objects of interest: original image.

Figure 39: Empirical quantitative validation of the approach for attentional selection of objects of interest: recorded human fixations.

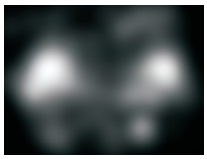


Figure 40: Empirical quantitative validation of the approach for attentional selection of objects of interest: human fixation map.

Figure 41 presents the obtained OOI's with delineated contour and masked surface on the original image, respectively. Figure 42 depicts the predicted computational fixation map. Additionally, Figure 43 visually illustrates the correlation between the computational obtained OOI's and the calculated human fixation map, by overlaying. Applying Equation 39, results in a correlation coefficient of 0.67. To summarize, the results

produced by the computational model of attentional selection and the human behavior exhibit a promising correlation for the selected image. The reduced number of subjects and the single test image do not allow to draw final conclusions. Further experiments are needed to investigate correlation between computational and human attention, in scenes where the amount of top-down information present, is varied.

Figure 41: Empirical quantitative validation of the approach for attentional selection of objects of interest: obtained OOI's.

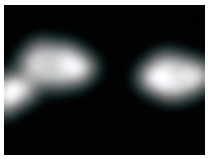


Figure 42: Empirical quantitative validation of the approach for attentional selection of objects of interest: predicted fixation map.

Figure 43: Empirical quantitative validation of the approach for attentional selection of objects of interest: overlaying the human fixation map on the delineated OOI's.

5.4. Segmentation Evaluation Criteria

In most hierarchical segmentation schemes, the desired degree (level) of segmentation remains interactive [6, 18]. In other words, the user may select visually and qualitatively, the hierarchical level at which the resulting segmentation is acceptable. The segmentation quality for each level of the hierarchical tree can also be quantified via an evaluation criterion, allowing for automatic hierarchical level selection. Segmentation quality measures are in general based on (i) intra-region uniformity, (ii) inter-region contrast and (iii) region shape (boundaries should be smooth and accurate). These characteristics serve as basis to design goodness measures for satisfying the human intuition on an "ideal" segmentation. In this section we shortly present multiple segmentation quality evaluation criteria.

Liu and Yang [76] suggested a segmentation evaluation function $\mathcal{LY}(.)$ which expresses the trade-off between the suppression of heterogeneity and preservation of details. $\mathcal{LY}(.)$ uses Euclidian distance in RGB color space for color variance estimation. In [144] a generalization of the Liu and Yang evaluation criterion $\mathcal{GLY}(.)$ is proposed, by taking into consideration different metrics for color distance estimation in any selected color space.

The segmentation evaluation functions $\mathcal{LY}(.)$ and $\mathcal{GLY}(.)$ do not incorporate directly the quality measures (ii) and (iii). In [144] an evaluation function $\mathcal{CH}(.)$ is proposed, which combines a region homogeneity measure H_i (local color error in region r_i) with an inter-region contrast measure C_i (color difference between adjacent regions). Similarly, in [60], the proposed evaluation criterion for a hierarchical level (k), denoted by $\mathcal{CH}^{(k)}$, is based on the intra-segment color homogeneity, and the inter-segment separability.

[159] proposes an objective segmentation evaluation method based on information theory. The method uses entropy as the basis both for measuring the uniformity of pixel characteristics (luminance) within a segmentation region, and for measuring the complexity of the division of the image into regions.

In [35] the segmentation quality is evaluated by an objective function that considers two requirements for the segmentation result: each of the resulting segments should be internally homogenous and should be distinguishable from its neighborhood. The function combines a spatial autocorrelation index, which detects separability between regions, with a variance indicator, which expresses the overall homogeneity of the regions.

6. Conclusion

In this Chapter, we have introduced, implemented, and qualitatively evaluated a novel biologically inspired region-based focus of attention mechanism simulating the middle stages of attention, with specific algorithmic details. From the viewpoint of modeling object-based visual attention, our approach, for proto-object detection and extraction, uses an innovative saliency driven perceptual grouping process, extending the pixel-based saliency map to salient groups/objects. Proto-objects are defined as blobs of uniform color in the image. The proposed method segments an image into regions (proto-regions) as an attentive process, during which only visually salient image regions are merged using perceptual organization criteria. At the same time, an attention region (AR) is selected from the region saliency map. A hierarchical perceptual grouping, involving multi-scale concepts, is used to select the salient (proto-)segments, which are then clustered into the proto-object, named Object Of Interest (OOI), using a new region merging criteria. Unlike other algorithms, the proposed method allows multiple OOIs to be segmented according to the saliency map.

The presented computational attentional systems inherits the main advantages of the human attention system: it is generally applicable to every scene, including artificial scenes (e.g. graphical displays used in perceptual psychology), natural scenes, indoor and outdoor images, office environments and traffic scenes. The proposed approach gives excellent results in extracting meaningful, perceptual objects when objects have limited internal color differences. However, when large color differences occur between object parts, which are not absorbed by the texture energy estimation, objects are selected in several steps.

The main contributions of the proposed model can be summarized in the following 3 aspects. Firstly, (proto-)region is chosen as the perceptive unit, which makes the method more effective in terms of perception. Secondly, compared with traditional attention models our model provides saliency maps with meaningful region information, by eliminating misleading high-contrast edges. Finally using both global effect and contextual difference the proposed focus of attention shifts in units of perceptual objects instead of spatial regions.

References

[1] 01, U., ???? ilab at the university of southern california. http://ilab.usc.edu.

[2] 02, U., ???? ilab image databases. http://ilab.usc.edu/imgdbs.

[3] Aziz, Z., Mertsching, B., 2007. Pop-out and ior in static scenes with region based visual attention. In: ICVS Workshop on Computational Attention & Applications - WCAA 2007.

[4] Backer, G., 2004. Modellierung visueller aufmerksamkeit im computersehen: Ein zweistufiges selektionsmodell für ein aktives sehsystem. Ph.D. thesis, Universität Hamburg, Germany.

[5] Backer, G., Mertsching, B., Bollmann, M., 2001. Data- and model-driven gaze control for an active-vision system. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 23 (12), 1415 – 1429.

[6] Bagon, S., Boiman, O., Irani, M., 2008. What is a good image segment? a unified approach to segment extraction. In: Forsyth, D., Torr, P., Zisserman, A. (Eds.), Computer Vision – ECCV 2008. Vol. 5305 of LNCS. Springer, pp. 30–44.

[7] Baylis, G., 1994. Visual attention and objects: two-object cost with equal convexity. Journal of Experimental Psychology: Human Perception and Performance 20, 208–212.

[8] Baylis, G., Driver, J., 1993. Visual attention and objects: evidence for hierarchical coding of location. Journal of Experimental Psychology: Human Perception and Performance 19, 451–470.

[9] Becker, L., Egeth, H., 2000. Mixed reference frames for dynamic inhibition of return. J. Exp. Psychol. Hum. Percept. Perform. 26, 1167–1177.

[10] Berkeley Segmentation Dataset, ., ???? http://www.cs.berkeley.edu/projects/vision/bsds.

[11] Beucher, S., Sept. 1994. Watershed, hierarchical segmentation and waterfall algorithm. In: Serra, J., Soille, P. (Eds.), Proc. Mathematical Morphology and its Applications to Image Processing. Kluwer Ac. Publ., Nld, pp. 69–76.

[12] Bister, M., Cornelis, J., Rosenfeld, A., 1990. A critical view on pyramid segmentation algorithms. Pattern Recognition Letters 11, 605–617, best paper award 1990-1991.

[13] Brandt, S. A., Stark, L. W., 1997. Spontaneous eye movements during visual imagery reflect the content of the visual scene. J. Cognitive Neuroscience 9, 27–38.

[14] Bundesen, C., 1998. A computational theory of visual attention. Phil. Trans. R. Soc. Lond. B 353, 1271–1281.

[15] Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8 (6), 679–698.

[16] Cave, K. R., 1999. The featuregate model of visual selection. Psychological Research 62, 182–194.

[17] Cave, K. R., Wolfe, J. M., 1990. Modeling the role of parallel processing in visual search. Cognitive Psychology 22 (2), 225–271.

[18] Chabrier, S., Emile, B., Rosenberger, C., Laurent, H., 2006. Unsupervised performance evaluation of image segmentation. EURASIP Journal on Applied Signal Processing 2006, 1–12.

[19] Chen, L., Xie, X., Fan, X., Ma, W., Zhang, H., Zhou, H., 2003. A visual attention model for adapting images on small displays. ACM Multimedia Systems Journal, 353–364.

[20] Cheng, H., Sun, Y., 2000. A hierarchical approach to color image segmentation using homogeneity. IEEE Trans. on Image Processing 9 (12), 2071–2082.

[21] Cheng, H.-D., Sun, Y., 2000. A hierarchical approach to color image segmentation using homogeneity. IEEE Transactions on Image Processing 9 (12), 2071–2082.

[22] Christopoulos, V., De Muynck, P., Cornelis, J., 1999. Contour simplification for segmented still image and video coding: Algorithms and experimental results. Signal Processing: Image Communication 14, 335–367.

[23] Chung, D., Hirata, R., Mundhenk, T. N., Ng, J., Peters, R. J., Pichon, E., Tsui, A., Ventrice, T., Walther, D., Williams, P., Itti, L., 2002. A new robotics platform for neuromorphic vision: Beobots. Lecture Notes in Computer Science 2525, 558–566.

[24] Congyan, L., De, X., Xu, Y., 2005. Perception-oriented prominent region detection in video sequences. Informatica 29, 253–260.

[25] Crespo, J., Schafer, R., Serra, J., Gratin, C., Meyer, F., 1997. The flat zone approach: A general low-level region merging segmentation method. Signal Processing 62 (1), 37–60.

[26] Davis, G., Driver, J., Pavani, F., Shepherd, A., 2000. Obligatory edge assignment in vision: the role of figure and part segmentation in symmetry selection. Vision Research 40, 1323–1332.

[27] Deklerck, R., Cornelis, J., Bister, M., 1993. Segmentation of medical images. Image and Vision Computing Journal (Special Issue: Medical Image Processing - Guest Editor: Jan Cornelis; Editorial pp. 458-459) 11 (8), 486–503.

[28] DeVries, S. H., Baylor, D. A., 1997. Mosaic arrangement of ganglion cell receptive fields in rabbit retina. J. Neurophysiol. 78, 2048–2060.

[29] Draper, B., Lionelle, A., Oct-Nov 2005. Evaluation of selective attention under similarity transformations. Computer Vision and Image Understanding 100 (1-2), 152–171.

[30] Driver, J., Baylis, G. C., 1998. The Attentive Brain. Cambridge, MA: MIT Press, Ch. Attention and visual object segmentation, pp. 299–325.

[31] Driver, J., Davis, G., Russell, C., Turatto, M., Freeman, E., 2001. Segmentation, attention and phenomenal visual objects. Cognition 80, 61–95.

[32] Duncan, J., 1984. Selective attention and the organization of visual in-

formation. J. Exp. Psychol. 113, 501–517.
[33] Egly, R., Driver, J., Rafal, R., 1994. Shifting visual attention between objects and locations: evidence for normal and parietal lesion subjects. Journal of Experimental Psychology: General 123, 161–177.
[34] Eriksen, C. W., James, J. D. S., 1986. Visual attention within and around the field of focal attention: a zoom lens model. Perception and psychophysics 40 (4), 225–240.
[35] Espindola, G. M., Camara, G., Reis, I. A., Bins, L. S., Monteiro, A. M., 2006. Parameter selection for region-growing image segmentation algorithms using spatial autocorrelation. International Journal of Remote Sensing 27 (14), 3035–3040.
[36] Fink, G. R., Dolan, R. J., Halligan, P. W., Marshall, J. C., Frith, C. D., 1997. Spacebased and object-based visual attention: shared and specific neural domains. Brain 120, 2013–2028.
[37] Foley, J., Dam, A., Feiner, S., Hughes, J., 1990. Computer Graphics, Principles and Practice (2nd ed.). New York, NY: Addison-Wesley.
[38] Forsyth, D. A., Ponce, J., 2003. Computer Vision: A Modern Approach. Prentice Hall, Berkeley.
[39] Frintrop, S., 2005. Vocus: A visual attention system for object detection and goal-directed search. Lecture notes in artificial intelligence (lnai), vol. 3899 / 2006, springer berlin/heidelberg, University of Bonn.
[40] Fritz, G., Seifert, C., Paletta, L., Bischof, H., 2004. Attentive object detection using an information theoretic saliency measure. In: Paletta, L., Tsotsos, J. K., Rome, E., Humphreys, G. W. (Eds.), WAPCV2004: 2nd international workshop on attention and performance in computational vision.
[41] Grimaud, M., 1992. New measure of contrast: dynamics. In: Proc. SPIE, Image Algebra and Morphological Processing III. San Diego, CA, USA.
[42] Hamker, F., Oct-Nov 2005. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance 100 (1-2), 64–106.
[43] Hamker, F. H., 2000. Distributed competition in directed attention. In: Proceedings in Artificial Intelligence, Vol. 9. Dynamische Perzeption. G. Baratoff and H. Neumann, Berlin, pp. 39–44.
[44] He, Z. J., Nakayama, K., 1995. Visual attention to surfaces in 3-d space. Proceedings of the National Academy of Sciences USA 92, 11155–11159.
[45] Heinke, D., Humphreys, G. W., diVirgilo, G., 2002. Modeling visual search experiments: Selective attention for identification model (saim). Neurocomputing 44, 817–822.
[46] Henderickx, D., Maetens, K., Soetens, E., 2008. Inhibition of return: A bottom-up routed attentional process. submitted.
[47] Hill, P., Canagarajah, C., Bull, D., 2003. Image segmentation using a texture gradient-based watershed transform. IEEE Trans. Image Process. 12 (12), 1618–1633.
[48] Hoffman, J. E., 1998. Attention. Psychology Press, Ch. Visual attention and eye movements, pp. 119–154.
[49] Hoffman, J. E., Subramaniam, B., 1995. Saccadic eye movements and visual selective attention. Perception and Psychophysics 57, 787–795.
[50] Hu, Y., Rajan, D., Chia, L., 2005. Robust subspace analysis for detecting visual attention regions in images. In: Proc. ACM Multimedia 2005. pp. 716–724.
[51] Hu, Y., Xie, X., Ma, W., Chia, L., Rajan, D., 2004. Salient region detection using weighted feature maps based on the human visual attention model. In: Proc. IEEE PCM 2004. pp. 993–1000.
[52] Humphreys, G. W., Müller, H. J., 1993. Search via recursive rejection (serr): A connectionist model of visual search. Cognitive Psychology 25, 43–110.
[53] Itti, L., 2000. Models of bottom-up and top-down visual attention. Ph.D. thesis, California Institute of Technology, Pasadena, California, USA.
[54] Itti, L., 2002. Real-time high-performance attention focusing in outdoors color video streams. In: (B. Rogowitz, T. N. P. E. (Ed.), In: Proc. SPIE Human Vision and Electronic Imaging VII (HVEI'02). pp. 235–243.
[55] Itti, L., 2003. Advances in Neural Information Processing Systems. Vol. 15 of Hardware Demo Track. MIT Press, Cambridge, MA, Ch. The Beobot Platform for Embedded Real-Time Neuromorphic Vision.
[56] Itti, L., 2005. Models of bottom-up attention and saliency. In: L. Itti, G. Rees, J. K. T. (Ed.), Neurobiology of Attention. Elsevier, San Diego, CA, pp. pp. 576–582.
[57] Itti, L., Koch, C., 2001. Computational modelling of visual attention. Nature Reviews Neuroscience 2 (3), 194–203.
[58] Itti, L., Koch, C., 2001. Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging 10 (1), 161–169.
[59] Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 20 (11), 1254–1259.
[60] Jabloun, M., Mihai, C., Vanhamel, I., Geerinck, T., Sahli, H., 2009. Multispectral data classification based on spectral indices and fuzzy c-mean. In: IGARSS.
[61] Jarmasz, J., 2001. Towards the integration of perceptual organization and visual attention: The inferential attentional allocation model. Ph.d. prospectus, Carleton University, Ottawa, Ontario.
[62] Jarmasz, J., 2003. Objects, pilots, and the act of attending: A conative account of visual attention. Ph.d. thesis in cognitive science, Carleton University, Ottawa, Ontario.
[63] Jarmasz, J. P., 2002. Integrating perceptual organization and attention: A new model for object-based attention. Tech. rep., Cognitive Science Program ad Centre for Applied Cognitive Research, Carleton University, Ottawa, Canada.
[64] Kanishwer, N., Driver, J., 1992. Objects, attributes, and visual attention: which, what, and where. Current Directions in Psychological Science 1, 26–31.
[65] Katartzis, A., Sahli, H., 2008. A stochastic framework for the identification of building rooftops using a single remote sensing image. IEEE Transactions on Geoscience and Remote Sensing 46 (1), 259–271.
[66] Klein, R., 2000. Inhibition of return. Trends Cogn. Sci. 4, 138–147.
[67] Koch, C., Ullman, S., 1985. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4, 219–227.
[68] Lavie, N., Driver, J., 1996. On the spatial extent of attention in object-based selection. Perception & Psychophysics 58, 1238–1251.
[69] Le Meur, O., Le Callet, P., Barba, D., Thoreau, D., 2004. Performance assessment of a visual attention system entirely based on a human vision modeling. In: Proc. IEEE Intl Conf. Image Processing.
[70] Le Meur, O., Le Callet, P., Barba, D., Thoreau, D., 2006. A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (5), 802–817.
[71] Li, F., Kolakowski, S. M., Pelz, J. B., 2007. Using structured illumination to enhance video-based eye tracking. In: ICIP. Vol. 1. pp. 373–376.
[72] Li, Y., Ma, Y., Zhang, H., 2003. Salient region detection and tracking in video. In: Proc. IEEE ICME 2003. Vol. 2. pp. 269–272.
[73] Liu, F., Gleicher, M., 2005. Automatic image retargeting with fisheye-view warping. In: Proc. ACM UIST'05. pp. 153–162.
[74] Liu, F., Gleicher, M., July 2006. Region enhanced scale-invariant saliency detection. In: Multimedia and Expo, IEEE International Conference on. pp. 1477–1480.
[75] Liu, H., Jiang, S., Huang, Q., Xu, C., Gao, W., 2007. Region-based visual attention analysis with its application in image browsing on small displays. In: Proc. MM'07. pp. 305–308.
[76] Liu, J., Yang, Y.-H., 1994. Multiresolution color image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence 16 (7), 689–700.
[77] Logan, G. D., 1996. The code theory of visual attention: an integration of spacebased and object-based attention. Psychological Review 103 (4), 603–649.
[78] Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2), 91–110.
[79] Luo, J., C., G., 2003. Perceptual grouping of segmented regions in color images. Pattern Recognition 36 (12), 2781–2792.
[80] Ma, Y., Lu, L., Zhang, H., Li, M., 2002. A user attention model for video summarization. In: Proc. ACM Multimedia 2002. pp. 533–542.
[81] Ma, Y., Zhang, H., 2003. Contrast-based image attention analysis by using fuzzy growing. In: Proc. ACM Multimedia 2003. pp. 374–381.
[82] Mack, A., Rock, I., 1998. Inattentional Blindness. Cambridge, MA: MIT Press.
[83] Mack, A., Tang, B., Tuma, R., Kahn, S., Rock, I., 1992. Perceptual organization and attention. Cognitive Psychology 24, 475–501.
[84] Mannan, S. K., Ruddock, K. H., Wooding, D., 1997. Fixation sequences made during visual examination of briefly presented 2d images. Spatial Vision 11 (2), 157–178.

[85] Marcotegui, B., Beucher, S., Apr 2005. Fast implementation of waterfall based on graphs. In: Ronse, C., Najman, L., Decenciere, E. (Eds.), Mathematical morphology: 40 years on; proceedings of the 7th international symposium on mathematical morphology. Vol. 30 of Computational imaging and vision. Springer, Fontainebleau-France, pp. 177–186.

[86] Martin, D., Fowlkes, C., Tal, D., Malik, J., July 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th Int'l Conf. Computer Vision. Vol. 2. pp. 416–423.

[87] Meyer, F., 2001. An overview of morphological segmentation. IJPRAI 15 (7), 1089–1118.

[88] Meyer, F., Oliveras, A., Salembier, P., Vachier, C., 1997. Morphological tools for segmentations: connected filters and watershed. Annals of Telecommunication, 367–380.

[89] Miau, F., Itti, L., 2001. A neural model combining attentional orienting to object recognition: preliminary explorations on the interplay between where and what. In: Proc. IEEE Engineering in Medicine and Biology Society (EMBS). pp. 789–792.

[90] Miau, F., Papageorgiou, C., Itti, L., 2001. Neuromorphic algorithms for computer vision and attention. In: Proc. SPIE 46 Annual International Symposium on Optical Science and Technology 4479, 12–23.

[91] Mihai, C., Vanhamel, I., Sahli, H., Katartzis, A., Pratikakis, I., May 30 - June 2 2007. LNCS: Proc. of First Inter. Conf. Scale Space and Variational Methods in Computer Vision (SSVM 2007). Vol. 4485. Springer-Verlag, Ischia, Italy, Ch. Scale selection for compact scale-space representation of vector-valued images, pp. 32–42.

[92] Milanese, R., 1993. Detecting salient regions in an image: From biological evidence to computer implementation. Ph.D. thesis, University of Geneva, Switzerland.

[93] Milanese, R., Wechsler, H., Gill, S., Bost, J., Pun, T., 1994. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In: Proc of CVPR. pp. 781–785.

[94] Najman, L., Schmitt, M., 1994. Watershed of a continuous function. Signal Processing 38 (1), 99–112.

[95] Navalpakkam, V., Rebesco, J., Itti, L., 2004. Modeling the influence of knowledge of the target and distractors on visual search. Journal of Vision 4 (8), 690a.

[96] Navalpakkam, V., Rebesco, J., Itti, L., 2005. Modeling the influence of task on attention. Vision Research 45 (2), 205–231.

[97] Neisser, U., 1967. Cognitive Psychology. New York: Appleton-Century-Crofts.

[98] Neisser, U., Becklen, R., 1975. Selective looking: attending to visually specified events. Cognitive Psychology 7, 480–494.

[99] Nothdurft, H. C., 1993. The rol of features in preattentive vision: Comparison of orientation, motion and color cues. Vision Research 33, 1937–1958.

[100] O'Callaghan, R., Bull, D., January 2005. Combined morphological-spectral unsupervised image segmentation. IP 14 (1), 49–62.

[101] Olshausen, B. A., Andersen, C. H., Essen, D. C. V., 1993. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neuroscience 13 (11), 4700–4719.

[102] Osberger, W., Maeder, A., 1998. Automatic identification of perceptually important regions in an image. In: Proc. ICPR. Vol. 1. pp. 701–704.

[103] Ouerhani, N., von Wartburg, R., Hügli, H., Müri, R., 2004. Empirical validation of the saliency-based model of visual attention. ELCVIA 3 (1), 13–24.

[104] Palmer, S. E., 1999. Vision Science-Photons to Phenomenology. Cambridge, MA: MIT Press.

[105] Pardo, A., 2002. Extraction of semantic objects from still images. In: Proc. ICIP. Vol. 3. pp. 305–308.

[106] Park, H. S., Ra, J. B., 1999. Efficient image segmentation preserving semantic object shapes. IEICE Trans. Fundamentals E82-A (6), 879–886.

[107] Parkhurst, D., Law, K., Niebur, E., 2002. Modeling the role of salience in the allocation of overt visual attention. Vision Research 42 (1), 107–123.

[108] Pashler, H. E., 1988. Familiarity and visual change detection. Percept. Psychophys. 44, 369–378.

[109] Phaf, R. H., van der Heijden, A. H. C., Hudson, P. T. W., 1990. Slam: A connectionist model for attention in visualseleciton tasks. Cognitive Psychology 22, 273–341.

[110] Posner, M. E., 1980. Orienting of attention. Q. J. Exp. Psychol. 32, 3–25.

[111] Pratikakis, I., 1998. Watershed-driven image segmentation. Phd thesis, Vrije Universiteit Brussel, Faculty of Engineering Sciences, Electronics and Informatics (ETRO).

[112] Pratikakis, I., Sahli, H., Cornelis, J., 1999. Low level image partitioning guided by the gradient watershed hierarchy. Signal Processing 75 (2), 173–195.

[113] Pratikakis, I., Sahli, H., Cornelis, J., 2005. Medical Image Analysis Methods. CRC Taylor and Francis Group, Ch. Three-Dimensional Multiscale watershed segmentation of MR images, pp. 271–314.

[114] Privitera, C., Stark, L., 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. Pattern Analysis and Machine Intelligence (PAMI) 22 (9), 970–981.

[115] Pylyshyn, Z. W., Storm, R. W., 1988. Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spatial Vision 3, 179–197.

[116] Ramström, O., Christensen, H. I., 2002. Visual attention using game theory. In: BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision. Springer-Verlag, London, UK, pp. 462–471.

[117] Rensink, R. A., 1998. Mindsight: visual sensing without seeing. Invest. Ophthalmol. Vis. Sci. 39, 631a.

[118] Rensink, R. A., 2000. The dynamic representation of scenes. Visual Cognition 7, 17–42.

[119] Rensink, R. A., 2002. Change detection. Annual Review of Psychology 53, 245–277.

[120] Rensink, R. A., O'Regan, J. K., Clark, J. J., 1997. To see or not to see: the need for attention to perceive changes in scenes. Psychol. Sci 8, 368–373.

[121] Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nat Neurosci 2, 1019–1025.

[122] Rock, I., Linnett, C. M., Grant, P., Mack, A., 1992. Perception without attention: Results of a new method. Cognitive Psychology 24, 502–534.

[123] Roerdink, J., Meijster, A., 2000. The watershed transform: Definitions, algorithms, and parallelization strategies. In: Fundamenta Informaticae. Vol. 41. IOS Press, pp. 197–228.

[124] Rubner, Y., Tomasi, C., 1996. Coalescing texture descriptors. Tech. rep., Computer Science Department, Stanford University, Stanford, CA, USA.

[125] Salembier, P., Serra, J., 1995. Flat zone filtering, connected operators and filters by reconstruction. IEEE Trans on Image Processing 4 (8), 1153–1160.

[126] Scholl, B. J., 2001. Objects and attention: state of the art. Cognition 80 (1-2), 1–46.

[127] Sears, C. R., Pylyshyn, Z. W., 2000. Multiple object tracking and attentional processing. Canadian Journal of Experimental Psychology 54, 1–14.

[128] Senders, J., 1997. Distribution of attention in static and dynamic scenes. In: Proc. SPIE. No. 3026. pp. 186–194.

[129] Setlur, V., Takagi, S., Raskar, R., Gleicher, M., Gooch, B., 2005. Automatic image retargeting. In: Proc. MUM. pp. 59–68.

[130] Singh, M., J.Scholl, B., 2000. Using attentional cueing to explore part structure. In: Poster presented at the 2000 Pre-Psychonomies Object Perception and Memory meeting. New Orleans, LA.

[131] Socrates, S., Vanhamel, I., Spyros, F., Cornelis, J., Sahli, H., 2005. A watershed-based multiscale segmentation method for color images using automatic scale selection. Journal of Electronic Imaging 14 (3), 1–16.

[132] Sofou, A., Maragos, P., 2003. PDE-based modeling of image segmentation using volumetric flooding. In: Proc IEEE Int. Conf. on Image Processing. Barcelona-Spain, pp. 431–434.

[133] Sun, Y., 2003. Hierarchical object-based visual attention for machine vision. Phd. thesis, University of Edinburgh.

[134] Sun, Y., Fisher, R., 2004. Object-based visual attention for computer vision. Artificial Intelligence, pp 77-123.

[135] Takacs, B., Wechsler, H., 1998. A dynamic and multiresolution model of visual attention and its application to facial landmark detection. Computer Vision and Image Understanding 70 (1), 63–73.

[136] Tipper, S., Driver, J., Weaver, B., 1991. Object-centered inhibition of return of visual attention. Q. J. Exp. Psychol. A 43, 289–298.

[137] Treisman, A., Gelade, G., 1980. A feature integration theory of attention. Cognition Psychology 12, 97–136.

[138] Treisman, A., Gormican, S., 1988. Feature analysis inearly vision: Evidence from search asymmetries. Psychological Review 95 (1), 15–48.
[139] Treisman, A. M., 1993. Attention, Selection, awareness and control. Clarendon Press, Oxford, Ch. The perception of features and objects, pp. 5–35.
[140] Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F., 1995. Modeling visual attention via selective tuning. Artificial Intelligence 78 (1-2), p 507 – 547.
[141] Tsotsos, J., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., Zhou, K., Oct-Nov 2005. Attending to visual motion. Computer Vision and Image Understanding 100 (1-2), 3–40.
[142] Vachier, C., Meyer, F., 1995. Extinction value : A new measurement of persistence. In: Proc IEEE Workshop on Non Linear Signal and Image Processing. Greece, pp. 254–257.
[143] Vachier, C., Vincent, L., 1995. Valuation of image extrema using alternating filters by reconstruction. In: SPIE Image Algebra and Morphological Processing. San Diego CA.
[144] Vanhamel, I., 2006. Vector valued nonlinear diffusion and its application to image segmentation. Phd thesis, Vrije Universiteit Brussel, Faculty of Engineering Sciences, Electronics and Informatics (ETRO).
[145] Vanhamel, I., Pratikakis, I., Sahli, H., 2003. Multiscale gradient watersheds of color images. IEEE Trans. Image Processing 12 (6), 617–626.
[146] Vanhamel, I., Pratikakis, I., Sahli, H., 8-11 October 2006. Multiscale graph theory based colour segmentation. In: International Conference on Image Processing, ICIP2006. Atlanta, USA, pp. 769–772.
[147] Vecera, S., Behrmann, M., McGoldrick, J., 2000. Selective attention to the parts of an object. Psychonomic Bulletin & Review 7, 301–308.
[148] Vecera, S., Farah, M., 1994. Does visual attention select objects or locations. J. Exper. Psychol.: General 123, 146–160.
[149] Walther, D., Itti, L., Riesenhuber, M., Poggio, T., Koch, C., 2002. Attentional selection for object recognition a gentle way. Lecture Notes in Computer Science (LNCS) 2525, 472–479.
[150] Walther, D., Koch, C., 2006. Modeling attention to salient proto-objects. Neural Networks 19, 1395–1407.
[151] Walther, D., Rutishauser, U., Koch, C., Oct-Nov 2005. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Computer Vision and Image Understanding 100 (1-2), 41–63.
[152] Walther, D., Rutishauser, U., Koch, C., Perona, P., 2004. On the usefulness of attention for object recognition. In: Paletta, L., Tsotsos, J. K., Rome, E., Humphreys, G. W. (Eds.), WAPCV2004: 2nd international workshop on attention and performance in computational vision. Prague, Czech Republic.
[153] Wolfe, J. M., 2001. Guided search 4.0: A guided search that does not require memory for rejected distractors. Journal of Vision, Abstracts of the 2001 VSS Meeting 1 (3), 349a.
[154] Wolfe, J. M., 2007. Integrated Models of Cognitive Systems. New-York: Oxford, Ch. Guided Search 4.0: Current Progress with a model of visual search, pp. 99–119.
[155] Wolfe, J. W., 1994. Guided search 2.0: A revised model of visual search. Psychonomic Bulletin and Review 1, 202–238.
[156] Wooding, D. S., 2002. Eye movements of large population: Ii. deriving regions of interest, coverage, and similarity using fixation maps. Behavior Research Methods, Instruments and Computers 34 (3), 509–517.
[157] Xie, X., Liu, H., Ma, W.-Y., Zhang, H.-J., 2006. Browsing large pictures under limited display sizes. IEEE Trans. on Multimedia 8 (4), 707–715.
[158] Yantis, S., 1998. Attention. Psychology Press Ltd., Ch. Control of visual attention, pp. 223–256.
[159] Zhang, H., Fritts, J. E., Goldman, S. A., 2004. An entropy-based objective evaluation method for image segmentation. In: SPIE: Storage and Retrieval Methods and Applications for Multimedia. Vol. 5307. pp. 38–49.
[160] Zhu, H., Basir, O., 2003. Proximity measure image based region merging for texture segmentation through gabor filtering and watershed transform. In: Proc. 2003 IEEE International Conference on Robotics, Intelligent Systems and Signal Processing. Vol. 2. pp. 742–747.

Face detection, tracking and recognition with online learning

Valentin Enescu and Hichem Sahli

***Abstract*—This paper describes our work on designing a face recognition module for the humanoid robot Nao, within the EU Aliz-e project. We regard this task as a pipeline consisting of face detection, face features tracking, image processing and, finally, face recognition. The first three stages are meant to increase resistance of recognition against adverse conditions such as illumination changes and in-plane rotations. As to the last stage, our goals are i) to achieve real-time operation with minimal data storage, and ii) to learn new faces or update the already stored ones in an online manner.**

I. Introduction

Person identification is a crucial task for robots that interact with people over long time periods. Indeed, getting to know the identity of the attended human enables the robot to engage in natural conversation with that person by calling on his/her name and, most important, to store/access/update the user profile. This profile contains data on the preferences of the user and the way the person reacted in the past to different interactions or events. While personalized greetings may increase the social acceptance of the robot, the user profile enables the robot to pursue long-term goals and generate behaviors that are supportive to the human, in a medical environment or just as an entertaining companion.

Visual and speech data represent important cues for person identification. Among vision-based techniques, we cite face and gait recognition. While the latter is based on dynamic data, the first is discriminatively enough to work with a static image. In this paper, we discuss the design of a face detector for the humanoid robot Nao. The principles guiding our work are real-time operation and online learning and adaptivity. The sequel of the paper is structured as follows. Section II presents the face detection stage, while Section III deals with the face tracking and image pre-processing stages. The proposed face recognition algorithm is described in Section IV. Finally, Section V concludes the paper.

II. Face detection

Face detection algorithms [1] can be largely grouped in two categories: i) based on skin detection, segmentation and geometric reasoning, and ii) based on binary classifiers. The first category is suited for images where a face occupies a large area and its geometrical profile is close to an ellipse. However, for small faces it is difficult to assess whether a detected skin region stems from a real face or from clutter (skin-like background). In this case, binary classifiers (face/non-face) such as AdaBoost [2] show excellent performance (high detection rate at a low false positive rate). One disadvantage of classifier-based detectors is that they entail thousands of samples for training.

We have tried to train other types of classifiers on a large face training sample of $10,000$ images, such as fast Support Vector Machines [3] and Random Forests [4]. Various features were used, including image templates, Haar features [2], and LBP features [5]. Since the number of Haar features could be large ($> 300,000$ per a 20×20 gray image), we applied a number of feature selection algorithms prior to learning, such as Adaboost, Random Forests, Forward Feature Selection [6], to enable the training data to fit into the computer's memory. To make the classifier robust against illumination changes, we have applied various techniques for illumination normalization prior to learning/classification: histogram equalization, gradient faces [7], and DoG filtering with gamma correction [8]. Despite all these efforts, the obtained false positive rates were much higher than the one of the AdaBoost-based detector [2], implemented in the OpenCV library [9]. After analysis, the conclusion was that the performance of a detector depends heavily on the "bootstrap" strategy [10]: start with a small number of non-face patterns, train a classifier and apply it on a set of new images containing no faces; use the wrongly detected faces as non-face patterns in another training session; repeat this procedure until the number of false positive is very low. Based on this procedure, we have trained a high performing variant of [2], based on [6], to detect faces at 95% detection rate and 10^{-6} false positive rate. However, there is a trade-off between the classifier size (number of stages or detection cascades) and its performance. For real-time operation, one might consider a less-performing but quicker classifier, and apply other techniques for false positives rejection, such as skin detection.

III. Features identification and tracking

An important issue in face recognition is how to normalize the eye locations when the face undergoes in-plane rotations. Toward this end, we have opted for tracking the face with an Active Shape Model (ASM) [11]. ASM builds a statistical model for the shape of the face, including its features (eye, nose, mouth, eyebrows), and iteratively deforms it to match the face in a new image. Once a face is detected in an image, we initialize the ASM tracker on the detected region. A positive aspect is that, after initialization, face detection is not needed anymore for tracking. Thus, detection failures when the face rotates will not affect the ASM tracker. A byproduct of ASM tracking is the locations of the eyes, from which we can derive a rotation matrix to warp the face image so that it becomes vertical (rectification). Further, before the recognition stage, the face pattern is cropped using the boundaries of the ASM contour in the warped image, scaled to 100×100 pixels, and transformed in a histogram equalized array of gray values. As such, the images are well aligned with respect of the eye locations (per person) and illumination-normalized, which greatly enhances the recognition accuracy.

IV. Face recognition

Two large categories of face recognition algorithms [12] exist in the literature: appearance-based and feature-based. Appearance-based approaches include template matching and various subspace projection techniques such as Principal Component Analysis (PCA), Linear discriminant analysis (LDA), Independent component analysis (ICA) and Non-negative matrix factorization (NMF). Feature-based technique design classifiers using various local features such as geometric features, Gabor jets, LBP, Local Ternary Patterns (LTP) [8], etc.

V. Enescu and H. Sahli are with the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Belgium, e-mail: {venescu,hsahli}@etro.vub.ac.be.

We have experimented with template-based, PCA-based and LTP-based classifiers and a database of 10 people, with 200 face samples per person. The samples are taken from videos where people talk, pose emotions, and move their heads, such that a large variety of poses and non-rigid deformations of faces is achieved. The objective of the experiments was to understand the performance of these algorithm when using nearest neighbor (NN) classifiers trained with only 20 samples per person. The PCA-based algorithm performed poorly, while an NN classifier based on state-of-the-art LTP features was outperformed by a template method using the L_2 distance (Euclidean distance).

The system we implemented is an extension of the template-matching approach of [12]. The template-based method involves the direct comparison of pixel intensity values taken from facial images. Each image is converted in a row-wise manner to a high dimensional vector ($10,000$ elements for a 100×100 face image). By measuring the Euclidean distance (L_2 norm) between these points, an indication of the image similarity is obtained. Similar images are located close together within the image space, while dissimilar images are spaced far apart. The nearest neighbor classifier recognizes a face as the class (identity) of the template that has the greatest similarity to the test image. *Our extension* of the algorithm include: replacing the L_2 norm with the L_1 norm for a better performance, designing two new template selection strategies, and proposing an *online learning/template-update* approach for fast learning (no batch learning step is necessary, learning is sequential) and long-term adaptation to appearance changes. In the following, we elaborate on these topics.

While experimenting with L_1 (sum of absolute differences) distances, we have noticed that L_1 makes less recognition errors than the original algorithm based on L_2, at a smaller computational cost. As such, we have based our NN classifier on the L_1 distance. The next step was to find the best strategy to select the image templates (exemplars) from a video, during the training stage. Out first strategy was directed towards the discovery of dissimilar face poses (face views with out-of-plane rotations). To this end, we start with the first video image as the first exemplar. Then we consider a time window of 20 frames from the last stored exemplar, in which we measure the current face image similarity with that of the last exemplar, using the L_1 distance. The face pattern with the highest dissimilarity within the time window is stored as a new exemplar until we reach a number of 20 exemplars per person. This is a simple and fast strategy, but it does not guarantee an exemplar is highly dissimilar to other exemplars than its close neighbors. It may therefore waist storage space. To remedy this problem, we have devised a second strategy, which stores an exemplar solely when the classifier using the current exemplars will not recognize the new pattern as belonging to the person in the training video. In fact, this latter strategy unifies the learning and recognition steps of the algorithm, which opens the possibility to provide online adaptability (re-learning), as discussed next.

In general, online learning refers to a learning technique whereby the training samples are presented in a sequential manner to the classifier, as opposed to the batch methods, and the classifier is updated after each sample. The proposed online learning approach proceeds as follows: the robot detects a face and tries to recognize it. If the lowest L_1 distance against the exemplar database is above a threshold, it declares the person unknown and engages a dialog with the person to ask her/his name. Subsequently, it triggers the learning procedure that associate 20 face exemplars to that name. Besides the 20 exemplars, the face recognition module reserves 5 exemplar slots per person in view of re-learning/updating the face pattern during future interactions, thereby adapting to changes in the human appearance. As such, after finishing the initial learning step, the recognition module immediately enters the update step. This also happens when the robot recognizes the person. During the update step, the face tracker follows the face in the video and provides the recognition module with the target identity. In case recognition fails (i.e. decides person is unknown or somebody else), but the tracker not, the online re-learning process is activated by filling in the 5 extra exemplars. When the extra exemplars are exhausted, 5 exemplars that were used less during interaction will be discarded, and the new exemplars will be stored in their place. Thus, a new re-learning session can immediately start. To our knowledge, this is the simplest online learning strategy for face recognition among just a few described in the literature.

Online learning is a great feature, but there is always a risk to re-learn the wrong person (if a person is mis-identified the first time is detected by the robot). To prevent this to occur two strategies can be considered: i) verbal feedback by having the robot asking the person to confirm his/her name, and ii) start the update procedure only when the lowest L_1 distance against the database is lower than a second threshold, lower than the one used for deciding whether a person is unknown.

V. Conclusions

In this paper, we presented a new algorithm for face learning and recognition, based on face detection, tracking, and a fast nearest neighbor classifier using the L_1 norm between face templates. The algorithm works in real-time and is able to update the learned face database over time. These features recommend the algorithm for the use with a robot interacting with humans during extended time periods.

References

[1] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, 2002.

[2] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[3] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR 2008*, 2008, pp. 1–8.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[6] J. Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg, "Fast asymmetric learning for cascade face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 369–382, 2008.

[7] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2599–2606, 2009.

[8] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, 2010.

[9] "Opencv (Open Source Computer Vision) library." [Online]. Available: http://opencv.willowgarage.com

[10] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, 1998.

[11] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38 – 59, 1995.

[12] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 10, pp. 1042–1052, 1993.