Proc. 2nd International Workshop on Attention and Performance in Computational Vision (WAPCV 04), 15 May 2004, Prague, Czech Republic pp. 83-89, 2004

AttentiRobot: A Visual Attention-based Landmark Selection Approach for Mobile Robot Navigation

Nabil Ouerhani, Heinz Hügli
Institute of Microtechnology
University of Neuchâtel
Rue A.-L. Breguet 2
CH-2000 Neuchâtel, Switzerland
Nabil.Ouerhani@unine.ch

Gabriel Gruener, Alain Codourey
Centre Suisse d'Electronique et Microtechnique
CSEM, Microrobotics Division
Untere Gründlistrasse 1
CH-6055 Alpnach Dorf, Switzerland
Gabriel.Gruener@csem.ch

Abstract

Visual attention refers to the ability of a vision system to rapidly detect visually salient locations in a given scene. On the other hand, the selection of robust visual landmarks of an environment represents a cornerstone of reliable vision-based robot navigation systems. Indeed, can salient scene locations provided by visual attention be useful for robot navigation? This work investigates the potential and effectiveness of the visual attention mechanism to provide pre-attentive scene information to a robot navigation system. The basic idea is to detect and track the salient locations, or spots of attention by building trajectories that memorize the spatial and temporal evolution of these spots. Then, a persistency test, which is based on the examination of the lengths of built trajectories, allows the selection of good environment landmarks. The selected landmarks can be used for feature-based localization and mapping systems which helps mobile robot to accomplish navigation tasks.

1. Introduction

Visual attention is the natural ability of the human visual system to quickly select within a given scene specific parts deemed important or salient by the observer. In computer vision, a similar visual attention mechanism designates the first low-level processing step that allows to quickly selecting in a scene the points of interest to be analyzed more specifically and in-depth in a second processing step.

The computational modeling of visual attention has been a key issue in artificial vision during the last two decades [9, 16, 15]. First reported in 1985 [10], the saliency-based model of visual attention is largely accepted today [8] and gave rise to numerous soft and hardware implementations [8, 14]. In addition, this model has been used in several computer vision applications including image compression [12] and color image segmentation [13].

In visual robot navigation, the detection, tracking and thus selection of robust visual-based landmarks represent the most challenging issues in building reliable navigation systems [3]. Numerous previous works have pointed to the visual attention paradigm in solving various issues in active vision in general [2, 1] and visual robot navigation in particular [7].

This work proposes a visual attention-based approach for visual landmark selection. The proposed approach relies on an extended version of Itti's *et al.* model of visual attention [8] in order to detect the most visually salient scene locations; the spots of attention. More specifically, these spots of attention are deduced from a saliency map computed from multiple visual cues including corner features. Then, the spots of attention are characterized using a feature vector that represents the contribution of each considered feature to the final saliency of the spot. Once characterized, the spots of attention are easily tracked over time using a simple tracking method that is based on feature matching. The tracking results reveal the persistency and thus the robustness of the spots, leading to a reliable criterium for the selection of the landmarks.

The navigation phase, which has not been tested yet consists in using the selected environment landmarks for feature-based Simultaneous Localization and Mapping (SLAM) on a mobile robot [4]. A schematic overview of the landmark selection approach as well as its integration into a general visual robot navigation system are given in Figure 1.

The remainder of this paper is organized as follows. Section 2 describes the saliency-based model of visual attention. Section 3 presents the characterization and tracking of spots of attention. The persistency test procedure is exposed in Section 4. Section 5 reports some experiments carried out

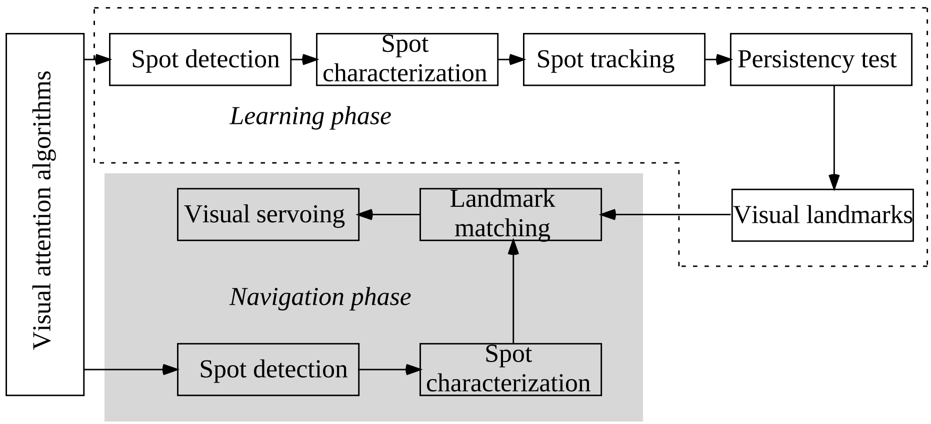


Figure 1. Overview of the attention-based landmark selection approach.

on real robot navigation image sequences in order to assess the proposed approach. Finally, the conclusions and some perspectives are stated in Section 6.

2. Attention-based landmark detection

2.1. Saliency-based model of visual attention

The saliency-based model of visual attention, which selects the most salient parts of a scene, is composed of four main steps [10, 8].
1) First, a number of features are extracted from the scene by computing the so called feature maps F_j. The features most used in previous works are intensity, color, and orientation. The use of these features is motivated by psychophysical studies on primate visual systems. In particular, the authors of the model used two chromatic features that are inspired from human vision, namely the two opponent colors red/green (RG) and blue/yellow (BY).
2) In a second step, each feature map F_j is transformed in its conspicuity map C_j. Each conspicuity map highlights the parts of the scene that strongly differ, according to a specific feature, from its surrounding. This is usually achieved by using a *center-surround*-mechanism which can be implemented with multiscale *difference-of-Gaussian*-filters.
3) In the third stage of the attention model, the conspicuity maps are integrated together, in a competitive way, into a *saliency map* $\mathcal{S}$ in accordance with equation 1.

$$\mathcal{S} = \sum_{j=1}^{J} \mathcal{N}(\mathcal{C}_j) \tag{1}$$

where $\mathcal{N}()$ is a normalization operator that promotes conspicuity maps in which a small number of strong peaks of activity are present and demotes maps that contain numerous comparable peak responses [8].
4) Finally the most salient parts of the scene are derived from the saliency map by selecting the most active locations of that map. A Winner-Take-All network (WTA) is often used to implement this step [10].

2.2. Extension of the model to corner features

In the context of vision-based robot navigation, corner features are considered as highly significant landmark candidates in the navigation environment [3]. This section aims at extending the basic model of visual attention to consider also corner features. To do so, a corner map C_c which highlights the corner points in the scene, is first computed. Then, this corner map is combined together with the color and intensity-based conspicuity maps into the final saliency map.

Multi-scale Harris corner detector [6, 11]. Practically, the proposed multiscale method computes a corner pyramid $\mathcal{P}_c$. Each level of the corner pyramid detects corner points at a different scale. Formally, $\mathcal{P}_c$ is defined according to Equation 2.

$$\mathcal{P}_c(i) = Harris(\mathcal{P}_g(i)) \tag{2}$$

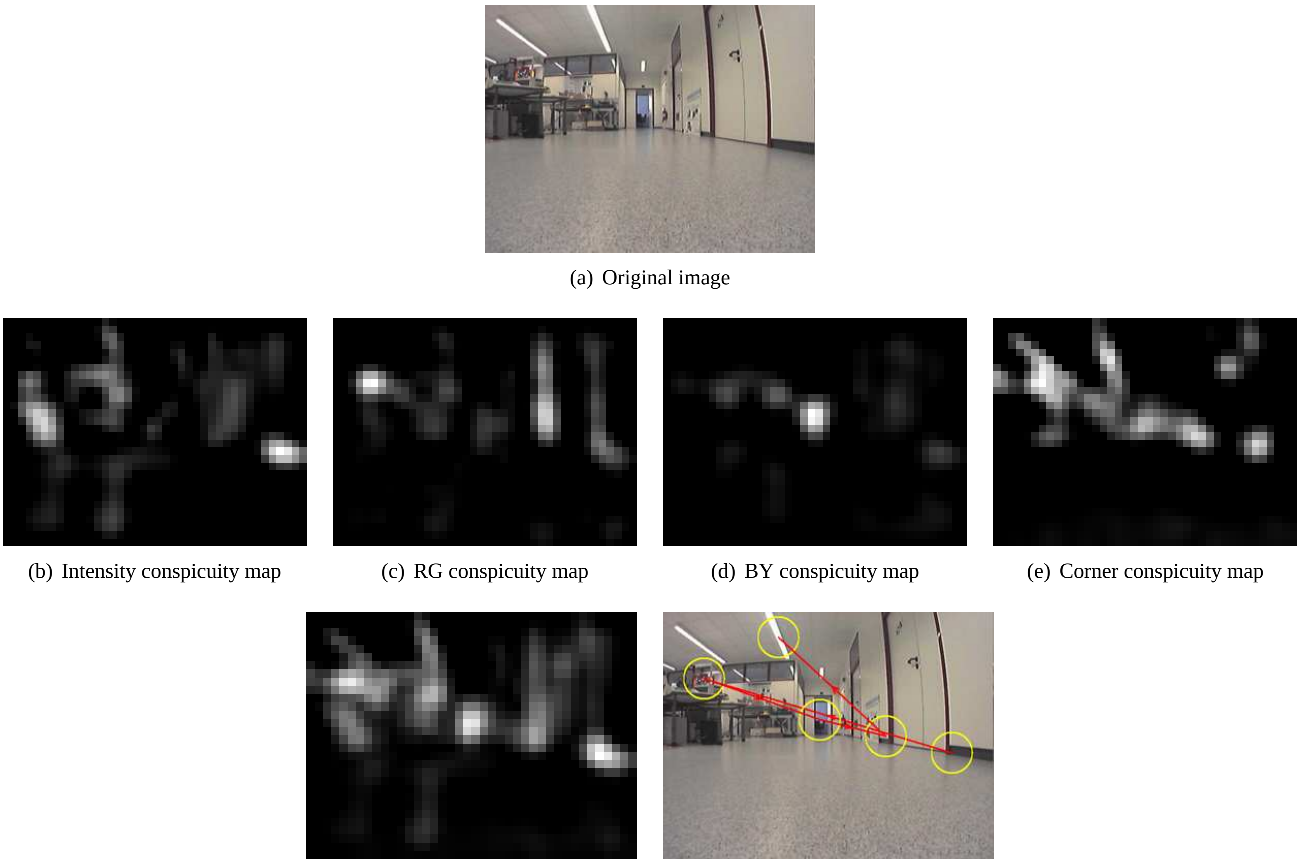
where $Harris(.)$ is the Harris corner detector as defined in [6] and $\mathcal{P}_g$ is a gaussian pyramid defined as follows:

$$\begin{aligned} \mathcal{P}_g(0) &= I \\ \mathcal{P}_g(i) &= \textcircled{\downarrow 2}(\mathcal{P}_g(i-1) * G) \end{aligned} \tag{3}$$

where I is a grey-scale version of the input image, G is a gaussian filter and $\textcircled{\downarrow 2}$ refers to the down-sampling (by 2) operator.

Corner conspicuity map C_c. Given the corner pyramid $\mathcal{P}_c$, C_c is computed in accordance with Equation 4.

$$C_c = \sum_{s=1}^{s_{max}} \mathcal{P}_c(s) \tag{4}$$



(a) Original image

(b) Intensity conspicuity map

(c) RG conspicuity map

(d) BY conspicuity map

(e) Corner conspicuity map

(f) Saliency map

(g) Spots of attention

Figure 2. Example of the Conspicuity maps, the saliency map and the corresponding spots of attention computed with the corner-extended model of visual attention.

Note that the summation of the multiscale corner maps $\mathcal{P}_c(s)$ is achieved at the coarsest resolution. Maps of finer resolutions are lowpass filtered and downsampled to the required resolution. In our implementation s_{max} is set to 4, in order to get a corner conspicuity map C_c that has the same resolution as the color- and intensity-related conspicuity maps.

Integration of corner feature into the model. The final saliency map $\mathcal{S}$ of the extended model is computed in accordance with Equation 5.

$$\mathcal{S} = \sum_{j=1}^{J+1} \mathcal{N}(C_j) \tag{5}$$

where

$$C_{J+1} = C_c \tag{6}$$

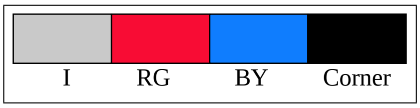
Selection of the spots of attention. The maxima of the saliency map represent the most salient spots of attention. Once a spot is selected, a region around its location is inhibited in order to allow the next most salient spot to be selected. The total number of spots of attention can be either set interactively or automatically determined by the activity of the saliency map. For simplicity, the number of spots is set to five in our implementation.

Figure 2 shows an example of the four conspicuity maps, saliency map and the spots of attention computed by the corner-extended model of visual attention.

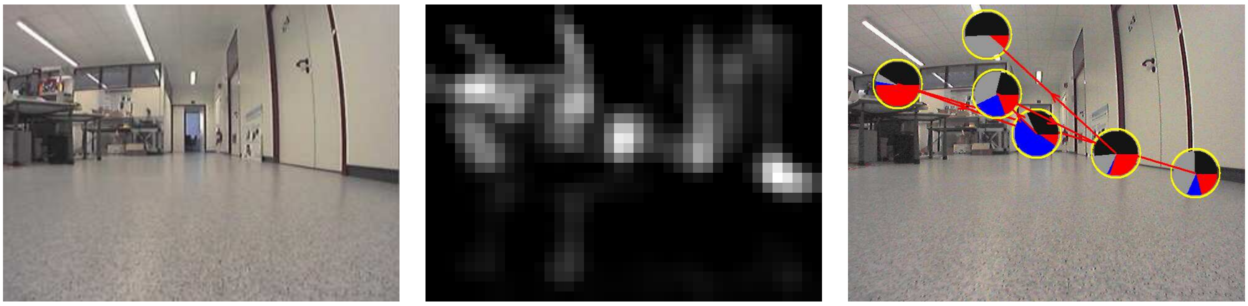
3. Spot Characterization and Tracking

3.1. Spot characterization

The spots of attention computed by means of the extended model of visual attention locate the scene features to be tracked. In addition to location, each spot $\mathbf{x}$ is also



(a) Feature representation



(b) Original image (c) Saliency map (d) Characterized spots of attention

Figure 3. Characterization of spots of attention. The five most salient spots of attention are detected and characterized using four visual features, namely intensity (I), red-green (RG) and blue-yellow (BY) color components, and corners.

characterized by a feature vector $\mathbf{f}$:

$$\mathbf{f} = \begin{pmatrix} f_1 \\ .. \\ f_J \end{pmatrix} \quad (7)$$

where J is the number of the considered features in the attention model and f_j refers to the contribution of the feature j to the detection of the spot $\mathbf{x}$. Formally, f_j is computed as follows:

$$f_j = \frac{\mathcal{N}(C_j(\mathbf{x}))}{\mathcal{S}(\mathbf{x})} \quad (8)$$

Note that $\sum_{j=1}^{J}(f_j) = 1$.

Let N be the number of frames of a sequence and M the number of spots detected per frame, the spots of attention can be formally described as
$P_{m,n} = (\mathbf{x}_{m,n}, \mathbf{f}_{m,n})$, where $m \in [1..M]$, $n \in [1..N]$, $\mathbf{x}_{m,n}$ is the spatial location of the spot, and $\mathbf{f}_{m,n}$ its characteristic feature vector. Figure 3 illustrates an example of the characterization of spots of attention.

3.2. Spot tracking

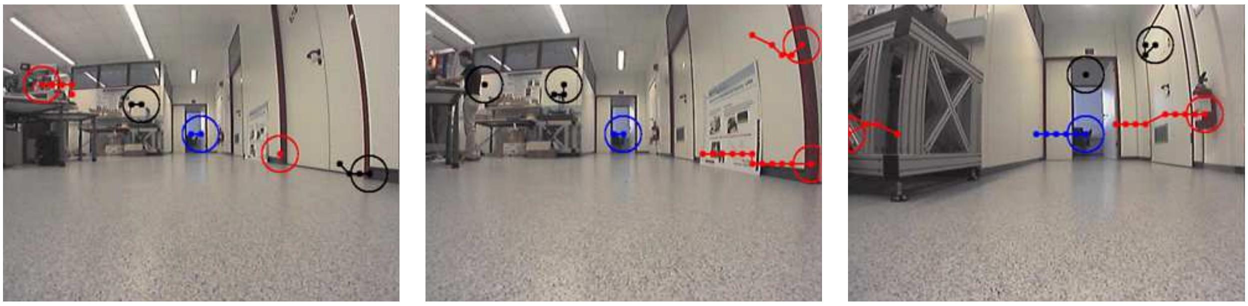
The basic idea behind the proposed algorithm is to build a trajectory for each tracked spot of attention. Each point of the trajectory memorizes the spatial and the feature-based information of the tracked spot at a given time.

Specifically, given the M spots of attention computed from the first frame, the tracking algorithm starts with creating M initial trajectories, each of which contains one of the M initial spots. The initial spots represent also the head elements of the initial trajectories. A new detected spot $P_{m,n}$ is either appended to an existing trajectory (and becomes the head of that trajectory) or gives rise to a new trajectory, depending on its similarity with the head elements P^h of already existing trajectories as described in Algorithm 1. Note that a spot of attention is assigned to exactly one trajectory (see the parameter *marked[]* in Algorithm 1) and a trajectory can contain at most one spot from the same frame. In a simple implementation, the condition that a spot $P_{m,n}$ must fulfil in order to be appended to a trajectory T with a head element $P^h = (\mathbf{x}_h, \mathbf{f}_h)$ is given by:

$$P_{m,n} \in T \;\; if \;\; \|\mathbf{x}_{m,n} - \mathbf{x}_h\| < \epsilon_{\mathbf{x}} \;\&\; \|\mathbf{f}_{m,n} - \mathbf{f}_h\| < \epsilon_{\mathbf{f}} \quad (9)$$

where $\epsilon_{\mathbf{x}}$ and $\epsilon_{\mathbf{f}}$ can be either determined empirically or learned from a set of image sequences. In a more advanced version of the tracking algorithm, the mean feature vector $\mathbf{f}_\mu$ of all spots of the same trajectory will replace the head element feature vector $\mathbf{f}_h$ in the similarity measure. In addition, the threshold $\epsilon_{\mathbf{f}}$ will directly depend on the standard deviation of the feature vectors of spots belonging to the same trajectory and $\epsilon_{\mathbf{x}}$ will be brought in relation with the robot motion deduced from odometry.

In the absence of ground-truth data, the evaluation of the tracking algorithm can be achieved interactively. Indeed, a human observer can visually judge the correctness of the trajectories, i.e. if they track the same physical scene constituents. Figure 4 gives some examples of trajectories built from a set of spots of attention using the tracking algorithm



(a) frame 14 (b) frame 33 (c) frame 72

Figure 4. Examples of trajectories built from a set of spots of attention.

described above.

4. Persistency test

This step of the approach is part of the learning phase and aims at selecting, among all detected spots of attention, the most robust as visual landmarks of the environment. The basic idea is to examine the trajectories built while tracking spots of attention. Specifically, the length of the trajectories reveals the robustness of the detected spots of attention. Thus, during the learning phase the cardinality ($Card(T)$) of a trajectory directly determines whether the corresponding spots of attention are good landmarks.

In addition, the cardinality of the trajectories can be used as measure to compare the performance of different interest points detectors, as stated in Section 5, but also of different tracking approaches.

5. Results

This section presents some experiments that aim at assessing the presented landmark selection approach. The tests have been carried out with four sequences acquired by a camera mounted on a robot that navigates in an indoor environment over a distance of about 10 meters (see Figure 3). The length of the sequences varies between 60 and 83 frames. Two groups of results are presented here. Qualitative results regarding the robustness of the detection and tracking algorithms and quantitative results that point to the superiority of the corner-extended model of attention over the classic one.

Regarding the first group of results, Figure 5 illustrates the trajectories built from each sequence. The trajectories are plotted in 3D (x, y, t) in order to better visualize their temporal extent.

In the first sequence (Figure 5(a)), the most robustly detected and tracked landmark is the entrance of the differ-

Algorithm 1 Attention-based object tracking

Image sequence $I(n)$ $(1..n..N)$
Number of detected spots of attention per frame: M
Boolean $appended$
Boolean $marked[\,]$
Trajectory set $\{T\} = \emptyset$

for $n = 1 \,..\, N$ **do**
  Detect & characterize the M spots of attention $P_{m,n} = (\mathbf{x}_{m,n}, \mathbf{f}_{m,n})$
  for $k = 1 \,..\, card(\{T\})$ **do**
    $marked[k] = 0$
  end for
  for $m = 1 \,..\, M$ **do**
    $appended = 0$
    for $k = 1 \,..\, card(\{T\})$ **do**
      if $(marked[k] == 0)$ **then**
        if $d(P_{m,n}, P_k^h) < \varepsilon$ * **then**
          $append(P_{m,n}, T_k)$
          $appended = 1$
          $marked[k] = 1$
          break
        end if
      end if
    end for
    if $(appended == 0)$ **then**
      $newTraject(T_{card(\{T\})+1})$
      $append(P_{m,n}, T_{card(\{T\})+1})$
      $\{T\} = \{T\} \cup \{T_{card(\{T\})+1}\}$
    end if
  end for
end for
* $d()$ is given by Equation 9

ently illuminated room toward which the robot is moving. The trajectory built around this landmark has a length of 83, which means that the spot has been detected in each frame of the sequence. In addition, the red-colored door frames (especially their corners) and a fire extinguisher have been tracked over a large number of frames. Ceiling lights figure also between the detected and tracked features.

Like the first example, the three others ((b) light switched off, (c) front door closed, and (d) other corridor) tend to show, qualitatively, the ability of the proposed approach to robustly detect and track certain visual features of the navigation environment over a large period of time and under different conditions. For instance, the door frames and the fire extinguisher figure among those features that can be considered as environment landmarks. A more quantitative and in-depth evaluation of the robustness of the proposed approach towards view angle changes and changing in lighting conditions is required, in order to definitely validate our landmark selection method.

Table 1, which resumes the second group of results, shows the advantage of the corner-extended model over the basic model regarding the stability of the detected spots of attention over time. For each of the four image sequences the total number of trajectories, their minimum, maximum, and mean cardinality (length) are represented. It can be seen that the integration of the corner features has leaded to more consistent trajectories.

6. Conclusions and future work

This work presents an attention-based approach for selecting visual landmarks in a robot navigation environment. An extended version of the saliency-based model of visual attention that considers also corners has been used to extract spatial and feature-based information about the most visually salient locations of a scene. These locations are then tracked over time. Finally, the most robustly tracked locations are selected as environment landmarks. One of the advantages of this approach is the use of a multi-featured visual input, which allows to cope with navigation environments of different natures, while preserving, thanks to the feature competition, a discriminative characterization of the potential landmarks. Qualitative results show the ability of the method to select good environment landmarks, whereas the quantitative results confirm the superiority of the corner-extended model of attention over the classic one, regarding the consistency of the detected spots of attention over time.

In future work, the rather simple tracking algorithm will be improved, essentially by introducing predictive filters such as Kalman and particle filters [5]. In addition, we are planning to apply the proposed approach to solve some problems related to Simultaneous Localization and Map building (SLAM) in real robot navigation tasks.

References

[1] K. Brunnstrom, J. Eklundh, and T. Uhlin. Active fixation for scene exploration. *International Journal of Computer Vision, Vol. 17, pp. 137-162*, 1994.

[2] J. Clark and N. Ferrier. Control of visual attention in mobile robots. *IEEE Conference on Robotics and Automation, pp. 826-831*, 1989.

[3] A. Davison. *Mobile Robot Navigation Using Active Vision*. PhD thesis, University of Oxford, UK, 1999.

[4] M. Dissanayakeand Gamini, P. Newman, S. Clark, H. Durrant-Whyte, and M. Csorba. A solution to the simultaneous localization and map building (slam) problem. *IEEE Transactions on Robotics and Automation, Vol. 17, pp. 229-241*, 2001.

[5] D. Fox, S. Thrun, F. Dellaert, and W. Burgard. Particle filters for mobile robot localization. *In A. Doucet, N. de Freitas, and N. Gordon, editors, Sequential Monte-Carlo Methods in Practice. Springer-Verlag, New York*, 2000.

[6] C. Harris and M. Stephens. A combined corner and edge detector. *Fourth Alvey Vision Conference, pp. 147-151*, 1988.

[7] L. Itti. Toward highly capable neuromorphic autonomous robots: beobots. *SPIE 47 Annual International Symposium on Optical Science and Technology, Vol. 4787, pp. 37-45*, 2002.

[8] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259*, 1998.

[9] B. Julesz and J. Bergen. Textons, the fundamental elements in preattentive vision and perception of textures. *Bell System Technical Journal, Vol. 62, No. 6, pp. 1619-1645*, 1983.

[10] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, Vol. 4, pp. 219-227*, 1985.

[11] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *European Conference on Computer Vision (ECCV), Vol.1, pp. 128-142*, 2002.

[12] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, and F. Pellandini. Adaptive color image compression based on visual attention. *International Conference on Image Analysis and Processing (ICIAP'01), IEEE Computer Society Press, pp. 416-421*, 2001.

[13] N. Ouerhani and H. Hugli. MAPS: Multiscale attention-based presegmentation of color images. *4th International Conference on Scale-Space theories in Computer Vision, Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. 2695, pp. 537-549*, 2003.

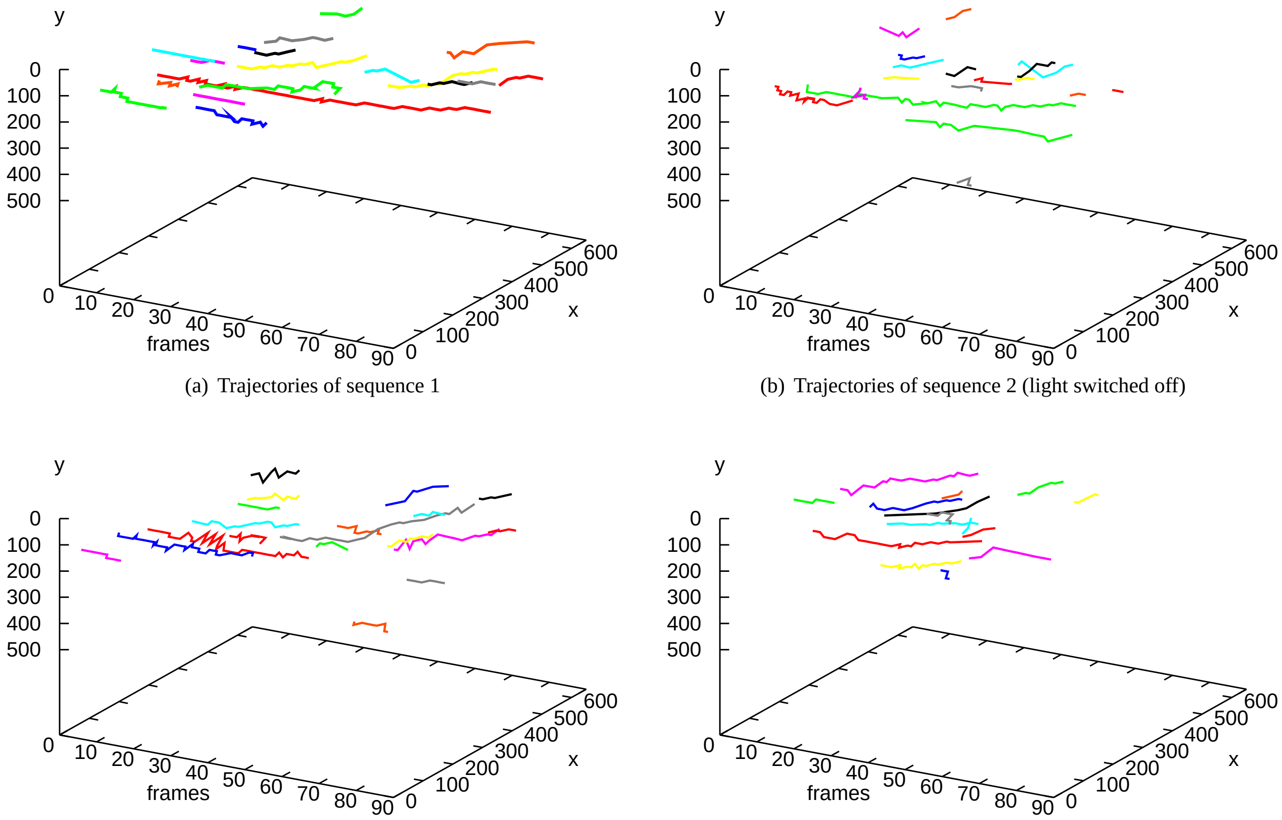
[14] N. Ouerhani and H. Hugli. Real-time visual attention on a massively parallel SIMD architecture. *International Journal of Real Time Imaging, Vol. 9, No. 3, pp. 189-196*, 2003.

[15] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence, Vol. 78, pp. 507-545*, 1995.

[16] J. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, Vol. 1, pp. 202-238*, 1994.

	number of spots	number of trajectories		Min / Max($Card(T)$)		Mean($Card(T)$)	
		without Harris	with Harris	without Harris	with Harris	without Harris	with Harris
Seq1	415	88	58	1 / 68	1 / 83	4.7	7.1
Seq2	320	136	80	1 / 19	1 / 56	2.3	4.0
Seq3	385	130	61	1 / 22	1 / 48	2.9	6.5
Seq4	280	123	56	1 / 22	1 / 31	2.2	5.0

Table 1. Impact of the integration of Harris corner features on the tracking algorithm. The total number of trajectories, the minimum, maximum, and mean cardinality of trajectories are computed for the classical (without Harris) and the corner-extended (with Harris) models.



(a) Trajectories of sequence 1

(b) Trajectories of sequence 2 (light switched off)

(c) Trajectories of sequence 3 (opposite door closed)

(d) Trajectories of sequence 4 (other corridor)

Figure 5. Trajectories of the tracked spots of attention from four different sequences ((a)..(d)). Note that only trajectories with $Card(T) > 3$ are represented here.