An Artificial System for Visual Perception in Autonomous Robots

Benjamin Auffarth[1,2], Yasumasa Muto[1], and Yasuharu Kunii[1]

[1]*Faculty of Science and Engineering, Department of Electronics and Communication Engineering Chuo University, Tokyo, Japan*
{yasu, kunii, benjamin@hmsl.elect.chuo-u.ac.jp}

[2]*Institute of Cognitive Science, University of Osnabrueck,Germany*
bauffart@uos.de

***Abstract* - Image processing of natural scenes is very processing-intensive. Through the localization of salient regions, later recognition processes can take place more efficiently, by focusing computation-intensive processing on several areas. Our approach to designing an artificial visual system is inspired by early filtering mechanisms in the human visual system. Our technique of calculating salient regions is computationally efficient and flexible, and can be extended to other applications.**

Index Terms – Intelligent Robotics, Image Processing, Robot Vision, Feature Extraction

I. INTRODUCTION

Field missions rely on an effective analysis of the environment. The greatest problem therein is to reduce the data to make image processing more feasible. The most important information is probably visual and for vision, the image processing of natural features in humans provides a source for many useful ideas for filtering mechanisms. In our research, we try to isolate some mechanisms used in the human visual system in search tasks and to implement them as filtering mechanisms, for an artificial visual system that is applied in autonomous robots.

Very roughly, the human visual system works by choosing regions in the visual field that are of potential interest. In our research, we call these points “attention points”, a word coined in Japanese research (compare [23]). Conceptions very similar to ours of attention points have been given various names before, e.g. human regions of interest (hRoI [15,26], saliency map [17,18], or location maps [1]. We refer by it generally to areas in the visual field that attract more processing capacities in human brains in overt (bottom-up) visual attention tasks than do other areas. Attention points may be indicating danger, food, goal relevance, etc. Based on the representation of attention points, urgent or vital target objects can be extracted. The speediness and accuracy of that process in humans originates probably in the characteristics of the distribution of spatial attention in these processes (compare [7]).

Treisman's feature integration theory (see [18]) has been the most influential model of human attention until recent years. According to Treisman, in a first step to visual processing, several primary visual features (such as color, orientation, and intensity) are processed and represented with separate feature maps that are later integrated in a saliency map that can be accessed in order to direct attention to the most conspicuous areas [17,18].

Psychological and computational models for visual attention, as in Wolfe's Guided Search [21] and Treisman's Feature Integration Model [18], conceive of the human visual system as having two stages in anatomy and physiology (compare [22], for a recent review of the research). Thus, the human visual system is devided into an initial processing stage (more or less data-driven) and a stage that includes more high level knowledge (compare the guided search paradigm [21]). The interaction of both stages leads to the selection of candidate points (attention points, saliency maps, or hRoI, etc.) and from them to the subsequent extraction of targets from the visual display for further analysis and image recognition. The first stage performs in a fast and parallel manner and carries out pre-attentive computations to detect and highlight conspicuous image locations for pre-attentive segmentation. It extracts visual features that in the subsequent stage are being analyzed. This second stage processes in serial and more slowly, and is characterized by a more conscious guidance.

Our model works in analogy to the human system by finding regions of high saliency in an image, attention points, that are to be analyzed in more detail. It is mimicking several aspects that are thought to make human attention so fast and effective. We will come to these aspects in sections III and IV and show some preliminary evaluation of the model in section V.

II. A SYSTEM FOR ARTIFICIAL VISION

In autonomous robots the natural environment is taken in by a camera. These complex data have to be filtered, in order to process only data, that are somehow important, ignoring data that are unnecessary.

We propose an architecture of artificial vision to make the recognition process both faster and less-error prone. We take a stance that is inspired by means by which the human visual system processes complex natural environments. Doing so we incorporated many psychological and neuroscientific findings in order to approach human performance in attention point selection and have more effective autonomous robots.

Our architecture, similar to Treisman's model, calculates attention points by a combination of feature maps. These attended ressources are exploited and finally targets can be extracted, objects be perceived and identified.

We distinguish conceptionally early and later processes (or stages) in visual processing, attention point selection (stage 1) and processes that set on attention points, i.e. processes for directing attention and higher-level processes for object recognition (stage 2). Fig.1 illustrates a conceived input-

output model of the visual information processing of our architecture.

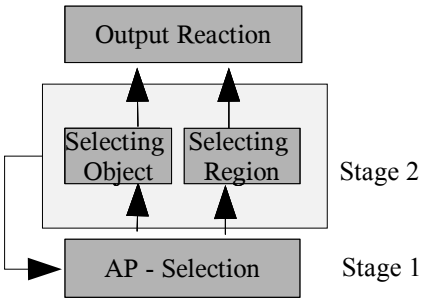

Fig. 1. Simplified sketch of our proposed visual system.
(AP=attention point)

By bottom-up perception attended locations are selected and then higher-level factors such as experiences and desires and the like influence (e.g. inhibit) the locations that are attended and thereby where the information is acquired from. In the following paragraphs stage 1 and stage 2 are briefly explained.

(i) Stage 1 mechanisms work data-driven, identify the locations which draw attention and store a representation of attention points (similarly to the saliency map). This representation includes an attentional gradient, i.e. it comprises regions that are perceived from highly salient over somehow salient, to not-salient-at-all. Furthermore, as the saliency map arises from the parallel activation relaxation process of neural networks, it has different preferences over time.

(ii) Stage 2 processes are the means by which targets (e.g. objects, or affordances) are extracted from the attention points and attention can be guided for information uptake from other visual areas.

In this article we are concerned with the bottom-up mechanisms that lead to the selection of attention points. Later on, we want to incorporate high-level processes in our model to achieve visual search and object recognition. We think that bottom-up and top-down processes perform in a highly interactive fashion, in a way that bottom-up processes lead to top-down processes and top-down processes guide bottom-up processes (compare [20]). Although we start by including bottom-up (stage 1) processes, we strive to include an interactive component in our architecture and we think that in a later step many parameters in the processing, described in this paper, can be tuned according to high-level processing (cf. [8]).

III. REALIZATION OF THE ATTENTION POINT SELECTION

We implemented feature dimensions, such as color (luminance), the area (i.e. size), shapes, position and others. Some of these dimensions apply to regions rather than single points. Therefore, before a selection of attention points is feasible, perceptional organization by segmentation and clustering has to occur, determining which points can be grouped together and which locations belong together and form a sufficiently coherent region.

Our model first performs an image segmentation by applying edge feature detectors (Laplacian of the Gaussian). There is strong evidence and it is commonly assumed that human center-off and center-on cells, held responsible for feature detection, work by a similar principle (see [5], or [3]). Afterwards we apply a growing regions algorithm for grouping neighboring points, based on luminance, in a pyramidal filtering approach [similar to 25]. Sample results of the segmentation are shown in fig. 2 and fig. 3.

Although the selection process relies heavily on knowledge and experience, we assume fixed stimulus variables (until implementation of the second level) and we define an evaluation function for attention points including several facts from psychological experiments, relying as broad a research as we can and experimenting with the features.

The research is very diversified and many results controversial. E.g. in many studies on visual features that correlate with saliency in overt attention, the stimuli have been simplified to geometric figures or Gabor patches, e.g., in order to exclude higher order properties of complex stimuli. Therefore, it is not clear, whether results, obtained in such experiments can be transferred to natural images. Here, we tried to concentrate on few visual features that had a good standing in research and on the other hand, achieved subjectively the best results.

For several features we used the formula for Stevens' power law,

$S = kI^a$,

which relates the intensity of a stimulus to its perceived strength. S is the perceived strength of sensation, k is a constant, I is the stimulus intensity, and a is an exponent. a is dependent on the type of stimulus, experimentally determinable. This we transformed to

$\log S = \log k + a \log I$

which comes as the form $y=m+bx$, with the original exponent a as the slope, k as the y intercept, and $\log I$ as the intensity of the physical stimulus (compare [16]).

We go on to discuss especially the importance of six stimulus dimensions for attracting bottom-up attention. These are are area, the degree of circular rate (shape), color hue and brightness, aspect ratio, centrality (and distance). Below, these six stimulus dimensions are explained and we comment on their implementation.

1.) Area (size): It has been repeatedly experimentally shown (e.g. [2]) that the bigger the area of a target the more intense the stimulus. Based on [16] we estimated magnitude saliency, with $a= 0.8$.

2.) Color hue and luminance: Saliency depends on color properties. Of the surface colors, red and yellow have been attributed high saliency. Shinzaku and others showed in

an experiment on saliency of colors (reported in [13]) that red and yellow are the most salient colors and we increased the intensity of red and yellow in our model. Many experiments in the past (e.g. [14]) have shown that luminance contrast correlated positively with stimulus intensity. However, Einhaeuser and Koenig showed recently that strong local reductions luminance contrast attract fixations [19]. We included grey-level difference with $a=1$ (assumed a perceptually linear space).

3.) Circular ratio (shape): According to Attneave (reported in [12]), humans attend more to irregular and complicatedly shaped figures rather than to simple geometric forms, such as circles and squares. We calculated the circular ratio that measures the regularity of shapes by

$R = \frac{4\pi A}{l^2}$, where R is the circular ratio of an area A and circumference l. We set $a=-0.43$ in Stevens' power law (compare [16]).

4.) Aspect ratio (ratio of length and breadth): According to a research by Clark (reported in [4]), the ratio of the edges of quadrilaterals a and b (where length and breadth are exchangeable)determines the saliency of such a figure. The more a figure stretches (i.e. the more a/b differs from 1) the more attention is attributed to the figure. Therefore, we took the ratio of the extension of a figure over the x and y axes. We set $a=0.26$ (compare [16]).

5.) Position (centrality): In an experiment conducted by Higuchi [2], the visual screen was divided into nine rectangles and subjects chose the targets in the middle rectangle with a higher frequency. Also Elias et. al. showed that TV-viewers eyes fixate on the center of the screen in most cases [24]. We calculate the euclidian distance from the center and set $a=-0.61$ (compare [16]).

6.) Distance from the observer: We are planning to include depth information later on.

Averaging or addition of factors would not allow extreme values for a particular feature to contribute proportionally more to the computed saliency. On the other side, excessive weight of some values, could bias the calculation and be affected by noise very easily. We therefore estimated reasonable bandwiths for each feature values, then squared and summed the so-obtained feature values.

IV. DISCUSSION

We built our system according to the principles that were stated above. We have tested our technique for a wide variety of images and results have been promising.

The computation works very efficient. After the segmentation, the localization of salient regions takes about 15 seconds in matlab on a Windows PC with 1.2 Ghz and 256 MB RAM.

In order to compare it to human attention point selection Muto and Kunii conducted a first experiment that is described in [9] and [10]. In the experiment, the stimuli were images of natural scenes (see fig. 8 for a sample). 60 human subjects had to choose three regions they thought to be most salient and as first, second, and third preference. In the appendix Fig. 8 shows an example for the stimuli used for the experiment. Fig. 7 shows how many people had particular regions among the first three choices for a particular. For many pictures, Muto and Kunii could obtain saliency gradients generated by the model that were very similar to the distribution of the results in experiment 1.

In a follow-up psychophysical experiment conducted by Muto [11], images depicting geometric figures or natural scenes were presented to subjects on a screen. 20 subjects were each looking at 400 pictures in total, 100 for each of the tested feature dimension. The tested feature dimensions were area (size), circular degree (shape), aspect ratio (length/breadth), and position (refer to section 4 for definitions and see figures 4,5,6 for sample stimuli). The pictures showed a target stimulus (comparison stimulus) next to a second figure that was a standard (standard stimulus) for which the subjects were told it had a saliency value of 100. The subjects had to weight the comparison stimulus according to how they thought it differed from the standard stimulus that was displayed next to it (compare figures 4,5,6).A comparison with the images showed that our system matches fairly well to the human attention point selection in the used set of images, though the data are not consistent enough to allow a prediction of matches for arbitrary images and more testing, probably using eye-trackers, is needed to give a reliable evaluation.

V. CONCLUSIONS

In this paper, we presented our model for robot vision that is inspired by many fascinating ideas about the human visual system. The model breaks down the problem of image processing by rapidly and efficiently finding conspicuous locations (attention points) to be analysed in more detail. We outlined how we identify salient regions. Further, in two experiments, for a set of pictures of natural scenes and geometric figures, attention points that were chosen by humans were evaluated, the attention points generated by our model compared. The attention points chosen by our model are generally very close to human intuition.

Our system could find future application in autonomous robots and could possibly facilitate tasks such as pathfinding, navigation, position sensing, and object recognition. It could further be used for active visual perception and attention in real-time, and automated target identification and acquisition systems.

VII. RESEARCH PERSPECTIVE

There are several directions in which we aspire to extend our research in future, several of them already indicated earlier in this paper. We want to improve our model and approach the human visual system in performance and experiment with different feature dimensions.

What we are working on in parallel is to combine our stage 1 with stage 2 processing that is being developed in our laboratory in order to build up towards a system which can perform many more tasks. We believe that the effectiveness of the human visual system originates in the characteristics of the attention point selection, with the interaction of bottom-up early processes and a top-down processes largely contributing (by changing parameters or selection weights or by inhibiting lower processes; c.f. [8]).

REFERENCES

[1] Cave, K.R. "Selection can be performed effectiveley without temporal binding, but could be even more effective with it", Psychology Press, part of the Taylor & Francis Group, Volume 8, Numbers 3-5/June, p. 467 – 487, 2001

[2] Higuchi, Y., "Object recognition from the actual image for autonomous movement of an intelligent robot", Japanese Society for Artificial Intelligence 2003

[3] Kandel, Eric R., Schwartz, James H., Jessell, Thomas M., Principles of Neural Science, McGraw-Hill/Appleton & Lange McGraw-Hill/Appleton & Lange, pp. 523-548, 2000

[4] Koyazu, T. "4th volume in memory of present age psychology", The University of Tokyo, research publication, 1982

[5] Kuffler, S. W., "Discharge patterns and functional organizationof mammalian retina", Journal of Neurophysiology, 16, 37-68, 1953

[6] Li, Zhaoping, "Contextual influences in V1 as a basis for pop out and asymmetry in visual search", Proc. Natl. Acad. Sci. USA, Vol. 96, pp. 10530–10535, August 1999

[7] Luck, S.J., Fan, S., Hillyard, S.A., "Attention-related modulation of sensory-evoked brain activity in a visual search task". Journal of Cognitive Neuroscience, 5, 188-195, 1993

[8] Moran, J. & Desimone, R., "Selective attention gates visual processing in the extrastriate cortex. Science 229. 782-4, 1994

[9] Muto, Y., Kunii, Y., "Observing attention point selection for natural feature recognition", the 22nd Robotics Society of Japan academic lecture meeting, 2004

[10] Muto, Y., Kunii, Y. "Study on attention point selection methods for natural feature recognition from a picture", a robotics mechatronics lecture meeting, 2004

[11] Muto, Y., Kunii, Y., "Examination of bottom-up visual attention for natural feature recognition", the 10th robotics symposium, 2005 (unpublished)

[12] Oyama, T., Psychology: 4 - perception, University of Tokyo Press 1977

[13] Oyama, T., Akita, S., "perceptual engineering", Fukumura publishing, 1989

[14] Oyama, T., Nanri, R., "The effect of hue and brightness of hue and brightness on the size of perception". Jap. Psychol. Res., 2, 13-20, 1960

[15] Privitera, C. M., Stark, L.W., "Evaluating Image Processing Algorithms that Predict Regions of Interest". Pattern Recognition Letters 19 (11), 1037-1043, 1998

[16] Into, Taro, Psychological measurement and learning theory, Morikita Publishing, 1977

[17] Treisman, A. M. and Gelade, G., "A feature-intergation theory of attention", Cognitive Psychology, Vol. 12, No. 1, pp. 97-136, 1980.

[18] Treisman, A., "Features and objects: the fourteenth Bartlett Memorial Lecture". Quarterly Journal of Experimental Psychology, 40A, 201-236, 1988

[19] Koenig, P., Einhaeuser, W., "Does luminance-contrast contribute to a saliency map for overt visual attention?", European Journal of Neuroscience, Vol. 17, pp. 1089-1097, 2003

[20] Vecera, S. P., O'Reilly, R.C., "Figure-Ground Organization and Object Recognition Processes – an interactive account", Journal of Experimental Psychology: Human Perception and Performance, Vol. 24, No. 2, 441-462, 1998

[21] Wolfe, J., Cave, K., Franzel, S., "Guided Search: An Alternative to the Feature Integration Model for Visual Search", Journal of Experimental Psychology in Human Perception an Performance 15, pages 419 – 433, 1989

[22] Wolfe, J.M., Horowitz, T.S, "What attribues guide the deployment of visual attention and how do they do it?", Nature Reviews Neuroscience, 5 1-7, 2004

[23] Yamamoto, H., Yeshurun, Y., and Levine, M. D., "A Foveated Vision System with Attentional Mechanisms", EiC(D-11), VoLJ77-D-11, No. 11, pp. 119-130, 1994.

[24] Elias, G., Sherwin, G., Wise, J., "Eye movements while viewing NTSC format television", SMPTE Psychophysics Subcommittee white paper, March 1984

[25] Mancas, M., Gosselin, B., Macq, B., "Segmentation using a region growing thresholding" ???

[26] Stentiford, F., "An estimator for visual attention through competitive novelty with application to image compression", Picture Coding Symposium 2001

APPENDIX

Continuum	Exp	Conditions
Loudness	0.6	Binaural
Brightness	0.33	5° target - Dark Adapted Eye
Lightness	1.2	Reflectance of Gray Papers
Smell	0.55	Coffee odor
Taste	1.3	Salt
Temperature	1.6	Warm-on arm
Vibration	0.95	60Hz - on Finger
Duration	1.1	White Noise Stimulus
Repetition Rate	1.0	Light,Sound,Touch,and Shocks
Finger Span	1.3	Thickness of Wood Blocks
Pressure on palm	1.1	Static force on skin
Heaviness	1.45	Lifted weights
Force of Handgrip	1.7	Precision Hard Dynamometer
Vocal Effort	1.1	Sound Pressure of Vocalization
Electric shock	3.5	60Hz through fingers

Table 1 - a compilation of some of the exponents measured for Stevens' power law (taken from [16])

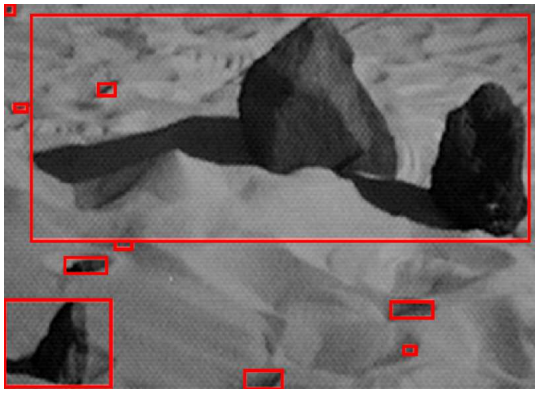

Fig.2: An illustration of the image segmentation of our model

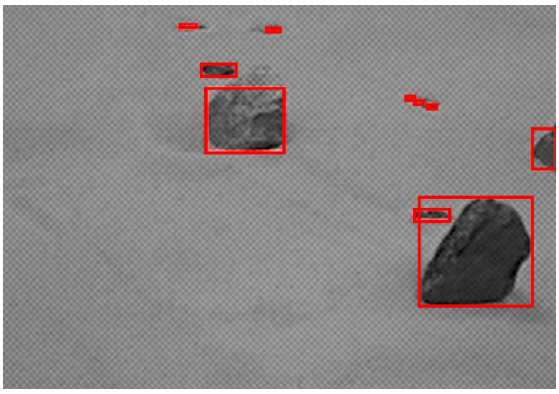

Fig.3: Another illustration of the image segmentation

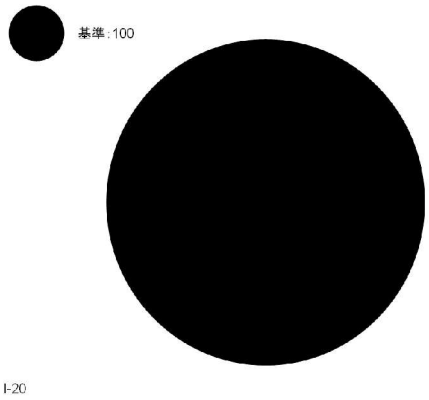

Fig.4: An example for the stimuli used in experiment 2. The independent variable was size.

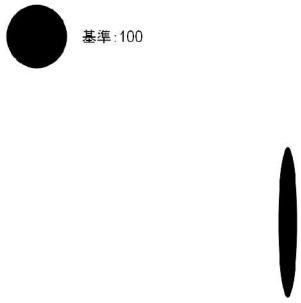

Fig.5: Another example of the stimuli from experiment 2. The independent variable was aspect ratio.

II-7

Fig.6: Example for the stimuli in experiment 2. Independent variable was circular ratio (shape).

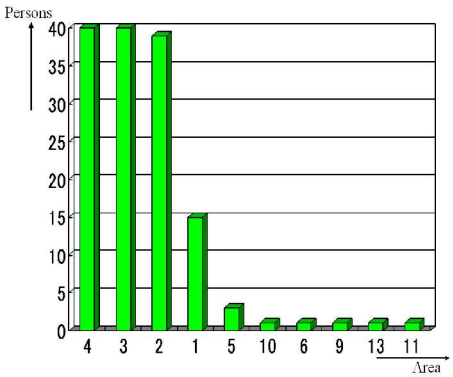

Fig.7: Sample answers in experiment 1. People were asked to circle on the paper the three regions they found most salient in the picture(shown below, Fig.8). The graph shows how many subjects chose a particular region (here called area) as one of the three most salient regions. Most subjects had similar choices. This was more or less consistent over the trials.

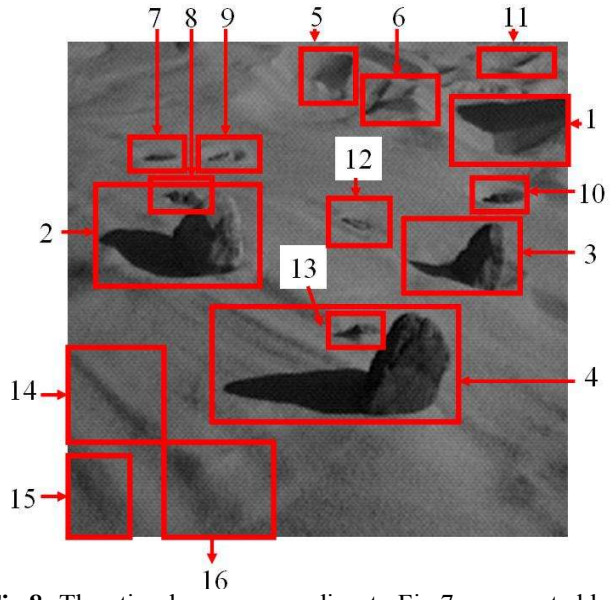

Fig.8: The stimulus corresponding to Fig.7, segmented by the model. Most people chose regions 4,3, and 2 to be most salient. For many pictures we obtained saliency gradients generated by the model that were very similar to the distribution of the results of experiment 1.