Computing Visual Attention from Scene Depth

Abstract

Visual attention is the ability to rapidly detect the interesting parts of a given scene. Inspired by biological vision, the principle of visual attention is used with a similar goal in computer vision. Several previous works deal with the computation of visual attention from images provided by standard video cameras, but little attention has been devoted so far to scene depth as source for visual attention. The investigation presented in this paper aims at an extension of the visual attention model to the scene depth component. A first part of the paper is devoted to the integration of depth in the computational model build around conspicuity and saliency maps. A second part is devoted to experimental work in which results of visual attention, obtained from the extended model and for various 3D scenes, are presented. The results speak for the usefulness of the enhanced computational model.

1. Introduction

Visual attention is the ability to rapidly detect the interesting parts of a given scene. Psychophysical studies show that it plays a fundamental role in human vision. Due to the biological structure of the retina, composed of a high resolution central part, the fovea, and a low resolution peripheral one, visual attention guides eye movements to place the fovea on the interesting part of the scene. The part of the scene imaged onto the fovea can then be processed in more details. Some biologically plausible models of attention have been presented [6] [2] [5].

Visual attention can be a useful preprocessing step in a computer vision task. Using such a step in computer vision systems permits a rapid selection of a subset of the available sensory information before further processing. The selected locations are supposed to represent the conspicuous parts of the scene, on which further computer vision tasks can focus. Therefore, an obvious application of visual attention is to reduce the computation cost of high level tasks like segmentation and object recognition, which are known to be complex, when achieved in a straightforward way. Fields of computer vision that can benefit from this task are, for example, industrial quality control, surveillance, autonomous mobile systems, etc.

Thus, several computational models of attention have been presented in previous works [4] [1]. Most of them are based on the *feature integration* principle [6]. Numerous features are extracted from the scene. According to each feature, conspicuous parts of the image are detected. A combination of the detected conspicuities gives rise to the final map of attention named saliency map. These saliency-based models apply to color images as input, as typically provided by a video camera. The considered features are, typically, *intensity, color* and *intensity gradient components* and may include multi-resolution.

Little attention has been devoted so far to scene depth as source for visual attention. This is considered a weekness of the models because depth or 3D vision is an intrinsic component of biological vision. As depth appears at an early stage in the visual system, it appears to contribute to visual attention. A further reason to include depth in the computational model of visual attention is the present availability of 3D range cameras.

A previous work where depth is considered as a source for attention is presented in [3]. A depth target mask, which corresponds to the depth conspicuity map in the saliency-based model of attention, is computed, based on histogramming. Peaks observed on the histogram of the disparity map are considered as conspicuous locations. This model has been used for example, to detect the closest, to the sensor, moving object. A limitation of the presented model is its task-dependency and the absence of a data-driven competitive mechanism to integrate the extracted features into the final map of attention.

The investigation presented in this paper aims at an extension of the bottom-up, task-independent, saliency-based model of visual attention to the scene depth component. A first part of the paper is devoted to the integration of depth in the computational model of attention developed for pure image vision. A second part is devoted to experimental work in which results of visual attention, obtained from the extended model and for various 3D scenes, are presented.

2. Visual attention model

2.1. Saliency-based model

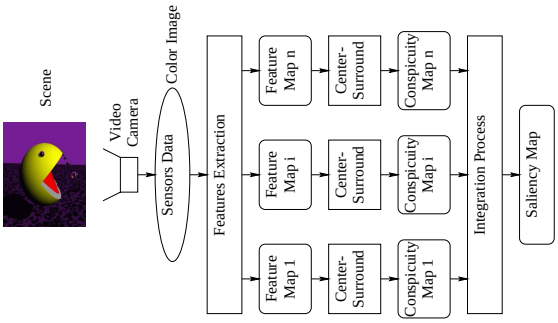

Figure 1. Scheme of a computational model of attention.

According to a generally admitted model of visual perception [4], a visual attention task can be achieved in three main steps (**Fig. 1**).
1) First, a number (n) of features are extracted from the scene by computing the so called feature maps. Such a map represents the image of the scene, based on a well-defined feature. This leads to a multi-feature representation of the scene. The features most used in previous works are intensity, color components, and intensity gradient components (norm and orientation).
2) In a second step, each feature map is transformed in its conspicuity map. Each conspicuity map highlights the parts of the scene that strongly differ, according to a specific feature, from its surrounding. In biologically plausible models, this is usually achieved by using a *center-surround*-mechanism. Practically, this mechanism can be implemented with a *difference-of-Gaussians*-filter, which can be applied on feature maps to extract local activities for each feature type.
3) In the last stage of the attention model, the n conspicuity maps are integrated together, in a competitive way, into a *saliency map* $\mathcal{S}$ in accordance with equation 1.

$$\mathcal{S} = \sum_{i=1}^{n} w_i \mathcal{C}_i \tag{1}$$

The competition between conspicuity maps is usually established by selecting weights w_i according to a weighting function w, like the one presented in [1]: $w = (M - \overline{m})^2$, where M is the maximum activity of the conspicuity map and $\overline{m}$ is the average of all its local maxima. w measures how the most active locations differ from the average. Thus, this weighting function promotes conspicuity maps in which a small number of strong peaks of activity is present. Maps that contain numerous comparable peak responses are demoted. It is obvious that this competitive mechanism is purely data-driven and does not require any a priori knowledge about the analyzed scene.

2.2. Multi-resolution visual attention

A visual attention task has to detect interesting objects, regardless of their sizes. Thus, a multi-scale conspicuity operator is required. It has been shown in [4], that applying variable size center-surround filter on fixed size images, has a high computational cost. An interesting method to implement the *center-surround*-mechanism, which is used to compute the conspicuity maps, has been presented in [1]. This method is based on a multi-resolution representation of images. For each feature, nine spatial scales are created using gaussian pyramids, which progressively lowpass filter and subsample the feature map. Center-Surround is then implemented as the difference between fine and coarse scales. The center is a pixel at scale $c \in \{2, 3, 4\}$ and the surround is the corresponding pixel at scale $s = c + \delta$ and $\delta \in \{3, 4\}$. Consequently, six maps $\mathcal{F}(c, s)$ are computed for each pyramid $\mathcal{P}$ (**Eq. 2**).

$$\mathcal{F}(c, s) = |\mathcal{P}(c) - \mathcal{P}(s)| \tag{2}$$

A weighted sum of the six maps $\mathcal{F}(c, s)$ results into a unique conspicuity map for each pyramid and, consequently, for each feature. The maps are weighted by the same weighting function w as described above.

3. Visual attention with depth

In addition to visual features like intensity, color components and intensity gradient components, this section introduces depth as new input. Two problems must be solved in order to integrate depth in the model of attention: *1)* Which features, related to depth, can be integrated in the model. *2)* How can these features be integrated.

3.1. Integration of Depth into the model of attention

The basic idea is to simply extend the multi-resolution model of visual attention, described above, to the scene depth component. Given m suitable features related to depth, the integration process can be achieved as follow. First, the feature maps are extracted from the depth data acquired by a range finder. The corresponding conspicuity maps are then computed. Hence, besides the n conspicuity maps computed from the color image, m additional ones, related to depth, are available. The integration module has to combine $n + m$ conspicuity maps in order to compute the saliency map and, consequently, $n + m$ features are

taken into account (**Fig. 2**). Equation 1, which has been used in classical models to compute the saliency map, can be adapted to:

$$\mathcal{S} = \sum_{i=1}^{n+m} w_i \mathcal{C}_i \qquad (3)$$

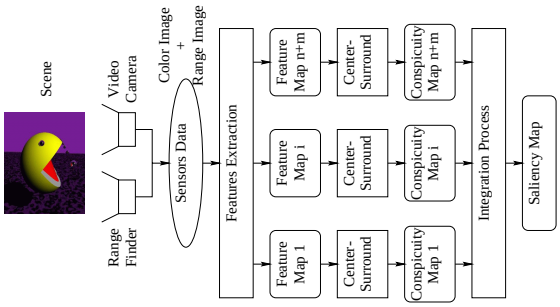

Figure 2. Scheme of a computational model of attention, considering depth.

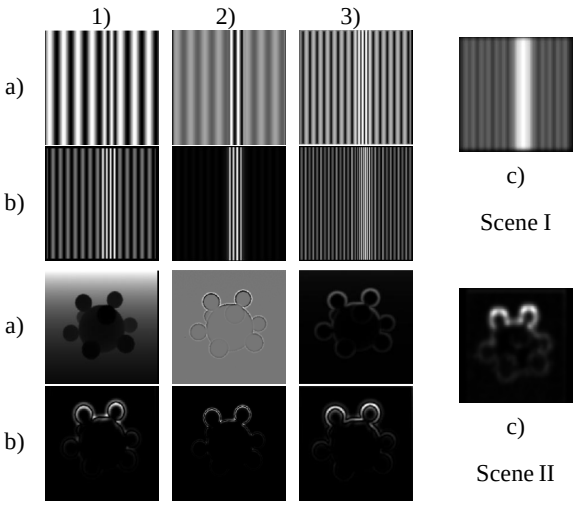

Figure 3. Computing the saliency map using *1) depth*, *2) mean curvature* and *3) depth gradient norm*. a) feature maps, b) corresponding conspicuity maps and c) saliency map.

3.2. Choice of features

This work considers three potential features related to depth.

1) Depth: measures the distance from camera to the objects in the scene. It is obviously relevant and directly available from the sensor data. Thus, no additional operators are needed to extract this feature.

2) Mean curvature: is an intrinsic surface feature that provides useful information about the geometry of the scene objects. As a second differential order feature, mean curvature has, however, a remarkable disadvantage, i.e. its sensitivity to noise and non significance on depth discontinuities. The disadvantage related to noise sensitivity can be overcome through applying smoothing operators on the range image. Depth discontinuities have to be detected in order to compute mean curvature only for continuous surfaces. Thus, integrating mean curvature in the computational model of visual attention requires some additional preprocessing operations.

3) Depth Gradient: This feature vector, based on first order derivative, can be an efficient means to detect important depth changes in the scene like angles and corners.

Experiments were carried on in order to assort the usefulness of these various features. Observations made with various real and synthetic range images tend to show a ranking of the features which is depth, mean curvature, gradient norm in an order of decreasing usefulness.

Figure 3 illustrates some observations. The first scene (I) contains no depth discontinuities, but important curvature variation. Thus, mean curvature contributes strongly to the computation of the saliency map. The second experiment (scene II) shows the sensitivity of mean curvature to depth discontinuities. It shows also the lower significance of gradient norm.

In the experiments presented in next section only the feature *depth* has been , finally, considered. Because of its low usefulness in the analysis of the considered scenes, depth gradient norm has not been considered and mean curvature has not been taken into account because of the current unavailability of the required additional operators.

4. Experiments and evaluation

This section presents some targeted experiments carried on in order to validate the depth enhanced computational model of visual attention and to show the usefulness of depth information in a visual attention task. Two features are considered in these experiments, *color* and *depth*.

Each scene considered in the experiments (**Fig. 4**) is represented by its color and its range image (left). Next to each feature map, the corresponding conspicuity map is represented. The two conspicuity maps are then combined, according to (**Eq. 3**), into the saliency map. Clearly, depth contributes significantly to visual attention since the depth enhanced model detects depth locations which stand out from their surrounding. In scene a) one attention spot is detected that stems from color contrast. In scene c) and d) two spots of attention are detected stemming from color contrast and depth contrast. Each spot is caused by a different fea-

ture. Thus, both features contribute, equally, to the saliency map.

Regarding the model's operation, let us analyze scene b). There, despite the presence of high color contrast, only depth contrast contributes to the detection of one significant spot of attention. This is due to the competition between the two features, which takes place during their integration process through the weight w_i that is assigned to each conspicuity map. The contribution of a feature to the saliency map increases with the assigned weight. The weighting function used in the saliency-based model of attention promotes conspicuity maps that contain one peak response. It, however, demotes maps that contain several peak responses. When both conspicuity maps contain, however, almost the same number of peak responses, they are both promoted by the weighting function. This explains the results of the experiments c) and d), where each conspicuity map contains one remarkable peak response.

By showing the clear usefulness of depth in scene analysis and its successful operation in a two-channel competitive task of visual attention, these experiments validate the depth enhanced computational model of attention.

5. Conclusion

This paper proposed an extension of the saliency-based computational model as a means to also consider scene depth as a feature for visual attention. In order to test the model, visual attention, computed from the two features color and depth, was analyzed for a number of scenes. The results validate the model, namely by showing the significance and the effectiveness of channel competition. This is considered a key element for approaching further applications involving a larger number of features.

References

[1] 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)*, 1998.

[2] C. Koch and S. Ullman. *Shifts in Selective Visual Attention: Towards the underlying Neural Circuity. In L.M. Vaina (edt), Matters of Intelligence, pp. 115-141.*, 1987.

[3] A. Maki and J. Eklundh. *A Computational Model of Depth-Based Attention. ICPR*, 1996.

[4] R. Milanese. *Detecting Salient Regions in an Image: from Biological Evidence to Computer implementation. Ph.D. Thesis, Dept. of Computer Science, University of Geneva*, Dec. 1993.

[5] C. Oldhausen, B. Anderson and D. Van Esseni. *A Neural Model of Visual Attention and Invariant Pattern Recognition. California Institute of Technology, Computation and Neural System Program, CNS Memo 18*, August 1992.

[6] A. Treisman and G. Gelade. *A feature-integration theory of attention. Cognitive Psychology, pp. 97-136*, 12, 1980.

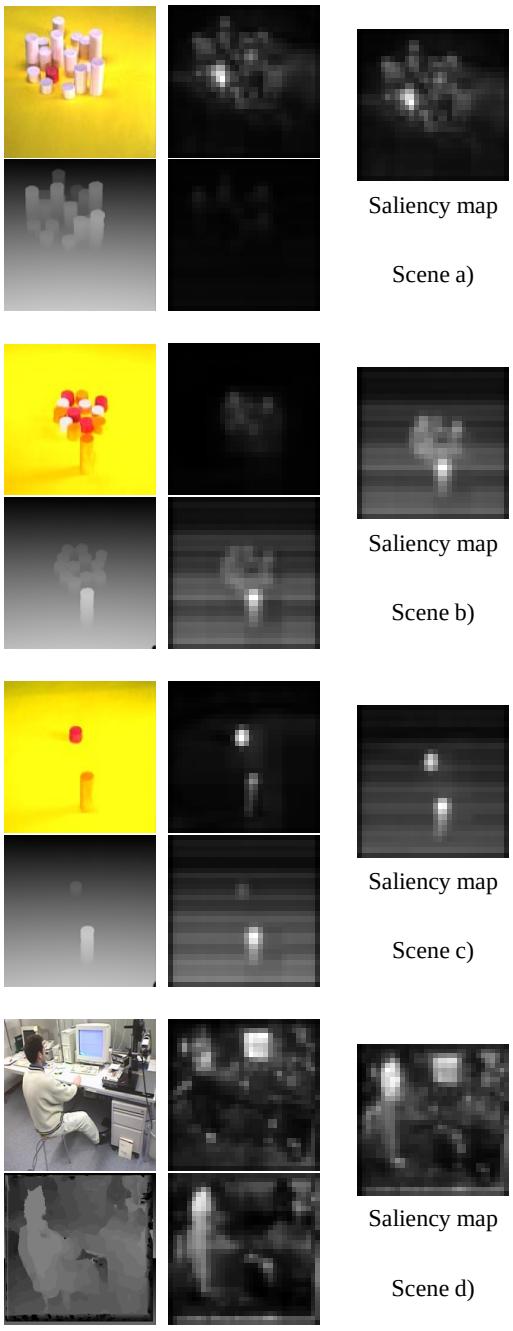

Figure 4. Detecting conspicuous objects from various 3D scenes, using *color* and *depth*.