DEPTH-BASED DETECTION OF SALIENT MOVING OBJECTS IN SONIFIED VIDEOS FOR BLIND USERS

Benoît Deville[1], Guido Bologna[2], Michel Vinckenbosch[2] and Thierry Pun[1]
[1]*Computer Vision and Multimedia Lab, University of Geneva, Route de Drize 7, Carouge, Switzerland*

[2]*Laboratoire d'Informatique Industrielle, University of Applied Science, Geneva, Switzerland*

Keywords: Focus of attention (FOA), visual saliency, depth map, stereo image processing, mobility aid, visual handicap.

Abstract: The context of this work is the development of a mobility aid for visually impaired persons. We present here an original approach for a real time alerting system, based on the use of feature maps for detecting visual salient parts in images. In order to improve the quality of this method, we propose here to benefit from a new feature map constructed from the depth gradient. A specific distance function is described, which takes into account both stereoscopic camera limitations and users choices. We demonstrate here that this additional depth-based feature map allows the system to detect the salient regions with good accuracy in most situations, even with noisy disparity maps.

1 INTRODUCTION

See ColOr is a system that aims at creating a mobility aid for people who lost their vision. With the use of spatialized musical instrumental sounds, visually impaired users will be able to hear an auditory reprensentation of the environment in front of them; parts of this system are in the testing phase (Section 2). However, as image points are represented by sound sources and typical cameras capture hundreds of thousand pixels, it is not feasible to transcribe the whole scene without risking to create a cacophony that would lead to miss important information. This is why we developed an alerting system to attract the user's attention towards regions of importance.

This alerting system is based on a visual attention model, built from conspicuity maps. A new feature map, i.e. the depth gradient, which to our knowledge has not been used before in this context, is added to the model. The purpose of this feature map is to detect objects that come towards the blind user, and that should be avoided. A distance function is integrated in our model. It lets users decide from which distance objects should be detected ; this function allows to take into account both stereoscopic camera limitations in distance computation, and user's choices.

The focus of attention (FOA) is a cognitive process that can be described as the decision to concentrate on one or more senses (e.g. both touch and vision) on a specific object, sound, smell, etc. This is for instance the case with the well known *cocktail party effect*, where one can follow a particular conversation out of a set of multiple and diverse sounds. This ability to focus on a specific point is guided by our senses, and our interests in a given situation. We are interested here in the dectection of salient areas from video sequences; we ground this detection on specific visual properties of the scene, namely distance and depth gradient.

It is shown here that the use of the depth gradient is an important feature in a mobility aid system. It obviously helps in the cases where objects might disturb the movements of a blind user. We also demonstrate that the combination of depth gradient and distance function improves the results in cases where objects in movement are not a threat for the user.

The article is organized as follows. In Section 2, we briefly describe the See ColOr system, and explain how colours are transformed into musical instrument sounds. The inherent limitation of this approach, that is the sonification of only parts of the scene, leads to the need for a visual saliency scheme, summarized in Section 3. In that section, we introduce some known methods that use visual attention models to infer a FOA. We then describe our approach based on distance function and depth gradient. Section 4 analyses the results provided by our method in comparison with the use of depth information only Section 5 con-

Deville B., Bologna G., Vinckenbosch M. and Pun T. (2008).
DEPTH-BASED DETECTION OF SALIENT MOVING OBJECTS IN SONIFIED VIDEOS FOR BLIND USERS.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 434-439
DOI: 10.5220/0001084104340439

cludes on these results, and offers some suggestions on improvements that are to be made.

2 SEE COLOR

The objective of the See ColOr project (Bologna et al., 2007a; Bologna et al., 2007b) is to develop a non-invasive mobility aid for blind users. The system uses auditive means to represent frontal image scenes which are recorded by a stereoscopic camera. Each colour is mapped into a set of up to three different musical instruments with varying pitch, depending on the three parameters of the colour in the HSL colour space. Thus, it allows a user who lost his/her vision to have access to the colour information of the environment through a synesthetic process.

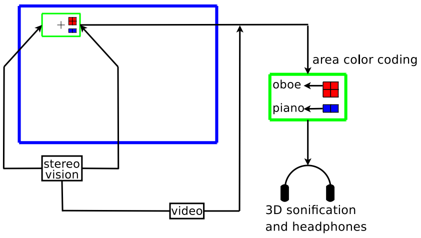

Figure 1: The See ColOr general framework.

The system does not sonify the recorded image as a whole. Indeed, this would create misunderstanding because a usual scene is composed of many different colours, which would lead to a high number of instruments playing at the same time. Since only a small part of the image is actually sonified, the risk of missing important parts of the scene is however not negligible. The effect of such a way of sonifying the scene can be considered similar to the tunnel vision. Furthermore, a previous experiment (Bologna et al., 2007b) showed that it was not so easy for blindfolded users to detect small regions on images, on the basis of hearing as well as with the help of a tactile display. In fact, they needed from 4 to 9 minutes to find a red door on a churchyard static image using only a tactile version of this picture and the auditive output depending on the touched part of the image. These doors represent about 1% of the total surface area of the picture (Figure 2), but are obviously visually salient. For this reason an alarm system based on the mechanism of visual saliency is being developed. This mechanism allows the detection of parts of the scene that would usually attract the visual attention of sighted people. Once the program has detected such saliencies, a new sound will indicate to the blind user that another part of the scene is noteworthy.

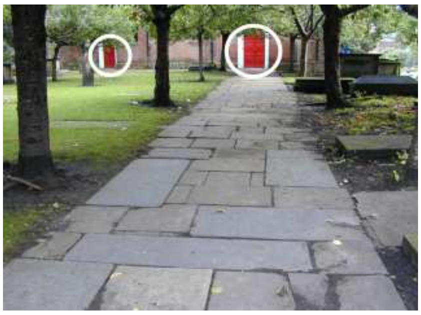

Figure 2: Blindfolded users had to find any of the two red doors (circled) using only sound and tactile information. Despite their saliency, the red doors only occupy 1.1% of the image surface.

3 VISUAL SALIENCY

Saliency is a visual mechanism linked to the emergence of a figure over a background (Landragin, 2004). During the pre-attentive phase of the visual perception, our attention first stops on elements that arise from our visual environment, and finally focuses the cognitive processes on these elements only. Different factors enter into account during this process, both physical and cognitive. The See ColOr project only focuses on physical factors. As a matter of fact, blind users will use their own cognitive abilities to understand the captured scene, given their personal impressions, their particular knowledge of the environment (e.g., if the user is inside or outside), and the sonified colours. Physical factors directly depend on the perceived scene and the characteristics of the objects that compose it. Lightness contrast, opposition of colours (e.g. red/green and blue/yellow), geometrical features, singularity in a set of objects or in an object itself (Hoffman and Singh, 1997) are some examples of these physical factors. Different computerized approaches have been designed to digitally reproduce this human ability, and we will briefly present one of them in the following.

3.1 Conspicuity Maps

In order to detect salient regions, some methods will center on specific characteristics of images like entropy (Kadir and Brady, 2001) or blobs, detected with *Difference of Gaussians* (DoG) (Lowe, 1999) or the speeded up robust features (SURF) (Bay et al., 2006), a simplified form of the *Determinant of Hessian* (DoH) approach. Other methods, based on conspicuity maps (Milanese et al., 1995; Itti et al., 1998), try to mimic the physical properties of the Human Visual System (HVS).

Features inspired by the HVS are analysed separately to build a set of maps called *feature maps*, denoted F_i. These maps are filtered so that the *conspicuity map* (*C-map*) of each feature map only contains information concerning the few regions that diverge the most from their neighbourhood. All C-maps are then combined by a *Winner-Take-All* approach in order to determine the most salient region of the recorded scene. Figure 3 summarizes the extraction of the saliency map S on a colour image using conspicuity maps.

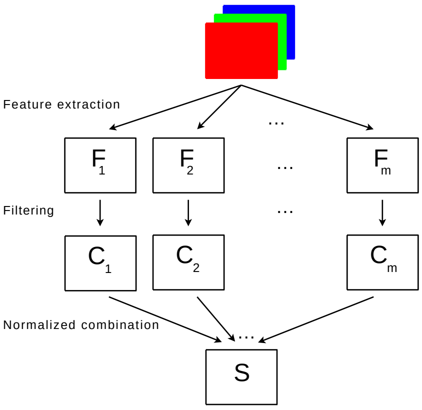

Figure 3: General overview of the conspicuity map approach for saliency detection. S being the final saliency map, and F_i, C_i the feature maps and their associated conspicuity maps, respectively.

Salient region detection using conspicuity maps have been proved to be efficient, both in terms of relevance of detected regions and speed, thanks to the increasing computing power of CPUs.

3.2 Depth Map

Depth is an important information when one has to decide whether an object is of interest or should be ignored. Close objects might be dangerous or interesting, thus implying an action from the user. Despite the importance of such an information for the FOA, the literature does not report much on approaches making use of depth. Very few methods takes depth into account to guide the FOA.

The first to propose the use of depth in an attention model (Maki et al., 1996) were only interested in determining the closest object from the user. Each time an object was closer than the previous match, the attention model would simulate a saccade, just as a human would do. The problem of such an approach is the absence of other important features, such as colour opposition, edge magnitude, illumination, etc.

It has been suggested later to use depth in the usual bottom-up approach of saliency-based visual attention models (Ouerhani and Hügli, 2000; Jost et al., 2004). They proved the interest of depth information as a new feature map, combined with common feature maps like colour opposition, intensity, and intensity gradient components. However, this does not give any information about objects movements, which is very important for a mobility aid, particularly when an object comes towards the user. This is why we propose here to combine depth and depth gradient. These features are computed and combined as follows.

Given a 3D point $\mathbf{p} = \{p_x, p_y, p_z\}$, we consider that close objects are more important than others. Nevertheless, an object distant from less than d_{min} (for instance $d_{min} = 1$ meter) should be detected by the blind user's white cane, which implies the following distance function when we compute the value of the feature map F_d:

$$F_d(\mathbf{p}) = \begin{cases} d_{max} - p_z, & \text{if } d_{min} < p_z \leq d_{max} \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

where d_{max} is the maximal considered distance from the user. This parameter depends on the environment, the stereoscopic device, and the user's choice.

The depth gradient, in order to contain the information of movement, is computed over time. Since we consider that the only objects that are noteworthy in term of gradient are the ones that get closer to the user, we obtain the following gradient function :

$$F_\nabla(\mathbf{p}) = \begin{cases} -\frac{\partial z}{\partial t}, & \text{if } \frac{\partial z}{\partial t} < 0 \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

To get the conspicuity map C_i from F_i, we look for maxima/minima of a DoG applied on F_i:

$$C_i = argminmax\left(F_i * g_{\sigma_1} - F_i * g_{\sigma_2}\right) \qquad (3)$$

where g_{σ_1} and g_{σ_2} are Gaussians at scale σ_1 and σ_2, respectively. For an image I, we consider the saliency map as the weighted sum of the conspicuity maps :

$$S_I = \lambda_d \cdot C_d\,(I) + \lambda_\nabla \cdot C_\nabla\,(I) \qquad (4)$$

where C_d and C_∇ are the conspicuity maps computed from the feature maps F_d and F_∇ respectively, and λ_d, λ_∇ ($\lambda_d + \lambda_\nabla = 1$), constants that determine the importance of each feature. The final point of interest, which will lead the user's FOA, will be the point of highest value in the obtained saliency map.

4 EXPERIMENTS AND EVALUATION

In this section we present some specific experiments carried out to validate the hypothesis that depth gradient is a useful information in order to guide the FOA. Depth itself having been proved to be of interest in such a task (Ouerhani and Hügli, 2000), we consider the combination of distance and depth gradient features relatively to the depth feature alone.

4.1 Hypotheses

Four different cases are studied, and for each case, we expect a specific result.

Completely Static Both camera and scene are static: depth gradient should not give any additional information because no movement is present.

Camera Moving The camera moves towards the scene: depth gradient should not give additional information because the depth gradient's feature map is uniform, and no conspicuity arises from it.

Object Coming An object comes from the background and crosses the camera field of view: depth gradient should indicate the user as the most salient region.

Object Leaving Same as previous, but the object leaves the scene, from foreground to background: depth gradient should first indicates the object as the most salient region if it is the closest object from the camera, then no more additional information should be given by the gradient.

4.2 Results and Discussion

In order to check the validity of the hypothesis, results obtained with the combination of distance and depth gradient have been compared to depth based results for each case.

The method described in the previous section has been tested on four different video sequences. These videos were recorded using a STH-MDCS2[1] stereoscopic camera, and its development library, which computes in real time the disparity map needed for the determination of the depth. The resulting disparity map is unfortunately far from perfect : the depth information is unaccurate or undetermined at many points of the scene. However, given the importance of real time in an alerting system, the presented method is performed with this raw information.

The results for each hypothesis describes previously are presented and analysed, and summarized in Table 1. It shows the percentage of *good match* for each sequence. The result on each picture of a video sequence is said to be a good match whenever the expected area (defined in Section 4.1) of the picture is determined as the most salient. This table also points out that the proposed method is almost real time, since it only takes around 90 ms to process each picture in a video sequence. A framerate of 11 image per second is sufficient for an alerting system, and for the See ColOr framework, where images are sonified around three times per second.

Table 1: Percentage of good matches on four different video sequences, with depth alone or with a combination of depth and depth gradient, and average computating time in ms.

Case	Depth alone	Depth and depth gradient	Average CPU time
1	72.8%	93.1%	97.3ms
2	70.0%	63.8%	94.8ms
3	26.7%	49.3%	85.0ms
4	56.4%	75.2%	84.1ms

4.2.1 Camera and Scene are Static

The sequence shows an office with different objects (Figure 4). The closest object that can be detected by the camera is a desk in the enlightened part of the scene. Due to limitations of the camera used for the recording, the closest object —the chair—, cannot be selected. Even if the computation of stereoscopic disparity creates many artifacts in the depth space, the desk is still determined as the most salient object 70% of the time. Here, the depth gradient is a mean to smooth these artifacts, and allows the detection of the desk in more than 90% of the sequence, which means an increase of more than 25% of accuracy than when depth alone is used.

4.2.2 Camera Moves

The camera moves towards some plants on the background (Figure 5). Again, even with the artifacts created by the disparity computation, we obtain 70% of good match. Including the gradient however reduces the accuracy of results, because the movement is not constant over the whole image. Thus some background objects have a higher gradient than foreground ones, because they are farther away from the center of the transformation the camera is subject to. An estimation of the camera motion could then give even better results than the ones we get.

[1] Videre Design: http://www.videredesign.com

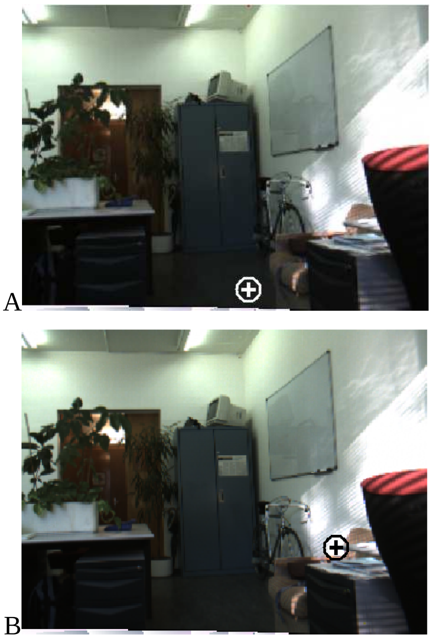

Figure 4: Results in the static case on the same image of the video sequence, using depth only (A) or depth and gradient (B). The detected FOA is indicated by a cross surrounded by a cirle.

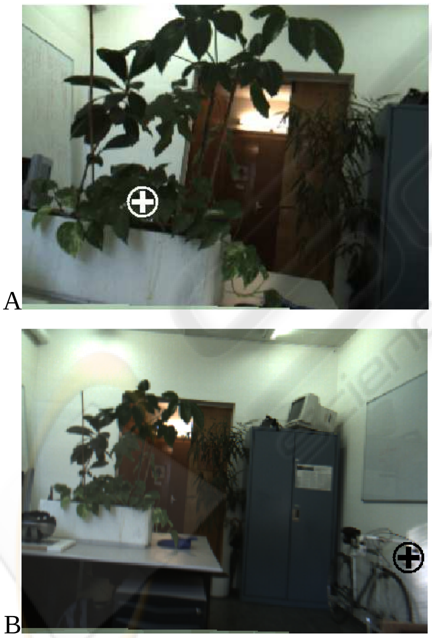

Figure 5: Good match (A) and false positive (B) when the camera moves. The detected FOA is indicated by a cross surrounded by a cirle.

4.2.3 Somebody Comes Towards the Camera

A person enters the office, walks towards the camera (Figure 6), arrives close to a chair in the foreground, then sits, partially out of the field of vision. In this case, when using only depth, the objects defined as salient are the closest ones. This means that along the video sequence, the character is determined as the most salient area only when touching the closest object, which is the case in approximately a quarter of the video sequence. The use of depth gradient is clearly of interest in this case, since it doubles the number of times the person is detected. Mostly, the incorrect selected object is the chair because it is the closest. Thus, giving more importance to the depth gradient in Equation 4 would increase the selection of the object of interest.

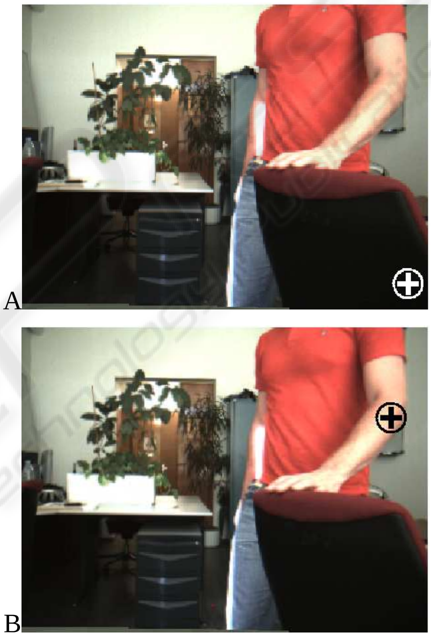

Figure 6: Results obtained when a person comes towards the camera, with depth only (A) or depth and gradient (B), on the same image of the video sequence. The detected FOA is indicated by a cross surrounded by a cirle.

4.2.4 Somebody Leaves the Scene

The character is sitting on a chair, which is the closest object, thus the one to find as the most salient (Figure 7). This person rises, walks towards the background of the scene, and finally disappears. Like in the first case (where scene and camera are static), the disparity computation creates a lot of artifacts. Then, depth gradient acts as a smoothing filter, and drastically improves results that are already interesting. In fact, in more than three images over four in the sequence, the expected object is selected as the most salient.

Finally, the results obtained with this framework

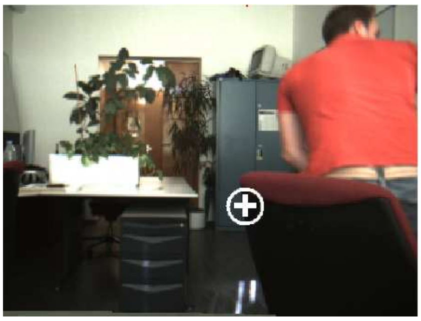

Figure 7: The detected FOA (indicated by a cross surrounded by a cirle) is the closest object from the camera.

are promising, knowing that the computation of disparity is a challenging task, specially in real time. Some optimizations on the distance function, the computation of the conspicuity map from the depth gradient feature map, and the combination of the different feature maps will likely lead to better results.

5 CONCLUSIONS AND PERSPECTIVES

In order to develop a mobility aid for blind people, we have presented in this article a new approach to detect salient parts in videos, using a depth based FOA mechanism. Depth gradient is introduced as a new feature map in an already known visual attention model. We have proved that this feature map allows for a better detection of objects of interest in video sequences when using depth only, as proposed in previous works (Ouerhani and Hügli, 2000; Jost et al., 2004). We have also proposed a specific distance function, in order to take into account both hardware limitations and user's choices ; this allows the user to decide if objects closer than his/her cane should be detected or not. The results we obtained with this simple framework are promising, and some optimizations such as a more realistic distance function or the determination of optimal coefficients in Equation 4, should lead to even better results.

Ongoing and future work concerns the following. First, we will integrate some usual feature maps like colour opposition, flicker, or motion, to ensure that the depth gradient brings useful information in a visual attention model. The presented method will then be integrated in the See ColOr framework. It is particularly important to decide how the salient area will be sonified ; we do not want the user to be confused by similar sounds meaning completely different things. Once this will be done, the system will finally be evaluated by blind and blindfolded users.

ACKNOWLEDGEMENTS

We gratefully acknowledge the financial support of the Swiss Hasler Foundation.

REFERENCES

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). Surf: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision*, pages 7–13.

Bologna, G., Deville, B., Pun, T., and Vinckenbosch, M. (2007a). Identifying major components of pictures by audio encoding of colors. In *IWINAC2007, 2nd. International Work-conference on the Interplay between Natural and Artificial Computation*.

Bologna, G., Deville, B., Pun, T., and Vinckenbosch, M. (2007b). Transforming 3d coloured pixels into musical instrument notes for vision substitution applications. *EURASIP Journal on Image and Video Processing*.

Hoffman, D. and Singh, M. (1997). Salience of visual parts. *Cognition*, 63:29–78.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machcine Intelligence*, 20(11):1254–1259.

Jost, T., Ouerhani, N., von Wartburg, R., Müri, R., and Hügli, H. (2004). Contribution of depth to visual attention: comparison of a computer model and human. In *Early cognitive vision workshop, Isle of Skye, Scotland*.

Kadir, T. and Brady, M. (2001). Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105.

Landragin, F. (2004). Saillance physique et saillance cognitive. *Cognition, Reprsentation, Langage*, 2(2).

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Seventh International Conference on Computer Vision (ICCV'99)*, volume 2.

Maki, A., Nordlund, P., and Eklundh, J. (1996). A computational model of depth-based attention. In *Proceedings of the International Conference on Pattern Recognition (ICPR '96)*.

Milanese, R., Gil, S., and Pun, T. (1995). Attentive mechanism for dynamic and static scene analysis. *Optical Engineering*, 34(8):2428–2434.

Ouerhani, N. and Hügli, H. (2000). Computing visual attention from scene depth. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 1, pages 375–378.