Guiding the focus of attention of blind people with visual saliency

Benoit Deville, Guido Bologna, Michel Vinckenbosch, Thierry Pun

Guiding the focus of attention of blind people with visual saliency

Benoît Deville[1], Guido Bologna[2], Michel Vinckenbosch[2], and Thierry Pun[1]

[1] University of Geneva,
Computer Vision and Multimedia Lab,
Route de Drize 7,
CH-1227 Carouge, Switzerland
Benoit.Deville,Thierry.Pun@cui.unige.ch
[2] Laboratoire d'Informatique Industrielle,
University of Applied Science,
CH-1202 Geneva, Switzerland

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 detection of visual salient parts in videos. The particularity of our approach lies in the use of a new feature map constructed from the depth gradient. A distance function is described, which takes into account both stereoscopic camera limitations and user's choices. We also report how we automatically estimate the contribution of conspicuity maps, which enables the unsupervised determination of the final saliency map. We demonstrate here that this additional depth-based feature map allows the system to detect salient regions with good accuracy in most situations, even in the presence of noisy disparity maps.

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

1 Introduction

In this paper we present an attention model applied to the See ColOr system [1] which aims at creating a mobility aid for people who lost their vision. The objective is to help them creating a mental representation of their environment using the auditory pathway, with colours being mapped into a 3D virtual environment of spatialized musical instrument sounds. In this manner, to a given colour is associated a set of musical instruments, played in the virtual sound space. Instruments' sounds originate from a location corresponding to the colour location in the scene: with a suitable Head Related Transfer Function (HRTF), one can hear the sound as if the sonified region was actually emitting sound. For instance, the user would hear a piano playing on the left if a blue region was on the left part of the recorded scene. 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 missing important information. This is why we have developped an alerting system to attract the user's attention towards regions of interest.

The focus of attention (FOA) is a cognitive process that can be described as the ability to concentrate one or more senses (e.g. both touch and/or vision) on a specific object, sound, etc. The *cocktail party effect* is a well known example: one can follow a specific 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 focus here on the dectection of salient areas from video sequences; we ground this detection on specific visual properties of the scene.

Besides its unusual context, i.e. the development of a mobility aid for blind users, the novelty of this work resides in the use of depth gradient as a feature, as well as in the unsupervised and near real-time determination of the saliency map. The alerting system is based on a specific model of bottom-up saliency-based visual attention, namely the conspicuity maps. A conspicuity map contains information about regions of an image that differ from their neighbourhood. Each conspicuity map is built according to a specific feature, which can consist of colours, orientations, edges, etc. In this paper we have combined the depth gradient feature with distance, illumination intensity, and two colour opponencies. To our knowledge, the depth gradient has never been used before to build conspicuity maps. The purpose of the depth gradient conspicuity map is to detect objects that come towards the blind user, and that need to be avoided. In order to let the user control the range of interest, a distance function is integrated in our model. This function also allows to take into account camera limitations in distance computation. 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.

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, and with other features. Section 5 concludes on these results, and introduces some suggestions on improvements that could be made.

2 See ColOr

The objective of See ColOr[3] [1, 2] is to develop a non-invasive mobility aid for blind users, based on visual substitution by the auditory channel. While there exist devices using other rendering modalities than sound, such as haptic and

[3] A video of the current prototype can be found at:
`http://cvml.unige.ch/downloads/socks-lines.wmv`

audio-haptic [3], sound seems nevertheless the most promising modality for rendering the image, because of its simplicity of use and implementation. Different contributions have deeply marked this area, like the *K Sonar-Cane* [4], *The Voice* [5], or a more musical one like the model introduced by Cronly-Dillon *et al.* [6].

Our system uses auditive means to represent frontal image scenes which are recorded by a camera returning both colour and depth information. This can be, for instance, a stereoscopic or a time-of-flight 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 vision to have access to the colour information of the environment through a synesthetic process. The system does not sonify the

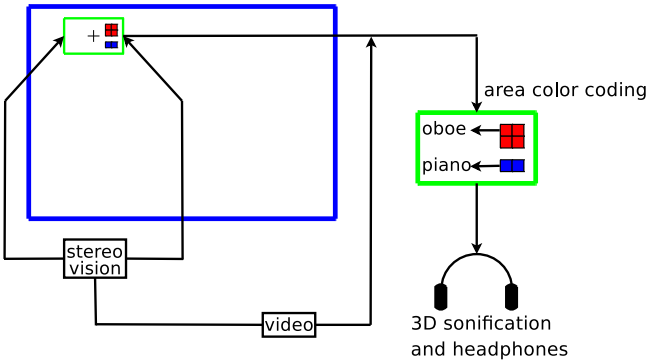

Fig. 1. The See ColOr general framework.

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 very large 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 tunnel vision. Furthermore, experiments carried out by Bologna [1] showed that the detection of small regions of interest was tricky for blindfolded users. In fact, they needed from 4 to 9 minutes to find a red door on a churchyard static image using both a tactile version of this picture and the auditive output of 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 specific sound indicates that another part of the scene is noteworthy. The sound being spatialized in a virtual sound space, the user knows where to focus his/her attention.

Fig. 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 an object over a background [7]. During the pre-attentive phase of the visual perception, the 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, i.e. linked to the iconic features of the scene, and cognitive. Several computerized approaches have been designed to digitally reproduce this human ability but only few methods [8, 9] combine these two type of factors. Methods that focus only on physical factors are called *bottom-up* approaches, while cognitive based ones are names *top-down* approaches.

The See ColOr project is based on a top-down attention model because the aim of the system is not to replace the blind user's 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, the users will be able to identify their environment. Physical factors directly depend on the perceived scene and the characteristics of the objects that compose it. Lightness contrast, colour opponencies (e.g. red/green and blue/yellow), geometrical features, singularity in a set of objects or in an object itself [10] are some examples of these physical factors. We now briefly present the theoretical framework which is the basis of the model we propose.

3.1 Conspicuity maps

In order to detect salient regions, some methods center on specific characteristics of images like entropy [11] or blobs, detected with *Difference of Gaussians* (DoG) [12] or the speeded up robust features (SURF) [13], a simplified form of the *Determinant of Hessian* (DoH) approach. Methods based on conspicuity maps [14, 15] 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 differ 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. Salient region detection using

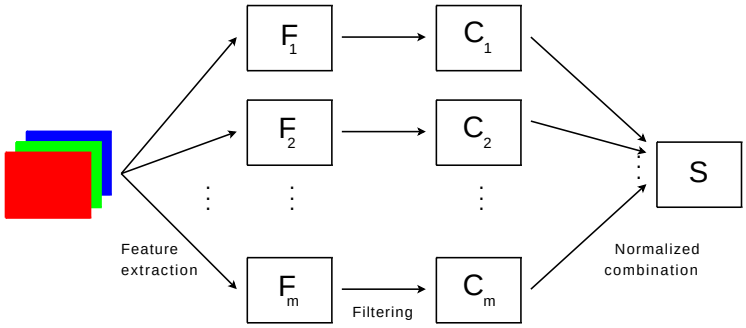

Fig. 3. General overview of the conspicuity map approach for saliency detection. S is the final saliency map, and F_i, C_i the feature maps and their associated conspicuity maps, respectively.

conspicuity maps has been proved to be efficient, both in terms of relevance of detected regions and speed. Particular attention however has to be paid to the choice of relevant feature maps, and how to combine them.

3.2 Depth based feature maps

Depth is a useful information when one has to decide whether an object of the environment is of interest or should be ignored. Close objects might be dangerous or noteworthy, thus implying an action from the user. Despite the interest of such an information, very few methods take depth into account to guide the FOA.

The first proposition [16] was to use depth in an attention model only to determine the object which was closest from the user. Each time an object was closer than the previous match, the attention model would simulate a saccade of the visual system. The limitation of such an approach is the absence of other important features, such as colour opposition, edge magnitude, illumination, etc. Furthermore, it does not give any information about movement in the recorded scene, especially movements towards the user.

It has been later suggested to use depth in the usual bottom-up approach of saliency-based visual attention models [17, 18]. The interest of depth information as a new feature map was proved, combined with common feature maps like colour opposition, intensity, and intensity gradient components. However, no

information about movements of objects is obtained either. In a mobility aid, this information is extremely important, especially when an object comes towards the user. This is why we combine both depth and depth gradient feature maps, which are computed as follows.

Close objects are considered to be more important than ones located farther away. Nevertheless, an object closer than distance d_{min} (for instance $d_{min} = 1$ meter) should be directly detected by the blind user with the white cane. Such objects are consequently not considered to be salient in the See ColOr framework. Given $\mathbf{p} = \{p_x, p_y, p_z\}$, with p_x and p_y the coordinates of the pixel in the image space, and p_z its distance from the camera, we base the feature map F_d on the following distance function:

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

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

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

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

To get any conspicuity map C_i from $F_i \in \left\{F_d, F_\nabla, F_{ill}, F_{r/g}, F_{y/b}\right\}$, i being the identifier of the sought feature, a Difference of Gaussians (DoG) is applied on F_i. In a given area, points of interest are those that differ from the neighbouring ones, and are local maxima of the resulting filtered map:

$$C_i(\mathbf{p}) = \begin{cases} F_i(\mathbf{p}) * g_{\sigma_1} - F_i(\mathbf{p}) * g_{\sigma_2}, & \text{if } \mathbf{p} \text{ is a local maxima} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where g_{σ_1} and g_{σ_2} are Gaussians at scale σ_1 and σ_2, respectively. It is common to use a multiresolution scheme so that conspicuous areas of different sizes can be detected more efficiently. *Blob* detectors like the *Laplacian of Gaussian* (LoG) or the *Determinant of Hessian* (DoH) can also be used.

3.3 Combination of feature maps

Given the image I, it is mostly agreed that the saliency map is the weighted sum of all the computed conspicuity maps:

$$S_I = \sum_i \lambda_i \cdot C_i \tag{4}$$

where C_i are the conspicuity maps computed from the feature maps F_i respectively, and λ_i ($\sum_i \lambda_i = 1$) are parameters 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 resulting saliency map. We denominate λ_d, λ_∇, λ_{ill}, $\lambda_{r/g}$, and $\lambda_{b/y}$, the weighting coefficients for the depth, the depth gradient, the illumination intensity, the red/green opposition, and the blue/yellow opposition features, respectively.

We propose to analyze the relative contribution of each conspicuity map in order to automatically determine the values of the λ_i coefficients. Because attention is naturally attracted towards large objects, and because these objects might be a higher threat for blind users than smaller ones, greater importance is given to feature maps containing large regions of conspicuity. To achieve this, we propose to set λ_i as a function of the mean size of conspicuous regions.

Unfortunately, the maps are not noise free, and many small regions are false positives. Therefore a morphological opening is performed with a 5×5 diamond structuring element, to ensure that only large regions are taken into account. Then, for each feature map F_i, the mean size $\overline{x_i}$ (in pixels) of conspicuity regions is computed on the opened map. Finally, each mean is weighted by the sum of all computed mean sizes, to ensure that the sum of all λ_i is equal to one:

$$\lambda_i = \frac{\overline{x_i}}{\overline{x_d} + \overline{x_\nabla} + \overline{x_{ill}} + \overline{x_{r/g}} + \overline{x_{b/y}}} \tag{5}$$

This definition allows the system to give more importance in the saliency map to objects that occupy a large part of the frontal scene.

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 users' FOA onto objects that might disturb their movements. Depth itself having been proved to be of interest in such a task [17], we consider the combination of distance and depth gradient features relatively to the depth feature alone.

4.1 Obstacle avoidance

First we define an *obstacle* as any static object that the user might walk in, and disturb his/her walk. These objects should typically be ones that a blind user might not detect with the white cane. The carried out experiments for that case consist of the followings.

Case 1 An indoor experiment where a table and a printer are possible obstacles. In that case, we expect the system to point the table or the printer, depending on which one is the closest to the user.

Case 2 This experiment depicts a real life situation: a street lamp and street flowerpots occupy the space on a sidewalk. In this sequence, the system is expected to first point at the flowerpots, and then the street lamp when the user goes near the obstacles.

4.2 Threat detection

We define a *threat* as any dynamic object in the recorded scene, that approaches the user and might cross his/her path. To validate our approach, we have recorded different scenes where pedestrians are walking in front of the user. Again, we have made a distinction between indoor and outdoor scenes.

Case 3 In a corridor, people cross the user's path. The objective of this experiment is to show that the different people are detected one after the other. This is why the area of interest is always the closest person to the user.

Case 4 Outdoor, a pedestrian is coming across the user, and does not change his path. He finally bumps into the user. We obviously want to see if the system considers this person as the major threat during the whole sequence.

4.3 Error evaluation

For each image, the ground truth is determined manually: a binary mask, like the one shown in Figure 4, was selected for each picture of each video sequence. In this case, the chair is considered to be close enough to be detected by the user's white cane because its distance from the camera is lower than d_{min}. Thus, the chair is ignore by the system. Moreover, the man is a possible threat, explaining why the system has to select him as the most salient region instead of the chair. To measure the error, the distance d between the ground truth region G and the pixel $\imath = (p_x, p_y)$ actually detected as the most salient is computed as follows:

$$d(\mathbf{p}, G) = min_{x,y}(d_E(\mathbf{p}, G(x, y))) \tag{6}$$

where d_E is the Euclidean distance.

In the case of stereoscopic sequences, the error at a given time is the mean of the errors obtained from both left and right pictures.

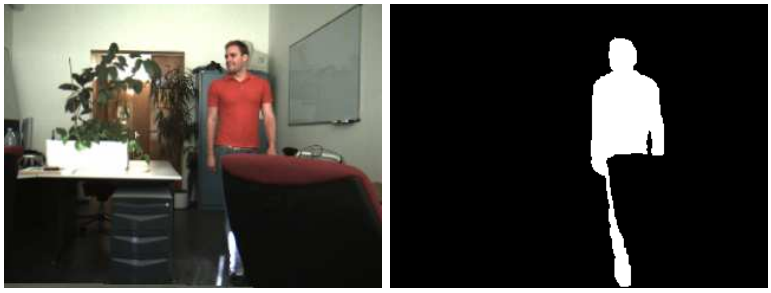

Fig. 4. Example of the ground truth mask (right) for a given image (left).

Videos were recorded using a STH-MDCS2[4] stereoscopic camera, and its development library, which computes in real time the disparity map needed for

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

the determination of depth. The resulting disparity map is unfortunately far from perfect: the depth information is unaccurate or undetermined at many points of the scene as it can be seen on Figure 5. However, given the importance of real time in an alerting system, the presented method is performed with this raw information.

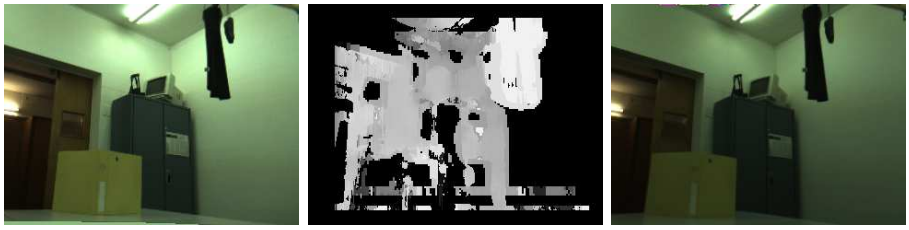

Fig. 5. Images of the left (left) and right (right) views of the stereocamera, and the resulting disparity map (centre). On the disparity map, the lighter the closer. A zero value (black) means that the distance is undetermined for that pixel.

4.4 Results

We have experimented with the case where illumination and colour opponency features are combined with the depth-based features. Video sequences contain from 150 to 500 frames. The results are summarized in Table 1. It shows the percentage of *perfect match* for each sequence and the average values of the λ_i computed with Equation 4. The result on each picture of a video sequence is said to be a *perfect match* whenever the expected area of the picture is determined as the most salient.

In these experiments, the scales for the Gaussian filters in Equation 3 were chosen in order to have simple 5×5 discrete kernels, i.e. $\sigma_1 = 0.7$ and $\sigma_2 = 1.0$. As it can be observed, the λ_i values do not exhibit large variations, which means that none of the conspicuity map is useless. Table 1 also shows the average distance in pixels between the detected points and the ground truth area. Compared to the size of the images and the regions of interest, this distance is relatively small. The focus of the blind user will then be attracted to a region close to the real salient area. Thus we then can assume that the sonified region will be, at least partly, the most salient one. Finally, this table points out that the proposed method operates in near real time, since it only takes around 130 ms to process each picture in a video sequence. A framerate of more than 5 images per second is sufficient for an alerting system, and specifically for the See ColOr framework, where images are sonified at an approximate rate of 3Hz.

Snapshots of each experiment with their respective detected saliency are shown on Figure 6.

Table 1. Average computation time in ms per image, mean error, and percentage of perfect matches on four different video sequences, using a combination of five feature maps.

Case	Average CPU time	Mean error	Perfect match (%)	Average λ_i values
1	136.9ms	26.23	68%	$\lambda_d = 0.21$, $\lambda_\nabla = 0.17$, $\lambda_{ill} = 0.22$, $\lambda_{r/g} = 0.17$, $\lambda_{b/y} = 0.23$
2	139.4ms	36.1	61.3%	$\lambda_d = 0.21$, $\lambda_\nabla = 0.16$, $\lambda_{ill} = 0.21$, $\lambda_{r/g} = 0.22$, $\lambda_{b/y} = 0.20$
3	115.9ms	30.31	58.6%	$\lambda_d = 0.21$, $\lambda_\nabla = 0.17$, $\lambda_{ill} = 0.23$, $\lambda_{r/g} = 0.19$, $\lambda_{b/y} = 0.21$
4	131.9ms	23.6	73%	$\lambda_d = 0.19$, $\lambda_\nabla = 0.18$, $\lambda_{ill} = 0.22$, $\lambda_{r/g} = 0.20$, $\lambda_{b/y} = 0.21$

Fig. 6. The four different experiments. Left: indoor. Right: outdoor. Up: obstacle detection. Down: threat detection. The surrounded cross indicates the selected saliency.

4.5 Discussion

The overall results obtained with this framework are promising, knowing that the computation of disparity is a challenging task, especially in real time. Some optimizations or simple modifications will likely lead to better results. For example, in Equation 1, a more realistic distance function could be defined. The current function not being differentiable at $p_z = d_{min}$ and $p_z = d_{max}$, the transitions at these points are steep. Using a sigmoid would create a smoother and more continuous variation at the extremas of the interval $[d_{min}, d_{max}]$.

Other approaches to automatically determine the weighting coefficient in Equation 3 will be investigated. With the current system, only feature maps with large regions are emphasized. Meanwhile, it is important to highlight features that can detect as much objects of interest as possible. A possibility is to use the correlation of conspicuities in different feature maps: an area detected as salient over more than one feature should be accentuated. With such an approach, the problem generated by regions not completelely covering each other is particularly challenging.

We have seen on Table 1 that the system's reliability is not perfect. This does not mean that saliency is not a useful information, even when it is not the expected area. As a matter of fact, we only want to guide the user's attention towards areas that might be relevant. Once he or she has listened to the sounds of musical instruments, it is the user's choice to decide whether or not this region needs more interest. To ensure that really important regions can be pointed out, we think that it would be important to keep track of regions that have not be selected as the most salient one. The idea is to memorize a set of the most salient areas in each frame and, using a tracking algorithm, to compute saliency on this point on the following frames. A region that remains salient in a long sequence might then be indicated as a threat or an obstacle.

Finally, we need to estimate the camera motion, because the accuracy of the system falls down when the camera moves non linearly. Given the constraint of near real-time –the system's feedback must be at least 3 Hz–, we intend to look for already existing methods.

5 Conclusion 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 in near real-time, using a depth based FOA mechanism. Depth gradient is introduced as a new feature map into the bottom-up visual attention model based on conspicuity maps. We have shown that this feature map allows for a better detection of objects of interest in video sequences than when using depth only, as proposed in previous works. 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 a specific distance, like the white cane's length, should be detected or not. The results we obtained with this simple framework are promising, and

some optimizations on the distance function for the determination of optimal weightings of conspicuity maps, should lead to even better results.

Ongoing and future work concern the following. First, the presented method will be integrated into the See ColOr prototype, and tested with blind users. 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. We also want to define a more adapted distance function, and investigate other ways of optimizing the automatic determination of the relative contribution of the different feature maps. Then, we will seek for a camera motion estimator that can be inserted in this framework without altering the computing time. Finally, we think that it is important to select more than one salient region, because a fully relevant detection is highly improbable. We will then track the most salient regions from one frame to one other, so that we can point out another region if the user considers that the first area was not of interest.

Acknowledgments. We gratefully acknowledge the financial support of the Swiss Hasler Foundation and of the SIMILAR European Network of Excellence. The authors would also like to thank the reviewers for their comments that helped improve the manuscript.

References

1. Bologna, G., Deville, B., Pun, T., Vinckenbosch, M.: Transforming 3d coloured pixels into musical instrument notes for vision substitution applications. EURASIP Journal on Image and Video Processing (2007) http://www.hindawi.com/getarticle.aspx?doi=10.1155/2007/76204.
2. Deville, B., Bologna, G., Vinckenbosch, M., Pun, T.: Depth-based detection of salient moving objects in sonified videos for blind users. In: VISAPP 2008, International Conference on Computer Vision Theory and Applications. (2008)
3. Pun, T., Roth, P., Bologna, G., Moustakas, K., Tzovaras, D.: Survey: Image and video processing for sight handicapped people. Eurasip International Journal of Image and Video Processing (2008) Special Issue: Image and Video Processing for Disability, to appear.
4. Kay, L.: A sonar aid to enhance spatial perception of the blind: Engineering design and evaluation. The Radio and Electronic Engineer **44** (1974) 605–627
5. Meijer, P.: An experimental system for auditory image representations. IEEE Transactions on Biomedical Engineering **39** (1992) 112–121
6. Cronly-Dillon, J., Persaud, K., Gregory, R.: The perception of visual images encoded in musical form: a study in cross-modality information. Proc. Biological Sciences **266** (1999) 2427–2433
7. Landragin, F.: Saillance physique et saillance cognitive. Cognition, Reprsentation, Langage **2** (2004) http://edel.univ-poitiers.fr/corela/document.php?id=142.
8. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2006) 2049–2056
9. Peters, R.J., Itti, L.: Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2007)

10. Hoffman, D., Singh, M.: Salience of visual parts. Cognition **63** (1997) 29–78
11. Kadir, T., Brady, M.: Scale, saliency and image description. International Journal of Computer Vision **45** (2001) 83–105
12. Lowe, D.: Object recognition from local scale-invariant features. In: Seventh International Conference on Computer Vision (ICCV'99). Volume 2. (1999)
13. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: Proc. 9th European Conference on Computer Vision. (2006) 7–13
14. Milanese, R., Gil, S., Pun, T.: Attentive mechanism for dynamic and static scene analysis. Optical Engineering **34** (1995) 2428–2434
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machcine Intelligence **20** (1998) 1254–1259
16. Maki, A., Nordlund, P., Eklundh, J.: A computational model of depth-based attention. In: Proc. International Conference on Pattern Recognition (ICPR '96). (1996)
17. Ouerhani, N., Hügli, H.: Computing visual attention from scene depth. In: Proc. 15th International Conference on Pattern Recognition. Volume 1. (2000) 375–378
18. Jost, T., Ouerhani, N., von Wartburg, R., Müri, R., Hügli, H.: Contribution of depth to visual attention: comparison of a computer model and human. In: Early cognitive vision workshop, Isle of Skye, Scotland. (2004)