# Towards 3D visual saliency modelling

Leroy Julien and Riche Nicolas

Saliency models in still images are numerous, while more and more video saliency models are available. A new branch of the saliency models which only begins to develop concerns 3D data. This new area of research comes from the 3D data which becomes available and uses 3D models for 3D printing or the novel low-cost RGBD sensors (as the Microsoft Kinect). The first results are very preliminary compared to the more mature algorithms for still images, but the new applications which can be foreseen using 3D data are very promising.

## 1 Understanding 3D Saliency

For visual attention, depth perception is just as fundamental as the perception of texture or movement. Human vision is an extremely complex process which by the nature of the organ (two eyes) is intrinsically linked to the perception of depth. Indeed, we do not bear the same interest in objects if they are near or far, structured or disorganized, big or small, etc. Although these features can be extracted from an image, raw access to spatial information greatly simplifies and increases the accuracy we can achieve using these characteristics. It is therefore essential if we want to be able to model to the nearest the human behaviour of visual attention to integrate this spatial information. If the literature is very rich on computational models, they are in the vast majority dedicated to 2D image analysis. Concerning the 3D saliency, it is unfortunately meagre. Even earlier than the appearance of well-known 2D saliency models, like Itti's model, [1], authors have studied the subject of 3D and began to propose integrating 3D information into their models. This subject will remain underprivileged until the last 5 years. Indeed, a new enthusiasm has taken hold

Leroy Julien
UMons, e-mail: julien.leroy@umons.ac.be

Riche Nicolas
UMons, e-mail: nicolas.riche@umons.ac.be

of the scientific community which takes advantages of the recent advances in 3D data acquisition. Three factors are likely to be considered:

1. The availability of 3D content visualization systems became increasingly democratic. 3DTV and 3D cinema have become essential in the media landscape. Indeed, research shows many new features between 3D and attentive behavior, such as cognitive overload-induced vision of 3D content. Understanding and modeling the careful visual process makes sense when we want to improve the technical aquisition and visualization with 3D systems to enable the user to make the most of 3D media.
2. The availability of accurate and inexpensive depth sensors influenced the field of 3D image analysis and subsequently the 3D saliency. If previously, getting a disparity or depth map demanded extensive work such as calibration or conversion of data for analysis with often expensive and poorly performing sensors. These new sensors, such as Microsoft Kinect or Asus Xtion, make easily accessible the use of depth information.
3. The modelling of human attention requires a validation step of algorithmic performances. In 3D, this step has long been difficult to perform by the lack of large databases that can be used. Nevertheless, as we said in the previous point, these new sensors have simplified the steps of acquiring and annotate 3D images for designing these databases dedicated to Saliency. Coupled with monitoring systems as efficient binocular eyes tracking system, it becomes easier to analyse the specific processes involved in 3D saliency but also to compare and validate the proposed models.

## 2 Why 3D features for attention ?

The 2D features extraction from videos can identify the relevant information within the (X,Y) plane. However, they show their limits when the information occurs on the Z (depth) axis. As shown in Figure 1, this is the case for motion features extraction. Indeed, the relevant motion is poolrly captured with 2D motion features as the main movement is along the Z axis.

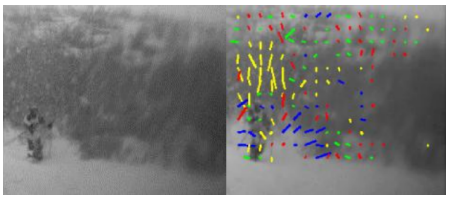

**Fig. 1** A frame with a skier coming towards the camera (depth - Z axis velocity): 2D motion features (optical flow for X and Y velocity).

The (X,Y) motion is properly captured: the snow falling vertically (Y axis) above the skier is detected (yellow vertical lines) and the snow moved by the skier on his right on the X axis (blue horizontal lines). But the motion of the skier himself is not well described: the image shows several lines of different colors (X,Y directions) on the skier while in reality he is coming towards the camera (Z axis). This example shows that detection of the motion on the Z axis would assign the skier with his real displacement. A better feature extraction will also enhance the attention model performance.

The availability of low-cost 3D sensors with active infra-red illumination (as the Microsoft Kinect described in [2]) is an opportunity to easily extract scene depth (Z) information along with classical videos providing (X,Y) information.

## 3 When using 3D features ?

In Fig. 2 from SMAMS's model [3], the speed and direction saliency maps by using a RGB final saliency map is represented.

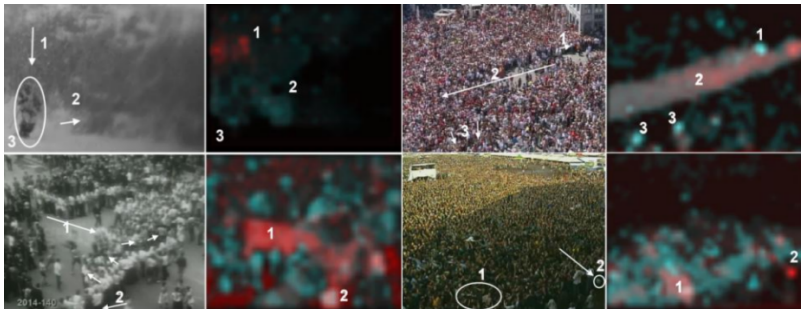

**Fig. 2** First and third column: annotated frames. Second and fourth column: color saliency maps from SMAMS's model. A red dominant means that the speed feature is the most interesting. A cyan dominant means that direction is the important feature.

A red dominant means that the speed feature is the most interesting. A cyan dominant means that direction is the most important. A white blob (which is a mix of red and cyan) means that both speed and directions may attract attention. Here we used only 2D motion features on complex real scenes. In the first two images from the first row there is a close scene with a frontal view. The other scenes contain wider and wider views with mostly top views.

On the second row, first and second images, we can see that people running towards the others are detected (1) and the person who is faster and with a different direction (2) is also highlighted. On the first row, third and fourth images, the two people walking against the main central flow (1) are well visible. It is also the case with some people having perpendicular directions (3). Finally in the second row,

third and fourth images one person carried by the crowd (1) and a thrown object (2) are also well detected with a higher speed compared with the other moving objects. Nevertheless, the results are very poor for the first row, first and second image in Fig. 2. While the rapidly falling snow (Y axis motion) is well detected (1) and the snow pushed by the skier (X axis motion) on his right (2) is also detected, the skier himself (3) is not detected at all! The skier is the only moving object on the Z axis, thus it is very salient, but as only 2D features are extracted, he is not well detected. This scene comparison in 2D shows that the more the scene is wide and the camera has a top view, the less important the Z axis motion is. Indeed, a top-view will map most of the motion on the (X,Y) plane and very small people doing gestures on Z (like jumping for example) are almost not detectable in those configurations. An interesting conclusion is that, while in videosurveillance-like situations (wide field of view, almost top-view) the knowledge of Z is important, for ambient intelligence and robot–like situations (smaller field of view, frontal view), the knowledge of the Z axis is crucial. This is convenient, as the Kinect sensor horopter is between 25 cm and 6 meters.

## 4 Chapter organization

Saliency can use any input features. 3D saliency is based on the new 3D sensors ouput. This ouput is twofold: 1) the automatic recognition of people silhouette and skeleton which can provide high-level features about people behaviour and 2) the 3D data ouput (RGB and depth maps or 3D point cloud) providing low-level 3D features.

This chapter first presents a model using the providing high-level information about people. In a second part, several models using RGB/depth data or point cloud, more generic on low-level features will be presented.

## 5 3D saliency model based on high-level features

In this section, the design of a new intelligent system capable of selecting the most outstanding user from a group of people in a scene will be discussed. This ability to select a user to interact with is very important in natural interfaces and in emergency-related applications where several people can ask to communicate simultaneousely. The proposed algorithm has three main steps: first, features are extracted from Kinect's sensor. In a second step a contrast-based approach is applied and finally, those constrast-based features maps are combined to focus the system attention on a specific user without complex rules.

### *5.1 Features extraction*

The first step is to extract features from the observed people. For that purpose, we use the Kinect sensor for its ability to extract smooth depth maps in complex illumination conditions. Libraries as OpenNI ( for example used in [4]) are available to detect human silhouettes and extract anatomical features from skeletons tracking.

Four features are extracted from the upper body part only as the legs are much less stable in our implementation. One of the four features is dynamic, namely the **motion index**. It is computed as the mean variation of the same skeleton points between two frames in 3D (on X, Y and Z). The barycenter point variation is extracted from the others (Eq. 2) in order to keep only the body relative motion which will describe an excitemenent degree or movement transition of the body without any assumption on the whole body speed.

$$D_{mk} = (\sum_{sk} |k_b - k_{sk}|)_t - (\sum_{sk} |k_b - k_{sk}|)_{t-1} \tag{1}$$

where $k = x, y, z$. The skeleton points are noted $sk$ and the barycenter $b$.

$$MI = \sqrt{Dmx^2 + Dmy^2 + Dmz^2} \tag{2}$$

A second feature extracted from the upper body part is a static feature, namely the **asymmetry index**. This feature is only computed on the X axis by differencing the distances between the barycenter point and the right shoulder, elbow and hand points with the left ones (Eq. 3). This index provides information about the symmetry of the upper body.

$$AI = \frac{\sum_{sk} |X_b - X_{sk_r}| - \sum_{sk} |X_b - X_{sk_l}|}{n_{sk}} \tag{3}$$

where $n_{sk}$ is the number of skeleton points.

The third extracted feature is the **contraction index**. This index is the ratio between the maximal distance between skeleton points on X axis and the maximal distance on the Y axis (Eq. 4). This index tells us if the person is more or less contracted.

$$CI = \frac{|max(X) - min(X)|}{|max(Y) - max(Y)|} \tag{4}$$

The fourth and final feature is the **player height**. That one is simply computed by measuring the player barycenter Y coordinate.

After normalization, those four features provide a quite complete description about the level of excitement, and the upper body configuration of each player.

### 5.2 Contrast-based mechanism

As stated in [5], a feature does not attract attention by itself: bright and dark, locally contrasted areas or not, red or blue can equally attract human attention depending on their context. In the same way, motion can be as interesting as the lack of motion depending on the context. The main cue, which involves bottom-up attention, is the contrast and rarity of a feature in a given context.

The approach here follows the one in [6]. In our case, as the group of players can be small, the rarity computation is not relevant. Therefore we only use the global contrast. Thus, the first step in this section is to calculate for the $i^{th}$ feature ( $f_{i,k}$) a contrast between the different users $k$.

$$C_{i,k} = \sum_{j=1}^{N} \frac{|f_{i,k} - f_{i,j}|}{N-1} \tag{5}$$

where $N$ is the number of users. Once all the contrasts for a given feature $C_{i,k}$ between each user and the others have been computed, they are ordered in ascending order $C_{i,k,o}$ with $o = [1 : N]$ from the maximum (o = 1) to the minimum (o = N). The difference between the two highest values is compared to a threshold $T$ which decides if the contrast is large enough to be taken into account as in Eq. 6.

$$\begin{cases} \alpha = 0 \;\; if \;\; |C_{i,k,1} - C_{i,k,2}| < T \\ \alpha > 0 \;\; if \;\; |C_{i,k,1} - C_{i,k,2}| \geq T \end{cases} \tag{6}$$

### 5.3 Fusion

Only the features being the largest and passing this threshold T are merged with different weights (Eq. 7)

$$C_k = \sum_{i=1}^{H} \frac{C_{i,k} * W_i * \alpha}{H} \tag{7}$$

where $H$ is the number of features and $\alpha$ is given in Eq. 6.

The weights $W_i$ are initially set to be the same for all the 4 features which are used here. Then, the number of times a feature is contrasted enough for a given user ( $\alpha > 0$), a counter is increased. The feature weight will be inversely proportional to its counter: if a feature i is often contrasted, its weight will be lower and lower, while a feature which is rarely contrasted enough will see its weight increased. This mechanism ensures a higher weight to novel behavior while too repetitive behavior will be penalized. As an example, someone who will sit down for the first time (different height feature compared to the others), the height will have the maximum weight. If this person thinks that a different height is enough to attract the system attention, he will try again, but the more he tries again, the more the height feature

weight will decrease as this behavior is no longer surprising. This approach allows the system to learn how much a feature is novel and provides higher weights to the most novel ones.

The contrast $C_k$ represents the bottom-up saliency for each user k. Saliency will be higher for the people exhibiting the most contrasted features within a given frame. The process of bottom-up attention is summarized on Fig 3 on a three-player scenario example. Each of the three players has its four features computed (in red for the asymmetry index, yellow for the contraction index, violet for the motion index and green for the height). The contrast computation and thresholded (Eq. 5 and 6) is displayed in the second column. Finally the contrasted features combination (Eq. 7) is explained in the third and fourth columns

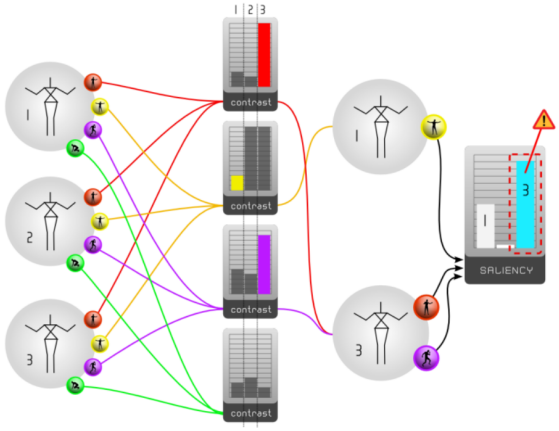


**Fig. 3** Example of bottom-up saliency computation for 3 players. For each of the four behavioral features, a contrast is computed between the players. A threshold will eliminate features which are not contrasted enough between players (here the fourth feature in green is eliminated). The player having more contrasted features with higher weights will be selected as the most salient (here the third player.

## 6 3D Saliency Models Based On Low-Level Features

Several algorithms taking into account 3D information were already set-up. However, this concept of 3D should be taken with caution. Indeed, under the aspect of integration of spatial information into a model of human attention, the concept of 3D is often used vaguely and should be refined. Indeed, it will be possible to find in the literature models that deal with 3D saliency but they will be difficult to compare. Take on one side [7], the precursor model on the use of a fused depth map with a

saliency map from a 2D analysis and on the other side [8], who proposed a saliency model on 3D meshes. These could be presented both as 3D saliency models, each using 3D information into their algorithm. However the inherent nature of the data imposes a differentiation. The 3D data processed differs greatly, a mesh can hardly be equated with an organized disparity map. It will be necessary to distinguish between methods using depth information and methods using all the available 3D information. Disambiguation of 3D in saliency can be submitted via the concept of 3D imaging, 3D data and their representation. When discussing a 3D image, the most common technique to reproduce the illusion of depth to an observer is stereoscopy. Stereoscopy brings together all the techniques used to reproduce a perception of depth from two planar images. Based on the acquisition of two slightly offset images similarly to human eyes, it is possible to generate a disparity map. From this disparity by using epipolar geometry, we can estimate a depth map. Indeed, the disparity and depth are inversely related. As the distance from the cameras increases, decreases the disparity. It is then possible to allocate to each pixel of the image a depth information. We then obtain a depth map where 3D information is represented by the triple (i,j,d) were i,j are coordinates in the image plane and d is the depth of the pixel. This is typically what RGBD sensors will get, often a depth map is also calibrated with a 2D color image. Let us now take a 3D model generated by any 3D modeling software. The 3D image or volume will consist of a set of vertex spatially represented by Cartesian coordinates.

We will do the following distinctions between different classes of saliency models:

1. The ones that we will call ”2.5D”. These methods are based on the use of spatial information using disparity or depth map. These methods takes as input, depth image or stereoscopic images. All these models have a step in which to calculate the final saliency map, 2D visual features and 2D saliency maps are estimated.
2. The 3D methods. These methods are based on all available spatial information and geometry. These method are heavily based on 3D geometry to extract salient information. These methods apply to 3D reconstructed objects and scenes or 3D modelled scenes. The result is a 3D saliency map.

This chapter discusses 3D salience as a whole. We begin with a review of the methods related to the modeling of salience in 3D. For this review, we will make a distinction between types of models presented based on their input data on which the algorithm is applied. Indeed, it is necessary to make the separation between the methods using depth information and methods using pure 3D structures. After this review, we will discuss this classification, which can be too simplistic and present other classification possibilities better able to take into account all the intricacies related to 3D salience. Finally, we will present a new model of salience in 3D with two particularity to process large amounts of 3D data regardless of their type (mesh, cloud point, or RGBD data) and integrate color texture information. Indeed, as the review will show, models capable of handling information on structures such as mesh or large point clouds systematically abandoning the color information.process

large amounts of 3D data regardless of their type: meshes, point clouds, or RGBD data.

## 7 3D Saliency: a review

In this section we will follow the proposed classification and present the most representative models of these categories. The first category of algorithms focuses on the integration of depth usually as a disparity map in their saliency model.

### *7.1 2.5D models or Depth-based Saliency*

#### 7.1.1 Depth and Disparity based Models

[7] is one of the first models incorporating a stereo disparity map as a pre-attentive characteristic which will be added to the information flow and motion detection. A depth target mask, which corresponds to the depth conspicuity map in the saliency-based model of attention, is computed. This research, however, are aimed at the integration of a saliency mechanism as an element of selection of a moving part and no validation of the algorithm performance with a human reference have been performed.
[9] propose an extension of the well-known Itti model based on the analysis in center-surround and extend with the introduction of a depth map. Their analyses also involved other features related to depth as the use of gradient or curvature extracts of depth but reject their use because of noise on the data.
[10] is to our knowledge one of the first research on the impact that depth information can have on attention and especially with an objective comparison with the computational counterpart. The authors demonstrate through two simple experiments the potential impact of the use of depth. First, they show with random dots stereograms (RDS), so pure disparity maps, that depth perception influences our attention. They draw conclusions as the objects with the greatest disparity attract first fixations. The generalization of this conclusion is that elements with a large disparity are more easily perceived, attract the earlier fixations and so are more salient. Second, they validate their observation with an objective comparison of human attention map, a saliency map based only on the color and a color+depth saliency map. Their conclusion is that the introduction of the depth drastically increases the measured performance. Although the experimental context can be discussed, given the metrics and the small size of the data base operated, this study was drawing the basic conclusions on the idea and the impact of the integration of depth in a saliency mechanism . The depth is an important characteristic capable of optimizing the similarity of saliency algorithms with human attention.
In [11],the authors offer a very similar approach to previous authors and realizes an

interesting comparison of the performance of their model through several validations. Their model is thus divided on the basis of its features: grayscale, color and depth. The analysis is interesting because they show not only the interest of color based features but also the impact of depth on the results, while discussing performance according to the nature of the analyzed scene.
In [12], the authors study the possibilities of using laser data for attention mechanisms. They propose a model (BILAS) based on that of Koch & Ullman but including here as input depth and reflectance images from a laser sensor.
In [13], the authors are interested in the use of attention mechanism exploiting 3D features to assist in the segmentation step preceding robotic tasks such as grasp and object manipulation. For this, they study several 3D features as a surface height, orientation and relative area occluded edges and merge them with 2D information (color, orientation, ...) through a probabilistic approach.
In [14], the authors focus on the exploitation of the depth map as support for extracting information related to motion. As they point out very few models operate depths of the data and to our knowledge they will be the first to integrate depth data from the stream of depth camera to constrain their model. Their premise is the limitation of movement of characterization possibilities in an RGB image, if one can easily define this movement in an XY plane, it becomes complex along a depth axis. A better feature extraction thus becomes trivial with the depth of information and will improve the performance of their attention model.
In [15], the authors suggest to study the differences of visual attention behaviour when the depth is involved. For this, the authors propose a first substantial database containing 600 fixation measures obtained on pairs of 2D and 3D images. These data come from a Kinect camera. The authors want to measure differences in fixation between 2D and 3D images and the impact that the introduction of depth data can have on well known 2D saliency models performance. The authors exhibit a set of priors related to the depth that are consistent with the attention process. Depth cues modulate visual saliency to a greater extent at farther depth ranges. Furthermore, humans fixate preferentially at closer depth ranges. A few interesting objects account for majority of the fixations and this behavior is consistent across both 2D and 3D. They also found that the relationship between depth and saliency is nonlinear and characteristic for low and high depth-of-field scenes. The additional depth information led to an increased difference of fixation distribution between 2D and 3D version, especially, when there are multiple salient stimuli located in different depth planes. Using their framework and approach on various models of 2D saliencies, the authors obtained a significant increase in the algorithms performance.
In [16], the authors focus on the extension of a model based on the contrast saliency allowing it to integrate the depth through a disparity map from a set of stereo images. The authors also show increased performance of their model by introducing the disparity data. They also offer a large database of 1,000 stereoscopic images to validate their method.
In [17] , the authors deal with saliency for 3D stereoscopic images. An interesting point of their approach is the comparison of the integration possibilities of depth either as a weighting element, either through the establishment of a depth saliency

map. The last method seems to give better results. The authors propose a detailed analysis of several 3D saliency models and defines a classification of the models based on hox depth information is integrated. Finally, they propose an adaptable framework for the existing 2D saliency model. The depth saliency map and 2D saliency map from a generic 2D saliency model saliency are merged to provide the final saliency map. Extensive validation is provided for the various models
In [18], the authors are interested in the role of depth in situations of competing saliencies due to appearance, depth-induced blur and center-bias. They propose a new saliency model by integrating first depth contrast then many other features like color histogram, contour compactness, dimensionality, etc. They create a feature vector of 82 elements that are fused by a learning algorithm (SVM). Their approach shows that 3D saliency outperforms the others 2D saliency models.
In [19], the authors propose a new model of saliency based on the depth. They propose not to use the depth measurements as another channel of an image but by explicitly constructing 3D layout and shape features from depth measurements. The main idea is that humans use coplanarity to guide their assessment of saliency. Their method is based on fitting planes to the 3D points of the depth image, allowing to associate each pixel with the dominant plane that contains it. Therefore they penalize points which lie on different depth planes and compute a dissimilarity measure between patches (locally adjacent pixels) to create a saliency map. The authors demonstrate the application of their algorithm on the segmentation of objects on RGBD data. To validate their approach the authors have made a new dataset for depth-based saliency, including pixel-level ground truth segmentation of salient objects.
In [20], the authors are interested in a particular element related to attention and for which they will introduce a new feature map called the ”depth-of-field map”. The idea is that the depth-of-field map functions works similarly to the depth of field effect of human vision by enhancing the saliency of the regions near the point of gaze in the direction of depth and reducing the gaze movements between regions widely separated from each other. Their model is based on that of Itti and Ozeki which they are adding their own constraint based on depth of field.
In [21], the authors had two major contributions: their primary objective is to offer a wide enough RGBD database to be a real benchmark for 3D saliency, secondly they proposed a model of saliency based on the depth that does not treat as an independent feature as in many models but simultaneously takes account of depth and appearance information from multiple layers. They based their approach on low-level feature contrast, mid-level region grouping and high-level priors enhancement. Thanks to their large database, they carry out a quantitative analysis of their method against other well known 2D augmented to 3D saliency models.

### 7.1.2 Stereoscopic based Models

In [22], the authors propose a saliency model in the context of the stereo vision. Their model is based on a biological approach and highlights the problems of binoc-

ular vision that have a direct impact on the attention as the concepts of binocular rivalry. Their model is based on an existing model, the Selective Tuning model, which extends naturally as they demonstrate to the binocular vision.
In [23] the authors are interested in stereoscopic vision and involvement that it can have on the design of a saliency model, based on a biological modeling of the human attentive process. Indeed, if we consider the binocular nature, the source of stimuli is double and redundant. This issue entitled The Attentional Stereo Correspondence Problem (ASCP). The authors propose a model of attention based on the depth that tends to consider this issue, proposing a model close to a model with relevant psychophysical characteristic of attention in depth.

## *7.2 3D Structure based Models*

### 7.2.1 Mesh based Saliency

In [8], the authors introduce for the first time the concept of mesh saliency as a measure of importance for regions of a mesh structure. Their mesh saliency is defined in a scale-dependent manner using a center-surround operator on Gaussian-weighted mean curvatures. The model is based on the assumption that for a 3D mesh, geometry is the largest contributor to saliency. Their method estimates the saliency in terms of mean curvature with a mechanism of center-surround.
In [24], the authors propose a new method for extracting salient critical points of a mesh combining saliency mesh with Morse theory. Their method is based on a center-surround mechanism but also Gaussian-weighted average of the scalar of vertices. It offers an extension of the previous model using as a weighting element, a bilateral filter rather than an absolute difference in weighted Gaussian.
In [25], the authors propose a variation of the model based on the difference of Gaussian for the extraction of salient points for the purpose of correspondence between various views of an object. Their method is based on measuring the displacement of vertices with respect to their original position after the various filters.
In [26], the authors propose a 3D object retrieval method based on the extraction of salient points in 3D. Their method of extracting salient points are based on classification by a SVM of the low characteristic histogram for each vertex of the model. The characteristic used is the absolute value of the curvature filtered by a gaussian. With this classification, each vertex is defined as salient or not with a confidence score. For a 2D projection, the method generates a two-dimensional map of salient points that will be used to perform the signature recovery object.
In [27], the authors have an approach based on the definition of an information channel between a set of views and polygons of an object. It is this mutual information channel expressed by the Jensen-Shannon divergence [28] which allows them to firstly define a measure of similarity between views and secondly to extract the saliency of the 3D object. The idea is to express the way in which the polygons are perceived as a function of a set of viewpoints. For this, they express the saliency of

a polygon as the average variation in the difference of Jensen-Shannon between this polygon with its neighbors. Based on these characteristics, the salient points are extracted with a classifier that detects points that have a combination of high curvature and low entropy values.
In [29], the authors propose an extension of their previously proposed method where they use now for the characterization of surfaces the absolute values of Gaussian curvature and Besl-Jain surface curvature characterization as the low-level surface properties.
In [30],the authors propose a new method for the detection of regions of interest on surfaces. Their method is local and global. It also takes into account the distance to the foci of attention. They use this method to determine the best possible view for a 3D object. Their method is based on a local approach and is based on the calculation of a descriptor on each vertex. Besides this local approach, method integrates two characteristics that are the extremity detection and definition of patch association to represent the fact that regions of interest the look that attracts more specific points.
In [31],the authors propose a rarity model on two levels: local and global. Indeed, they make the observation that all the models presented before them were based only on a local analysis of mesh but a human observer also had a global vision. They thus introduce global saliency calculated on the mesh. The local part is based on the calculation of a heightmap to encode local structure. The global rarity is achieved using the same characteristic but where the comparison to a vertex is no longer local but global. To reduce the necessary calculation time, the authors uses clustering to group the vertices with similar properties.

### 7.2.2 Point Cloud based Saliency

In [32], the authors propose one of the first models of saliency that apply specifically to point clouds. The major interest in extracting geometric features for each point of the cloud and based on its neighborhood. The final saliency map is the composition of two intermediate maps: the first one is obtained based on what the authors call the local properties surface (LSP) based in particular on normal surfaces or curvature. A second map is generated based on the distance of the points in the camera. Both cards are then linearly combined to produce the final saliency map.
In [33], the authors propose a 3D object detection framework based on the saliency. Although the system is intended to extract in the 3D environment salient objects, the employed attention mechanism is only based on 2D color image from a RGBD sensor. The interesting contribution is the idea of inhibition of return mechanisms (IOR) that inhibit the currently attended region in 3D.
In [34], the authors propose for the first time a saliency model specifically designed to handle large unorganized point clouds. The proposed method is based on the concept of global ”distinctiveness”. Their method employs the use of a descriptor for each point of the cloud. The authors define the Simplified Point Feature Histogram (SPFH), variation of Fast Feature Point Histogram (FPFH). A dissimilarity measure based on Chi-square will be used to estimate the saliency.

In [35], The authors propose a method of estimating the saliency of 3D point cloud based on gaussian normal vector estimation. The proposed method defines the salient critical points in a scalar function space using a center-surround filter operator on Gaussian-weighted average angle of normal vectors.

### *7.3 Discussion*

The classification of different methods solely based on their input data is a bit simplistic. Indeed, a system based on a RGBD sensor although providing a depth map, it may very well be converted into 3D point cloud by knowing the intrinsic properties of the sensor. Models design for point cloud can be adapted to depth map processing and vice versa. For models based on meshes, data can be converted into a point cloud based on tesselation, opening the possibility to use a method for point cloud. There is thus a possibility of conversion and interoperability between methods.

An interesting classification approach was proposed by [17] to differentiate how is integrated spatial information into each saliency model. Three categories are made on integrating the spatial aspect:

1. Depth-weighting models. This type of models (e.g. [7],[36]) does not contain any depth-map-based feature-extraction processes. Apart from detecting the salient areas by using 2D visual features, these models share a same step in which depth information is used as the weighting factor of the 2D saliency. The saliency of each location (e.g. pixel, target or depth plane) in the scene is directly related to its depth. Both 2D scene and depth map are taken as input.
2. Depth-saliency models. The models (e.g. [13] and [9]) in this category take depth saliency as additional information. This type of models relies on the existence of depth saliency maps. Depth features are first extracted from the depth map to create additional feature maps, which are then used to generate the depth saliency maps. These depth saliency maps are finally combined with 2D saliency maps (e.g. from 2D visual attention models using color, orientation or intensity) by using a saliency map pooling strategy to obtain a final 3D saliency map. This type of model also takes the 2D scene and the depth map as input.
3. Stereo-vision models. Instead of directly using a depth map, this type of models (e.g. [22]) takes into account the mechanisms of the stereoscopic perception in the HVS. Bruce and Tsotsos extend the 2D models that use a visual pyramid processing architecture [37] by adding neuronal units for modelling the stereo vision. Images from both views are taken as input, from which 2D visual features can be considered. In addition, the model takes into account the conflicts between two eyes resulting from occlusions or large disparities.

The classification proposed here by [17] was proposed on a set of 2.5D model, it could be extended this class to all models extracting geometrical characteristics

from their input data. The 3D features are indeed not only limited to the depth map, richer features can be extracted from 3D data such as mesh.

The idea of depth saliency and depth weighting are interesting because of the abstract data type on which we work. The concept of depth saliency can be extended and be generalized to the extraction of salient geometric features related to spatial structure regardless of the data. We can therefore speak of spatial saliency. The depth weighting also introduced another idea for classification. Weighting by the depth map is fundamentally linked to the sensor and its field of vision. It therefore depends on the viewpoint. Inversely, if we apply this idea to methods on mesh, the methods are not viewpoint dependent.

This distinction between models linked or not with a point of view is fundamental. In particular, it raises issues related to the validation and the link with the human visual system. Could we still consider a saliency model, an algorithm running on a mesh processing 3D volumes independently from the viewpoint. Such a system operates basically beyond human visual capabilities.

The classification we propose is to make the following distinction:

1. 2.5D saliency, that process 3D data but is dependent on the viewpoint.
2. 3D saliency, that process 3D structure as a whole and is not dependent of the viewpoint.

This separation makes abstraction of the data type being processed while taking into account the nature of the data dependency or not to a point of view. The depth weighting, discussed previously as a category in itself, is ultimately a bias that is applied to a saliency map. This bias in itself defines the idea of 2.5D saliency, a thought method for 3D saliency constraint to a specific point of view and depth biased fall into the 2.5D category.

## 8 SuperRare3D: A new model of Point Cloud 3D Saliency based on Supervoxels Rarity

We propose a novel object-oriented algorithm of bottom-up attention dedicated to analyze colored point clouds. This model builds on the one proposed in [37]. One contribution is the use of a rarity-based approach not based on superpixels as in [37] but on supervoxels. Supervoxels consist of an over-segmentation of a point cloud in regions with comparable sizes and characteristics (in terms of color and other 3D features). More details on supervoxels and the method used here are provided in the next sections. Our approach has four major interests:

1. Supervoxels let us reduce the amount of processing and allow our method to work on organized or unorganized clouds. Thus, it can analyze point clouds or even fused point clouds coming from various sensors.
2. Supervoxels allow us to have an object-oriented approach in the 3D space.
3. Supervoxels multi-level decomposition allows us to maintain detection performance regardless of the size of the salient objects present in the data.

4. This approach provides a bottom-up 3D saliency map which is viewer-independent. It is than possible to add viewer-dependent top-down information as a viewer-dependent centered Gaussian and depth information. In our paper we only use the centered Gaussian that all the other models also use to remain fair in our comparison.

Our method only uses one feature of the point cloud: the color. Other features like supervoxels orientation or other specific 3D features will be taken into account in future work. As the color feature is the only one we use, this approach is subject to the influence of the choice of the color representation. To provide a first solution to this influence, we propose to fuse the saliency maps computed on several color spaces. Our algorithm can be divided into three major stages: (1) supervoxels decomposition, (2) supervoxel rarity-based saliency mechanism, (3) fusion. We present in the following sub-sections the three main steps of our algorithm.

## *8.1 Super Voxels Cloud Segmentation*

The superpixels are the result of over-segmentation of an image into regions of pixels having similar perceptual properties. This is a step commonly used in computer vision as a preprocessing stage to reduce the amount of information to be processed while still minimizing the loss of information.

We build our system, on the same idea by using supervoxels instead of processing at the point level. We use the Voxel Cloud Connectivity Segmentation method (VCCS) [39] that extracts supervoxels from an organized or unorganized point cloud. The supervoxels can replace the structure of the voxel-based original point cloud by a set of atomic regions that capture the local redundancy of information. They provide a convenient way to summarize the point cloud and thus greatly reduce the complexity of the following analysis process. But if there is a major difference between the size of supervoxels and the size of the salient object to be detected, this one can by merged with a nearby supervoxel and its information is lost 8.1. To avoid this situation, the rarity mechanism is applied to different levels of supervoxels decomposition so that at some level of detail the salient object is well captured. At this point the pathway of the algorithm is split between the different levels of supervoxels decomposition. This separation is made to capture all the information of salients objects by adjusting the size of supervoxels. Indeed,like shown in Figure 8.1, if a supervoxel is too large, it may not stick properly to an object and it is seen disappearing into an adjacent supervoxel. To remedy this, the algorithm works on several levels in parallel that will then be merged into a final saliency map, to maintain both the information of large objects and smaller ones, while refining the segmentation of salient regions.

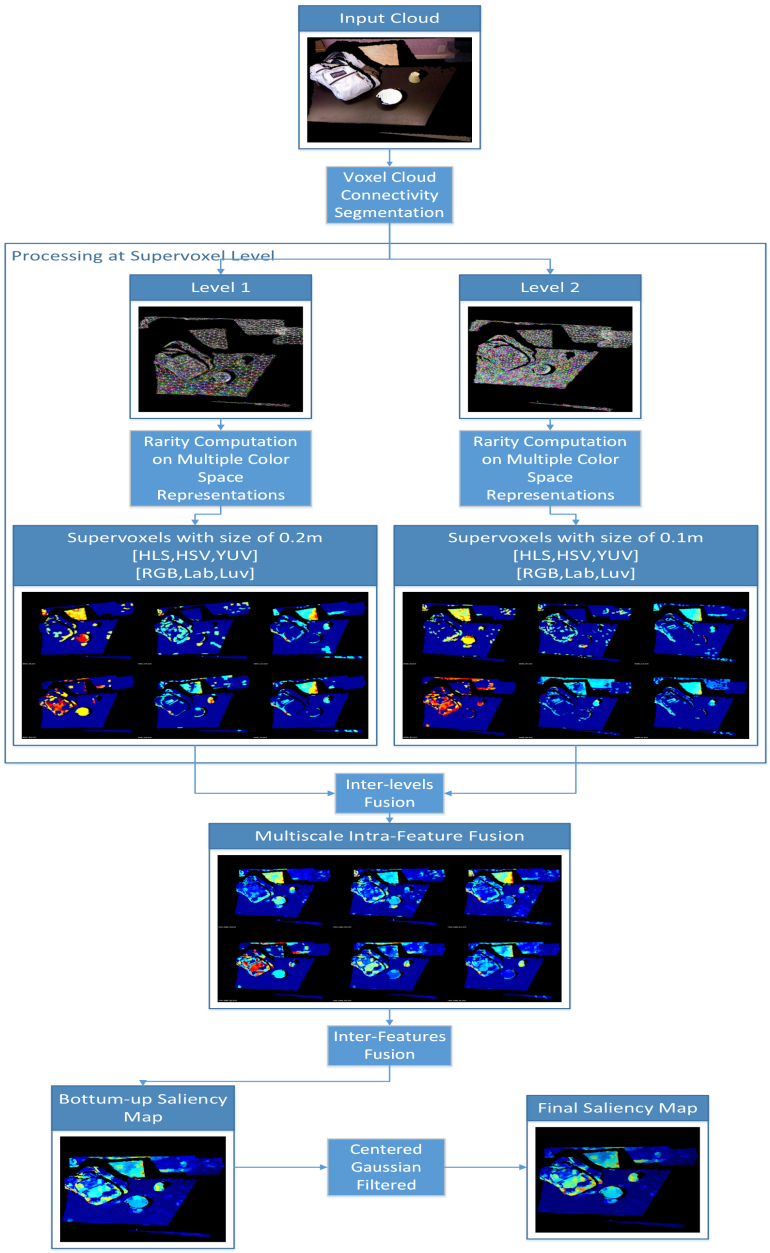


**Fig. 4** Our method is divided in 3 major steps: (1) multiscale supervoxels decomposition, (2) color rarity applied on multiple color spaces, (3) inter-level and inter-feature fusion. A top-down centered gaussian can be use to simulate the human centric preference [38].

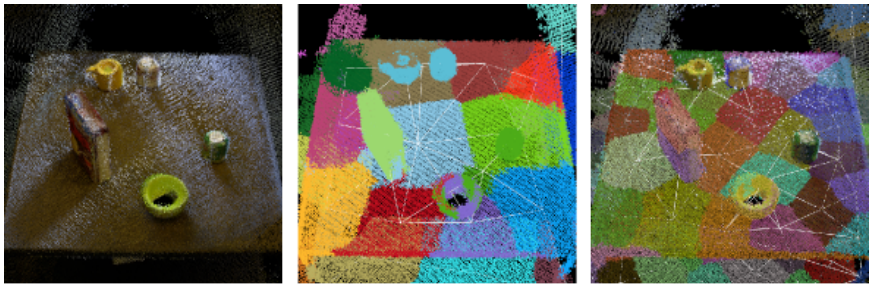


**Fig. 5** (a) Example of table point cloud with a box and cups, (b) supervoxels segmentation using VCCS with a seed size of 0.5m, (c) supervoxels segmentation using VCCS with a seed size of 0.25m. The size of the supervoxels is essential to extract the information of all the objects. If the seed is too large, like in (b), we see the object being absorbed in an adjacent supervoxel, losing the information for the rarity mechanism.

## *8.2 Rarity based saliency*

The rarity mechanism consists, for each supervoxel vector, to compute the cross-scale occurrence probability of each of the $N$ supervoxels. At each color component $i$, a rarity value is obtained by the sekf-information of the occurrence probabilities of the supervoxel as shown in Eq.(8). $P_i$ is the occurrence probability of each supervoxel $Sv_i$ value in respect with the empirical probability distribution represented by the histogram within the $i$th color channel.

$$Rarity(Sv_i)_i = -log(P_i/N) \tag{8}$$

Then, the self-information is used to represent the attention score for the supervoxel region. This mechanism provides higher scores for rare regions. The rarity value falls between 0 (all the supervoxels are the same) and 1 (one supervoxel different from all the others).

### 8.2.1 Intra and inter-supervoxels level fusion

The rarity maps obtained from the rarity mechanism on each color channel (in this case, we select 6 color space representations: HSV, HLS, YUV, RGB, Lab, Luv) are first intra-color combined. In our example, we empirically select a 2 levels decomposition using supervoxels seed of 0.05m and 0.02m for balance between accuracy and computation time. A fusion between same color rarity maps is achieved at each decomposition level by using the fusion method proposed in Itti et al. [41]. The idea is to provide a higher weight to the map which has important peaks compared to its mean (Eq. 9).

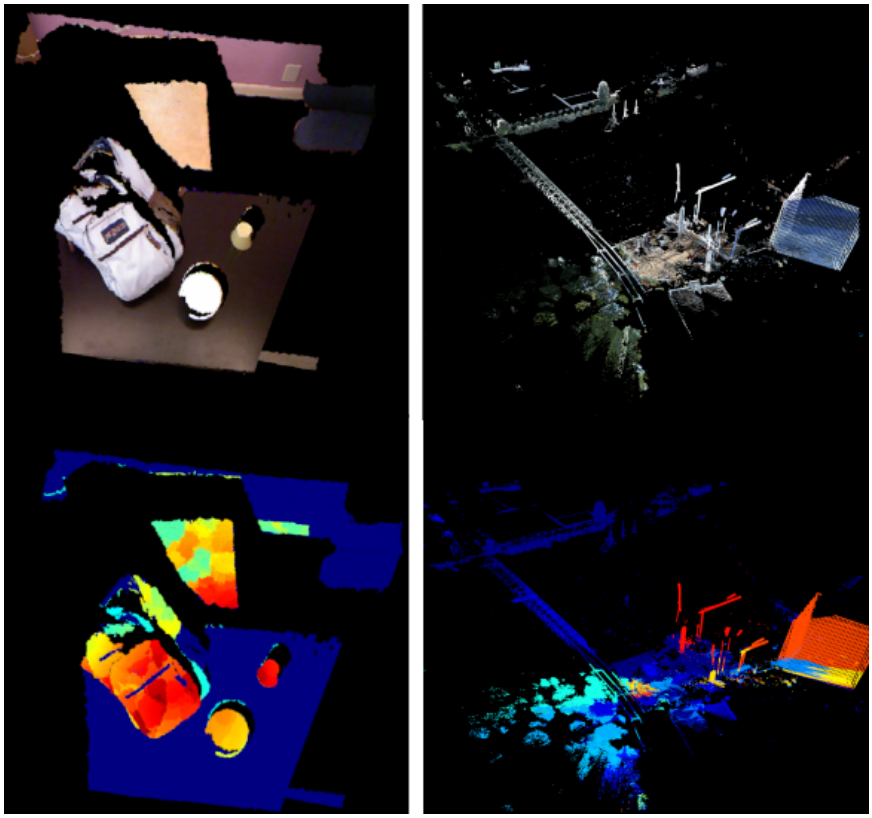

**Fig. 6** Examples of results obtained with our method on 3 different point clouds: (left) a Kinect cloud (organized, 307200 points, three levels of decomposition with 92, 32 and 13 supervoxels); (right) a point cloud recorded using a Riegl VZ-400 and a co-calibrated Canon 1000D camera with 10 Megapixels [40] (unorganized, 5976977 points, 1 level of decomposition with 271 supervoxels).

$$S = \sum_{i=1}^{N} EC_i * map_i \tag{9}$$

where $EC_i$ is the efficiency coefficient for each channel and is computed as in Eq. 10.

$$EC_i = (max_i - mean_i)^2 \tag{10}$$

These coefficients let us sort the different maps ($map_i$) based on each map efficiency coefficient $EC_i$. Each map is than multiplied by a fixed weight defined as $i = 1...K$ where $K$ is the number of maps to mix (here $K = 3$) and $i$ the rank of the sorted maps as shown in the first line of Eq.11. $T$ is a empirical threshold defined in [42].

$$\forall i \in [1,K] \begin{cases} saliency_i = 0 & \text{if } \frac{EC_i}{EC_K} < T \\ saliency_i = \frac{i}{K} * map_i & \text{if } \frac{EC_i}{EC_K} \geqslant T \end{cases} \tag{11}$$

At the end of this first fusion process, the model provides one saliency map per color space representation. The second fusion step, an inter-color feature fusion between each map coming from the different color space representation, is achieved using the same method as the one explained for the inter-decomposition level fusion (Eq. 9).

### *8.3 Color Space influence*

Our method estimates saliency using the rarity only on color feature. The accuracy of this feature is very important and our method is strongly influenced by the choice of the color space representation. If we observe independently saliency maps for the different color modes, we can see that the performance is highly dependent on the mode2, ranging from excellent to poor, but in all cases at least one map provides good performance. For this reason we have chosen to apply the rarity on several color spaces and merge the different rarity maps.

### *8.4 Final Saliency Map*

Finally, in this case, we work with an organized point cloud, we apply a Gaussian centered filter to represent the central preference that people exhibit in images [38]. In the case of object avoidance, this centered human preference makes also sense in the context of robotics as one wants to correct the path of a robot to avoid collisions with objects in front of it.

## 9 VALIDATION

### *9.1 Database*

The database that we used to validate our method was published by [19]. It has 80 shots obtained using a Microsoft Kinect sensor mounted on a Willow Garage PR2 robot. The database consists of RGB images, depth maps and point clouds associated with pixel level ground truth segmentation masks. The 80 scenes are very complex both in terms of number and shape of objects, colors, illumination but also in terms of depth differences. Indeed, there are a lot of bjects which have little depth difference with those objects.

## 9.2 Metric

Several measures like the Area-Under-the-Curve (AUROC) the Precision-Recall, have been suggested to evaluate the accuracy of salient object detection maps. However, as shown in [43], these most commonly-used measures do not always provide a reliable evaluation. The authors start by identifying three causes of inaccurate evaluation: 1) interpolation flaw 2) dependency flaw and 3) equal-importance flaw. By amending these three assumptions, they propose a new reliable measure called $F_\beta^w - measure$ and defined in Eq. 12.

$$F_\beta^w = (1+\beta^2)\frac{Precision^w * Recall^w}{\beta^2 * Precision^w + Recall^w} \tag{12}$$

with

$$Precision^w = \frac{TP^w}{TP^w+FP^w}$$
$$Recall^w = \frac{TP^w}{TP^w+FN^w}.$$

The weight $w$ has been chosen to resolve the flaws. This metric provides better evaluation than previous measures. We will use this new method to validate SuperRare3D on the database in order to be as fair and precise as possible.

## 9.3 Method

We made the validation of our SuperRare3D model (called SR3D) in two steps. First, we computed a 2D saliency map as a view of the 3D saliency map (2D projection). We compared SR3D to 5 other depth-extended (2.5D) models. The Weighted F-measure is used to compare SR3D with 2.5D saliency methods given a pre-segmented ground truth. Models of visual attention can be split in two main categories based on their purpose. The first category of models aims to predict the human eye gaze distribution. The second category focuses on finding interesting objects. Our model fits in the second category and intends to segment complex scenes into an object hierarchy based on the objects of interest. Some of them are extended to use depth feature maps (called further in this paper 2.5D models). Those models are the ones also used to asses our method in this section.

In [9], the authors aim at an extension of the visual attention model with the integration of depth in the computational model built around conspicuity and saliency maps. This model is an extension of center surround 2D saliency with depth proposed by [1]. In [19], the method constructs 3D layout and shape features from depth measurement that they integrate with image based saliency. This method an extension of center surround 2D saliency with depth proposed by [44].

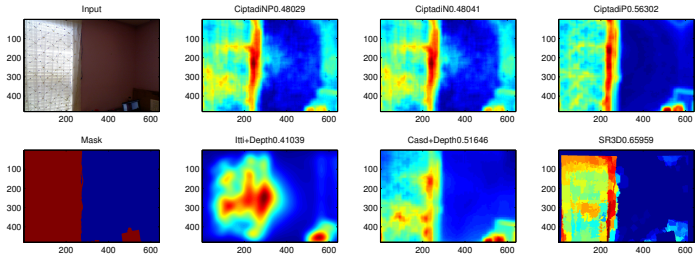


**Fig. 7** Qualitative comparison of our model.

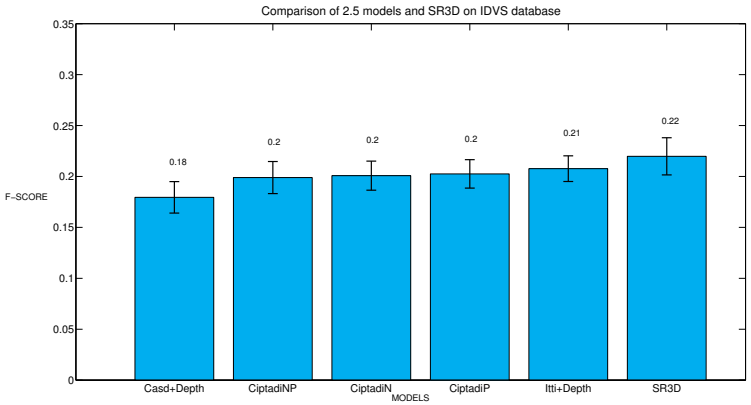


**Fig. 8** Quantitative comparison of our model with 5 state-of-the art 2.5 saliency models from the database [19]. SR3D outperforms with the other models.

## *9.4 Results*

Our full-3D model (SR3D) provides a 3D viewpoint independent saliency map of any kind of organized or unorganized point cloud. Figure 6 shows 2 examples of results for 2 different type of point clouds. First column shows the results from a single Kinect point cloud. Second column, an example of result on a point cloud obtained using a co-calibrated laser scanner is displayed. First row shows the input colored point clouds and second row the full-3D bottom-up viewpoint-independent saliency maps. This figure shows two crucial advantages of the proposed model over any existing 2D or 2.5D saliency model: (1) the ability to work on any kind of structured or unstructured point cloud and (2) the ability to provide viewpoint-free 3D saliency maps which might be adapted to any given viewpoint.

Figure 9 shows the results of the validation. Concerning the comparison with the 2.5D models, SR3D clearly outperforms all the other models. However,like shown on Figure , this performance difference is not statistically significant.

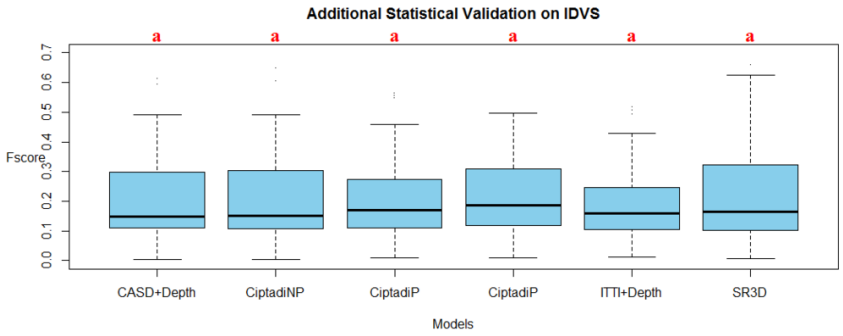


**Fig. 9** If our model outperforms the others, however, it is not significantly above.

## 10 Summary

3D is a fundamental element of the human vision system and it is as much for visual attention mechanisms. If the study and integration of 3D features in the design of computational models of attention began early, it is only in recent years they have really grown. This 3D information plays an essential role in a 3D Attention mechanism whatsoever in bottum-up than top-down. For bottum-up, the use of spatial information not only weight conventional salience map for giving importance to the regions according to their proximity but also by extracting 3D features, could improve significantly the performance of a saliency model. For the top-down, 3D data offer many opportunities to extract information on the environment, the scene or objects, leading to a more detailed or semantic analysis of the environment to constrain the saliency. In this chapter, we made a review of multiple methods of "3D saliency". Indeed, it is necessary to distinguish between models based on their dependence on a point of view, this is what prompted us to redefine our vision of saliency models by classifying them according to the notions of 2.5D and 3D. Given this classification, we have proposed a new model of salience based on rarity, effective in both categories, capable of handling large amounts of 3D data while taking into account the color information. Surprisingly few models are interested in full 3D that integrates this color information. Our model is efficient, 3D saliency is an early field but now re-emerges thanks to the appearance of numerous 3D sensors. This area is rich and complex and still offers many challenge in modeling human attention.

## 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 . . .*, pp. 0–4, 1998.

2. Zhengyou Zhang, "Microsoft kinect sensor and its effect," *MultiMedia, IEEE*, vol. 19, no. 2, pp. 4–10, 2012.
3. Matei Mancas, Nicolas Riche, Julien Leroy, and Bernard Gosselin, "Abnormal motion selection in crowds using bottom-up saliency," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 229–232.
4. Norman Villaroman, Dale Rowe, and Bret Swan, "Teaching natural user interaction using openni and the microsoft kinect sensor," in *Proceedings of the 2011 conference on Information technology education*. ACM, 2011, pp. 227–232.
5. Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, and Thierry Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 642–658, 2013.
6. Matei Mancas, Nicolas Riche, Julien Leroy, Bernard Gosselin, and Thierry Dutoit, "Toward a social attentive machine.," in *AAAI Fall Symposium: Robot-Human Teamwork in Dynamic Adverse Environment*, 2011.
7. A. Maki, P. Nordlund, and J.-O. Eklundh, "A computational model of depth-based attention," *Proceedings of 13th International Conference on Pattern Recognition*, vol. 4, no. 1, pp. 132–141, 1996.
8. Chang Ha Lee, Amitabh Varshney, and David W. Jacobs, "Mesh saliency," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 659, July 2005.
9. N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 1, pp. 375–378, 2000.
10. Timothee Jost, Nabil Ouerhani, and R Wartburg, "Contribution of depth to visual attention: comparison of a computer model and human behavior," *Proc. Early Cognitive . . . , ,* no. May, pp. 1–4, 2004.
11. H Hügli, T Jost, and Nabil Ouerhani, "Model performance for visual attention in real 3D color scenes," *Artificial Intelligence and Knowledge . . .* , pp. 469–478, 2005.
12. Simone Frintrop, Erich Rome, Andreas Nüchter, and Hartmut Surmann, "A Bimodal laser-based attention system," *Computer Vision and Image Understanding*, vol. 100, no. 1-2 SPEC. ISS., pp. 124–151, 2005.
13. Ekaterina Potapova, Michael Zillich, and Markus Vincze, "Learning what matters: Combining probabilistic models of 2D and 3D saliency cues," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6962 LNCS, pp. 132–142, 2011.
14. Nicolas Riche, Matei Mancas, Bernard Gosselin, and Thierry Dutoit, "3D Saliency for abnormal motion selection: The role of the depth map," *Computer Vision Systems*, 2011.
15. Congyan Lang, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan, "Depth matters: Influence of depth cues on visual saliency," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7573 LNCS, pp. 101–115, 2012.
16. Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu, "Leveraging stereopsis for saliency analysis," *Computer Vision and Pattern . . .* , 2012.
17. Junle Wang, Matthieu Perreira Da Silva, Patrick Le Callet, and Vincent Ricordel, "Computational model of stereoscopic 3D visual saliency.," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 22, no. 6, pp. 2151–65, June 2013.
18. Karthik Desingh, Madhava Krishna K, Deepu Rajan, and Cv Jawahar, "Depth really Matters: Improving Visual Salient Region Detection with Depth," *Procedings of the British Machine Vision Conference 2013*, pp. 98.1—-98.11, 2013.
19. Arridhana Ciptadi, Tucker Hermans, and James Rehg, "An In Depth View of Saliency," *Procedings of the British Machine Vision Conference 2013*, pp. 112.1—-112.11, 2013.
20. Takahiro Ogawa, Motoyuki Ozeki, and Natsuki Oka, "A Visual Attention Model Using Depth Information from the Point of Gaze," *ii.is.kit.ac.jp*, pp. 125–130, 2014.
21. Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji, "RGBD Salient Object Detection : A Benchmark and Algorithms," in *C_ECCV*, 2014, number 1, pp. 92–109.

22. N.D.B. Bruce and J.K. Tsotsos, "An attentional framework for stereo vision," *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, 2005.
23. Neil D B Bruce and John K Tsotsos, "Attention in Stereo Vision : Implications for Computational Models of Attention," *Developing and Applying Biologically-Inspired Vision Systems: Interdisciplinary Concepts*, pp. 65–88, 2012.
24. YS Liu, Min Liu, Daisuke Kihara, and Karthik Ramani, "Salient critical points for meshes," *... of the 2007 ACM symposium on ...*, 2007.
25. U. Castellani, M. Cristani, S. Fantoni, and V. Murino, "Sparse points matching by combining 3D mesh saliency with statistical descriptors," *Computer Graphics Forum*, vol. 27, no. 2, pp. 643–652, Apr. 2008.
26. Indriyati Atmosukarto and Linda G. Shapiro, "A salient-point signature for 3d object retrieval," *Proceeding of the 1st ACM international conference on Multimedia information retrieval - MIR '08*, p. 208, 2008.
27. Miquel Feixas, Mateu Sbert, and Francisco González, "A unified information-theoretic framework for viewpoint selection and mesh saliency," *ACM Transactions on Applied Perception*, vol. 6, no. 1, pp. 1–23, Feb. 2009.
28. Jianhua Lin, "Divergence measures based on the shannon entropy," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991.
29. Indriyati Atmosukarto and Linda G. Shapiro, "3D Object Retrieval Using Salient Views," *Proceedings of the international conference on Multimedia information retrieval - MIR '10*, p. 73, 2010.
30. G. Leifman, E. Shtrom, and a. Tal, "Surface regions of interest for viewpoint selection," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 414–421, June 2012.
31. Jinliang Wu, Xiaoyong Shen, Wei Zhu, and Ligang Liu, "Mesh saliency with global rarity," *Graphical Models*, vol. 75, no. 5, pp. 255–264, Sept. 2013.
32. Oytun Akman and Pieter Jonker, "Computing saliency map from spatial information in point cloud data," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6474 LNCS, no. PART 1, pp. 290–299, 2010.
33. GM Garca and Simone Frintrop, "A Computational Framework for Attentional 3D Object Detection," *Proc. of the Annual Conf. of the Cognitive Science ...*, pp. 2984–2989, 2013.
34. Elizabeth Shtrom, George Leifman, and Ayellet Tal, "Saliency Detection in Large Point Sets," *2013 IEEE International Conference on Computer Vision*, pp. 3591–3598, Dec. 2013.
35. R WANG, C GAO, J CHEN, and W WAN, "Saliency Map in 3D Point Cloud and Its Simplification Application," *Journal of Computational ...*, vol. 8, no. 61301027, pp. 3553–3560, 2014.
36. Yun Zhang, Gangyi Jiang, Mei Yu, and Ken Chen, "Stereoscopic visual attention model for 3d video," in *Advances in Multimedia Modeling*, pp. 314–324. 2010.
37. Julien Leroy, Nicolas Riche, Matei Mancas, Bernard Gosselin, and Thierry Dutoit, "Superrare:an object-oriented saliency algorithm based on superpixels rarity," 2013.
38. Tilke Judd, Frédo Durand, and Antonio Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.
39. Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter, "Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2027–2034, June 2013.
40. Borrmann Dorit, Elseberg Jan, Houshiar HamidReza, and Nchter Andreas, "Robotic 3d scan repository," .
41. Laurent Itti and Christof Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, pp. 161–169, 1999.
42. Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, and Thierry Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Sig. Proc.: Image Comm.*, vol. 28, no. 6, pp. 642–658, 2013.
43. Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal, "How to evaluate foreground maps?," *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
44. Stas Goferman, "Context-aware saliency detection," *Pattern Analysis and ...*, vol. 34, no. 10, pp. 1915–26, Oct. 2012.