Assistive Visual Framing in 3D Dense Points Cloud

Zaynab Habibi Guillaume Caron El Mustapha Mouaddib
Université de Picardie Jules Verne, MIS laboratory
33 rue Saint Leu, 80039 Amiens Cedex 1, France
zaynab.habibi, guillaume.caron, mouaddib @u-picardie.fr

***Abstract*—Recent progress in digital technology and its popularity in the context of cultural heritage are contributing to the emergence of several applications using a very large and dense digitization of archaeological sites or historical monuments. Navigation in such virtual environment is difficult especially for novice users. In this paper, we propose a new approach in order to make easier the navigation process by performing an assistive visual framing. This approach exploits a new visual feature: the image saliency-based Gaussian mixture. We applied our method on many different environments and we present the example of the Saint Sébastien chapel of the cathedral Notre Dame of Amiens. In order to evaluate our method, we present user evaluation and comparison between manual, assistive and automatic framing.**

Index Terms —visual framing; saliency; Gaussian mixture

I. Introduction

During the last decade, the increased use of digital technologies is growing day by day specially in the cultural heritage preservation giving the opportunity to develop new user experiences in several virtual reality applications as games and virtual navigation.

Our work is focused in the exploitation of a digital 3D model of the Cathedral Notre Dame of Amiens. This work is in the context of the ASSIDUITAS project where the aim is to provide an assistive system for virtual navigation in the 3D digital environment, coupling user inputs and automatic camera control. In this paper, we are interested in the assistive visual framing of a 3D points cloud model introduced in the literature [1] by the concept of reaching a relevant camera viewpoint. Starting from an initial camera pose with some image rendering information, we define the assistive framing by the action of moving the camera, to get more relevant image rendering information in its field of view, adequately with users inputs.

Viewpoint relevance can be modeled as the image brightness [2] or sharpness [3]. Information theory has also been exploited in [1] by Vasquez using the Shannon entropy, in order to compute the viewpoint quality. Polygonal meshes are considered in the latter work. To compute the probability distribution for the entropy expression, [1] uses the relative area of 3D mesh projected faces over the sphere of directions centred in the viewpoint. The maximum entropy is reached when a certain viewpoint can see all the faces with the same projected area. However, the latter approach cannot be applied to 3D point clouds.

In the same context of viewpoint relevance computation, to detect the most interesting part of an image or a scene, we identify the works based on the exploitation of visual attention. In bottom-up approaches of visual attention modeling, saliency algorithms were widely used to detect some object in cluttered environments. The most known work [4] integrates different maps and compute the center-surround difference in order to detect the most salient region on the image. The saliency computation algorithms were also extended to 3D environments, of which the aim is to compute the saliency for a mesh [5] or a 3D point cloud [6].

Since framing also includes the camera motion control, the visual relevance must be linked to the camera degrees of freedom. For doing so, we exploit the image-based visual servoing framework by iteratively updating the camera pose, optimizing a criterion defined in the image plane. For instance, Marchand [2] presents a visual servoing approach that maximizes the image brightness. Even if the latter work was not designed for framing, it can be seen, from a methodological point of view, as an automated framing approach. Indeed, it defines a relevance measure, that is the image brightness, and the control algorithm to reach the viewpoint at which that relevance is maximum, thanks to visual servoing. Then, we propose a new relevance criterion that is suitable for the relevance computation itself and for the camera control too: a new modelling of the image saliency, that is the image saliency-based Gaussian mixture. This represents our key contribution, detailed in Section II, including constraints, such as maintaining a minimum distance between the camera and the object in addition to the user interaction consideration.

II. Proposed Method

A. Relevance Modelling

A good framing in our context, implies to give relevant information in the rendered image and to center it in all the image. [7] [4] demonstrate that the saliency is a good candidate as relevant information for visual human exploration. That's why we propose to maximize the saliency. However, the usage of the saliency as visual feature in its raw state is insufficient and limited (the algorithm falls easily in a local maximum). That is the reason why we introduce a new image feature: saliency-based Gaussian mixture (GM). This new modelling is independent to the saliency estimation method.

The maximisation process will be done by an iterative non linear optimization. Applied to the camera displacement problem, from an initial camera pose, the estimation of the

978-1-5090-0048-7/15/$31.00 ©2015 IEEE

camera displacement is computed by an iterative scheme in order to find the camera pose optimizing a cost function $f(\mathbf{r})$:

$$\mathbf{r}^* = arg\min_{\mathbf{r}}(f(\mathbf{r})) \tag{1}$$

$f(\mathbf{r})$ depends on the saliency. The iterative scheme consists in estimating the pose increment $\dot{\mathbf{r}}_k$ to modify the current pose $\mathbf{r}_k$ (step k):

$$\mathbf{r}_{k+1} = \mathbf{r}_k + \dot{\mathbf{r}}_k \tag{2}$$

We consider $I(\mathbf{r})$ the image at the pose $\mathbf{r} = (t_X, t_Y, t_Z, \theta w_X, \theta w_Y, \theta w_Z)^\top$, where the three translations are $[t_X, t_Y, t_Z]^\top$ and the rotations are represented by an angle θ and an unit vector (axis of rotation) $\mathbf{w} = [w_X, w_Y, w_Z]^\top$. n the virtual environment, the image $I(\mathbf{r})$ is obtained from the rendering of the 3D model. The sixth camera degrees of freedom (d.o.f) will not be considered to avoid unrealistic rotations around the virtual camera optical axis. I

From $I(\mathbf{r})$, we compute the saliency map I_s of the image at the current pose $\mathbf{r}_k$ using [7]. Then, each salient pixel is considered as a Gaussian. Thus, the saliency distribution in the image is modelled with a saliency-based GM.

GM is a parametric probability density function constituted as a sum of Gaussian component densities. Thus, an image is considered as:

$$G(r) = \sum_i g_i(\mathbf{u}_i, \mu_i, \sigma_i) \tag{3}$$

where $\mathbf{u}_i = (u, v)$ are the pixel coordinates of index i and $g_i(\mathbf{u}_i, \mu_i, \sigma_i)$ are the component mixture densities for the pixel i. We choose the saliency pixel coordinates for the mean vector μ_i to avoid more treatments.

To avoid complex treatments and considering the image representation, we propose that σ is proportional to the image intensity $I_s(u, v)$ at the pixel (u, v) :

$$\sigma_i = \lambda I_s(u, v) \tag{4}$$

Where the coefficient λ is the same for the whole image. This parameter allows to adapt the Gaussians with and to center the saliency area in the image. But so that the maximization converges, the following condition must be respected:

$$-\lambda I_s(u, v) < \mathbf{u}_i - \mu_i < \lambda I_s(u, v) \tag{5}$$

The goal is to frame the salient information in the image, then $2\lambda I_s(u, v)$ should be equal to the image width or height.
If we choose to perform the framing in proportion to the width W, then we take $2\lambda I_s(u, v) = W$. Thus, we obtain:

$$\lambda = \frac{W}{2I_s(u, v)} \tag{6}$$

where W is the image width.
However the image is represented by a mixture of Gaussian then we propose to approximate I_s by the mean of saliency values in the image $\bar{I}_s$. Finally:

$$\lambda = \frac{W}{2\bar{I}_s} \tag{7}$$

As detailed in the experimental part, this approximation is available for ever considered situations.

B. Camera Control

1) Maximizing the Saliency Based GM: The goal is to maximize $G(\mathbf{r})$ (eq. (3)) which is equivalent to minimize its opposite $-G(\mathbf{r})$. Applying the Taylor expansion of the function G at the new pose $\mathbf{r}_{k+1}$, we solve:

$$\frac{\partial(-G(\mathbf{r}_{k+1}))}{\partial \mathbf{r}} = 0 \tag{8}$$

We note the partial derivative of the function $-G$ with respect to $\mathbf{r}$ by $d_{\mathbf{r}}(-G)$ and the second partial derivative by $d^2_{\mathbf{r}}(-G)$:

$$d_{\mathbf{r}}(-G(\mathbf{r}_{k+1})) = d_{\mathbf{r}}(-G(\mathbf{r}_k)) + d^2_{\mathbf{r}}(-G(\mathbf{r}_k))\dot{\mathbf{r}}_k \tag{9}$$

After the computation of the Gradient and the Hessian, solving $d_{\mathbf{r}}(G(\mathbf{r}_{k+1})) = 0$ using the Taylor development leads to the saliency-based GM control law given by:

$$\dot{\mathbf{r}}_k = (d^2_{\mathbf{r}}(-G(\mathbf{r}_k)))^+ d_{\mathbf{r}}(-G(\mathbf{r}_k)) \tag{10}$$

where $()^+$ is the pseudo-inverse.

2) Managing Distance Control Law: To manage the distance between the camera and the object, we consider all the vertices constituting the 3D object and we maximise the distance between them and the camera which is equivalent to minimize:

$$D(\mathbf{r}) = \sum_k \frac{1}{2||{}^w\mathbf{p}_c - {}^w\mathbf{p}_{o_l}||^2} \tag{11}$$

such that l is the index of a vertex, ${}^w\mathbf{p}_c = (x_c, y_c, z_c)^T$ is the camera position and ${}^w\mathbf{p}_{o_l} = (X_{o_l}, Y_{o_l}, Z_{o_l})^T$ is the position of one vertex in the world frame $\Re_w$. As we want to avoid that the camera goes near the considered object, only the position (specially the position on the camera Z axis) is considered and not the orientation because we suppose that our camera is a point (optical center).

Using again the first order Taylor development, we solve:

$$D(\mathbf{r}_{k+1}) = 0 \tag{12}$$

We obtain the distance based control law given by:

$$\dot{\mathbf{r}}_k = (d_{\mathbf{r}}(D(\mathbf{r}_k)))^+ D(\mathbf{r}_k) \tag{13}$$

3) Automatic Control Law: The two previous control laws are combined to obtain the final pose increment $\dot{\mathbf{r}}_k$:

$$\dot{\mathbf{r}}_k = {}^G\dot{\mathbf{r}}_k + \beta\, {}^D\dot{\mathbf{r}}_k \tag{14}$$

where ${}^G\dot{\mathbf{r}}_k$ the pose increment obtained from Equation (10), ${}^D\dot{\mathbf{r}}_k$ the pose increment resulting from Equation (13) and $\beta > 0$ is a parameter to fuse together the both control laws (β is the ratio between the two pose increments norm ${}^G\dot{\mathbf{r}}_k$ and ${}^D\dot{\mathbf{r}}_k$).
We considered a weighted linear combination of the two pose increments, this choice was made empirically and experimentally validated. However, in our perspectives a deeper work will be provided on the combination.

4) Assistive Control Law: In addition to the automatic control laws presented above, we give the user some freedom in the navigation while considering an on-line process, which consists in interactively modifying the distance between the virtual camera and the object. Then, using keyboard keys for instance, he can decide of the distance which separate him from the considered object. The pose increment generated by the user ${}^{U}\dot{\mathbf{r}}_k$ is combined with the automatic control law $\dot{\mathbf{r}}_k$ (eq (15)) to obtain the assistive control law ${}^{A}\dot{\mathbf{r}}_k$:

$$ {}^{A}\dot{\mathbf{r}}_k = \dot{\mathbf{r}}_k + \gamma {}^{U}\dot{\mathbf{r}}_k \tag{15} $$

where $\gamma > 0$ is a parameter adjusting the convergence speed.

The algorithm presented in Fig. 1 sums up all the process to compute the final control law taking into account the two control laws and the user input.

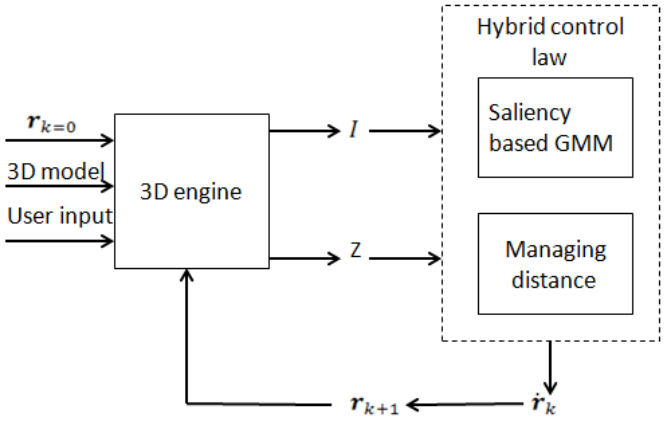

Fig. 1: The process loops until convergence

III. Experimentation

A. Visual Framing

We applied our approach on a 3D model composed from 3D colored point clouds: Saint Sébastien chapel of the cathedral of Amiens. The model is precisely measured with Leica C-10 laser range-scanners. For the visualization of the 3D model, we use the OGRE 3D graphics engine. This section demonstrates our results on this type of data.

Starting from an initial camera pose which is poor in visual information we perform an automatic visual framing using every camera d.o.f except the rotation around the camera optical axis (Z axis). This rotation is blocked to avoid unrealistic rotations. The visual framing is performed using the final control law (Equation (15)). It was important to add the constraint maintaining a minimum distance between the camera and the 3D object and also the user interaction. Indeed, in one hand the saliency-based GM control law alone can be performed while getting very close to the object, so the object will take all the camera field of view (not always relevant, it depends on the nature of the object), and in other hand we give the user the freedom to get closer or farther to the considered object. According to the evaluation achieved in the next subsection, the majority of users may agree that a good framing is the one that will ensure to center the salient information in the rendered image, but moving towards the scene details is something difficult to automate.

From Figure 2, we observe that the virtual camera moves in the environment while successfully positioning in front of the chapel including in its field of view all the "relevant" information (the painting and the statue in this case). Furthermore, when we observe the Gaussian mixture map, in the initial camera pose the GM main peak is positioned in the left side and slightly in the top of the map while it is centred when the convergence is reached.

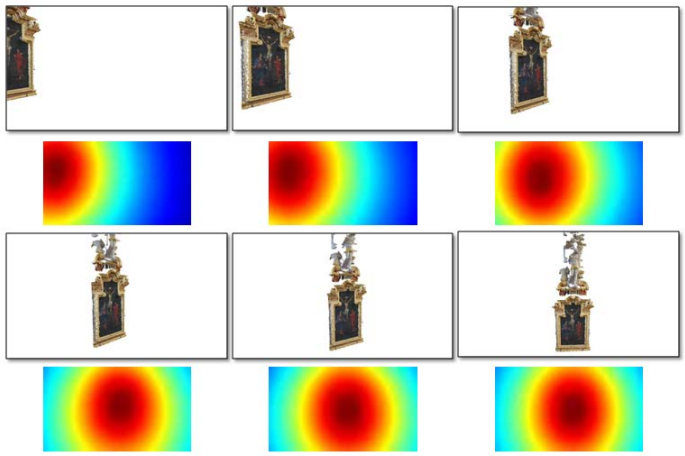

Fig. 2: Resulted images sequence for the visual framing and their corresponding saliency based GM maps

We observe in Fig. 3 a top view of the 3D model with the resulting camera path in red for the visual framing task.

Fig. 3: Top view of the 3D model and the camera path (the yellow pyramid correspond to the initial camera pose while the blue pyramid to the final camera pose)

We conducted a second experimentation, but this time starting form a different camera pose. In this example, the chapel is centred regarding the image width, this is why the initial GM map is centred, but the chapel is shifted up in the initial image hence a small part of the GM map is hidden (in the bottom). The goal in this experiment is to correct the camera position regarding the relevant information. This is well done when we observe the resulting images sequence in Fig. 4 and the final camera path in Fig. 5. We observe in these two experiments that GM surface is uniform in the map, thanks to the used method (Equation (7)) to compute the λ parameter.

B. System Evaluation

We conducted a formal evaluation in order to determine the behaviour of the visual framing task in three navigation modes:

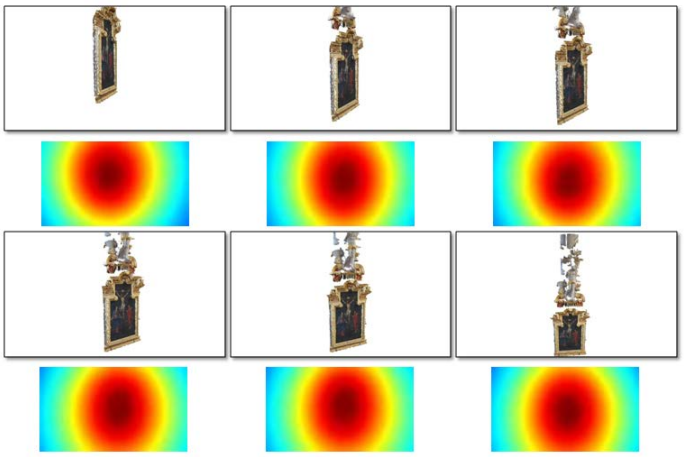

Fig. 4: Resulted images sequence for the visual framing and their corresponding saliency based GM maps

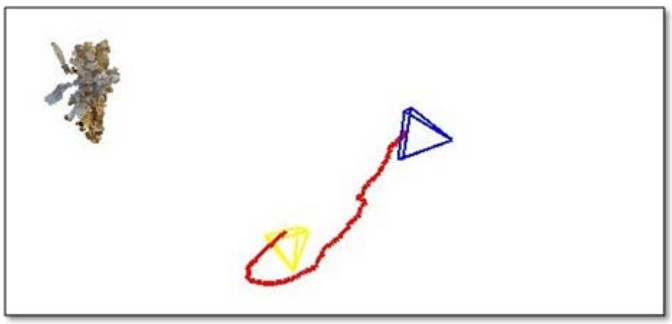

Fig. 5: Top view of the 3D model and the camera path (the yellow pyramid correspond to the initial camera pose while the blue pyramid to the final camera pose)

manual, assisted and automatic. The aim of the experiment is to compare between these three modes when the user is asked to perform a visual framing of the complete model of the chapel. The evaluation was achieved by 10 participants (6 males, 4 females) aged from 23 to 50 participated in the experiment. To compute the final control law, we use 60×30 for the image resolution, which enabled us to perform the navigation in real time without reducing the robustness.

In Figure 6a, we give the user the 3D model of the chapel with an initial camera pose and the goal was to frame the chapel by moving the camera manually using keys keyboard. The controlled factors is the framing quality which corresponds to the final obtained image in the automatic mode, in the manual mode and the resulted video. We observe that the users were more attracted by the result provided by the automatic mode because the virtual camera allows to center the chapel and to give a relevant viewpoint. Concerning the video quality, the users were more impressed by the resulting video of the automatic mode because the camera d.o.f are coupled, so the camera movement is smooth and not jerky contrary to the manual mode where the camera d.o.f are not coupled (see the video accompanying this paper). In Figure 6b, we conduct a satisfaction study to compare the three mode. It is clear from the diagram that the users are more interested by the assisted mode.

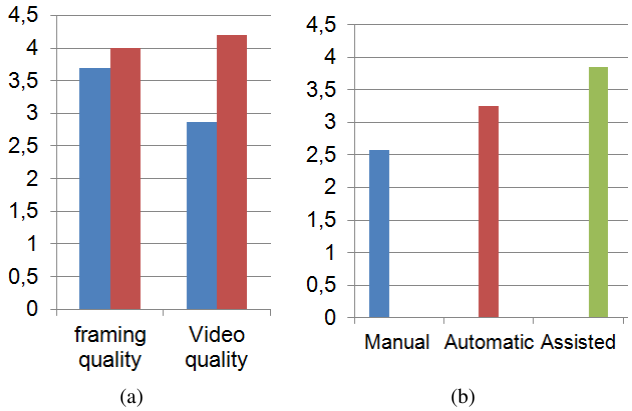

Fig. 6: (a) User evaluation of the framing and the resulted video quality (blue column corresponds to the manual mode while the red one to the automatic mode); (b) User evaluation of the three navigation modes

IV. Conclusion

The addressed problem in this paper is to control the 3D displacement of a virtual camera in order to perform an assistive relevant visual framing of a 3D object in the context of cultural heritage. A new photometric image feature, ie saliency-based GM, was proposed to tackle this problem combined with a managing distance constraint and adequately coupled with user interaction. We validated this new approach on the Saint Sébastien chapel of the cathedral Notre Dame of Amiens. Results show that the proposed method is robust and meets our expectations while using images of very low resolution to compute the control law. The assistive framing is performed in real-time (using low image resolution decreased significantly the computation time). A user study was conducted to highlight our contribution.

Acknowledgments

This work was supported and financed by Région Picardie and fonds FEDER (ASSIDUITAS project).

References

[1] P. Vázquez and M. Feixas and M. Sbert and W. Heidrich, *Viewpoint Selection using Viewpoint Entropy*, Workshop on vision, modelling and visualisation. vol. 1 p. 273–280, 2001.

[2] E. Marchand, *Control camera and light source positions using image gradient information*, IEEE International Conference on Robotics and Automation. p. 417–422, 2007.

[3] K. Tomihisa and K.Satoru, *A simple method for computing general position in displaying three-dimensional objects*,Computer Vision, Graphics, and Image Processing. vol. 41 p. 43–56, 1988.

[4] L. Itti and C. Koch and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis*, IEEE Transactions on pattern analysis and machine intelligence. vol. 20 p. 1254–1259, 1998.

[5] C H. Lee and A. Varshney and D W. Jacobs, *Mesh saliency*, ACM Transactions on Graphics. vol. 24 p. 659–666, 2005.

[6] S. Elizabeth and L. George and T. Ayellet, *Saliency Detection in Large Point Sets*, IEEE International Conference on Computer Vision. p. 3591–3598, 2013.

[7] A. Radhakrishna and H. Sheila and E. Francisco and S. Sabine, *Frequency-tuned salient region detection*, IEEE Conference on Computer Vision and Pattern Recognition. p. 1597–1604, 2009.