Proceedings of the 18th Mediterranean Electrotechnical Conference
MELECON 2016, Limassol, Cyprus, 18-20 April 2016

Monoscopic Inpainting Approach using Depth Information

Dylan Seychell [1], Carl James Debono [2]

Department of Communications and Computer Engineering,
University of Malta, Msida, MSD 2080, Malta
[1] dylan.seychell@ieee.org
[2] c.debono@ieee.org

***Abstract*—Cheap depth sensors that can be integrated in consumer cameras provide additional data that can be used for improved post-processing results of the captured images. Removal of objects in a scene is one such editing procedure that demands inpainting techniques that limit noticeable artifacts generated in the process. In this paper, a monoscopic inpainting technique that uses depth information to process results is presented. It allows users to select an object from the foreground that needs to be removed and then inpaints this region from the neighborhood. This approach uses texture and depth information and is pipelined in a way that allows for parallelization. Results are returned in 0.034 seconds on average. A mean opinion score evaluation was carried out and the current technique scored an average of 3.24 from a scale of 5 on the quality of inpainted regions. This exercise was held to identify the attributes that need to be improved in future implementations.**

I. Introduction

Natural multimedia such as photographs and video streams can be captured from a range of devices that allow consumers to capture a record of different moments. The need of editing an image and removing objects often arises and users have to resort to inpainting techniques, such as the patch-match [1], that can be found in different commercial software packages. Since such an algorithm is generally based on a randomized approach of the closest regions, it might perform badly in certain situations such as the example shown in Figure 1.

The stereoscopic approach of using a selection of Texture (color) and Depth images of the same scene from multiple points allows us to process any given region of interest in terms of pixel values within the structural context provided by the depth information [2]. On the other hand, such stereoscopic techniques, take advantage of epipolar geometry to predict the structure of objects that are occluded in the view of a camera but are visible in another. The monoscopic technique presented in this paper makes use of a single viewpoint and takes advantage of the depth information to address challenges that were exposed in techniques such as [1] that make use of only texture information. With the proliferation and availability of depth cameras, such techniques are becoming more feasible and attractive. Recent projects such as Google's Project Tango [3] are showing that depth cameras will soon be available on mobile devices. This means that inpainting techniques such as the one presented in this paper should be as efficient as possible to run on portable devices while making use of monoscopic information.

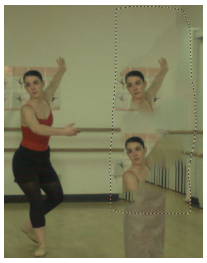

Fig. 1. Case where the Patch-Match algorithm performs badly [1]

The first part of this paper presents a brief background of the techniques being used to optimize the system. Subsequently, the proposed solution is presented and explained. This is then followed by a report of an evaluation exercise that was carried out to shed light on how to further improve the current technique.

II. Background

A. Inpainting

Inpainting is the process of modifying an image such that the editing is not perceived, or its visual impact is minimized, by filling the region of interest with texture that is known from another location within the image [2]. Inpainting approaches may be categorized into 2: Structural and Textural. Structural inpainting is when geometric information from the target area is used in order to reproduce the missing content. Textural inpainting takes advantage of repetitive patterns while disregarding any structural information in the surrounding pixels.

$$\bigtriangledown L \cdot \overrightarrow{N} = 0 \quad (1)$$

978-1-5090-0058-6/16/$31.00

Modern computerized inpainting techniques combine these two types. The traditional approach, based on differential equations, follows Equation 1 where the change of information happens perpendicular to the edge of the mask towards the center of the region to be inpainted. This means that Image **I** will deform until $\bigtriangledown L \cdot \overrightarrow{N}$ reaches an equilibrium. This can be implemented either along edges or a line of luminance.

B. Segmentation

In the processing procedure of the depth information, image thresholding can be carried out to reduce the grayscale image into a binary image for faster processing. Otsu's method assumes that the image contains two classes of pixels following a bi-modal histogram [4]. This results in the separation of foreground pixels and background pixels without the need of carrying out more complex operations for segmentation. Equation 2 shows the general equation that minimizes the intra-class variance. The weights q_1 and q_2 are the probabilities for the respective classes separated by threshold *t* and variance ${\sigma_i}^2$. The expansion of Equation 2 may be found in [4].

$$ {\sigma_w}^2(t) = q_1(t){\sigma_1}^2(t) + q_2(t){\sigma_2}^2(t) \tag{2} $$

where

$$ q_1(t) = \sum_{i=1}^{t} P(i) \tag{3a} $$

$$ q_2(t) = \sum_{i=t+1}^{I} P(i) \tag{3b} $$

and

$$ \mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_1(t)} \tag{4a} $$

$$ \mu_2(t) = \sum_{i=t+1}^{I} \frac{iP(i)}{q_2(t)} \tag{4b} $$

then

$$ {\sigma_1}^2(t) = \sum_{i=1}^{t} [1 - \mu_1(t)]^2 \frac{iP(i)}{q_1(t)} \tag{5a} $$

$$ {\sigma_2}^2(t) = \sum_{i=t+1}^{I} [1 - \mu_2(t)]^2 \frac{iP(i)}{q_2(t)} \tag{5b} $$

III. Proposed Monoscopic Inpainting Algorithm

The solution presented in this paper focuses on the processing of a single frame for simplification purposes. It can be however applied to a stream and its advantages will be presented in later sections of this paper. This technique requires the depth map and texture frames for the scene. Furthermore, it requires the user to identify the foreground object that needs to be inpainted by means of a bounding box. The *Ballet* and *Balloons* sequences were used to evaluate this technique.

A. Preliminary Modules

In the first instances of the technique, the user is presented with the texture image and is allowed to draw a bounding box around the object that is required to be inpainted. The initial and final sets of coordinates are preserved. These coordinates are then used to identify the region of interest on the depth map that are used for processing.

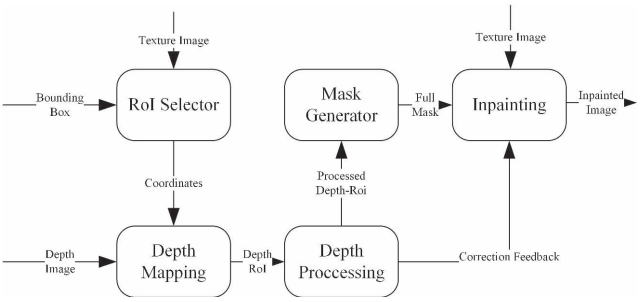

Fig. 2. Data-flow representation of the technique

B. Depth Processing

In the depth processing module, thresholding using Otsu's method is carried out on the region of interest. The image is then transformed into a binary single channel image. After this transformation, two sequential routines of region growing are carried out mainly to compensate for the difference in translation between the depth and texture camera. In each region growing routine, all neighboring pixels of a foreground pixel are assigned as foreground pixels.

Subsequently, a distance transform operation is carried out on the region of interest. This is a representation of an image where every pixel on the foreground takes the value of its distance from the background. This therefore means that the further the pixel is from the background, the higher its intensity value is. This representation is then used to pass feedback to the inpainting module.

The algorithm of the Depth Processing module can be outlined accordingly:

1) Generate set of working images with the dimensions of the Region of Interest, RoI
2) Create roiD; the working image with reduced dimensions of the RoI and its respective depth information
3) Covert roiD to greyscale
4) Use Otsu's method to find the threshold on the roiD image, apply 255 as the pixel value to the pixels whose value in the source is greater then the computed threshold level
5) Using fast-marching, compute a distance transform and store separately
6) Repeatedly apply two runs of region growing on the resultant roiD. For each run, the neighbor of each pixel with value 255 is also set to 255.

Prior to the interviews explained in the evaluation process below, laboratory experiments were carried out to fine-tune the parameters of this algorithm. Region Growing is the key parameter of this method and it is applied twice. Experiments showed that a single run was not very effective and more than two runs were redundant.

C. Mask Generation and Inpainting

The processed depth information is passed as a single channel image to the mask generator. The mask generator also has the coordinates of the RoI. A new empty image with the same dimensions of the original image is created. The set of coordinates is used to stitch the processed depth RoI onto the empty black image and thus generate the mask for the object to be inpainted. The mask is then passed to the Inpainting module. The inpainting module carries out a Navier-Stroke inpainting [5] procedure using the mask and minor corrections are then carried out with the feedback given from the depth processing module. The feedback procedure allows for significant future improvements to this technique following the evaluation presented in the next section.

The algorithm of the Mask Generation and Inpainting module can be outlined accordingly:

1) Create a blank black image, B, with the dimensions of entire image frame
2) On B, using the coordinate data of the RoI, identify the position where the resultant image from the Depth Processing module is to be stitched
3) Stitch roiD on B.

The only parameters used that affect the performance of this module are the coordinates where the RoI needs to be stitched. On the other hand, the parameter of Inpainting affects the quality of the resultant image. The radius of pixels considered by the inpainting algorithm was set to 1. This is the minimum value for such parameter since experiments showed that greater values reduce the quality of the inpainted region.

IV. Evaluation

The evaluation of the proposed algorithm was first carried out in laboratory conditions where experiments were carried out to compare it to existing techniques. The challenge at this phase was that similar algorithms use stereoscopic sequences while this algorithm is monoscopic. In order to address this situation, the evaluation procedure explained below was undertaken.

A. Results

During subjective evaluation tests, the proposed algorithm was evaluated using Microsoft's *Ballet* sequence [6] and Nayoga's *Balloons* [7] sequence. In the laboratory tests, the proposed technique was tested using a larger set of sequences. Eventually, 2 sequences were selected in order to keep the interviews manageable and avoid test subject fatigue. From previous projects, we learnt that the effectiveness of interviews falls drastically when they are longer than 7 minutes. A Mean Objective Score (MOS) as specified by the ITU-T BT.500 [8] was used. 23 non-expert individuals were interviewed while being presented with artefacts from the technique in question. 59% of the respondents were males and the average age was 31. There were no observers with serious visual impairments. The images presented had a resolution of 1024x768 pixels.

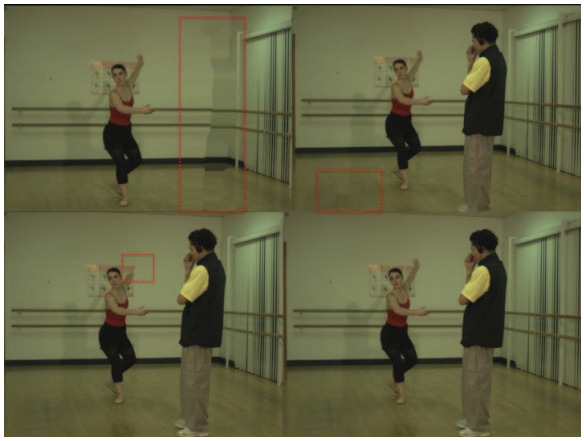

Fig. 3. Selection from the *Ballet* Sequence presented to observers. Impairment regions are indicated by the red dotted boxes.

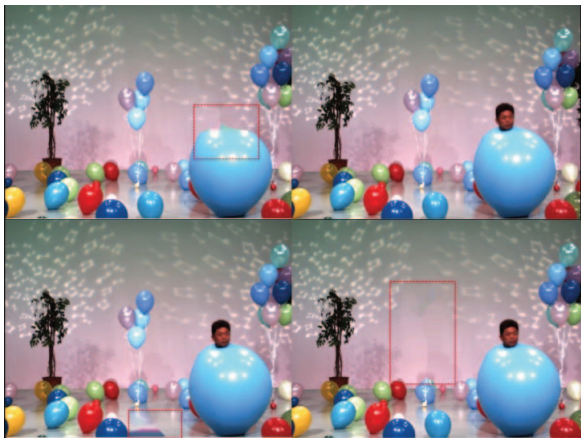

Fig. 4. Selection from the *Balloons* Sequence presented to observers. Impairment regions are indicated by the red dotted boxes.

A frame from each sequence was chosen. The observers were presented with 4 variations of each frame. In these sets, there were 3 frames that were edited using this technique while the other one was the original frame. The frames presented to each observer are shown in Figures 3 and 4. Observers were asked to rate the severity of the impairment in the image provided together with their perception of quality of the same images. Each session took around 6 minutes and the observers were encouraged to elaborate on their votes, stating what attribute of the impairment was found to be most annoying.

It was observed that in general, observers tend to either comment about issues related to texture or issues related to

structure. Texture issues refer to defects in change of gradients, smudges and also blurring among others. On the other hand, Structure issues refer to cases where the user expected a certain structure to be maintained where in reality it was not. Examples of this are the misaligned banisters in the first frame of the *Ballet* frame and the interrupted circular nature of the balloon in the first frame of the *Balloons* frame. In 32% of the cases, observers constantly commented about both texture and structure while the rest were equally divided.

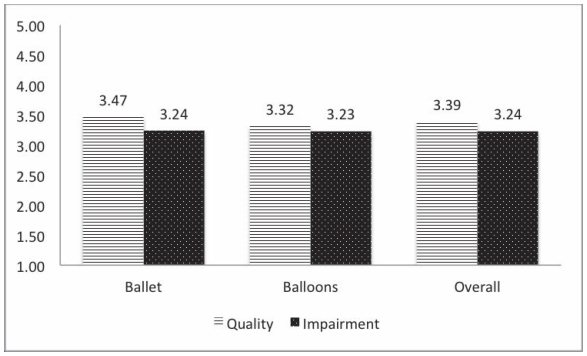

Fig. 5. Scores for Quality and Impairment of the two scenes presented above.

Figure 5 shows the MOS for both sequences, followed by an overall score. Each score denotes Quality (left) and Impairment (right). The overall mean opinion score for impairments of both scenes was that of 3.24. The worst impairment score was obtained in the first top left image from the *Ballet* set shown in Figure 3, where the man on the right hand side was inpainted. The MOS for the impairment of this frame was that of 2. Remarks vary from texture to structure and observers explained that the background had structural properties that they felt had to be respected. 39% of the observers underlined this importance of structural attributes for this frame and an equal number of observers commented about the texture. Only one observer found the image to be fine while another 17% complained about the center-line where the inpainting procedure met at the center of the inpainted object. The inpainting algorithms used within this technique involve a distance transform. Prior to this evaluation, work had already started to mitigate the weight of this transform and this should therefore improve results in future experiments. After the experiment was concluded, the observers where presented with the results of the Patch-Match algorithm presented in Figure 1. The disadvantage of the Patch-Match algorithm of not having depth information was explained but all observers remarked that they would prefer the results of the technique presented in this paper after seeing Figure 1, even though they gave a MOS of 2 to the 'preferred' figure.

Another interesting result was the importance of semantic context in impaired locations of the image. There were a significant number of cases where observers explained that they found the impairment annoying because they were expecting something in the image and it was actually inpainted. The clearest example is the inpainted arm of the ballerina as presented in the top-right image of Figure 3. While this frame had an impairment MOS of 3.04, 68% of the observers mentioned that they were mostly annoyed because the arm was not complete. Some observers also went into the depth of explaining how they tried to match the arm with the shadow to see whether it was actually an impairment or the arm was actually occluded. These comments in themselves imply that the technique performed well especially when considering that only 8% of the respondents complained about the defects in structure that are mainly given away by the distorted poster in the background. The importance of semantic context was then confirmed when the observers were presented with the top-right frame of Figure 4 after seeing the top-left frame. When seeing the top-left frame, observers complained about the uneven structure of the balloon that gave away that there was an intervention in this region. However, after seeing the head of the man in the top-right frame, 27% thought that the image was edited and the head was placed on the balloon. These results show that in the consideration of impairment, a significant number of observers also consider the semantic context of the objects being observed.

B. Performance Results

Following the procedure described above, the observers were presented with a working demonstration artefact of this technique. They were asked to try it out and give feedback about its performance. There were no perceivable delays and many described the result as 'instant'. In the background, the time taken for the program to carry out the operation with the user's selection was being recorded. On average, the program returned a result in 0.035 seconds. The worst time was recorded when a large object is selected and this was 0.074 seconds. On the other hand, when the area selected was smaller the program returned a result in 0.009 seconds. There were no complaints about the performance.

The computational complexity depends on the size of the Region of Interest. For a RoI w pixels wide and h pixels high, the complexity of this algorithm is $3wh \Rightarrow \mathcal{O}(wh)$.

V. Conclusion

The key findings in this paper show that while there is room for improvement, it is important that the performance is kept as close as possible to the one in the current results. On the other hand, a MOS of 3.24 indicates there are issues that need to be addressed. The most annoying impairment was the centreline in the inpainted region and the way it affected the surrounding texture. This can addressed by applying focused filters on the affected area to even out the disparity of the background textures. However, as supported by this exercise, structure is very important and needs to be preserved as much as possible. This can be addressed by tracing the contour of the object to be inpainted and applying techniques such as those used in [9] to preserve structure.

Another interesting finding was the issue of the location of the impairment on the screen. Observers complained mostly about impairments that were visible in the central regions of

the images and in most cases did not notice the impairments in the peripheral regions. This may be used to preserve the performance of the technique at hand. Effort should be made to improve the quality at the central regions of the image and the algorithm should be set to focus its resources on these regions rather than the entire region that is required to be inpainted. The last category of observer remarks was that related to the semantic context. As described earlier, there were cases where the observers identified an impairment because they found something in the image that did not make sense for them, such as the missing hand or the head on the balloon. These were selected on purpose to assess the weight of the semantic context and these results show that the algorithm should also cater for the semantic context in order to mitigate the effect on the context prior to presenting the final result.

REFERENCES

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," in *Proc. of ACM SIGGRAPH 2009 Papers*, ser. SIGGRAPH '09. New York, NY, USA: ACM, 2009, pp. 24:1–24:11.

[2] L. Wang, H. Jin, R. Yang, and M. Gong, "Stereoscopic inpainting: Joint color and depth completion from stereo images," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[3] "Google Project Tango," [Online]. Available: https://www.google.com/atap/project-tango/, accessed: 2015-05-22.

[4] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[5] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2001, pp. 355–362.

[6] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *ACM SIGGRAPH 2004 Papers*, ser. SIGGRAPH '04. New York, NY, USA: ACM, 2004, pp. 600–608. [Online]. Available: http://doi.acm.org/10.1145/1186562.1015766

[7] "Nagoya university sequences," [Online]. Available: http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/, accessed: 2015-06-1.

[8] "ITU-T BT.500. Methodology for the subjective assessment of the quality of television pictures," Aug. 2012.

[9] A. Criminisi, P. Peréz, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.