Sewing Photos: Smooth Transition Between Photos

Tzu-Hao Kuo*, Chun-Yu Tsai*, Kai-Yin Cheng*, and Bing-Yu Chen**

National Taiwan University
*\{kakukogou,apfelpuff,keynes\}@cmlab.csie.ntu.edu.tw, **robin@ntu.edu.tw

Abstract. In this paper, a new smooth slideshow transition effect, *Sewing Photos*, is proposed while considering both of smooth content transition and smooth camera motion. Comparing to the traditional photo browsing and displaying work, which all focused on presenting splendid visual effects, *Sewing Photos* emphasizes on the smooth transition process between photos while taking the photo contents into account. Unlike splendid visual effects, the smooth transition tends to provide more comfortable watching experience and are good for a long-term photo displaying. To smooth the content transition, the system finds the similar parts between the photos first, and then decides what kind of camera operations should be applied according to the corresponding regions. After that, a virtual camera path will be generated according to the extracted Region of Interests (ROIs). To make the camera motion smoother, the camera path is generated as a cubic interpolation spline.

1 Introduction

People right now tend to take more photos at the same place [14], because of the advanced camera technologies and cheap storage. Hence, owing to the changed behavior, we do not only have more photos in one photo album, but also have more redundant information of a photo set.

To deal with the digital photos, traditional approaches usually utilize the extracted features from photos through a series of image analysis. The duplicated photos will be eliminated first before further processing [4, 9]. Though some research work like Microsoft Photosynth[1] [18, 17] and Photo Navigator [8] utilized the redundant information of a huge amount of photos to generate the 3D effects for photo browsing and presentation, they are not very suitable for creating a photo slideshow for casual usages.

According to the results from our field study, we found that though splendid slideshow visual effects like Animoto[2] can attract people's eyes, it is not suitable for playing repetitively and watching for a long duration. Hence, for different purposes, the needs will be quite contrary. However, the traditional researches for creating the slideshow visual effects usually focused more on the short-time

[1] http://photosynth.net/

[2] http://animoto.com/

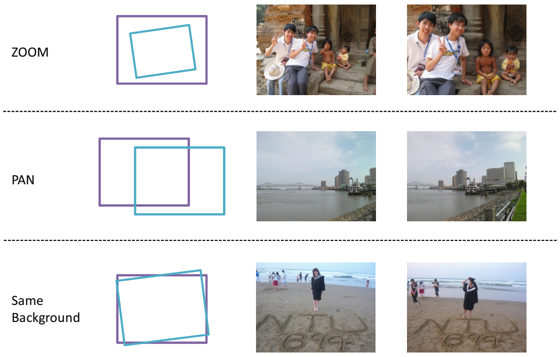


Fig. 1. Three major types of normal users' photo taking behaviors.

eye catching effects, rather than the needs for a long time playing. Besides advertising, there are many situations needed for a long-term playing including home photo displaying or repetitively playing photos in a digital photo frame. However, the traditional fade-in and fade-out effects do not take the photos' content into consideration, and just arbitrarily run through the photos, and the result is actually not really acceptable.

Through the observations of normal users' photo sets, we found that most of the photos can be categorized into following three types, as shown in Fig. 1: (1) Global and detail: while taking pictures, users might zoom out to take an overview with the target object or scene, and zoom in to focus on some interesting details. (2) Pan to whole view: sometimes, the target object or scene is too large, and users might pan their cameras to capture the whole view, due to the limited view angle of the digital camera lens. (3) Same background: for some reasons, while users have events together, they might take pictures in turns in front of the landmarks or something representative.

Hence, if we can recover the camera operations, a simple but smooth transition might be achieved. However, rather than dealing with each case separately, a conceptual framework, *Buffer Region*, is proposed. Generally, we can find some similar parts between two photos. Then the found similar part(s) can be treated as the transition area between the two photos, where the virtual camera will pan or zoom through. Besides the smooth transition, to make the motion of camera smoother, the camera path is modeled as a cubic interpolation spline curve. Through this way, a simple but smooth visual effect can be produced.

2 Related Work

Generally, the photo slideshow displaying techniques can be categorized into 2D-based and 3D-based approaches.

In 2D-based photo slideshow, the most often seen and common used displaying effect is the Ken Burns effect[3], which applies only zoom and pan operators to a given photo. Based on the Ken Burns effect, Microsoft Photo Story[4] and their related research project, Photo2Video [9], developed a method to automatically extract ROI(s) in each photo and then the Ken Burns effect is applied to the camera motion according to the extracted ROI position(s) in the photo. Besides the Ken Burns effect, the other common used approach is the fade-in and fade-out effect. However, the normal fade-in/fade-out effect does not take the content into account, so the transition may not be smooth enough due to the difference between the two photos.

The ROI feature is also used by Tiling Slideshow [4], which clusters photos by time and content, and then arranges the photos within the same cluster into one frame with predefined tiling layouts. To browse several photos as whole, AutoCollage [15] combines each photo's ROI by blending them at boundaries. Though Tiling Slideshow and AutoCollage can show lots of information (photos) in one frame and can provide pretty exciting visual effects, they may not be suitable for a long-term displaying photo slideshow.

In addition to the 2D-based methods, some research work presents the photos in the 3D manner. Horry *et al.* proposed Tour Into the Picture (TIP) [7] which constructs a 3D space from one input photo by utilizing the proposed spidery mesh method. To construct the 3D space, Hoiem *et al.* proposed another approach, Automatic Photo Pop-up [6], which is a system that automatically constructs the 3D model from one single image and provides users a 3D geometrical view about the photo.

Instead of using a single photo, Hsieh *et al.* proposed Photo Navigator [8] which utilized the TIP method to create the 3D models of each photo and then connects them to generate a sequence of "walk-through" viewing path from one photo to another. Microsoft Photosynth[1] and their relative research projects, Photo Tourism [18, 17], also utilized a number of photos to reconstruct a 3D space model and put the photos at the relative positions in it.

Though the above 3D-based methods aim to reconstruct the original 3D scene according to the extracted geometry features in the photos, they all need a huge amount of photos to construct the 3D space or need to take the pictures carefully. Hence, they may not be suitable for creating a photo slideshow for casual usages.

3 Field Study

3.1 Participants and Method

In order to know people's preference about slideshow effects, an interview was performed with 7 participants, who are 5 males and 2 females. For each participant, we let him/her watch a photo slideshow with some special transition

[3] http://en.wikipedia.org/wiki/Ken_Burns_Effect/

[4] http://www.microsoft.com/windowsxp/using/digitalphotography/photostory/default.mspx

effects and normal fade-in/fade-out effects. After watching the video, we asked them the following questions:

- What kinds of slideshow effects do you prefer?
- What will you do, if you will need to produce a photo slideshow?
- For different purposes, will the desired effects be differed?

3.2 Results and Findings

For the first question, most of them do not have specific preference. While asking the next two questions, the preference was revealed. For different purposes, people would use different kinds of slideshow effects, depending on the target audiences or the purposes of the photo slideshow. If the purpose is for advertising or the target audiences are strangers, they would prefer using the eye-catching effects. On the contrary, if the produced photo slideshow is for home video displaying, they would prefer using lightweight visual effects like the Ken Burns effect and simple fade-in/fade-out effect.

We quote some participants' explanations here: "Each different transition visual effect should have its own semantic meaning." "If the photos belong to the same topic, they should apply the same effect." (The mentioned effect here is the Ken Burns effect.) "The special transition effects can be used between two different topics as a hint to the audiences." "Too many special effects cause the video clutter and makes me feel bored finally, though it arouses my interest at first." "Special effects are some kinds of emphasis. You can't emphasize everything." "Rather than using special effects often, I prefer using lightweight effects, which makes the video watching comfortable."

The interview results can be concluded into three major findings. First, for different purposes and different audiences, the required transition visual effects are quite different. Second, compared with the special eye-catching effects, lightweight transition can be used all the time and would not cause any illness. Third, for the photos of the same topic, the transition should be smooth to make them as integral.

4 Framework of Buffer Region

4.1 Observation

Through the observation of users' photo sets, the photos can be generally classified into four major categories, which are *global and detail*, *pan to whole view*, *same background*, and *no-relationship*. The first three ones are shown in Fig. 1.

Global and detail - This kind of photos is always bound with camera's zoom operation. Users might zoom out to take an overview with the target object or scene, and zoom in to focus on some interesting details.

Pan to whole view - Sometimes, since the target object or scene is too large, in order to capture the whole view, users might need to pan their cameras. For this case, the usually used approach is to stitch the photos [19]. However, in our case, we just need to know the affine transformation between the two photos and generate a virtual camera path crossing through the two photos by registering them first.

Same background - While people go travel together, they might take pictures in turns in front of the landmarks or something representative. To demonstrate those photos, one possible approach is to use a photomontage method [2]. However, the content of the photo is altered. Hence, rather than altering the content, the similar parts between the two photos are detected and treated as the *Buffer Region* to let the virtual camera smoothly pass through.

No-relationship - For the photos have no above three relationships, they will be treated as no relationship.

4.2 Buffer Region

The core idea is to utilize the overlapped parts or similar regions as the smooth transition bridge between two photos. To achieve this goal, a similar system was proposed by Sivic *et al.* [16]. With the calculated GIST features [12], the system can automatically find the best "next" image for users to go forward in the virtual space endlessly. However, for people's casual photos without careful filtering, it may be hard to find so many similar photos.

Hence, to smoothly transit arbitrarily two photos, a so-called *Buffer Region* method is provided. As illustrated in the right lower two boxes in Fig. 2, the red rectangles are the extracted ROIs in both of purple and blue images and the green ones are the detected similar parts which are treated as the *Buffer Region.* Because the Rectangle C in the blue image is the most similar part with the Rectangle B in the purple one, therefore, while panning through the Rectangle B, the content inside the Rectangle B is smoothly replaced with the content inside the Rectangle C with alpha blending. Therefore, Rectangle B and Rectangle C are treated as the buffer regions while transiting from the purple image to the blue one. Through this way, the smooth visual effect can be generated.

4.3 General Method

First, the aspect ratio should be the same with the original image. Second, the width and height of the region should not lower than W/n and H/n, where W and H are the width and height of the original image, and n is the scale factor. To make things simple, we let n be the number of power of two.

To calculate the similarity, we transfer the color space into $L^*a^*b^*$ model. Then the two norm operation is applied. To be general, each contribution can be weighted and the measurement of difference for one pixel can be written as:

$$\sqrt{\alpha \Delta L^2 + \beta(\Delta a^2 + \Delta b^2)}, \tag{1}$$

where $\alpha + \beta = 1$. Moreover, to remove the noise, before calculating, Gaussian smooth is applied first. To deal with the zoom case, the multi-scale approach is provided. When detecting the buffer region, the Image B will be downsized or enlarged with S levels to build a pyramid.

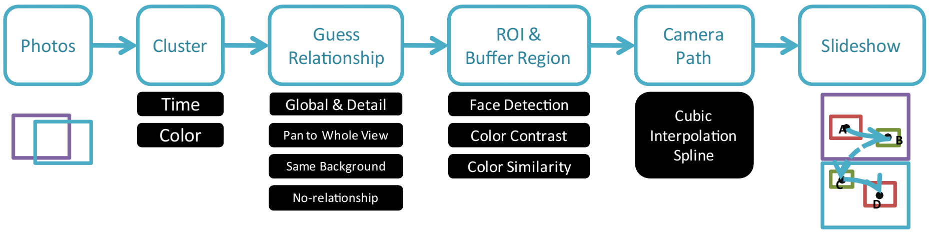


Fig. 2. System architecture.

5 System

5.1 System Overview

The system structure is illustrated in Fig. 2. While inputting a set of photos, the system first clusters the photos based on the color similarity among them. After clustering the photos, the system matches each two adjacent photos in the same cluster by SIFT [11] features to compute the transition matrix of these two photos to guess the original camera operation. Due to the guessed camera operation, the relationship between the two photos can be determined. Then the ROI(s) and *Buffer Region* are extracted in each photo and the smooth camera path is calculated based on them. Finally, the photo slideshow is generated by playing through the ROIs and *Buffer Regions* with smooth transition.

5.2 Clustering Photos

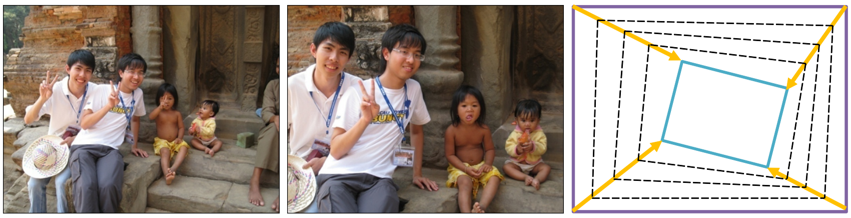
Since photos in the same folder are usually captured consecutively at a certain location in a short period of time. Therefore, a two-step clustering method is used in this paper. The input photos are first clustered by the taken time, and then the clustered photos are further categorized according to the color similarity between each other.

Clustered by time - To cluster the input photos according to the taken time, we used the algorithm proposed in PhotoToc [13], which can dynamically decide the cutting threshold by using a sliding window.

Clustered by color - Though the photos in the same group clustered by the taken time, the content might be still varied. In our system, the color-based content similarity is used to categorize the photos in the same group, by assuming the color in the same theme is usually similar.

5.3 Guessing Camera Operations

To guess the original camera operations, the SIFT [11] features with RANSAC algorithm is used to match two adjacent images and then based on the matched corresponding points, the two images' transformation matrix can be calculated.



(a) (b) (c)

Fig. 3. Camera move in the *global and detail* case.

If the two images are not similar enough, they will be regarded as the *no-relationship* case. Generally, the camera operation between the two photos in the same group containing both zoom and pan. Therefore, to discriminate the two types of *global and detail* and *pan to whole view*, the system will check whether one photo can be contained into another one or not.

5.4 Extracting ROIs and Buffer Regions

For each relationship, the ROIs and *Buffer Regions* are different. Generally, each photo itself will be defaultly treated as one ROI and will be put into the camera path. To extract the ROI(s) of each photo, the top-down and bottom-up approaches are both used [10].

For the top-down approach, the OpenCV face detector is used.For the bottom-up approach, the color contrast is taken as the measurement. First, the saliency map of the image is calculated by using the method proposed by Achanta *et al.* [1]. Then, the segmentation of the image is calculated by using the method proposed by Felzenszwalb and Huttenlocher [5]. Based on the segmented results, for each segmented region, the saliency scores are aggregated and the average score is calculated. If the average score of a segment is larger than a defined threshold, the segment will be treated as a candidate of the ROIs. The threshold is defined as two-fold of the average score of the whole saliency map here.

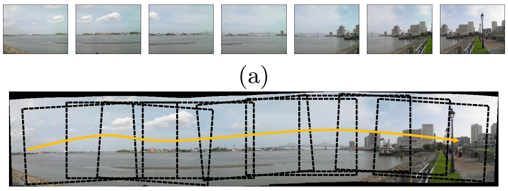
5.5 Calculating Camera Path

To calculating the camera path, the basic idea of Photo2Video [9] is adapted. However, two aspects are modified to meet our goal. First, we also model the areas of the ROIs into the spline calculation.

Second, the start and end points of the path inside a photo is bound by the two *Buffer Regions*, one for connecting the previous photo and the other for connecting the next one.

5.6 Generating Slideshow

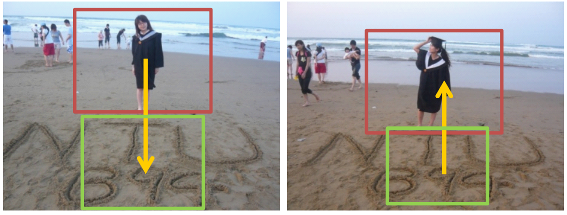
To generate the smooth slideshow transition, the camera operations are simulated by the determined relationship between the adjacent two photos. Fig. 3



(a)

(b)

Fig. 4. Camera move in the *pan to whole view* case.



(a) (b)

Fig. 5. Camera move in the *same background* case.

illustrates the *global and detail* case, where the Fig.s 3 (a) and (b) are two adjacent photos captured through a zoom camera operation. Fig. 3 (c) illustrates the zoom relation of the two photos.

Fig. 4 is the illustration of the *pan to whole view* case. Fig. 4 (a) shows a series of consecutive photos captured by panning the camera from left to right. The system first tries to find the panorama [3] relationship of the set of photos under the pan operation and the result is as Fig. 4 (b). Then, according to the order of the input photos, the system applies a pan camera operation to replay the whole scene again.

Fig. 5 is the illustration of the *same background* case. However, the process is also used for the *no-relationship* case. After panning and zooming from the ROI region (red rectangle) to the *Buffer Region* (green rectangle) in Fig. 5 (a), the content inside the *Buffer Region* is smoothly replaced with the content inside the *Buffer Region* of Fig. 5 (b). Then, while finishing the transition, the camera continues panning and zooming to the ROI in Fig. 5 (b).

Besides the smooth transition, we also designed how long the camera should stay for each ROI according to the determined relationships. In the *global and detail* case, while the relationship is zooming in, it means that the next photo might be more important and more interested than the present one. Then, the camera should stay longer while playing the next photo comparing to the current one, and vice versa. While in the *pan to whole view* case, it means that the users care more about the whole view, not the detail in each transited photo. Then, the camera should not stay, but play through all the photos steadily. While in

Table 1. Evaluation Data

	Photo Set 1	Photo Set 2
Number of Photos	8	14
Composition	7 Pan	2 Zoom 1 Same Background 10 Others

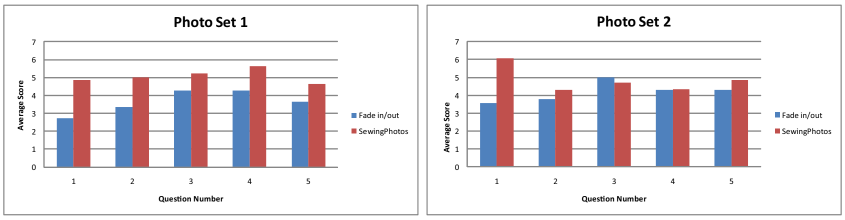


Fig. 6. Evaluation Result.

the *same background* case, of course, the users are interested in the foreground ROI(s). Therefore, it should stay on the foreground ROI(s) for a while. For the *no-relationship* case, the total display time for one photo is the same. Hence, if there are more than one ROI in a photo, the display time for each ROI is divided equally. Finally, while passing through the *Buffer Regions*, the camera should not stay. Through this way, the smooth slideshow video is generated.

6 Evaluation

Sewing Photos emphasizes on the smooth transition processes between photos, so we compared our result with the basic fade-in/fade-out transition effect. In the following evaluation, two photo sets are used. Table 1 listed the details of them, where Photo Set 1 is a panorama view of a river, and Photo Set 2 is captured in a reunion at a rabbit restaurant. In the following user testing, fifteen evaluators are invited. Their ages are from 22 to 28. The two slideshows of the three photo sets are played randomly.

We adopted the 7-points Likert-scale (1 is the worst, and 7 is the best) to design our evaluation scoring form. The evaluators were asked to score each video with the following five questions in different perspective:

1. **Fun**: Do they think it is an interesting presentation?
2. **Smoothness**: How do they think the smoothness at the transition?
3. **Experience**: How does the transition effect help them experience the trip?
4. **Camera Operation**: Do they think the applied camera operations suitable?
5. **Acceptance**: How about the willingness of adapting the transition effects?

The results are illustrated in Fig. 6, which show that the photo slideshows generated with *Sewing Photos* are generally better than the basic fade-in/fade-out transition effect, especially in Photo Set 1 which originally contains more zoom and pan relationships between the photos.

In Photo Set 2, the result of Question 4 (camera operation) is not significantly better, and even worse in Question 3 (experience). We think that it is because Photo Set 2 mainly contains photos taken at the same background or with no specific camera operation, and such the advantages of adapting the original camera operations are not used, in the same time, the camera path applied by the system to those photos might not be suitable for this photo set which leads to the worse result in experience.

7 Conclusion

In this paper, we propose a novel method, *Sewing Photos*, which utilizes the original camera operations among the photos to generate a photo slideshow with smooth transition effects. The three major types of photo relationships in users' daily life photos are also defined. However, rather than dealing with each case individually, a general framework, *Buffer Region*, is proposed by utilizing the similar regions between the adjacent photos as the smooth transition tunnel. To smoothly play the photos, the motion of the camera is also taken into consideration. A cubic interpolation spline with the smallest curvature is calculated by using the extracted ROI(s) and *Buffer Regions* in one photo.

Though the alpha blending is good enough while transiting between two photos, a better method can be explored in the future by considering the human attention and camera direction. Though our system is treated as an engine to provide the smooth visual effects, to be an authoring tool, a user interface might be helpful, which can help the users to refine the results of the extracted ROIs and *Buffer Regions*, because some semantic ROIs are hard to be retrieved through low level contrast-based operation.

Concluding our work, the provided new technique is not going to replace the splendid visual effects in present the photo slideshow, but to enhance the traditional fade-in/fade-out effects to meet the needs. Nevertheless, the proposed smooth effect can provide more comfortable watching experience and are good for a long-term photo displaying.

8 Acknowledgments

This paper was partially supported by the National Science Council of Taiwan under NSC98-2622-E-002-001-CC2 and also by the Excellent Research Projects of the National Taiwan University under NTU98R0062-04.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of the 2009 IEEE International Conference on Computer Vision and Pattern Recognition. pp. 1597–1604 (2009)
2. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive digital photomontage. ACM Transactions on Graphics 23(3), 294–302 (2004), (SIGGRAPH 2004 Conference Proceedings)
3. Brown, M., Lowe, D.G.: Recognising panoramas. In: Proceedings of the 2003 IEEE International Conference on Computer Vision. vol. 2, pp. 1218–1225 (2003)
4. Chen, J.C., Chu, W.T., Kuo, J.H., Weng, C.Y., Wu, J.L.: Tiling slideshow. In: ACM Multimedia 2006 Conference Proceedings. pp. 25–34 (2006)
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
6. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. ACM Transactions on Graphics 24(3), 577–584 (2005), (SIGGRAPH 2005 Conference Proceedings)
7. Horry, Y., Anjyo, K.I., Arai, K.: Tour into the picture: using a spidery mesh interface to make animation from a single image. In: ACM SIGGRAPH 1997 Confernece Proceedings. pp. 225–232 (1997)
8. Hsieh, C.C., Cheng, W.H., Chang, C.H., Chuang, Y.Y., Wu, J.L.: Photo navigator. In: ACM Multimedia 2008 Conference Proceeding. pp. 419–428 (2008)
9. Hua, X.S., Lu, L., Zhang, H.J.: Automatically converting photographic series into video. In: ACM Multimedia 2004 Conference Proceedings. pp. 708–715 (2004)
10. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
12. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
13. Platt, J.C., Czerwinski, M., Field, B.A.: PhotoTOC: Automatic clustering for browsing personal photographs. In: Proceedings of the 2003 IEEE Pacific Rim Conference on Multimedia. vol. 1, pp. 6–10 (2003)
14. Rodden, K., Wood, K.R.: How do people manage their digital photographs? In: ACM CHI 2003 Conference Proceedings. pp. 409–416 (2003)
15. Rother, C., Bordeaux, L., Hamadi, Y., Blake, A.: Autocollage. ACM Transactions on Graphics 25(3), 847–852 (2006), (SIGGRAPH 2006 Conference Proceedings)
16. Sivic, J., Kaneva, B., Torralba, A., Avidan, S., Freeman, W.T.: Creating and exploring a large photorealistic virtual space. In: Proceedings of the First IEEE Workshop on Internet Vision. pp. 1–8 (2008)
17. Snavely, N., Garg, R., Seitz, S.M., Szeliski, R.: Finding paths through the world's photos. ACM Transactions on Graphics 27(3), Article No.: 15 (2008), (SIGGRAPH 2008 Conference Proceedings)
18. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Transactions on Graphics 25(3), 835–846 (2006), (SIGGRAPH 2006 Conference Proceedings)
19. Szeliski, R.: Image alignment and stitching: a tutorial. Foundations and Trends in Computer Graphics and Vision 2(1), 1–104 (2006)