Diminished reality for privacy protection by hiding pedestrians in motion image sequences using Structure from Motion

Kentaro Yagi* Keio University
Kunihiro Hasegawa† Keio University
Hideo Saito‡ Keio University

ABSTRACT

We present a method for generating images in which people are hidden from image sequences taken with a hand-held camera. Our method is basically used for privacy protection of people whose images are unintentionally captured in image sequences.

We hide people from images by reconstructing a 3D model of background and projecting it to 2D images. By detecting the area in which people are present beforehand, we can reconstruct a 3D model of the background without people. In the experiment, We compare our method with some conventional approaches for diminished reality.

Keywords: Diminished reality, structure from motion, privacy protection, multi-view stereo, inpainting.

Index Terms: H.5.1 [INFORMATION INTERFACES AND PRESENTATION]: Multimedia Information Systems—Artificial, augmented, and virtual realities; I.5.1 [IMAGE PROCESSING AND COMPUTER VISION]: Models—Structural

1 INTRODUCTION

In recent years, the research to remove objects or people in images has been gathering remarkable attention. This kind of research is called Diminished reality. It can be used for many kinds of purposes. For example, it is useful for privacy protection. When we see TV shows, movies, or some images uploaded on the Internet, we can find many people who are not conscious of being in the pictures. We have to protect their privacy because they may be totally unaware that they are in the images.

The easiest way to solve this problem is by mosaicking or masking people's faces. However, such processing sometimes spoils the original image because it makes it look strange; in addition, we must rely on the uploaders' ethics, as the images must be altered before uploading. Thus, a system should be developed that can automatically detect people in images and hide them, while ensuring that the output image looks natural.

We developed the required system using Structure from Motion. The input for our method comprises image sequences taken with a hand-held camera. The output is composed of images in which the people are hidden. Our method can be divided into four processes. The first is a human-detection process that detects people in an image. The second is a 3D reconstruction process that reconstructs a 3D model of the background using Structure from Motion. The third process is a 3D to 2D projection process that is carried out to obtain 2D background images using a projection matrix. Finally in the blending process, we mix the original images and projected images to generate output images.

*e-mail: yagi@hvrl.ics.keio.ac.jp
†e-mail:hiro@hvrl.ics.keio.ac.jp
‡e-mail:hs@keio.jp

2 RELATED WORK

Many studies have been carried out on hiding objects in images. The simplest way to do this is to transform another image taken from different view points and superimpose it on the target objects.

In Enomoto's research, the researchers solved a problem in which users could not see a plane in front of them because of obstacles located between users and the plane [1]. They used an AR tag [2] to estimate the target plane and warped another image to the target plane using a homography matrix. Flores's method solved privacy invasion problem in terms of people in Google Street View's images [4]. This method also uses a homography matrix under the situation that objects in the images do not move, and the background of the target person is captured in other frames. Hasegawa and Saito's method hides people in image sequences [7]. This method can also be applied to a situation in which a target person is moving while a video is being taken. According to these researchers, the background can be considered as a plane, and camera must stay in the same position. Zhuwen's method hides people in pictures using images uploaded on the Internet [12]. The researchers store a lot of images taken at the same place using geo-tag information attached to each image. They estimate 6 DoF of each image using Structure from Motion. Then, they warp other images and superimpose the most appropriate image on target person. These studies all assume that the background is a plane or can be considered as a plane; therefore, they have a problem in that they can not be applied to a scene in which the background has a complicated 3D shape.

Unlike the studies mentioned above, Granados's method considers about the 3D shape of the background [6]. The researchers divide the background into small planes. Then, they estimate homography matrix for each plane and remove target person by warping other images. Kawai's research remove people in an image sequence taken with an omnidirectional camera [10]. This method removes moving objects in it using energy minimization. Since they do not consider the meaning of each object, not only people in it but also other moving objects are removed. In addition, people might not be removed if the background is similar in color to the clothes that people wear. Some other researches did not use other images taken from different points of view; rather, these researchers used estimations of pixel values based on the consistency of the background texture in the invisible part [11] [8]. This method works well if the background has a continuous texture; however, if this is not the case, the invisible region is represented differently from how it is in the real world. In addition, this method can be used for objects that remain in the same place, but not for objects that move in the image sequences.

In contrast to these researches, our method aim to the dynamic scene that both the target person and camera are moving while the video is being taken. In addition, we do not use homography to remove people in images but use 3D model of the background so that we can apply our method to scenes in which background has arbitrary shape. There are some researches doing 3D reconstruction in dynamic scenes [15] [17]. These researches propose methods to reconstruct moving objects using multiple cameras. However, we use only one hand-held camera. Furthermore, we reconstruct only static background in dynamic scene. Therefore, we reconstruct 3D shape of the background using Structure from Motion and Multi-View Stereo.

978-0-7695-6327-5/17 $31.00 © 2017 IEEE
DOI 10.1109/ISMAR-Adjunct.2017.101

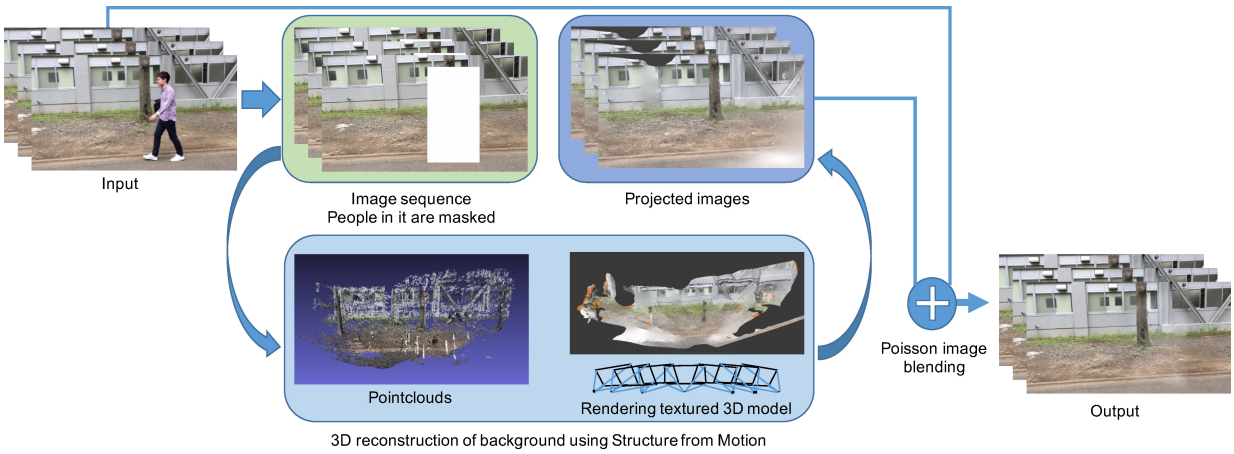


Figure 1: Flow chart of our method.

3 THE PROPOSED METHOD

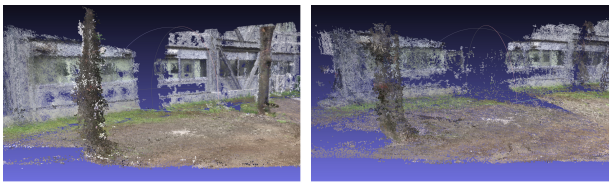
Figure 1 shows the flow of our method. It is applicable to dynamic scenes in which both the camera and target move, and a hand-held camera is used to take the images. Our method can be divided into four processes, namely human detection, 3D model reconstruction, 3D-to-2D projection, and blending.

3.1 Human detection

The first process is human detection. By detecting where people are in images and hiding them by painting them with white, we can obtain only background 3D model that does not include people when we reconstruct the model. First, we split a given input video into each frame and obtain an image sequence. Next, we detect people who are in those images using Liu et al's method [13], which can detect specific objects quickly and robustly compared to other object-detection methods. When we find people, we set a rectangular region that is large enough to cover each person completely. However, sometimes we can not detect people correctly because of their posture, the brightness of the lighting, or other factors. Therefore, we set the same rectangle region in the next three frames if fewer people are detected than in the previous frame, assuming that a person in a frame does not move a lot in successive frames. After detecting people in the images, we erase them by painting the rectangular region that encloses them in white. In this way, we can reconstruct only 3D points in the background since few feature points are found in the region painted same color.

3.2 3D model reconstruction

The second process is 3D model reconstruction. We use Structure from Motion to reconstruct 3D sparse pointclouds and estimate camera position of each frame. In Structure from Motion, we use the SIFT feature value [14] to find feature points, and we compare near 150 frames to find corresponding points. We also optimize the 3D points and camera pose using bundle adjustment.We do not have to achieve camera parameters by doing camera calibration beforehand. After getting sparse pointclouds and camera positions, we obtain a 3D textured model based on pointclouds given by Structure from Motion. We use Multi-View Stereo [5] to obtain a dense pointclouds from sparse pointclouds. Next, we connect pointclouds located close to each other and make a mesh. The density of pointclouds is dependent on input image sequence; therefore, we used Jancosek's method which is able to make a mesh independently of density of pointclouds [9]. Then, we refine the 3D mesh model with Vu's method to remove noises by smoothing 3D model [19]. Finally, we put texture in the 3D mesh model and obtain a dense 3D model with color. When we add the texture, we do not use input images of Structure from Motion, as this would make the texture a little bit whiter, since we painted the human region in white. To solve this problem, we have generated another image sequences for texturing with Telea's inpainting method [18] to complement the rectangular regions. We compared two inputs for reconstructing 3D model of background, (a) image sequence in which masked regions are painted white, (b) image sequence in which masked regions are inpainted. As you can see in Figure2, pointclouds generated with (b) has more noises. For example, a tree which is located right side in the picture is not reconstructed in (b), and also the wall of the building is reconstructed sparser. This is because 3D structure cannot correctly be recovered by detected corresponding points in the inpainted area.To achieve our goal, pointclouds should be dense and all objects except for target person should be reconstructed. Therefore we used image sequence in which masked regions are painted white as input for Structure from Motion.



(a) pointcluds generated from masked image sequence

(b) pointcluds generated from inpainted image sequence

Figure 2: Pointclouds generated using Structure from Motion and Multi-View Stereo [5].

3.3 3D-to-2D projection

The third process is projection from the 3D model to 2D images. Since we have already estimated the camera positions, we can calculate the projection matrix of each frame. By rendering the 3D model to each camera position with the projection matrix, we can obtain an image sequence does not include people.

3.4 Blending

The final process is blending the original images and projected images. First, we mask the projected image with rectangular region

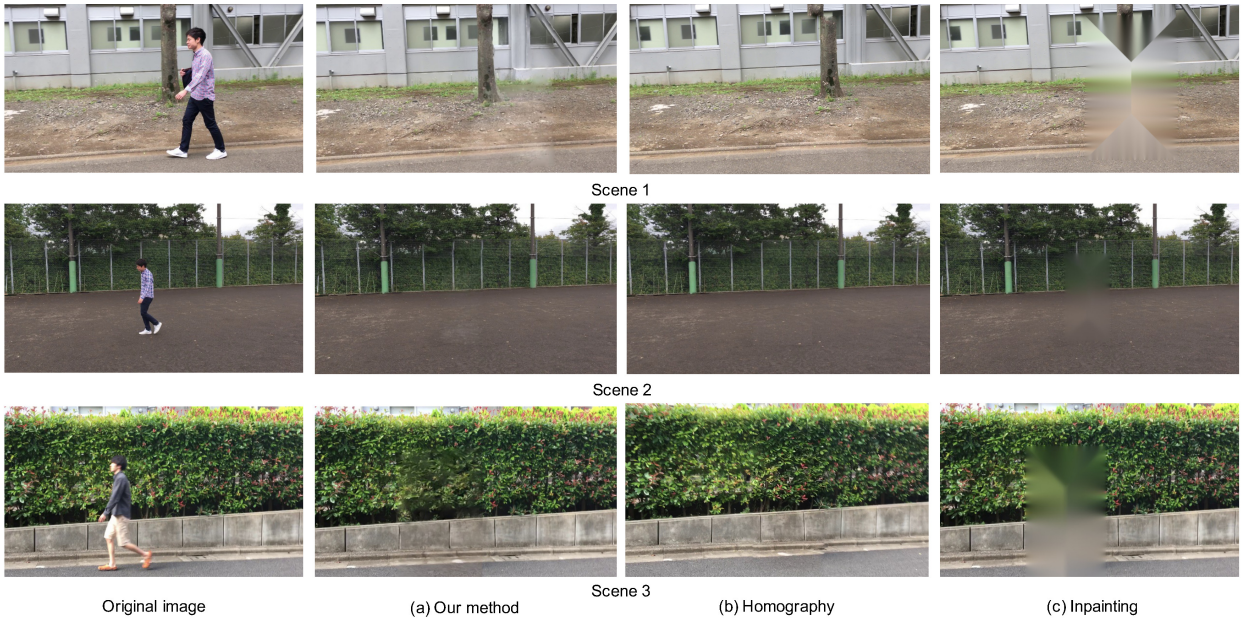


Figure 3: Results of (a) our method, (b) homography, (c) inpainting [18].

given by Liu et al.'s object-detection method [13] and superimpose it on original image. However, this shows the boundary between the original image and projected image clearly. Accordingly, we apply Perez et al.'s method [16] to adjust the pixel values of the neighboring pixels of the boundary and make the output image appear more natural. In Perez et al.'s method [16], we set loop count 1800. We adjust target pixel value depends on neighbor 4 pixels.

4 Results

We apply our method to some scenes with different backgrounds. Moreover, we compare (a) our method with (b) homography and (c) Telea's inpainting method [18] to illustrate the efficiency of our method. We choose (b) homography and (c) inpainting as comparison because those methods are suitable for the dynamic scenes which we aim.

Figure 4 shows how we analyze videos in the experiments. As the Figure shows, as a camera is moving in the opposite direction a pedestrian and aims the camera towards the pedestrian while walking. We used an iphone6s (Apple) to shoot the video. The resolution is 1920 $\times$ 1080 and the frame rate is 30 fps. We used VisualSfM ver 0.5.26 [20] [21] [22] for Structure from Motion, OpenMVS for Multi-View Stereo and making the textured 3D model, and Blender ver 2.78b for rendering the 3D model to 2D images. Each input video is 10-15 seconds long and comprises about 300-450 frames.

4.1 Experiment

Figure 3 shows different frames of input image sequences and the results of (a) our method, (b) homography and (c) Telea's inpainting method [18] for three different scenes. Scene 1 is the 172nd frame out of 312 frames, scene 2 is the 240th frame out of 460 frames, and scene 3 is the 190th frame out of 421 frames. For method (b), we used SIFT feature value [14] and RANSAC [3] to obtain the homography matrix. For method (c), we carried out 500 repetitions to calculate the appropriate pixel values.

We could hide the people in the images in all three methods. However, our method is apparently superior to the other methods in terms of how natural it looks. Method (b) works well when the background can be considered as a plane; however, when the background has 3D shape, it has geometric inconsistency. For example, in scene 1, the tree is misaligned because it is located in front of the wall. In the result of method (c) in scene 1, the tree was removed which is supposed to be not removed. This is because method (c) complements a masked region by propagating outside pixel values. Therefore, the bigger the target person exists is, the harder to compliment a masked region because more objects in the background are hidden.

In contrast, our method works well in any situation, and the output images are appear more natural than those of the other methods do.

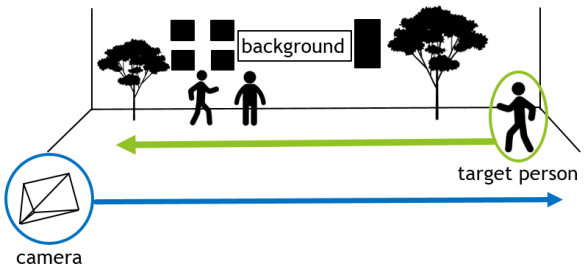


Figure 4: Experiment environment.

5 Discussion

In contrast to Figure 3, we show some results that our method did not work well in Figure 5 and Figure6.

In Figure 5, there appears to be a hole in the area in which we applied our method. This is because the target person had stayed at the same position in the first 20 or 30 frames in the input video, so that we could not obtain the 3D shape of the area. It is difficult to hide people and generate images naturally with our method if we could not capture some part of the background from multiple viewpoints, since Structure from Motion requires a lot of images taken from different viewpoints to obtain correct 3D pointclouds and camera pose estimations. Accordingly, to hide the target person from

Figure 5: Failure case 1.

Figure 6: Failure case 2.

all frames, we need to mix our method with another image-recovery method, such as inpainting or AutoEncoder. In Figure 6, we can see geometrical inconsistency in the area where people were originally present. The edges of the steps and sidewalk are inconsistent. These inconsistencies arise because noise is generated in Structure from Motion. Noise may occur in two main ways. First, a pair of feature points may be wrongly matched, causing distortion in the 3D model; if the background's texture is continuous, as in case 2, mismatching is more likely to occur. The second issue is that we generated a mesh even in areas where the point clouds were not dense enough, as the mesh was created based on sparse pointclouds; insufficient density causes distortion of the 3D shape. We need to consider ways to eliminate noise, such as by plane fitting of pointclouds or a better matching method that is less likely to result in mismatches.

6 Conclusion

We proposed a method of hiding people who are in images or videos without their intension to protect their privacy. In our method, we first detected people in images and paint them white. Second, those images were as input for Structure from Motion to obtain the 3D shape of background and camera positions. Third, the 3D model was projected to 2D images based on the estimation of the camera positions. Finally, the original and projected images were blended, and images were generated in which the people were hidden. We compared our method with other ways to hide objects in images, such as the homography and inpainting [18], and we demonstrated our method's superiority. In future work, we will mix our method with other image-recovery methods such as inpainting or AutoEncoder, to make it possible to hide people from all frames even when we can not obtain a complete 3D model.

Acknowledgement

This research presentation is supported in part by a research assistantship of a Grant-in-Aid to the Program for Leading Graduate School for Science for Development of Super Mature Society from the Ministry of Education, Culture, Sport, Science, and Technology in Japan.

References

[1] A. Enomoto and H. Saito. Diminished reality using multiple handheld cameras. In *Proc. ACCV*, vol. 7, pp. 130–135, 2007.

[2] M. Fiala. Artag, a fiducial marker system using digital techniques. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 2, pp. 590–596. IEEE, 2005.

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, vol. 24(6):381–395, 1981.

[4] A. Flores and S. Belongie. Removing pedestrians from google street view images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010)*, pp. 53–58. IEEE, 2010.

[5] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *IEEE Computer Vision and Pattern Recognition (CVPR 2010)*, pp. 1434–1441. IEEE, 2010.

[6] M. Granados, K. I. Kim, J. Tompkin, J. Kautz, and C. Theobalt. Background inpainting for videos with dynamic objects and a free-moving camera. In *European Conference on Computer (ECCV 2012)*, pp. 682–695. Springer, 2012.

[7] K. Hasegawa and H. Saito. Diminished reality for hiding a pedestrian using hand-held camera. In *IEEE International Symposium on Mixed and Augmented Reality Workshops (ISMARW 2015)*, pp. 47–52. IEEE, 2015.

[8] J. Herling and W. Broll. Pixmix: A real-time approach to high-quality diminished reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2012)*, pp. 141–150. IEEE, 2012.

[9] M. Jancosek and T. Pajdla. Exploiting visibility information in surface reconstruction to preserve weakly supported surfaces. *International scholarly research notices*, vol. 2014, 2014.

[10] N. Kawai, N. Inoue, T. Sato, F. Okura, Y. Nakashima, and N. Yokoya. Background estimation for a single omnidirectional image sequence captured with a moving camera. *Information and Media Technologies*, vol. 9(3):361–365, 2014.

[11] N. Kawai, T. Sato, and N. Yokoya. Diminished reality based on image inpainting considering background geometry. *IEEE Transactions on Visualization and Computer Graphics (TVCG 2016)*, vol. 22(3):1236–1247, 2016.

[12] Z. Li, Y. Wang, J. Guo, L.-F. Cheong, and S. Z. Zhou. Diminished reality using appearance and 3D geometry of internet photo collections. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2013)*, pp. 11–19. IEEE, 2013.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision (ECCV 2016)*, pp. 21–37. Springer, 2016.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60(2):91–110, 2004.

[15] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton. General dynamic scene reconstruction from multiple view video. In *IEEE International Conference on Computer Vision (ICCV 2015)*, pp. 900–908, 2015.

[16] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *ACM Transactions on Graphics (TOG 2003)*, vol. 22, pp. 313–318. ACM, 2003.

[17] A. Taneja, L. Ballan, and M. Pollefeys. Modeling dynamic scenes recorded with freely moving cameras. In *Asian Conference on Computer Vision (ACCV 2010)*, pp. 613–626. Springer, 2010.

[18] A. Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, vol. 9(1):23–34, 2004.

[19] H.-H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI 2012)*, vol. 34(5):889–901, 2012.

[20] C. Wu. Towards linear-time incremental structure from motion. In *IEEE 3DTV-Conference (3DTV-CON 2013)*, pp. 127–134. IEEE, 2013.

[21] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pp. 3057–3064. IEEE, 2011.

[22] C. Wu et al. VisualSFM: A visual structure from motion system. 2011.