Creating a 3d Face Model from Multiple Images

Robert Hakobyan junior

National Polytechnic University of Armenia
Yerevan, Armenia
e-mail: hakobyan.rob@gmail.com

***Abstract*—This paper investigates the possibility of generating a 3D model of a human face from multiple images of the face from different angles. The program will analyze the given images and camera parameters, generate cloud points using the images. The result can be used for authentication purposes, to determine if the person is authorized to access the location or not.**

***Keywords*—Panorama, object panorama, homography, 3D reconstruction, multi-view stereo, depthmap reconstruction, point cloud reconstruction.**

I. Introduction

With the help of modern progress, things that were previously considered unattainable are becoming daily routine nowadays. As sad as it may sound, it is a good thing, because in return all of us get the opportunity to benefit from these great achievements. Facial recognition is one of the fields that has greatly evolved thanks to this progress.

From the early days people recognized each other by the human face. The person's face is used to identify him; besides it can give a lot of information to the perceiver, including his mood, attentiveness, and intention. Clothing, voice, and other characteristics can also be used to recognize the person, but the face is the most distinctive part of the human [1].

Nowadays, people are recognized not only by other people, but even by using computers. Face recognition is still a new and less studied field. Despite it, face recognition is already a significant part of our life. It in great demand and is used in security, marketing, and even phone unlocking.

The aim of this work is to create a 3D face model from given images of a person. The result can be used to authenticate him, restrict access to unauthorized personnel and make authorized persons entry easier.

II. Image Processing

Image processing is used to make changes or extract some useful information from images. In image processing, the image is given as an input, and the result might be a part of the image, the color scheme of the image, or any other info that the consumer might need. Image processing has two main directions: image enhancement and image analysis.

Image enhancement improves certain image characteristics. For example, removing noise, increasing dark image's contrast, etc. It is worth mentioning that enhancements useful for cosmic images might not be useful for indoor images [2].

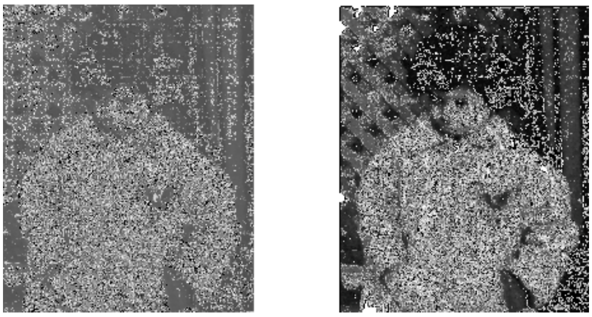

Figure 1. Image contrast enhancement

Image analysis is performed by a computer to extract the needed information about an image by different techniques: 2D and 3D object recognition, image segmentation, etc. [3]

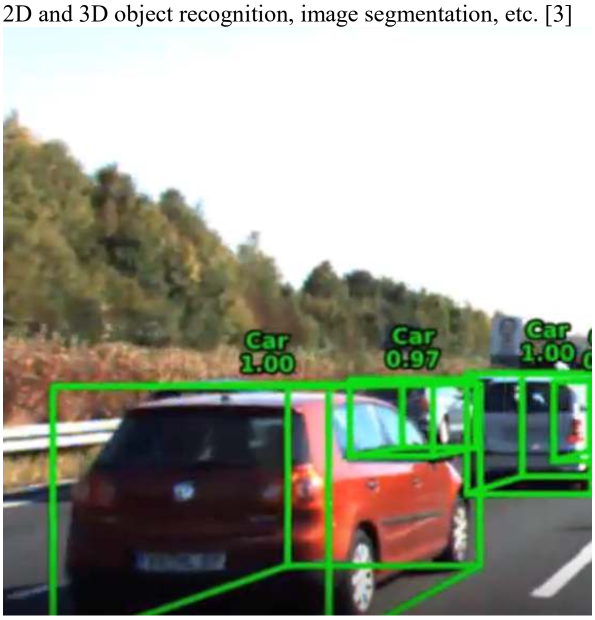

Figure 2. 3D object recognition

Image analysis is the extraction of meaningful information from images; mainly from digital images by means of digital image processing techniques.

Image processing is a rapidly growing area and is being used in more and more spheres, such as engineering and computer science.

Image processing is composed of these three steps:

- Input the image for analyzing
- Process and modify the image
- Output the modified image or the necessary information obtained while processing the image.

One of the most useful tools for image analysis is homography.

A. Homography

Consider two images on a plane, for example, a book cover shown in Figure 3. On the left part of the image 4 points are shown, red point is marked as (x1, y1). Homography maps the points in one image to the corresponding points in the other image by using a 3x3 matrix and shows the same corresponding points in the right part of the image, where the same red point is marked as (x2, y2).

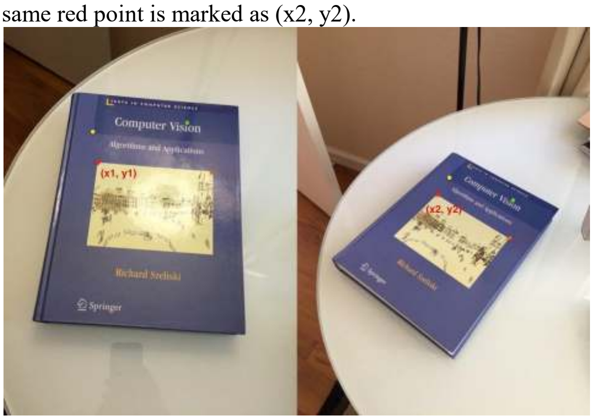

Figure 3. Two images of the book cover on a plain are homographically connected

Now, since homography is a 3x3 matrix, we can show it like this:

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix}$$

Equation 1. Homography

Using Equation 1, we can get the equation for homographic transformation from (x1, y1) to (x2, y2), which will look like this:

$$\begin{pmatrix} x_1 \\ y_1 \\ 1 \end{pmatrix} = H \times \begin{pmatrix} x_2 \\ y_2 \\ 1 \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \times \begin{pmatrix} x_2 \\ y_2 \\ 1 \end{pmatrix}$$

Equation 2. Homographic transformation for two points

Given equations are true for any two corresponding points if they are located on the same plain. After applying the above equation to the images in the Figure 3, we get the following result:

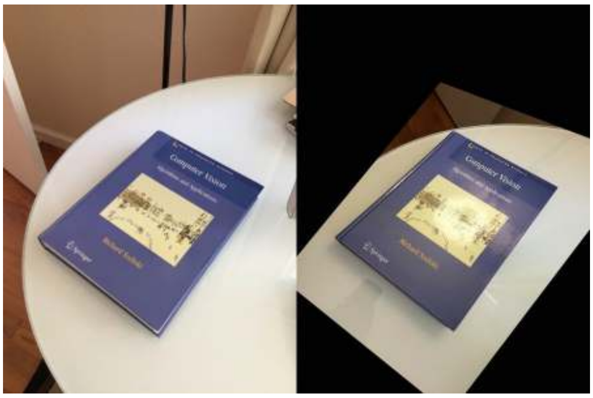

Figure 4. Book images aligned using homography

As visible in Figure 4, the points that were not located on the same plain are removed from the transformed version. And if there are two plains of the same image that are suitable for homography, the latter can be done in two ways, retrieving two different results [4].

B. Creating panoramas

Using homography, one image can be transformed into another, but there is a catch: images must have a common plain, and only that plain can be used for transformation. If a picture of some scene is taken, then the camera is rotated, and the picture is taken again, homography is still possible and a panorama can be created. The pictures will have some common objects, using which the pictures can be aligned and stitched together. The initial result might not be excellent, as the lightning, contrast, etc. in two images might not coincide, so the stitching line will be visible in the panorama.

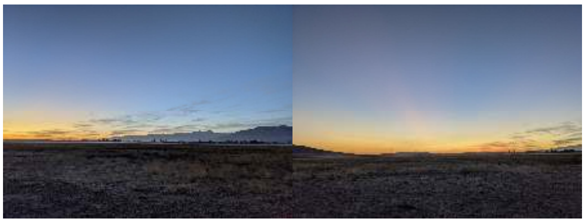

Figure 5. Two images used to create a panorama

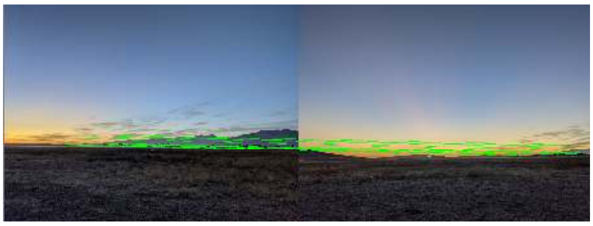

Figure 6. Matching points in two images

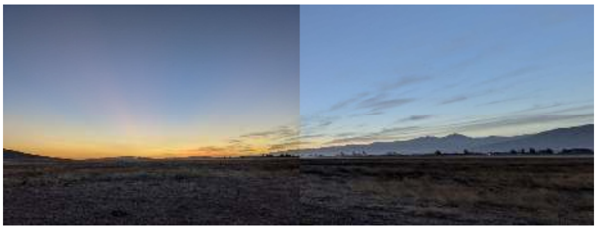

Figure 7. Result

To create this panorama, different computer vision and image processing techniques have been used: keypoint detection, keypoint matching, RANSAC, and perspective warping.

To get the matching keypoints, first we need to detect keypoints and extract local invariant descriptors (i.e., SIFT). SIFT is the process of extracting the interesting points from the image to provide a “feature description” of the object. Matching keypoints is done by looping over the descriptors of both images, computing the distances, and finding the smallest distance for each pair [5].

After getting the matching keypoints (Figure 6) and running them through the RANSAC algorithm, the result is the homography matrix, which is then used to stitch two images together to create the result shown in Figure 7.

III. Reconstruction of 3D Model

The difference between making an object panorama and a scene panorama is that in the case of a scene panorama only the camera angle changes. While making an object panorama, also the camera itself moves. As the cameras that are taking the images are fixed in position, we know the distance from the object, and the angle, at which the object is photographed.

To be able to authenticate the face from multiple images, the shape of the face must be known. 3D reconstruction is used for this purpose.

3D reconstruction is performed using point clouds. Point clouds are a set of data points located in space. Each point has several measurements, such as the coordinates in 3D plain, the luminosity of the given point and sometimes the color of the given point, stored in RGB format. Cloud points together represent the 3D shape of the object.
Point clouds can be rendered but converting them to polygon meshes makes it is easier to inspect the result. This process is called surface reconstruction.

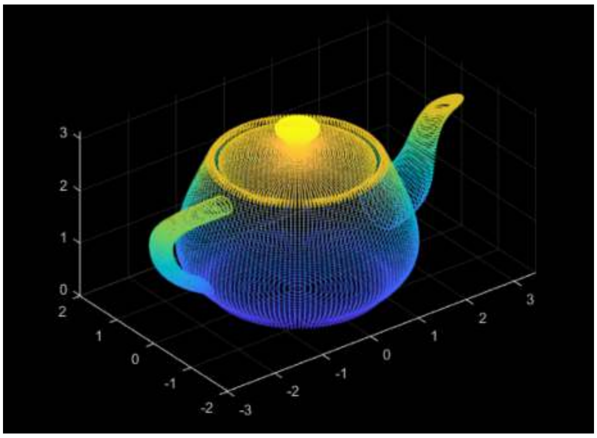

Figure 8. Point clouds representing a teapot

Point clouds are mostly generated using LiDAR, which stands for Light Detection and Ranging. LiDAR works much like radar, but instead of transmitting radio waves to calculate the distance, it uses light rays, which are reflected by the objects in the surroundings and detected by the LiDAR sensor. From this data cost map or point cloud outputs are generated. [6]

A. *Multi-view stereo*

Multi-view stereo (MVS) is a type of 3D reconstruction, which is used to reconstruct the 3D shape of the object using multiple photographs. It uses images from different angles to generate cloud points and construct an object.

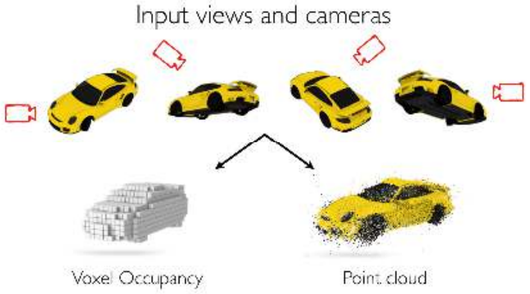

Figure 9. Creating 3D model using multiple photographs from different angles

The main goal of MVS is to estimate the most likely 3D shape, that is given in a set of photographs, assuming the system will have the knowledge about the materials, viewpoints, and lighting conditions.

The main difficulty of this algorithm is to know the viewpoints, materials, and lighting conditions, but as we need to construct a face from multiple photographs, and we are the ones placing the cameras, all the parameters mentioned above will be known. But even in that case some extra assumptions must be made to fully construct the 3D shape of the object.

Many hints exist that can be used to extract the object from photographs: defocus, texture, contours, shading, and stereo correspondence. Recently, the last three have gained the greatest fame, as they seem to be the most accurate. The techniques that use stereo correspondence as their main hint and use more than two images to extract the object are called multi-view stereo [7].

All MVS algorithms expect a set of images and their corresponding camera parameters. The result of the MVS algorithm will be as good as the quality of the input images and camera parameters. The generic pipeline of an MVS algorithm is shown in Figure 10 and looks like this:

- Image collection,
- Computation of camera parameters for each image,
- Reconstruction of the 3D object of the scene from the set of images and the corresponding camera parameters,
- Reconstruction of the materials in the scene, if possible.

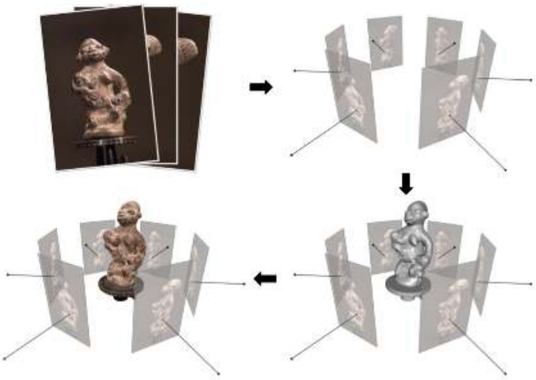

Figure 10. Example of multi-view stereo pipeline

Multi-view photometric consistency or photo-consistency in short is the main signal in any multi-view stereo algorithm. It measures the consistency and agreement between the set of input photographs and camera parameters. If the right assumptions are made, then the result of the multi-view stereo algorithms will be very accurate and of high quality.

A mandatory requirement for photo-consistency is that all images in the set see the same 3D object. But the problem is that to know if images see the same 3D object is to have the needed 3D object creating a circular dependency. To break this dependency, space-carving is used.

a) *Space-carving*

The invention of space-carving was very influential in photo-consistency. Given a 3D object partitioned into a 3D grid of voxels, voxels that are not photo-consistent and are step-by-step removed from the result.

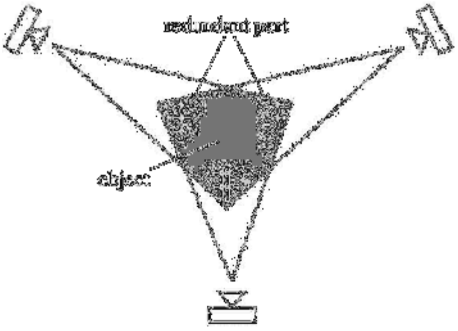

Figure 11. Space-carving example

As shown in Figure 11, an object is photographed from different angles, and after each photo the redundant parts are cut off from the 3D object. Using this technique, the system will be able to 'imagine' what shape the 3D object has. To make the process easier, the background of the object in question might be made black or white to make the object recognition easier for the system [8].

B. Representation of 3D model

Throughout time many different algorithms have been suggested. Two algorithms have proved to be most effective:

- Depthmap reconstruction
- Point cloud reconstruction

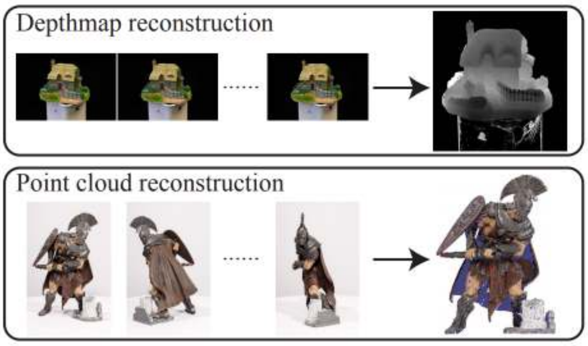

Figure 12. 3D restoration algorithms

a) Depthmap reconstruction

Depthmap algorithm is one of the most flexible and scalable algorithms and is very popular for that. If given thousands of images and camera parameters, it is possible to reconstruct a depthmap for each image and use neighboring images for photo-consistency evaluation. By viewing a depthmap as a 2D array of 3D points, multiple depthmaps can be considered as a 3D point cloud model. It is easy to process images in this way, and the amount of input images can be increased to thousands for more accurate results.

In short, the Depthmap algorithm takes a set of images and their camera parameters, analyzes the images, and converts the images into a finite set of depth values, and then reconstructs a 3D object [9].

b) Point cloud reconstruction

Depthmaps are effective for scene analysis and visualization, but it is difficult to create a 3D object from depthmaps. Depthmap loses quality if some parts of the object are visible only from one image or if there are depth discontinuities.

The point cloud algorithm resolves these problems since it uses all the input images to construct a single point cloud 3D object.

Corresponding points from all the input images are collected into a point cloud. When the 3D object is being constructed, the point cloud is used to place the points in the right places. Although some points might be wrongly corresponded creating minor mistakes in the result [10].

C. Final result

The mentioned algorithms have both strong and weak points, but they work best together supplementing each other. For faces, this process becomes even harder, as a human face has many small details, which might be missed when taking a picture.

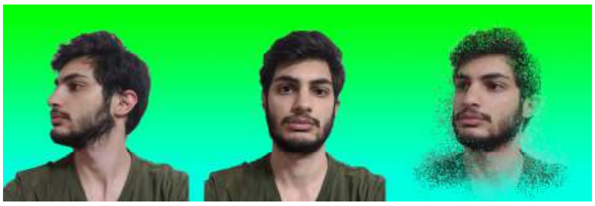

Figure 13. Result of 3D reconstruction

In Figure 13, a 3D reconstruction result is shown. The ear was reconstructed poorly, and the right cheek overflowed into the side, as not enough images passed to the program showing that parts of the head. The program is not perfect, so more than two pictures were used for reconstruction. But the result shows the capability of the mentioned algorithms, which can be used to reconstruct 3D models from multiple images. The more pictures are passed to the program, the more detailed and with better quality result is obtained.

IV. Conclusion

Multi-stereo view algorithms solve many problems regarding 3D reconstruction, including 3D face reconstruction. Even so, many problems remain, as the algorithms are still not humans, they can never use their intuition to assume some important details, which humans do without thinking.

Mistakes and inaccuracies will always take place with 3D face reconstruction and will never be totally trustworthy. Therefore, face authentication will always have the chance to fail or authenticate someone by mistake.

References

[1] V. Bruce, A. Young, "Understanding face recognition", *British Journal of Psychology*, pp. 305-327, 1986.

[2] W. Ren et al, "Low-Light Image Enhancement via a Deep Hybrid Network", *IEEE Transactions on Image Processing*, pp. 4364-4375, 2019.

[3] J. Besag, "Digital Image Processing", *Journal of Applied Statistics*, pp. 395-407, 1989.D

[4] G. Roth, "Homography", available at *http://people.scs.carleton.ca/~roth/comp4900d-12/notes/homography.pdf*, 2011.

[5] S. Tripathi, "Panorama Formation using Image Stitching using OpenCV", available at *https://medium.com/analytics-vidhya/panorama-formation-using-image-stitching-using-opencv-1068a0e8e47b*, 2020.

[6] B. Sharma, "What is LiDAR technology and how does it work?", available at *https://www.geospatialworld.net/blogs/what-is-lidar-technology-and-how-does-it-work/*, 2021.

[7] S. Seitz et al, "A comparison and evaluation of multi-view stereo reconstruction algorithms", IEEE Conference on Computer Vision and Pattern Recognition, pp. 519-528, 2021.

[8] K. Kutulakos, S. Seitz "A Theory of Shape by Space Carving", International Journal of Computer Vision 38(3), 199–218, 2000.

[9] F. Li, E. Li, M. J. Shafiee, A. Wong and J. Zelek, "Dense Depth Map Reconstruction from Sparse Measurements Using a Multilayer Conditional Random Field Model", 12th Conference on Computer and Robot Vision, pp. 86-93, 2015

[10] I. Reisner-Kollmann, "Reconstruction of 3D Models from Images and Point Clouds with Shape Primitives", Technische Universität Wien, 2013.