ACIT'2017
The International Arab Conference on Information Technology
Yassmine Hammamet, Tunisia
December 22-24, 2017

# *Multi-camera 3D modeling of a human body using the Shape from Silhouettes technique*

Ahlem Youssef
LATIS- Laboratory of Advanced Technology and Intelligent Systems,
ENISo, Sousse University - Tunisia
ahlamyoussef92@live.fr

Sami GAZZAH
LATIS- Laboratory of Advanced Technology and Intelligent Systems,
ENISo, Sousse University - Tunisia
sami.gazzah@gmail.com

***Abstract*—3D modeling is becoming increasingly popular in many areas such as computer-aided design, biomedical imaging, video games, 3D reconstruction of real objects and 3D special effects. The domain of three-dimensional reconstruction is a subject of several approaches, which can be classified into three main categories: mono, stereo and multi reconstruction. In this paper, we are interested in the reconstruction of 3D multi-camera and in particularly the technique Shape from Silhouette (SFS). The main objective is the 3D modeling of a dynamic human body by a multi-camera vision system using the Shape from Silhouette technique. The first task, our system takes captures of the images from the video sequences by the OpenCV library and then it segments each image using the technique of segmentation by thresholding. To show the camera synchronization, the mapping is used to establish a point-to-point relationship between two images. Our contribution is made to the level of point of detection. The second task of our system is the 3D reconstruction from the images coming from the several cameras of a published database.**

***Keywords—3D reconstruction, multi-sources, multi-camera calibration, SFS, 3D model.***

## I. INTRODUCTION:

3D reconstruction of objects from multiple images is one of the older problems in the field of computer vision. 3D reconstruction from images is the technique used to obtain a three-dimensional representation of an object or a scene from a set of images taken from different points of view of the object or the scene. In fact, the problem called 3D reconstruction is as follows: one or more 2D representations of an object are available, and it aims to determine the coordinates of the elements visible on these representations in a reference frame of the object 3D real space.

The applications are diverse and include back projection, quality control, virtual reality, augmented reality, pattern recognition, archeology, medicine and robotics, collaborative interactions, 3D reconstruction of real objects. For the domain of 3D reconstruction from calibrated images (i.e. the geometric relationship between 3D space is the image reference of the camera is known), several techniques have been proposed.

There is a distinction between mono-camera systems and multi-camera systems. The monocular approaches allow the processing of sequences from film scenes, collected videos from the Internet, or interpreted images from surveillance cameras. The estimation of a 3D pose from a single camera is particularly a complex problem due to the ambiguities associated with the 3D to 2D projection. The methods of multi-cameras vary according to the number of cameras, their position, and their calibration. For multi-cameras, as more as several images taken from slightly different deviation of view are available, it becomes possible to calculate the spatial position of the observed points in at least two images. The multi-camera reconstruction designates the process that allows combining several images of the same scene to extract three-dimensional information.

The approaches were developed based on the number of views available. Monocular approaches that estimate the geometry of the scene from a single view. From two points of view, stereovision represents the two most used concepts. Multi-view approaches built on the use of several views taken simultaneously, or the approaches built on the analysis of one or more views taken at different times.

In our work, we are interested specifically in 3D reconstruction using the Shape-From-Silhouette approach, which is able to reconstruct only an object fully visible by all cameras. A 3D modeling covers the object from their silhouettes. Therefore, Shape from silhouette (SFS) is a 3D reconstruction modality where the available data are the 2D binary projections of the shape in different views.

The paper is organized as follows. We first start with a state of the art where the different approaches

to three-dimensional reconstruction are presented. The second section will be devoted to the description of the tools to be used to develop our system, which must be able to reconstruct a 3D model from several images. Different experiments are conducted as well as the results obtained in Section 3. The paper is then concluded in which we discuss the presented study and state eventual improvements.

## II. STATE OF THE ART

Reconstruction of a complex three-dimensional (3D) rigid object from its two dimensional images (2D) is a challenging computer vision issue under general imaging conditions. We deal with the state of the art of the so-called optical methods (reconstruction from acquisitions with cameras).
The calibration is the first step in the 3D reconstruction process from multiple cameras. This step determines the relation between a point of the three-dimensional space and its point projection on the image given by the camera. Reconstruction from acquisitions with cameras has two sub-categories: passive methods and active methods. For the domain of 3D reconstruction from calibrated images, several techniques have been proposed. The 3D reconstruction mainly consists of the differences existing between the images acquired from different points of view, is dividing into 3 fundamental axes: Mono-Reconstruction 3D, Stereo-Reconstruction 3D, Multi-Reconstruction 3D. The monocular approaches use a single camera support on the extraction of certain characteristics of the image in order to determine the depth information [1]. This characteristic information can be Shape From Shading, Shape From Texture, Shape From Focus (Defocus), etc. The aim of the stereovision is to compute the 3D coordinates of each point of the scene from the coordinates of their projection on the two images provided by the two cameras.

By replacing one of the cameras of a stereo vision system with a device that projects a known pattern on the scene, matching is equivalent to compute the correspondences between the pixels and the points of the projected pattern.

The addition of a structured light source to the vision system greatly reduces the ambiguities associated with the problem of passive stereo vision matching. It is also easier to digitize non-textured objects (i.e. of uniform color). Indeed, the stereo sensor will have difficulty matching the points of the different images in the zones where the intensity (ie the color) is constant, since no visual index (characteristic point) can be extracted on the surface of the image, object. By projecting light according to a known structure, the system is no longer dependent on the characteristic points of the scene. It knows the type of visual index which he must be extracted from the images, which is an advantage.

The Shape From Silhouette (SFS) method [2] allows to reconstruct a 3D model from several images of an object taken from different points of view. Each image is then segmented to separate the object from the background. Shape From Silhouette, estimate the visual hull [3] (Visual Hull) of objects of interest (a human), which is an encompassing estimate of their 3D shape [4].

According to the literature, as first proposed by Baumgart [5], and then formalized by Laurentini [6], the form is computed as the maximal volume compatible with all silhouettes. In practice, it is calculated as the intersection of the visual cones formed by the silhouettes and their corresponding optical centers, the said visual hull (VH) [2]. From a set of n views, the silhouette information corresponding to the projection of the objects of interest is extracted from the images captured by a so-called silhouette extraction approach. The silhouette cone of a camera is defined by the set of half-lines coming from the optical center of the camera through the pixels belonging to the silhouette.

The visual hull is defined by a three-dimensional shape generated by the intersection of the silhouette cones of all the cameras. The form produced provides a volume encompassing objects of interest. Shape-From-Silhouette approaches have been used for various applications, such as monitoring, 3D modeling. There are different implementations of Shape-From-Silhouette approaches to estimate the visual hull. Some offer an estimate of the real-time visual hull, either volumetric or surface-based.

Since the visual envelope defines an enclosing 3D shape, this reconstruction proves to be inaccurate when the number of views is reduced. Other approaches include the use of additional color information to reduce the amount of artifacts constructed. The silhouette of the object is obtained for each image. The combination of these different silhouettes allows generating a 3D model. Laurentini [3] characterized the best approximation, obtained within the limit of an infinite number of silhouettes captured from all points of view outside the convex envelope of the object, such as the visual hull.

## III.OUR APPROACH

The main goal [7] of this system is to define a robust system that can used for 3D multi-camera reconstruction of a human body. For this, the Shape-From-Silhouette technique is used. Several cameras placed in the environment of the object were used to capture the images. The use of multiple cameras offer several different viewpoints on the object.

The complete procedure of Shape From Silhouette is broken down into the following steps:

• Calibration of the cameras to determine the position, orientation and intrinsic camera settings.

• Segmentation of the object from the background in the captured images.

• Silhouette matching.

• Intersection of all vision cones (visual Hull).

• 3D reconstruction.

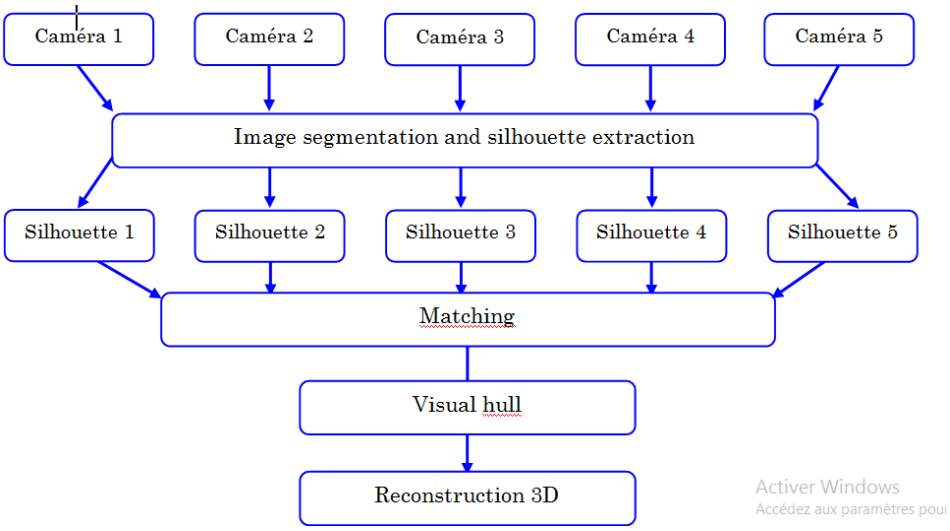


Fig. 1. Description of the proposed processing steps for 3D reconstruction

When it is desired to reconstruct a 3D object (or objects) belonging to a scene, it is necessary to distinguish them in the images. In each image, we determine the pixels corresponding to these objects of interest according to the rest of the scene. The set of these pixels forms the silhouette of these objects. In our work, we used the segmentation technique by thresholding.

The next step is matching. Matching is widely used in the field of computer vision. In fact, it is one of the most important steps during the 3D reconstruction from a stereo pair or a sequence of images. The correspondence of two images consists in establishing a point-to-point relationship between them. It requires a condensed representation of the images to reduce the calculations and allow applying similarity criteria. The ORB algorithm, which is based on FAST and BRIEF algorithm, is a method for describing feature points using the binary string. Since the characteristic point of the ORB is detected by the improved detection of the FAST function, which is described using an improved descriptor of the BRIEF function. Since FAST and BRIEF are very fast, ORB has an advantage absolute in speed. The huge features of this algorithm are rapidly and sensitivity to noise reduces.

Our contribution focuses on the level of detection of points of interest in the detected characteristic points. To sort the detected point and to take the most convened N points the most suitable, and also to calculate the distance between two points of interest, and to eliminate false points using the RANSAC algorithm with RANSAC is a technique to estimate a mathematical model from a data set, as the least-squares method could do. The latter works performs well if all the data are very close to the expected model.

The visual hull is the intersection of the different cones created from the different silhouettes used [8]. Because of the reconstruction from the several silhouettes, the well adapted visual envelope constitutes an upper bound of the real object to be represented. From *n* views, the well adapted silhouette information for the projection of the objects of interest is extracted from the images captured by a so-called silhouette extraction approach.

Finally, The Shape From Silhouette approach estimates in real time a 3D shape encompassing objects from their silhouettes from each camera. This method is able to reconstruct the objects fully visible by all the cameras, which imposes constraints of placement of the cameras with respect to the objects. The 3D shape calculated by Shape-From-Silhouette depends on the input silhouettes. Thus, any silhouette extraction error directly affects the quality of reconstruction. The 3D reconstruction consists in retrieving the three-dimensional information of the objects lost during the acquisition of images, that is to say the projection of the scene on a plane. Indeed, a camera moving in a 3D world collects a set of images that are two-dimensional data. However, both the trajectory of the camera and the world in which it evolves are in 3D.

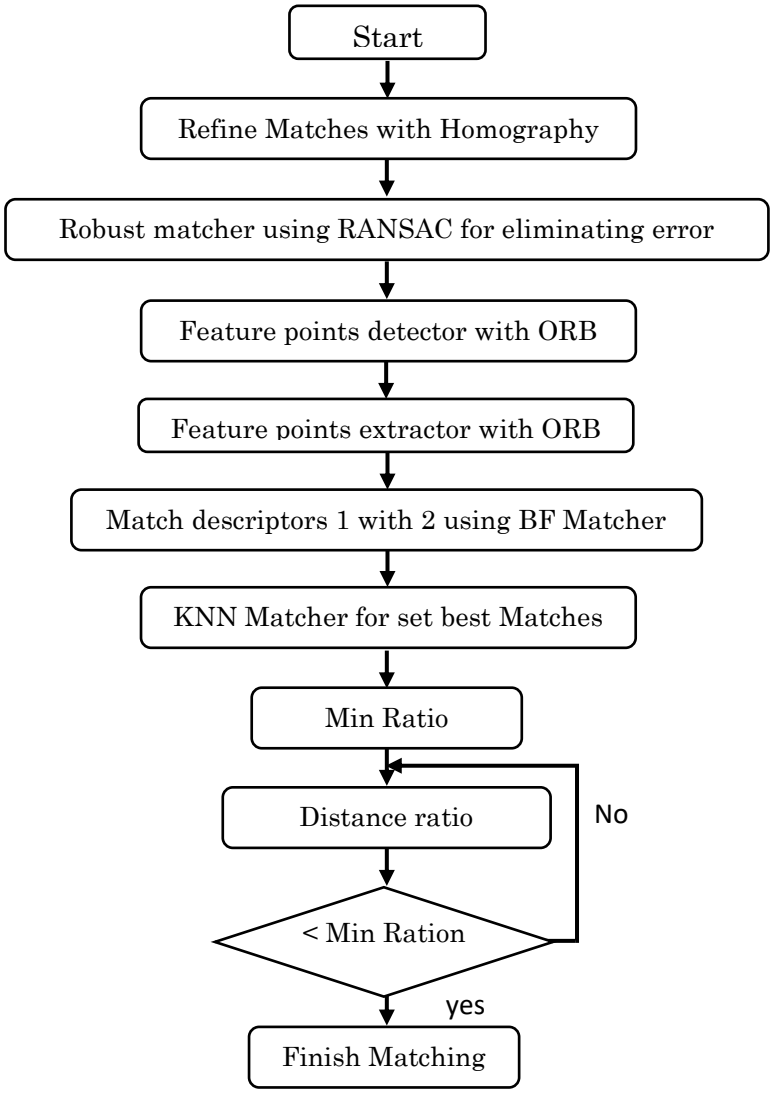


Fig. 2. Descriptive schema for image matching algorithm

## IV. EXPERIMENTATION

### A. *Implementation details*

In this section, we represent the different steps of 3D reconstruction using the Shape From Silhouette method. To show the features of the application, we take screen prints and describe each step. In our work, we used the Visual Studio 2010 software and then the OpenCV library.

In our project, we are interested to use the Multiple Cameras Full Data Set. The database presents one of the main risks for seniors living alone at home, causing serious injury is falls. In this paper, he is able to manage occlusions in 24 realistic scenarios showing 22 fall events and 24 confounding events (11 crouching positions, 9 sitting positions, 4 in sofa position) under several camera configurations. With eight cameras placed in the room, each scene contains 13 s, and the resolution of 720 × 480. An example (fig. 3) of fall seen by the eight cameras of the system.

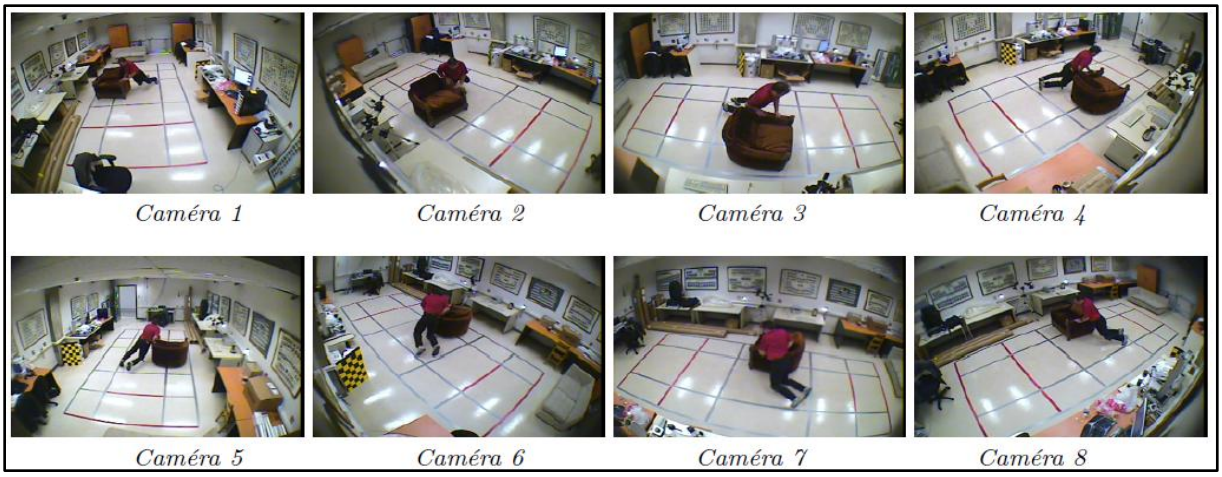


Fig. 3. An example of fall seen by the 8 cameras of the system

Then, to validate our work, we used another database "Multi-camera pedestrians video" [10]. In [28], the authors addresses the problem of tracking people who get confused by using a small number of synchronized videos such as those shown in Figure 3.3, which were taken at the head and under very different angles. This is important for the type of installation is very common for video surveillance in public places. Two scenarios called campus were contains three cameras. Up to four people walk simultaneously in front of them. The frame frequency is 25 frames per second. Each scenario contains 3 cameras and the duration of each scene 1:20 s and the resolution of 360 × 288.


Fig. 4. Campus sequences

In our work, we consider five video sequences acquired by five cameras moving in a specific environment. The aim of this work is to determine the 3D reconstruction of a human body by the Shape From Silhouette method. Five calibrated and synchronized cameras are available. The developed algorithm relies on the image segmentation that it is captured at a specific time t from each video, and then mapping images from the point of interest detection. Finally, the reconstruction of a 3D modeling from the silhouettes is performed.

We obtained a fully calibrated scene. Since the calibration parameters are known, it is necessary to be able to isolate the person using the segmentation of the image for the reconstruction of the volume.

### *B. Subtraction image*

The segmentation operation consists of determine whether a pixel of an image belongs to the foreground (i.e. the person) or to the background. In our study, we opted to the method of segmentation by thresholding, which aims to isolate the silhouette of the person.

In the following, we use thresholding image segmentation on which each object is selected and then analyzed. To separate the regions of interest from the background in an image, it is necessary to take the silhouette of each ones. When it is desired to reconstruct a 3D object belonging to a scene, it is necessary to distinguish them in the images.

In each image, we want to determine the pixels corresponding to these objects of interest among the rest of the scene. The retained set of pixels is extracted from the silhouette of objects and therefore not labeled ones are called background pixels.

The silhouette of the objects from a point of view can be defined by the projection of all the 3D points of these objects on the 2D image plane. We recall that one of our objectives is to estimate the 3D shape of objects of interest in real time. Thus, the silhouette extraction requires being precise with a strong real-time constraint.

Fig. 5. Ssilhouette ectraction

*C. Matching image*

It is necessary to extract the points of interest using the matching between two images. Mapping is the first step towards reconstruction. To well perform the reconstruction, it is necessary to find the homologous features in two images that correspond to the same physical entity of the real world.

To ensure high accuracy in image reconstruction, the algorithm must be applied to pairs of images describing the same event. If the right and left images of the pair were desynchronized, matching and reconstruction would be of poor quality.

We use for matching the mathematical operation which is the method of subtraction between two image capture times to facilitate the point of interest detection (fig. 7). For this, all the above figures show the difference of image to capture from camera 1, camera 2.

In a first place, we used the detector / extractor ORB to match two images of camera 1 and camera 2. We obtained the following result:

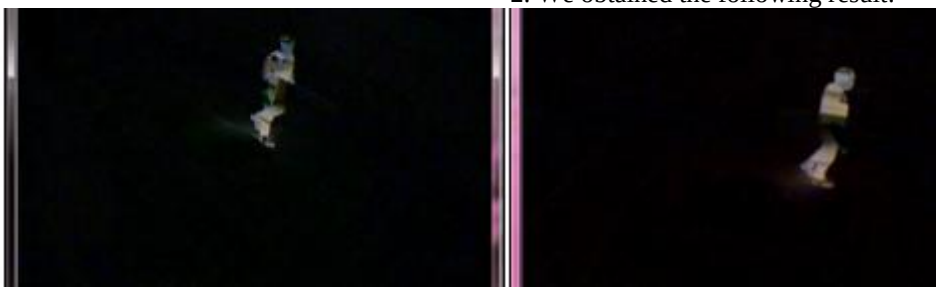

Fig. 6. Difference of two frame from different time


Fig. 7. Matching two frame without contribution

We found a difficulty in matching the image. For this, we propose to make our contribution to the level of the detection of point of interest. In a second place, we used the ORB (Oriented FAST and Rotated BRIEF) detector / extractor with a BFMatcher matcher, which uses the knnMatch, which finds the best numbers of matches. We have added to the code instructions for creating a file.txt whose values for each point of interest are saved during execution.

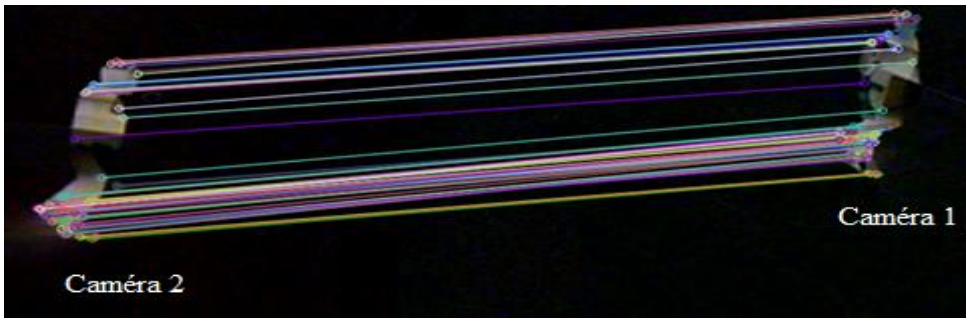


Fig. 8. Matching two frame with our contribution

Now we get all the necessary images and we compute the projection matrix parameters to do the 3D reconstruction of our model. Finally, to validate our matching contribution, we tested another published database.

Fig. 9. Difference de frame from another scene

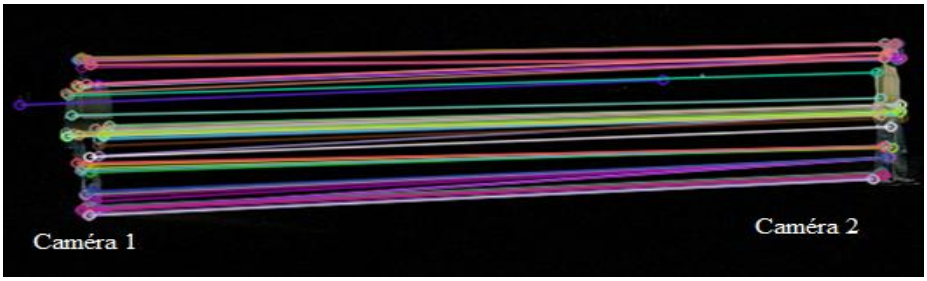


Fig. 10. Validation of matching

*D. Tree-D reconstruction*

Thee-D reconstruction is performed using the volumetric pixel "voxel". The voxel approaches are generally based on the projection of the voxels in the image plane of each camera. A voxel belongs to the extended visual hull if it is projected in all the silhouettes of the cameras that see this voxel, if the color consistency of the pixel of each projection is correct, the pixel is then defined as occupied.

Otherwise, it is set to blank and is deleted. Iteratively, all the voxels belonging to the external surface are then tested and after several iterations, only the minimal volume remains. For the volume reconstruction of the person, we will be used the voxel method. Consequently, by constructing a series of infinite cones coming from the optical center and traversing through the contour of the silhouette of each camera, the volume of person is found at the intersection of all these cones. Moreover, the first part of this method corresponding to the identification of the silhouette of the person in the image is facilitated in this context. The reconstruction of the volume is then immediate using the projection of these silhouettes. For the second part, we used the projection matrix of each camera which is modeled by the intrinsic parameters and the extrinsic parameters to identify and specify the position of person.

In this context, we have computed the projection matrix of our model with all the parameters that exists in the database article [11]. In this part, the processing of all the images gives a set of 3D voxel representing the person. We calculate the coordinates of each pixel and we save the result in a file ".ply" and then displayed it in software that is MeshLab. Therefore, using our matching file gives a clearer result but with noises.

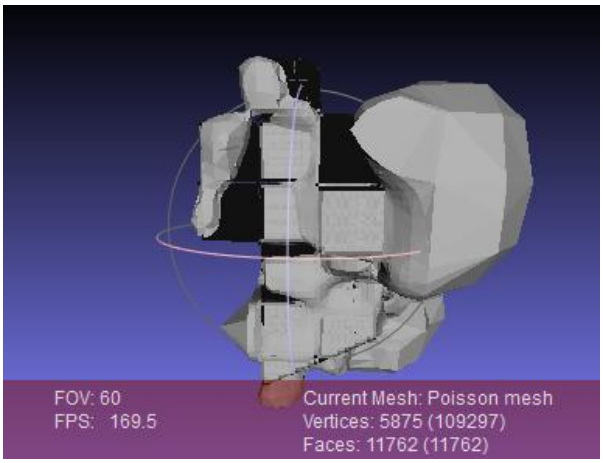


Fig. 11. Tree-D reconstruction

## V. Conclusion

In this paper we have presented 3D reconstruction approach based on "Shape From Silhouette". In this method, we have devoted a first step of image segmentation using thresholding segmentation after taking the image captures. To do this, we operated five cameras from a database.

In the second step, we used the mapping to show the synchronization of the captured images and to detect the points of interest by the ORB descriptor. Our contribution is based on the point of interest detection and calculates the distance between two points by KNN matcher.

Finally, we showed the development of each step of the Shape From Silhouette approach, through the choice of program in C / C ++ and verification of results of this code with a free OpenCV graphical library. Then, we were interested in reconstructing a 3D model of a human being with several cameras. We were able to retrieve the voxel of the real scene from the several images; afterwards, we saved the coordinates of the obtained points in a file of the format ".ply".

As a prospect for this final dissertation, one of the main tasks is the reconstruction of a 3D model generated from our images coming from the database. Secondly, we also generate a skeletal-based 3D reconstruction from a set of calibrated views.

## VI. References

[1] B. Michoud, « Reconstruction 3D à partir de séquences vidéo pour l'acquisition du mouvement de personnages en temps réel et sans marqueur», PHD thesis, 'Université de Lyon délivre par l'Université Claude Bernard Lyon 1, 2009.

[2] G. Haro, « Shape From Silhouette Consensus », Department of Information and Communication Technologies, Barcelona, Spain, Pattern Recognition, volume 45(9), pages 3231-3244, 2012.

[3] A. Laurentini, « The Visual Hull Concept for Silhouette-Based Image Understanding », IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Volume 16 Issue 2, 1994.

[4] R. Guerchouche, O. Bernier et T. Zaharia, « Reconstruction volumétrique multi-résolution d'objets 3D », 16ème Congrès Francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle, 2008.

[5] B. G. Baumgart, « A polyhedron representation for computer vision », national computer conference and exposition, page 19, 1975.

[6] A. M. Loh, R. Hartley, « Shape from non-homogeneous, non-stationary, anisotropic, perspective texture » ,Proceedings of the British MACHINE VISION CONFERENCE BMVC ,PAGES 69-78, 2005.

[7] G. K. M. Cheung, T. Kanade, J.-Yves Bouguet, M. Holler, « A Real Time System for Robust 3D Voxel Reconstruction of Human Motions », IEEE Conference on Computer Vision and Pattern Recognition, volume 2, , pages 714-720 CVPR , 2000.

[8]. J. S. Franco, E. Boyer, « Une Approche hybride pour calculer l'enveloppe visuelle d'objets complexes », In Actes des Journées ORASIS, Gerardmer, France, page 67-74, 2003.

[9] http://cvlab.epfl.ch/data/pom#calibration consulted at 28th august 28/08/2017.

[10] F. Fleuret, J. Berclaz, R. Lengagne, P. Fua. «Multi-Camera People Tracking with a Probabilistic Occupancy Map », IEEE transactions on pattern analysis and machine intelligence, volume 30(2), pages 267-282, 2008.

[11] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, Jacqueline Rousseau. « Multiple cameras fall data set », Research Center of the Geriatric Institute, DIRO-University of Montreal, volume 1350, 2010.