Multi-Stereo 3D Object Reconstruction

Carlos Hernández Esteban Francis Schmitt

Ecole Nationale Supérieure des Télécommunications
Signal and Image Processing Department
46, rue Barrault 75634 Paris Cedex 13
{carlos.hernandez,francis.schmitt}@enst.fr

Abstract

We present a method for the reconstruction of a 3D real object from a sequence of high-definition images. We combine two different procedures: a shape from silhouette technique which provides a coarse 3D initial model followed by a multi-stereo carving technique. We propose a fast but accurate method for the estimation of the carving depth at each vertex of the 3D mesh. The quality of the final textured 3D reconstruction models allows us to validate the method.

1. Introduction

As computer graphics and robot vision become more and more performing, attention is being focussed on complex high quality 3D models and the way they can be acquired from real objects. There exist a lot of different 3D object reconstruction methods but they can be classified into two different groups: active methods and passive methods. Laser range scanners and encoded light projecting systems use active triangulation to acquire precise 3D data. However they remain expensive and require special skill for the acquisition process itself. Furthermore only few scanners are capable of recording concurrently the 3D shape information with the color texture. Compared to active scanners, passive methods work in an ordinary environment with simple devices. The target object is pictured by a digital RGB camera from different view points, for example as it rotates on a turntable. The 3D information is then extracted from the sequence of 2D color images by using different techniques. To do so, we need the image sequence to be calibrated, and even if this is an open subject, we will not develop it in the paper and we will assume that all the images have been calibrated [8]. Our reconstruction approach consists of two different and complementary methods: a shape from silhouette [7, 1, 2, 10] and a multi-stereo method [13, 14]. The former constructs an initial 3D model by volume intersection from multiple views as described in section 2. The coarse 3D model obtained is then used as the initialization of a multi-stereo reconstruction method described in section 3.

2. Shape from silhouette

Let S_i be the i^{th} image silhouette, P_i the corresponding camera projection matrix and p a 3D point. We can define the cone C_i generated by the silhouette S_i as the set of lines l_{iv} which verifies

$$C_i = \{l_{iv} = P_i^{-1}p, p \in S_i\}$$

The visual hull E defined by the silhouette set S_i (see Fig. 1) can be written as:

$$E = \bigcap_{i=1,\ldots,n} C_i$$

Silhouettes enable us to define an implicit function in each 3D point. Different approaches for the construction of the visual hull have been developed [11, 12]. We have chosen to use an octree-based technique to construct it [15], the zero-level octree being the object bounding box. This bounding box is deduced from silhouette bounding boxes. The final visual hull resolution is only function of the octree depth level. For each new level, the resolution is multiplied by 4. Once the octree has been built, we get a 3D mesh of the

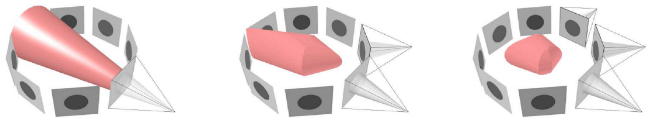


Figure 1. Visual hull construction by cone intersection.

visual hull by *marching cube* triangulation. We can see in Fig. 2 the visual hull obtained for a statuette with an octree depth of 8. The resulting 3D mesh contains 97223 vertices and 194442 triangles.

3. Multi-stereo carving

Since the visual hull is an upper-bound on the surface, we propose to carve it by a multi-stereo technique which will deform the initial model in order to adjust it to the true surface. To do so, we will use information contained in the object texture. Evidently, if the object has no texture or if its information is too weak, the method will fail. In this case, there exist alternative solutions such as paint projection (which is not acceptable for many objects) or structured light projection to *create* the information.

The multi-stereo carving procedure can be decomposed into 4 different steps:

- carving candidate detection,
- carving direction selection,
- carving depth estimation,
- carving depth filtering.

We describe these 4 different steps in the following subsections.

3.1. Carving candidate detection

To detect candidates for carving among the vertices of the visual hull mesh, we have to evaluate the depth quality of the surface at each one of its vertices. This quality measure is based on colour coherence of the projection of a vertex into the different images. To measure image coherence, we need to find a way to extract vectors of information from original data (i.e. the images). This is discussed in section 3.1.2. We need also to define a multi-vector likeness criterion as follows.

3.1.1 Multi-vector cross-correlation criterion

Let $\vec{v}_i, i \in \{1, \ldots, n\}$ be n different vectors. We would like to measure their likeness. If $n = 2$, a very well known criterion is the normalized cross-correlation. We can define the normalized vector $\vec{n}_i$ as

$$\vec{n}_i = \frac{\vec{v}_i - \vec{m}_i}{||\vec{v}_i - \vec{m}_i||}, i \in \{1, \ldots, n\},$$

with $\vec{m}_i$ being the vector whose components are the mean value of the $\vec{v}_i$ components. Cross-correlation between two vectors $\vec{m}_i$ and $\vec{m}_j$ is just defined as the scalar product of the normalized vectors $\vec{n}_i$,$\vec{n}_j$:

$$C_{ij} = \vec{n}_i \cdot \vec{n}_j.$$

If $n > 2$, a possible measure of likeness is the mean cross-correlation between all the possible vector pairs:

$$\begin{aligned} \mathcal{C}_1 &= \frac{1}{n(n-1)} \sum_{i,j=1, i \neq j}^{n} C_{ij} \\ &= \frac{\sum_{i,j=1}^{n} C_{ij} - n}{n(n-1)}. \end{aligned}$$

If $\vec{s}$ is the total mean sum vector,

$$\vec{s} = \frac{1}{n} \sum_{i=1}^{n} \vec{n}_i,$$

and $\vec{s}_i$ is the partial mean sum vector,

$$\vec{s}_i = \frac{1}{n-1} \sum_{j \neq i} \vec{n}_j = \frac{n\vec{s} - \vec{n}_i}{n-1},$$

we have

$$\begin{aligned} \sum_{i,j=1}^{n} C_{ij} &= \sum_{i,j=1}^{n} \vec{n}_i \cdot \vec{n}_j \\ &= \sum_{i=1}^{n} \vec{n}_i \sum_{j=1}^{n} \vec{n}_j \\ &= n\vec{s} \cdot n\vec{s} = n^2 ||\vec{s}||^2, \end{aligned}$$

and $\mathcal{C}_1$ can be written as a function of $\vec{s}$:

$$\mathcal{C}_1 = \frac{n||\vec{s}||^2 - 1}{n-1}.$$

The major problem with this criterion is that it does not allow eliminating vectors which are very different from others. This can happen for example when there is a highlight in one of the correlation windows. A possible solution can be found if we write $\mathcal{C}_1$ as the mean correlation between each vector $\vec{n}_i$ and its partial mean sum vector $\vec{s}_i$:

$$\mathcal{C}_1 = \frac{1}{n} \sum_{i=1}^{n} \vec{s}_i \cdot \vec{n}_i.$$

We can compute the likeness between every vector and the partial mean sum vector:

$$p_i = \vec{s}_i \cdot \vec{n}_i, \forall i,$$

compute the maximum of p_i,

$$p_{max} = \max_{i \in \{1, \cdots, n\}} p_i,$$

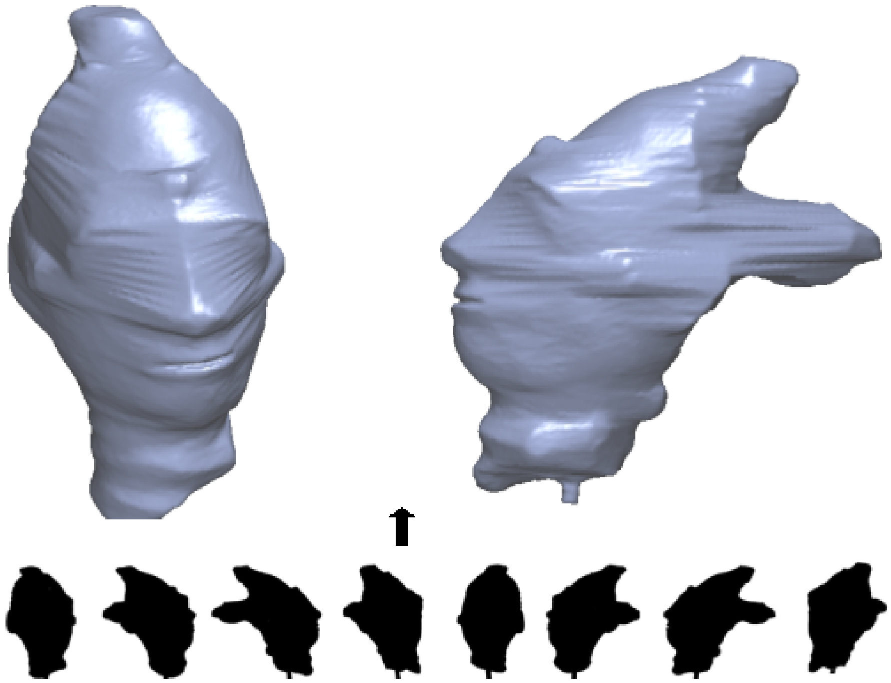


Figure 2. Visual hull of a statuette. Some samples of the 72 silhouettes are shown at the bottom.

and only use those vectors $\vec{n}_i$ which verify

$$p_i > T \cdot p_{max}, T \in [0,1].$$

If $\#(K)$ is the number of vectors that verify the preceding equation, we can write a second criterion $\mathcal{C}_2$:

$$\mathcal{C}_2 = \frac{1}{\#(K)} \sum_{i,p_i > T \cdot p_{max}} \vec{s_i} \cdot \vec{n}_i.$$

3.1.2 Evaluation methods

Having chosen a multi-vector criterion, we need to define the way information is extracted from images. Different authors have used various approaches but they always use the same basic idea: if a 3D point belongs to the object surface, its projection into the different cameras which really see it (i.e. there is no occlusion) will be closely correlated. A first method consists of projecting the 3D point into the image of each camera, sampling the image color at each point of projection and measuring the consistency of all color samples[3]. But this measure is very sensitive to noise, since we only have one sample per image. To increase robustness, it is preferable to take several samples per image. A way of doing this is to fit a quadric surface to the model [5], and to use the surface-induced mapping between cameras to compute coherence. This procedure has been successfully tested. But thanks to the high resolution of our 3D model, we can assume the surface to be locally flat. This improves speed as the mapping induced by a plane between two cameras is a homography.

We show in Fig. 3 the result of using the $\mathcal{C}_2$ criterion in a plane-based correlation. We can see the relation between the regions with a weak correlation (black and blue colour) and the concave regions, in particular in the area close to the ears and the nose.

3.2. Carving direction selection

For every carving candidate we need to define the direction in which to carve. This direction is chosen as the optical ray passing through the optical center of one of the cameras and the 3D point. Ideally, the best camera would have its image plane as perpendicular to the object normal as possible, but unfortunately, at this stage we know neither the right shape of the object nor its normal. As a first try, we could use the visual hull model as an object estimation and consider the visual hull normal as the object normal. If the visual hull is close to the true shape, estimation will be good enough. Otherwise, the visual hull normal may be completely wrong (Fig. 4). A more efficient solution consists of taking the median camera among the set of cameras which actually see the candidate. This method does not depend on the normal estimation and thus is more stable. In some cases, such as in Fig. 4, the method may not work as the above set of cameras can be void at the bottom of some folds. This is an intrinsic problem of the visual hull.

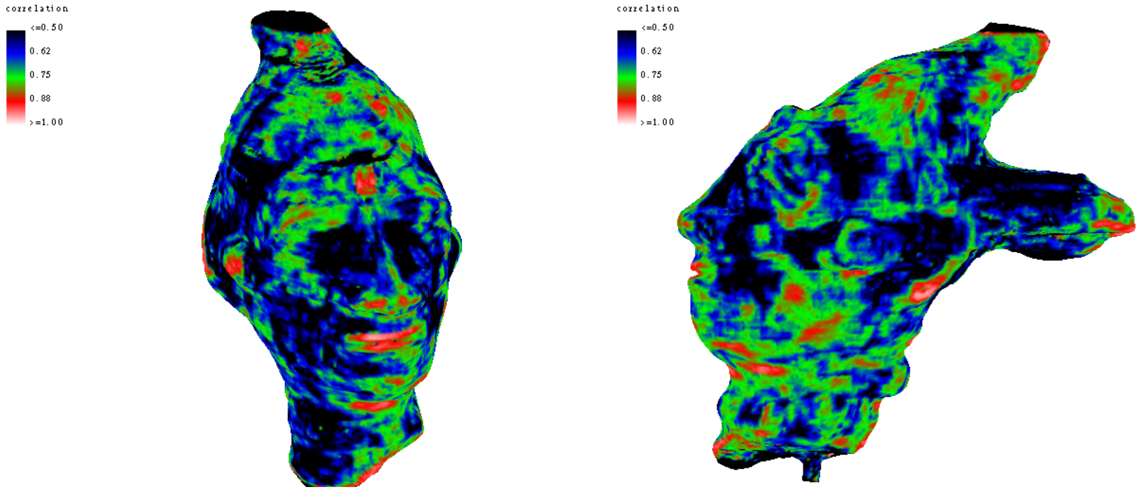


Figure 3. Plane-based correlation results

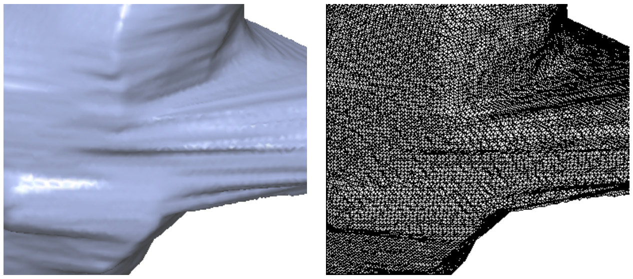


Figure 4. Back side of the visual hull (detail). Left: The surface is very folded with strong normal variations. Right: The corresponding high resolution mesh.

3.3. Carving depth estimation

For every point, we have to estimate the depth along the carving direction to attain the object surface. Unlike the carving candidate detection method, which was an evaluation method, the depth estimation cannot benefit of a mapping-induced surface. For the correlation method we have chosen the easiest solution: a rectangular window based correlation, also called front parallel correlation. Even if it does not exist any surface whose projection is always a rectangular window, the method is still robust since it does not assume any a prioristic assumption about the surface of the object. In addition, it is quicker to calculate. The algorithm used for evaluation of a depth along the carving direction is similar to that of [13]. For every 3D depth, the algorithm computes the projected 2D point for every image, extracts the centered window by resampling the image, and measures coherence of the resulting windows. One question that arises is the way depth sampling has to be done. For a given carving direction, its projection into the different images are also lines[1]and they are all related by the epipolar constraint. It is easy to establish the geometric relationship between a 2D deviation along a projected line d_{2D} and the corresponding 3D depth d_{3D}. This relationship depends only on the 3 parameters A, B, C as follows:

$$d_{3D}(d_{2D}) = \frac{A \cdot d_{2D}}{B - C \cdot d_{2D}},$$

$$d_{2D}(d_{3D}) = \frac{B \cdot d_{3D}}{A + C \cdot d_{3D}}.$$

Using these two equations, we can deduce the relationship between two different 2D deviations d^i_{2D} and d^j_{2D}:

$$d_{3D}(d^j_{2D}) = \frac{A_j \cdot d^j_{2D}}{B_j - C_j \cdot d^j_{2D}},$$

and

$$d^i_{2D}(d_{3D}) = \frac{B_i \cdot d_{3D}}{A_i + C_i \cdot d_{3D}},$$

we can substitute

$$\begin{aligned} d^i_{2D}(d_{3D}(d^j_{2D})) &= \frac{B_i \cdot A_j \cdot d^j_{2D}}{A_i(B_j - C_j \cdot d^j_{2D}) + C_i \cdot A_j \cdot d^j_{2D}} \\ &= \frac{E_{ij} \cdot d^j_{2D}}{F_{ij} + G_{ij} \cdot d^j_{2D}}, \end{aligned}$$

$$\begin{aligned} E_{ij} &= B_i \cdot A_j, \\ F_{ij} &= A_i \cdot B_j, \\ G_{ij} &= C_i \cdot A_j - C_j \cdot A_i. \end{aligned}$$

This relation is exact whereas in [13] the approximation $d^i_{2D} = K d^j_{2D}$ is used. With these equations we can relate in a single coordinate system all the correlation curves

[1] In a pinhole camera model context.

generated by a set of cameras. For a given image, we only need to calculate the depth steps which correspond to a local image variation of one pixel. This means that for a given depth interval, cameras near to the median camera need to compute a low number of correlations and cameras far from the median camera need a higher number. Logically, the bigger the parallax is, the higher the depth resolution. This property has led us to split the carving procedure into two different and complementary stages as follows:

- a first stage with a reduced number of cameras near the median camera for a quick rough initial search in all the depth interval,
- a second stage with the full set of cameras within a smaller interval for a high-accuracy search.

3.3.1 Initial search

For the initial search, we use only the 4 nearest cameras to the median one. The maximum depth to carve is directly given by the visual hull because we know that points must stay inside the visual hull volume. Using cameras with a small parallax allows us to scan a big depth range with a small number of correlations. For a given depth range and for every camera, we compute the correlation with the median camera, which gives us 4 correlation curves. By using the equations from the preceding section, we can represent all the 4 curves in the same coordinate system. Ideally, all the curves should have an absolute maximum at the depth of the object surface (Fig. 5.a). Unfortunately, it is not always like that (Fig. 5.b), for example when there is a highlight present or when not enough texture information is available (we correlate only image noise). The method used to estimate the initial depth is basically a voting approach. We search the depths corresponding to the local and global maxima for each curve whose global maximum is bigger than a threshold and we give to each depth a weight. If the maximum cumulated weight exceeds a threshold, the corresponding depth is validated as initial estimation and we proceed to the next stage. Otherwise, the estimation is rejected and the point remains uncarved.

3.3.2 Accurate estimation

Once the initial carving depth estimation has been obtained, we can use all the available cameras to improve the precision. We choose a little interval around the initial estimation and compute all the possible correlations with the median camera (Fig. 6.a). The curves whose maximum does not attain a threshold are rejected and the remaining ones are transferred to a single coordinate system for the computation of the mean curve (Fig. 6.b). The maximum of the mean curve gives us the final position of the point. The results of the multi-correlation carving are shown in Fig. 7.

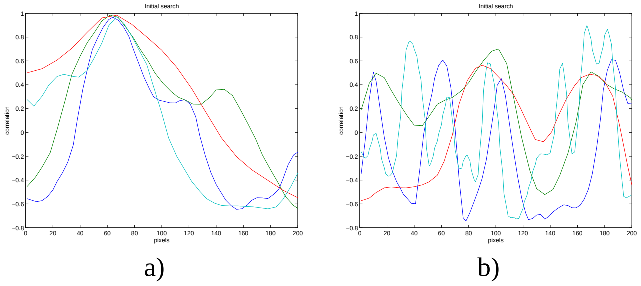


Figure 5. Initial search with the 4 closest cameras to the main camera. a) Valid case. b) Non-valid case.

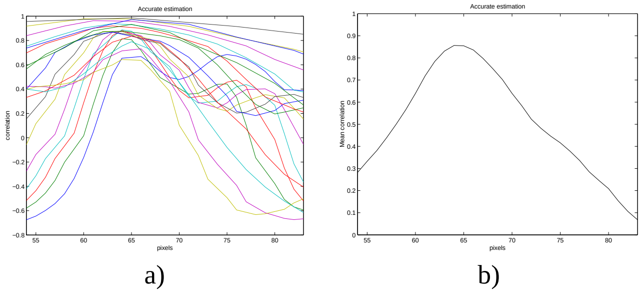


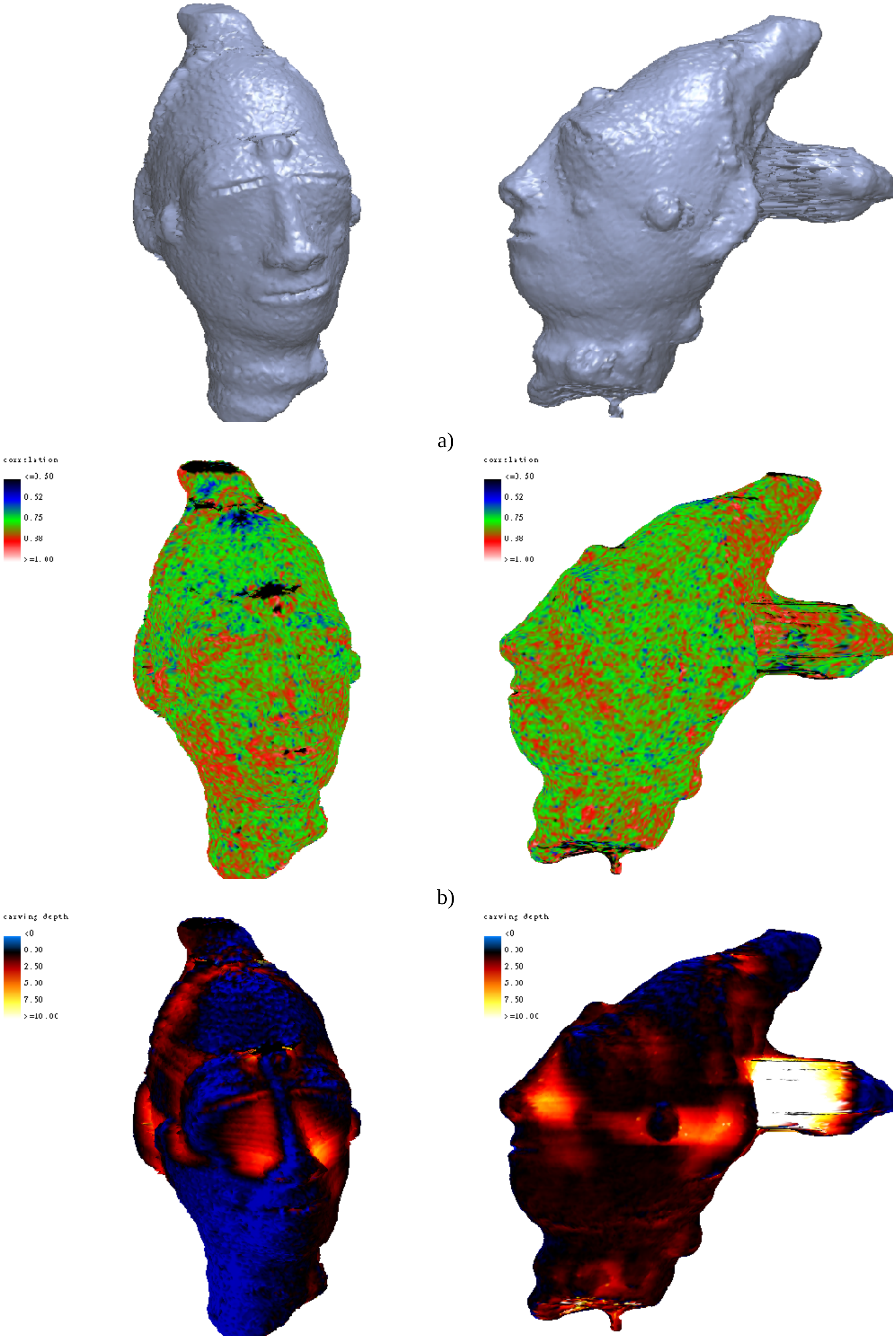
Figure 6. Carving depth accurate estimation. a) Set of valid correlation curves. b) Mean of correlation curves.

Regarding the depths map in Fig. 7.c, it can be highlited that there are points whose carving depth is negative (in blue). This has been due to a poor segmentation of the silhouettes which has caused a slight erosion at some regions of the visual hull. The multi-correlation technique has enabled us to precisely recover these regions.

3.4. Carving depth filtering

Let consider the 3D mesh resulting from the deformation of the visual hull mesh after carving. As the carving depth estimation is a local estimation, the results are noisy, which leads us to a new stage of depth regularisation. For a given point p of this mesh, we project every neighbour onto the carving direction, which gives us a depth for each of them according this direction (Fig. 9). Each depth is given a weight equal to its correlation value. The final position of p is chosen as the weighted median value.

A mean filter can also be applied to reduce irregularities due to the noise and the residual imprecision of the method. The results of both filters are presented in Fig. 8. Finally a texture mapping is applied with the method described in [15] (Fig. 10). The total CPU time required



a)

b)

c)

Figure 7. (a) Carved model. (b) Multi-correlation results. (c) Carving depths.

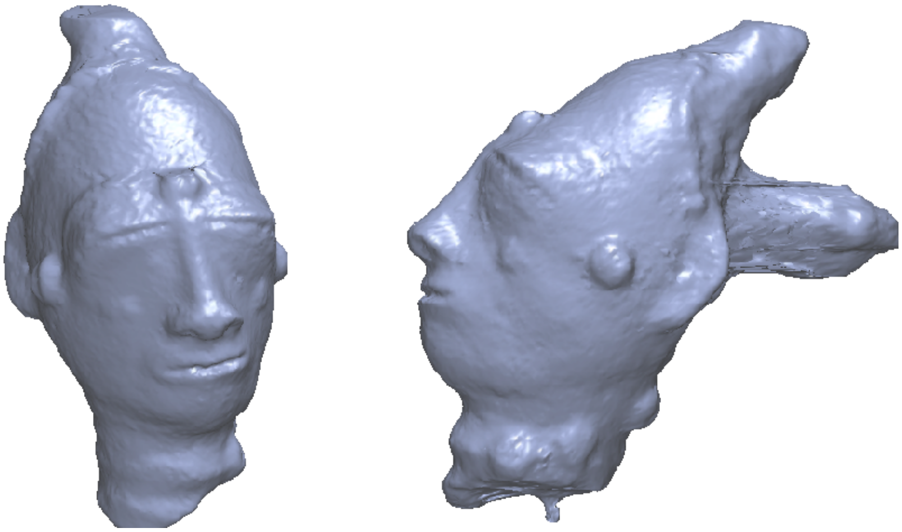


Figure 8. Carved model with filtering.

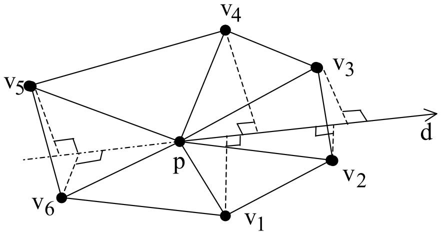


Figure 9. Depth regularisation.

for the full 3D model reconstruction (visual hull construction+carving+texture mapping) is around 10 minutes in a PC with a Pentium 4 processor at 1,4GHz.

4 Conclusions

We propose a 3D reconstruction method of real objects from a sequence of images. After the construction of the visual hull defined by the object silhouettes, we carve it with a multi-correlation technique. This method works well if the shape of the visual hull is not very different from the object, even if the carving depth is big (ex. the face of the statue). If the mesh deformation is very important (ex. the back appendix in the model) the results are noisy because there are many points that cannot be carved. We are considering a more global method as the level set approach [6, 4, 9] which we would like to combine with the present technique in order to control its computational cost.

5. Acknowledgements

We would like to thank the “Centre de Recherche et de Restauration des Muses de France (C2RMF)” for allowing us to use the Anyi sequence acquired for the European project IST ACOHIR.

References

[1] B. G. Baumgart. *Geometric Modelling for Computer Vision*. PhD thesis, Standford University, 1974.

[2] E. Boyer. Object models from contour sequences. In *Proc. ECCV*, pages 109–118, 1996.

[3] H. B. C.-E. Liedtke and R. Koch. Shape adaptation for modelling of 3d objects in natural scenes. In *IEEE Proc. Computer Vision and Pattern Recognition*, 1991. Hawaii.

[4] A. Colosimo, A. Sarti, and S. Tubaro. Image-based object modeling: a multiresolution level-set approach. In *IEEE International Conference on Image Processing*, volume 2, pages 181–184, 2001.

[5] G. Cross and A. Zisserman. Surface reconstruction from multiple views using apparent contours and surface texture. In A. Leonardis, F. Solina, and R. Bajcsy, editors, *NATO Advanced Research Workshop on Confluence of Computer Vision and Computer Graphics, Ljubljana, Slovenia*, pages 25–47, 2000.

[6] O. Faugeras and R. Keriven. Variational principles, surface evolution, pdes, level set methods, and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):336–344, march 1998.

[7] A. Laurentini. The visual hull concept for silhouette based image understanding. *IEEE Trans. on PAMI*, 16(2), 1994.

[8] J. M. Lavest, M. Viala, and M. Dhome. Do we really need an accurate calibration pattern to achieve a reliable camera calibration? In *Proc. of the 5th European Conf. on Computer Vision*, volume 1, pages 158–174, 1998. Germany.

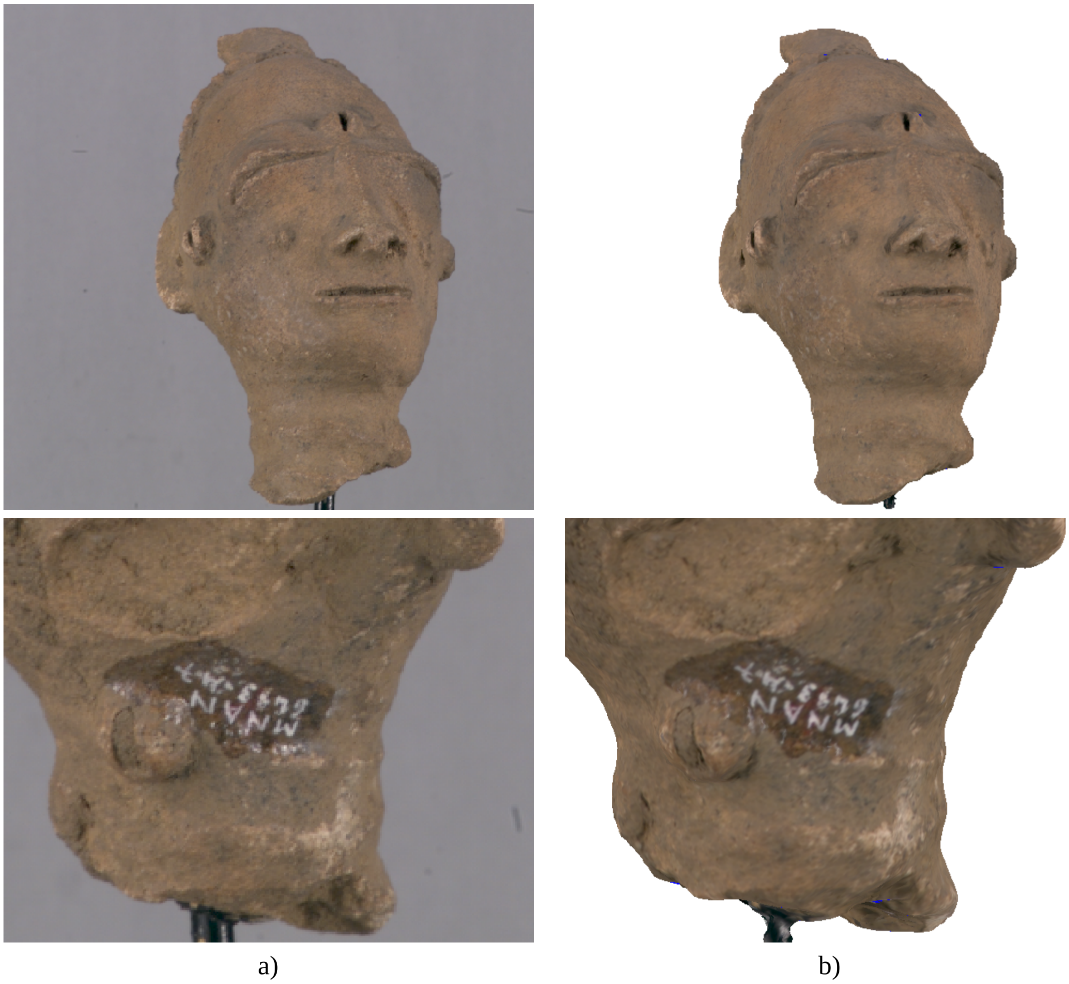


Figure 10. a) Original image. Below: Detail of the neck in a posterior view. b) 3D model with texture mapping in the same point of view.

[9] R. Malladi, J.A.Sethian, and B. Vemuri. Shape modelling with front propagation: A level set approach. *IEEE Tr. on PAMI*, 17(2):158–175, February 1995.

[10] Y. Matsumoto, K. Fujimura, and T. Kitamura. Shape-from-silhouette/stereo and its application to 3-d digitizer. *j-LECT-NOTES-COMP-SCI*, 1568:177–188, 1999.

[11] Y. Matsumoto, H. Terasaki, K. Sugimoto, and T. Arakawa. A portable three-dimensional digitizer. In *Int. Conf. on Recent Advances in 3D Imaging and Modeling*, pages 197–205, 1997. Ottowa.

[12] W. Niem and J. Wingbermuhle. Automatic reconstruction of 3d objects using a mobile monoscopic camera. In *Int. Conf. on Recent Advances in 3D Imaging and Modeling*, pages 173–181, 1997. Ottowa.

[13] M. Okutomi and T. Kanade. A multiple-baseline stereo. In *IEEE Proc. Computer Vision and Pattern Recognition*, pages 63 –69, 1991.

[14] P.Fua and Y. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *International Journal of Computer Vision*, 16:35–56, September 1995.

[15] F. Schmitt and Y. Yemez. 3d color object reconstruction from 2d image sequences. In *IEEE Interational Conference on Image Processing*, 1999. Kobe.