Actor 3D reconstruction by a scene-based, visual hull guided, multi-stereovision framework

Muhannad Ismael[1,2], Raissel Ramirez Orozco[1], Céline Loscos[1], Stéphanie Prevost[1],Yannick Remion[1],
[1] CReSTIC-RVM lab,University of Reims Champagne-Ardenne,France,<firstname.lastname>@univ-reims.fr
[2]WISIMAGE, Clermont Ferrand, France,<firstname.lastname>@wisimage.com

Abstract

This paper proposes a novel framework to produce 3D, high-precision models of humans from multi-view capture. This method's inputs are a visual hull and several sets of multi-baseline views. For each such view set, a surface is reconstructed with a multi-baseline stereovision method, then used to carve the visual hull. Carved visual hulls from different view sets are then fused pairwise to deliver the intended 3D model. The contributions of this paper are threefold: (i) the addition of visual hull guidance to a multi-baseline stereovision method, (ii) a carving solution to a visual hull from an interpolated and smooth stereovision surface, and (iii) a fusion solution to merge differently carved volumes differing in several areas. The paper shows that the proposed approach helps recovering a high quality carved volume, a 3D representation of the human to be modelled, that is precise even for small details and in concave areas subjected to occlusion.

Keywords

3D reconstruction, shape from silhouette, multi-baseline stereovision, visual hull

1 INTRODUCTION

This paper presents a solution to 3D reconstruction with constraints set by the broadcast industry with economically sustainable 3D post-production capabilities [1]. It aims at providing a new "virtual cloning" system of actors based on multi-video capture, natively delivering full 4D textured models of actors' performance.

Modelling of 3D objects from multiple views remains a major research problem in computer vision. Several techniques such as multi-stereovision, shape-from-silhouette, shape-from-shading, and structured-light 3D scanner have been proposed for 3D reconstruction. They are usually classified as active or passive reconstruction. Active reconstruction requires controlled illumination such as a laser or a structured light, which enable high precision 3D modelling. Whereas passive reconstruction relies only on the information contained in captured images, is less restrictive on the movement of the actors, and offers the possibility of capturing actual textures. In our case, passive reconstruction is preferable as live shooting of actual performances makes controlled illumination not desirable for our 4D textured model reconstruction.

In this paper, we propose a new passive multiview approach which aims at reaching the visual quality and the precision of active approaches. Our method merges results from shape-from-silhouette and multiple multi-baseline stereovision reconstructions. Multiview or multiocular stereovision methods such as [2, 3] conveniently reconstruct surface details and concave regions. However, they fail for textureless surfaces or repetitive textures because their core computational process relies on image texture. Shape-from-silhouette methods such as [4, 5] are very useful for real time applications in multi-camera environments [6] and handle conveniently textureless and specular surfaces. However, their reconstruction quality is somehow limited as the produced visual hull (VH) cannot recover concave regions laying inside every silhouette beam. Thus, multi-stereovision and shape-from-silhouette are complementary to each other and numerous hybrid methods have already been published (see section 2).

This paper is organised as follows. Relevant previous work related to silhouette-based and stereovision reconstruction is described in section 2. Our solution builds upon a multi-baseline stereovision framework overviewed in section 3.1. The acquisition system and geometry are presented in section 3.2. The main contributions are developed in the following sections. Section 3.3 describes an adapted version of a multi-baseline stereovision framework [7] to encompass VH guidance in order to enhance its performances. Section 3.4 exposes the VH carving process, from the necessary interpolation and smoothing of raw multi-baseline results with integer disparities to the VH carving from a float-

ing point disparity map. Section 3.5 explains our process for merging the carved volumes obtained from all multiscopic units into a 3D omnidirectional model of the actor. Finally, experimental results and conclusions are discussed in sections 4 and 5.

2 RELATED WORK

3D reconstruction methods combining shape-from-silhouette with stereo can be sorted into three groups.

(i) Stereovision methods guided by visual hull Seitz and Kutulakos [8, 9] propose to build a VH and carve it according to the photo consistency of each voxel on its external surface. After a VH process, surface voxels are iteratively eliminated if they project for each view on pixels of different color. It has the benefits to model occlusions and to achieve real time reconstruction. Unfortunately, the regular space discretizing scheme leads to sampling, aliasing artifacts and partial voxel occlusions. Matsuda *et al.* [10] propose direct carving to avoid local optima. They classify the points extracted from stereovision as either credible and not credible. The VH is then carved by the credible point cloud that verifies properties. For instance, credible point normals should not significantly differ from the VH normal of the nearest point on VH surface. However, this condition is not reliable for objects with steep concavities. Li *et al.* [11] propose to use polyhedral VH to improve a stereovision-based 3D reconstruction by quality, deleting outliers. However, this method does not handle known stereovision difficulties: textureless or specular surfaces and repetitive textures.

(ii) Energy function guided, model deforming using information provided by shape-from-silhouette and stereovision This class is concerned by deformation methods (e.g. snake) exploiting concurrently the information derived from silhouette-based reconstruction and stereovision [12]. Hilton *et al.* [13] optimize the deformation of a generic mesh model of a human shape to minimize an energy function encompassing the constraints of the VH and stereovision. The main drawback of it lies in its chosen model shape and topology dependent reconstruction. It does not consider actual performance specificities such as posture (self contacts), garment (loose clothes) or physical interactions with objects or other actors. Such restrictions in shape and topology assumptions are not desirable for our project.

(iii) Collaborative methods applying simultaneously criteria borrowed from VH and stereovision techniques Song *et al.* [14] adjust a point cloud extracted from stereovision using VH information. Their method groups the VH voxels into three classes: (1) voxels containing stereovision point(s), (2) voxels intersecting a segment between such a point and the optical centre of the stereovision reference image, (3) all remaining voxels, which are assumed to represent low texture or occluded areas. A point cloud is built from first and third voxel groups and only voxels which occlude stereovision points in the reference image (group 2) are carved out. The methods in [15] and [16] are relying on Kinect sensor which has a practical limiting range of (1.2 to 3.5 m) distance.

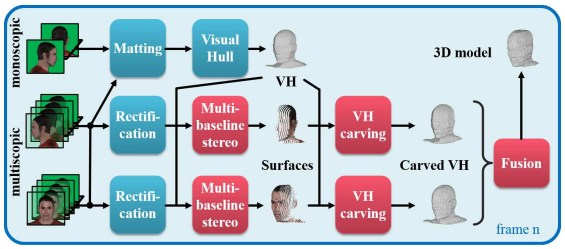


Figure 1: Proposed 3D reconstruction pipeline. Red blocks refer to specific contributions of the paper

3 PROPOSED FRAMEWORK

The proposed framework is summarized in figure 1 and borrows ideas from classes (i) to (iii). After its computation, the VH guides each multi-stereovision process per multiscopic unit. Then VH carving from stereovision is performed for each multiscopic unit similarly to class (i) but relies on a more global stereovision result close to the class (iii) concept. Finally, multiple (one per multiscopic unit) VH/multi-stereovision results are merged into a single global 3D model. Beyond its cross classification, our framework is innovative among each class. For each multiscopic unit a global scene-based multi-baseline stereovision process is run in disparity space which totally avoids partial occlusions and yields a robust stereovision result replacing more local and noisy photo-consistency usually used in class (i) carving. The proposed VH guidance class (i) is dedicated to our multi-baseline stereovision framework [7] which it enhances in terms of domain size, outliers avoidance, and, more innovatively, robustness in multi-stereovision similarity. The class "VH carving from stereovision" (iii) relies on voxel classification for voxels occluding the stereovision solution (group 2 in [14]). This classification is usually based on rays from the surface to the reference image. Replacing this image-based classification by a volumetric one in disparity space brings more precision and robustness to our solution. Furthermore, the multiple carved VH are merged at final stage. A smart handling reconstruction of inconsistencies from separate multiscopic units, conveniently corrects some residual stereovision mismatches.

3.1 Underlying multi-stereovision framework

3.1.1 Overview

This paper builds upon the VH guided multi-baseline stereovision process of Ismael *et al.* [7] for multi-

ISBN 978-80-86943-41-1

baseline stereo-vision, illustrated in figure 1. It relies on the assumption that the n views provided are synchronized images respecting the simplified multi-epipolar geometry (parallel optical axes and converging lines of sight, see [7]). The views are numbered 0 to $n-1$ from left to right facing the scene. The main features are twofold. Firstly, the solution is searched upon its natural domain in the disparity space (DS) introduced by [17], an efficient scene sampling scheme available thanks to simplified epipolar multiscopic geometry (see figure 2 that will be detailed next section). Secondly, this solution is formulated as a *materiality* map defined on this domain, expressing for each sample point its likelihood (in range $[0,1]$) of lying on a visible object surface as a perceived (indirect) light emitter. In the following, we reformulate specific parts of this method [7] important to understand the remaining of the present paper.

3.1.2 Scene space sampling scheme

Contrarily to numerous image-based approaches, this framework is deliberately scene-based as it works wholly and directly in a discrete 3D space laid in front of a multiscopic unit. This *workbench* space expresses directly the solution domain (see figure 2) on which several relevant properties are mapped. It is defined as a set of 3D points called *target points* and defined as the intersections of pixel rays from views of adjacent cameras of the multiscopic unit in a simplified geometry configuration.

Such points are aligned on constant depth planes as illustrated in figure 2 where f is the common virtual focal length of the cameras (the actual focal length divided by the horizontal pitch) and b their interocular distance. In any such plane, every point projects on any successive views on pixels of a same horizontal shift. This common column index shift of the projections is called disparity δ and is specific to the plane. A disparity δ is an integer value, defined as the common difference of column indices of the projections and is related to the depth z of the plane by $\delta = f \cdot b/z$. In figure 2, see the point circled in red whose pixels in images 2 and 3 are distant of $b-\delta$.

Any 3D target point may thus be defined by the intersection of a plane π_δ with a constant disparity δ with the ray which goes through its pixel projection $\mathbf{p}_i$ of any image i. Hence, each target point $\mathbf{T}$ may be indexed by a DS index $\mathbf{t} = (\mathbf{p}^t, \delta)^t$. $\mathbf{p} = (u,v)^t$ is the index of the pixel on which $\mathbf{T}$ projects in a chosen reference view of index i_0 (we usually choose $i_0 = 0$). According to simplified geometry, a target point indexed by $\mathbf{t} = (u,v,\delta)$ projects into any image i of the multiscopic set on $\mathbf{p}_i \in \mathbb{Z}^2$ identified by equation 1:

$$\mathbf{p}_i \equiv (u_i, v_i)^t = \mathbf{p} + (i_0 - i)\delta.\mathbf{u} = (u + (i_0 - i)\delta, v)^t \qquad (1)$$

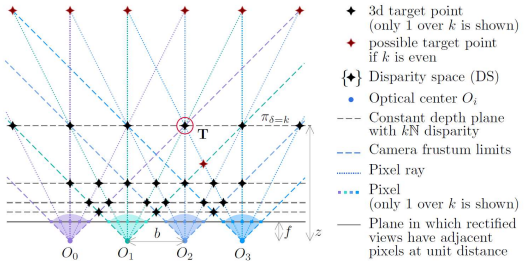


Figure 2: Disparity space: an efficient discrete reconstruction space. For clarity, only 1 over k pixels, associated rays, and constant depth planes are actually drawn.

3.1.3 Main framework concepts

Visibility: visibility reasoning evaluates for each target point with the function proposed by [2] and used in [18, 19]. This function is defined in the framework as the product of non-materiality of potentially occluding samples (see [7] for more details about the visibility function formula). DS ensures that each 3D sample point (target point) precisely lies on a genuine pixel ray in each image of the multiscopic unit for which it is inside the frustum. It thus intrinsically describes semi-occlusions ($\mathbf{p}_i$ in image i domain) and totally avoids complex treatment of partial inter-sample occlusions.

Similarity and confidence: the materiality and visibilities of target points are compared to input views, according to pre-computed similarity scores of neighbourhoods of their projections in some couples of views. This rather classical similarity computation includes (i) confidence computation typically based on variances of the neighbourhoods and (ii) a normalizing step of similarities along pixel rays which yields final similarity scores in range $[0,1]$.

Optimization and binarization: the materiality map is shaped by an optimization process, minimizing a dedicated energy penalizing deviation from intended map properties (such as completeness, smoothness, thinness) and inconsistencies between materialities, visibilities, and similarities (see [7] for more details). A binarization process delivers the final result, a binary materiality map. It is standing as a volumetric direct model of the intended solution, whereas image-based methods usually deliver disparity/depth maps that have to be processed to yield the reconstructed scene.

3.2 Shooting system and geometry

3.2.1 Studio layout and processing

Our method relies on a studio [1] composed of many synchronised and time stamped cameras with a green background (see figure 3), scattered around the observed scene in order to build the VH. Several groups laid as *multiscopic units* dedicated to multi-baseline stereovision. These units are composed of four aligned

ISBN 978-80-86943-41-1

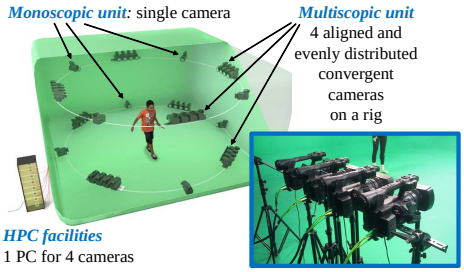


Figure 3: Dedicated multiview studio

and evenly distributed cameras. We choose a group of four which seems, according to experience, a good compromise between robustness, relying on views' redundancy, and computational efficiency (see [20]). The camera set is calibrated [21] in geometry and colorimetry in a pre-shooting step. For each time stamp, every image is matted thanks to pre-computed Chromakey and resulting silhouettes are used to compute the VH. For each multiscopic unit, captured images are then rectified to match simplified epipolar geometry.

3.2.2 VH-DS geometrical mapping

Hybridizing VH and multi-baseline stereovision implies mapping results of both methods in a same coordinate frame. Natively, VH is expressed in a regular grid in the scene frame, whereas multi-stereovision results are given in local disparity spaces, irregular in actual 3D space because their samples are not evenly spaced on fan-spread pixel rays (see figure 2).

This section presents, for any multiscopic unit, the mathematical relationship between voxel grid index $\mathbf{g} = (w,h,d)^t$ in VH, coordinates $\mathbf{r} = (x,y,z)^t$ in the frame of the rectified reference camera i_0, and coordinates $\mathbf{t} = (u,v,\delta)^t$ in DS.

This involves using (i) the VH grid parameters (scene frame reference and cell size $sw \times sh \times sd$) chosen at VH extraction step, (ii) calibration results for rectified cameras of the chosen multiscopic unit, and (iii) conversion of local depth z from reference camera i_0 to disparity δ. More precisely, we use the matrix $\mathbf{G}$ (mentioned previously in equation 2) positioning the VH grid in scene space, and the extrinsic $\mathbf{E}$ and intrinsic $\mathbf{I}$ matrices of the rectified reference camera i_0. Those matrices and their usage are exposed in equations 2, 3 and 4, where $\mathbf{R}_\Omega$ and $\mathbf{O}_\Omega$ are respectively orientation (rotation) matrices and origin points of frames Ω expressed in scene frame, and $\mathbf{Diag}(a,b,c)$ is the diagonal matrix with a,b,c values:

$$\mathbf{G} = \begin{pmatrix} \mathbf{R}_g & \mathbf{O}_g \\ & 1 \end{pmatrix} \times \begin{pmatrix} \mathbf{Diag}(sw,sh,sd) & \\ & 1 \end{pmatrix} \quad (2)$$

$$\mathbf{E} = \begin{pmatrix} \mathbf{R}_{i_0} & \mathbf{O}_{i_0} \\ & 1 \end{pmatrix} \quad \mathbf{I} = \begin{pmatrix} \alpha_u & s & & 0 \\ & \alpha_v & & 0 \\ & & 1 & 0 \end{pmatrix} \quad (3)$$

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \sim \mathbf{I} \times \begin{pmatrix} \mathbf{r} \\ 1 \end{pmatrix} \qquad \begin{pmatrix} \mathbf{r} \\ 1 \end{pmatrix} \sim \mathbf{E}^{-1} \times \mathbf{G} \times \begin{pmatrix} \mathbf{g} \\ 1 \end{pmatrix} \quad (4)$$

The DS index $\mathbf{t}$ is thus obtained from equations 3 and 4 by adding a convenient row in $\mathbf{I}$ (in red) which adds δ, defined in section 3.1.2, to its usual $(u,v,1)^t$ output. This yields the intended equations and matrices **DSfV** and **VfDS**, transforming respectively coordinates from VH to DS (equation 5) and backwards (equation 6):

$$\begin{pmatrix} \mathbf{t} \\ 1 \end{pmatrix} \sim \underbrace{\begin{pmatrix} \alpha_u & s & & \\ & \alpha_v & & \\ & & f \cdot b & \\ & & 1 & \end{pmatrix}}_{\mathbf{DSfV}} \times \mathbf{E}^{-1} \times \mathbf{G} \times \begin{pmatrix} \mathbf{g} \\ 1 \end{pmatrix} \quad (5)$$

$$\begin{pmatrix} \mathbf{g} \\ 1 \end{pmatrix} \sim \mathbf{VfDS} \times \begin{pmatrix} \mathbf{t} \\ 1 \end{pmatrix} \quad \text{with } \mathbf{VfDS} \equiv \mathbf{DSfV}^{-1} \quad (6)$$

3.3 Stereovision guidance by VH

This section exposes how VH guidance is added and enhances performances of the multi-baseline stereovision framework [7] presented in section 3.1.

3.3.1 Core principle

In the classical VH guidance, the reconstruction solution is necessarily included in the visual hull. Indeed, any point projected outside of at least one silhouette is labelled out. It requires mapping VH and target point spaces to bounded discrete 3D grid (*cf.* eq. 5 and 6, which give real coordinates). Thus, evaluating a map $\mathbf{M}$ defined in one space for a sample of the other space is achieved via trilinear interpolation. As shown in equations 7 and 8, the interpolation is noted with angular bracketing $\langle\ \rangle$ and the mapping by round bracketing$(\ )$, whereas direct map sample evaluation uses usual square bracketing $[\ ]$:

$$\mathbf{M}(\mathbf{t}) = \mathbf{M}\langle \mathscr{U}(\mathbf{VfDS} \times (\mathbf{t}^t,1)^t)) \rangle \text{ with } \mathscr{U}((\mathbf{v}^t,a)^t) = \mathbf{v}/a \quad (7)$$

$$\mathbf{M}(\mathbf{g}) = \mathbf{M}\langle \mathscr{U}(\mathbf{DSfV} \times (\mathbf{g}^t,1)^t)) \rangle \quad (8)$$

3.3.2 Bounding DS domain

The multi-baseline stereovision framework [7] works on a 3D grid laid on disparity space DS and indexed by $\mathbf{t} = (u,v,\delta)^t$. As such, this grid has to be bounded as close as possible to useful areas where the solution is expected to stand. Without any such prior information, which is usual in purely multi-stereovision, some lateral limits are easily set in u and v according to image frustums. The disparity range is usually asked for as an input parameter delivering the missing DS boundaries. VH, defined in a bounded 3D grid, may be seen as a superset of the actual solution. Thus, the solution is in a finite and closed area of scene space generally close to the actual solution, yielding opportunities to automate and optimize the delimitation of the DS.

Projecting in DS the eight corners $\mathbf{g}_i$ of the VH grid and keeping minimal and maximal DS coordinates gives a first axis-aligned bounding box (usually abbreviated

ISBN 978-80-86943-41-1

AABB) in DS in which the solution is necessarily included. This AABB is identified by its min and max indices $\mathbf{t}_m$, $\mathbf{t}_M$ in DS as follows:

$$\left.\begin{array}{l}\mathbf{t}_m = floor(min_{i=0,\ldots,7}\ \mathbf{t}_i)\\ \mathbf{t}_M = \ ceil(max_{i=0,\ldots,7}\ \mathbf{t}_i)\end{array}\right\} \text{ with } \mathbf{t}_i = \mathscr{U}\left(\mathbf{DSfV} \times \begin{pmatrix}\mathbf{g}_i\\ 1\end{pmatrix}\right) \quad (9)$$

With no user input, this step automatizes the DS bounding. It may even optimize in lateral dimensions as the VH bounding box may appear thinner than the available views. This first AABB is further optimized according to VH information. A sweeping process is run on each of its six faces, moving them inwards as long as they contain only target points whose interpolation in VH are considered *out*. This supposes (i) that the VH is defined on the grid as a numerical map **VH** with values monotonically (let us suppose increasingly) associated to $in, surf, out$ semantics and (ii) that some interpolation threshold out_t is set. A target point indexed by $\mathbf{t}$ is thus considered out of VH according to its interpolation in **VH**:

$$\mathscr{O}ut(\mathbf{t}) = \mathbf{VH}(\mathbf{t}) \geq out_t \quad (10)$$

This double process reduces the DS domain on which the different maps are laid (allocated) which thus optimizes computational efficiency.

3.3.3 Target point filtering according to VH

Despite its computational interest, the previously described VH guidance for DS bounding eliminates only some of the potential outliers outside the final AABB. Much more outliers are to be avoided if we remember that solution samples have to lie inside VH.

A simple preprocessing step labels every target point in the optimized AABB as undoubtedly outside or possibly inside the solution according to its VH interpolation $\mathscr{O}ut(\mathbf{t})$ (equation 10). Target points labelled as outside are not given similarity scores, nor considered for matching in the multi-baseline stereovision process. They are only used as conclusively non material points for visibility reasoning purposes. This target point labelling enhances computational efficiency. It also restricts the solution domain and avoids the evaluation of some more potential outliers which directly impact the reconstruction quality as illustrated in figure 5.

3.3.4 Enhancing similarity quality

Similarity scores are computed between similar rectangular neighbourhoods in couples of views. It relies on the assumption that neighbouring pixels usually have equal disparities, and thus, that the solution is locally at constant disparity. Adaptive windowing helps to modulate this assumption according to some heuristics which may be evaluated from known data (usually pixel values) statistically expressing the assumption quality for each neighbour. We use symmetrical bilateral filtering encompassing a neighbour weight factor computed according to the colorimetric similarity to the reference pixel. For each neighbour, this weight factor is the maximum of a computation on both views. As classically stated, this enhances similarity quality.

Furthermore, the similarity computation for a target point is also enhanced by a target point labelling: as this computation implies a local constant disparity assumption, it is reasonable to exclude target points, neighbouring in the constant disparity plane, labelled outside the VH. Such neighbouring samples are filtered out of the adaptive window before similarity computation. This ensures that target points known as irrelevant do not hinder the similarity scores computation. Those similarity scores are thus more relevant, enhancing the reconstruction quality and robustness.

3.4 Carving VH from stereovision

Our visual hull voxels are labelled as *in*, *out* and $surf$. However, multi-baseline stereovision yields a surface composed of the 3D points valued 1 in the binary materiality map. Each such point also bears a final confidence score related to its confidence scores associated to its similarities and possibly, its comparison to other target points on its pixel rays. Therefore, merging both models results in the intersection between the VH and the complement of the space between the multiscopic unit and the reconstructed surface. This corresponds to the subtraction or carving from VH of the multiscopic unit to surface space.

3.4.1 Stereovision surface coding

Precisely defining the space *"between"* the reconstructed surface and the multiscopic unit is not straightforward. It is a continuous space containing and interpolating, for every view of the unit, every part of a ray going from the optical centre to any solution point which is not occluded in this view. Most of those rays are redundant across the different views. We chose, for the sake of simplicity, to replace all these view dependent segments by others, far less numerous and redundant, attached to the same solution points but coming from a single centre located at the middle of the multiscopic unit. A drawback of this simplification may lie in a loss of solution points which could become occluded in this *virtual central view*. However, as a solution point has to be seen in at least a couple of successive views, this loss does not occur when $n < 5$ because the occluding rays of a solution point are limited to 0 to $n-2$ extreme views. As such the central ray cannot be flanked by two actually occluding rays ($n = 4$) or be itself occluding the solution point ($n = 3$). This remark enforces our choice to compromise using $n = 4$.

Our surface representation is built according to a central disparity space, abbreviated as CDS, indexed in reference of the (virtual) central view. This central view is

 ISBN 978-80-86943-41-1

less biased in actual 3D space than any other and, thus, interpolation in CDS will be more relevant. According to the multiscopic geometry (see 3.1.1), it corresponds to a camera indexed $i_c \equiv (n-1)/2$. Hence, a target point of index (u,v,δ) in DS would project in the central view at $(u_{i_c}, v_{i_c}) = (u + (i_0 - i_c)\delta, v)$ (see equation 1). In order to keep integer indices for n even, we multiply the horizontal coordinate in CDS by $\gamma = 2 - n \bmod 2$. This leads to new matrices managing transformation between coordinates $\mathbf{t} = (u,v,\delta)^t$ in DS and $\mathbf{c} = (c,v,\delta)^t$ in CDS and between VH and CDS:

$$\begin{pmatrix}\mathbf{c}\\1\end{pmatrix} = \underbrace{\begin{pmatrix}\gamma & & \gamma(i_0-i_c) & \\ & 1 & & \\ & & 1 & \\ & & & 1\end{pmatrix}}_{\mathbf{CfR}} \times \begin{pmatrix}\mathbf{t}\\1\end{pmatrix} \quad (11)$$

$$\begin{pmatrix}\mathbf{c}\\1\end{pmatrix} \sim \underbrace{\mathbf{CfR}\times\mathbf{DSfV}}_{\mathbf{CDSfV}} \times \begin{pmatrix}\mathbf{g}\\1\end{pmatrix} \quad \begin{pmatrix}\mathbf{g}\\1\end{pmatrix} \sim \underbrace{\mathbf{CDSfV}^{-1}}_{\mathbf{VfCDS}} \times \begin{pmatrix}\mathbf{c}\\1\end{pmatrix} \quad (12)$$

In this central space, we decide to represent the solution surface as a disparity map **DM** tagged by a confidence map **CM**. This is achieved by assigning for each solution point in DS, from far to near, at its CDS *pixel coordinates* (c,v), its disparity δ to **DM** (initialized to $-\infty$) and its associated final confidence score to **CM**. When n is even, in order to fill gaps induced by the horizontal stretching in CDS, if two successive target points on a row of CS are both solutions, their middle point is also assigned their common disparity in **DM** and mean confidence in **CM**. No other gap may occur because the solution in CS is computed in a way to ensure that its intersection with any (u,δ) plane is a continuous sequence of adjacent target points which are of same or adjacent disparities.

3.4.2 Carving VH from disparity map

The algorithm to carve the VH according to the stereovision surface coded by **DM** and **CM** is described in 1. It aims at filling a carved volume defined as a map **CV** laid over the VH grid and valued $in, surf_0..surf_q, out$. The different $surf_i$ values refer to increasing quantified confidence levels for surface voxels. The lowest confidence level $surf_0$ is reserved for $surf$ voxels of VH that are either occluded or out of frustum for the current solution. Other levels are associated with voxels identified as $surf$ in the stereovision solution: the effective level *i* is quantified according to the interpolated **CM** value of the voxel. A key feature of this step for the latter fusion process is to yield a coherent topology to the carved volumes: *in* and *out* sets are considered in a 6-connected space while $surf_{\{0..q\}}$ is considered in a 27-connected space. With such topological evaluation, no direct 6-connection should occur between *in* and *out* voxels.

In order to handle the grid sampling while responding to the previous intended topological property, point

```
1  c ≡ (c,v,δ)   N4 = {(−1,0),(1,0),(0,−1),(0,1)}
2  foreach g in VH domain do
3      if VH[g] is in or surf then
4          c = 𝒱(CDSfV × (g^t,1)^t)
5          if (c,v) in DM domain then
6              δ_s = DM⟨(c,v)^t⟩   g_s = 𝒱(VfCDS × (c,v,δ_s,1)^t)
7              if (‖g_s − g‖_∞) ≤ 1 then
                   CV[g] = surf_{Quant(CM⟨(c,v)^t⟩)}
8              else if δs < δ then CV[g] = out
9              else
                   if VH[g] is in then CV[g] = in
10                 else CV[g]=surf_0
11                 foreach n ∈ [0,4[ do
12                     lg =
                           ‖𝒱(VfDS × (c^t + (N4[n],0,0))^t) − g‖_∞
                       n_c = (c,v)^t + N4[n]/lg
13                     if n_c in DM domain and
14                       (δ_n = DM⟨n_c⟩) < δ then
15                         cnf = (CM⟨(c,v)^t⟩(δ − δ_n)+
                               CM⟨n_c⟩(δ_s − δ)) / (δ_s − δ_n)
                           CV[g] = surf_{Quant(cnf)}
                   end
               end
16         else if VH[g] is in then CV[g] = in
17         else CV[g]=surf_0
18     else CV[g] = out
19 end
```

Algorithm 1: Carving VH by central disparity map

comparison in CDS is related to actual axis-aligned distance $\|\ \|_\infty$ in VH. Hence, the interpolated solution point $\mathbf{c}_s = (c,v,\delta_s)^t$ is projected back in VH to measure its distance to thinitial voxel $\|\mathbf{g} - \mathbf{VfCDS}\times\mathbf{c}_s\|_\infty$, with $\mathbf{g} = (w,h,d)^t$. When this distance is less than 1 (*line 7*), $\mathbf{g}$ is labelled $surf$ in the carved volume with a confidence level quantified from $\mathbf{CM} < (c,v)^t >$. If the voxel $\mathbf{g}$ is in front the surface ($\delta > \delta_s$), it is labelled *out* in **CV**. Otherwise, the voxel is *a priori* labelled *in* but could be labelled $surf_i$ if it lies close enough of a steep slope of the surface. To check this possibility, we evaluate (*line 11*) if any of its 4 neighbours in CDS of same disparity δ, at unitary distance in VH, are to be considered *out* (with interpolated disparity lower than δ). This evaluation consists in measuring the distance lg in VH of the initial voxel to a neighbour $\mathbf{n}_0$ at a unitary distance in CDS and interpolating disparity δ_n in **DM** at a neighbour $\mathbf{n}_c$ in same direction but distance lg^{-1}. If $\delta_n < \delta$ this neighbour is considered *out* and the initial voxel is re-labelled $surf_i$ where the confidence level *i* is quantified from the linear interpolation at δ of $\mathbf{CM}\langle\mathbf{n}_c\rangle$ at δ_n and $\mathbf{CM}\langle(c,v)^t\rangle$ at δ_s.

3.4.3 Improving surface smoothness

The result of multi-stereovision method leads to discontinuous surface divided into frontal planar patches with constant and integer disparity, one for each multiscopic unit (see figure 6). Removing this effect is required for the visual quality of the result (see figure 6) and for a more accurate management of reconstruction inconsistencies between different multiscopic units. To deal with

ISBN 978-80-86943-41-1

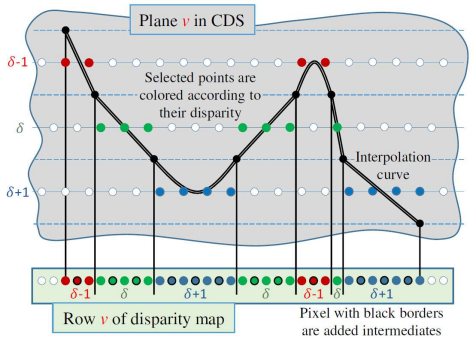


Figure 4: Disparity interpolation: relation between disparity map **DM** (coloured points) and interpolated disparity map $\mathbf{DM}_r$ illustrated in CDS by the interpolation function (black double lined curve)

this problem coming from the integer disparities quantification, we propose to represent the solution surface previously saved in **DM** by a floating point derivative version $\mathbf{DM}_r$. The map $\mathbf{DM}_r$ is computed to ensure continuous transitions between adjacent horizontal segments of constant disparities with a disparity gap of 1. Computing $\mathbf{DM}_r$ consists in looping over rows of **DM**.

Every row v of **DM** is thus scanned from one end to the other to identify disparity steps between adjacent pixels of finite disparity. When the disparity step is of magnitude $(-1,+1)$, a contact point (black point in figure 4) is placed in CDS in the middle of the two pixels with the mean of their disparity values as illustrated in figure 4, and serves as end point of both segments. Otherwise, one end point is placed for each adjacent segment in the middle of the two pixels at the segment disparity. When one of the pixels is of infinite disparity as well as for first and last pixels, a single end point is generated on the relevant pixel at its finite disparity. This process yields two end points per segment expressed in CDS $(c0,v,\delta_0)$ and $(c1,v,\delta_1)$. When a right end point $(c1,v,\delta_1)$ is generated, the corresponding segment of initial constant disparity δ is filled in $\mathbf{DM}_r$ by a dedicated interpolation scheme between the end points.

$$\mathbf{DM}_r[(c,v)^t] = \delta + (1-t)(2t-1)(\delta-\delta_0) + t\cdot(2t-1)(\delta_1-\delta) \quad , \quad t = \frac{c-c_0}{c_1-c_0} \tag{13}$$

The interpolation function in equation 13 ensures that both end points are respected (see figure 4 where the black double lined curve expresses the interpolation function producing the interpolated disparities in $\mathbf{DM}_r$). When δ_0 and δ_1 are both under or above δ, or if one equals δ, this interpolation is parabolic. When one is above and the other under, they are equal and the interpolation is linear.

3.4.4 *Smoothing using bilateral filter*

The result of the disparity interpolation described in the section 3.4.3 is a floating point disparity map more *continuous* or smooth on each row but still presenting vertically numerous depth steps. A bilateral filter is applied on the disparity map $\mathbf{DM}_r$ to compute a smoothed disparity map $\mathbf{DM}_s$ as described in equation 14 and demonstrated in figure 6. The centred operating window is chosen rectangular as regulating transitions between segments implies a rather low width $2ww+1$ but reducing vertical depth steps involves a much taller height $2wh+1$.

$$\mathbf{DM}s[\mathbf{q}] = \frac{\sum_{\mathbf{n}\in W} \mathbf{DM}r[\mathbf{p}+\mathbf{n}]\ \mathscr{W}(\mathbf{p},\mathbf{n})}{\sum_{\mathbf{n}\in W} \mathscr{W}(\mathbf{p},\mathbf{n})} \tag{14}$$

with $\mathbf{n} = (dc,dv)^t$, $W = [-ww,ww]\times[-wh,wh]$ and

$$\mathscr{W}(\mathbf{p},\mathbf{n}) = \mathscr{G}_{\sigma_c}(dc)\ \mathscr{G}_{\sigma_v}(dv)\ wd(\mathbf{DM}_r[\mathbf{p}+\mathbf{n}] - \mathbf{DM}_r[\mathbf{p}])$$

$$\mathscr{G}_\sigma(t) = \gamma_\sigma \cdot exp(-t^2/(2\sigma^2)) \qquad \gamma_\sigma = (\sigma\sqrt{2\pi})^{-1}$$

wd a function decreasing from 1, for example

$$wd(\Delta\delta) = \sigma_\delta^2/(\sigma_\delta^2+\Delta\delta^2)$$

3.5 Omnidirectional 3d modelling

3.5.1 *Merging difficulty*

The final step of the 3D reconstruction consists in merging carved VH volumes $\mathbf{CV}_m$ from multi-baseline stereovision results for all multiscopic units m in order to obtain a single 3D model representing the 3D pose of the reconstructed actor.

Figure 6 illustrates that the result of each multiscopic unit provides information only on visible surfaces facing the unit while other surface areas are left to VH result. Multiple carved VH from different multiscopic units spread around the scene thus yield stereovision details for almost every surface area of the model.

However, parts of the model surface are to be seen and reconstructed by multiple multiscopic units and those independent reconstructions are usually inconsistent one to another. Therefore, in such common areas, we have to decide which reconstruction is locally kept in the final solution. This decision is based on the confidence attribute of surface voxels: as stated in section 3.4.2, surface voxels in $\mathbf{CV}_m$ bear different labels $surf_i$ indicating their quantified confidence level according to the stereovision process.

3.5.2 *Merging process*

The overall principle of this final step is to initialize the final merged volume **FV** to one of the carved VH ($\mathbf{FV} = \mathbf{CV}_{m_0}$) and then iteratively merge each other carved VH $\mathbf{CV}_m$ into **FV** according to surface confidence decisions in differently labelled areas. As VH is known to be a superset of the solution, the process only evaluates voxels labelled *in* or *surf* in **VH**. It thus loops over every voxel **g**, treating each one for which $\mathbf{VH}[\mathbf{g}]$ is not *out* according to its labels $\mathbf{FV}[\mathbf{g}]$ and $\mathbf{CV}_m[\mathbf{g}]$:

- both *out*: voxel **g** is kept *out* in **FV**
- both *in*: voxel **g** is kept *in* in **FV**

ISBN 978-80-86943-41-1

- $surf_i$ and $surf_j$: voxel **g** is kept $surf$ with the highest confidence level $\mathbf{FV}[\mathbf{g}] = surf_{max(i,j)}$
- all other cases: voxel **g** bears inconsistent labels, the global loop is suspended while an inconsistency resolution process is run from **g**.

To decide which solution is to be kept in the last case, we propose a global evaluation of the 6-connected area implied in the detected inconsistency rather than a per voxel decision. Thus, when a voxel **g** is detected as inconsistent in the global loop, a two-pass process starts in order to make a decision.

The first pass aims at making the right decision. It goes from **g** through its inconsistent 6-connected area in order to compute the per-confidence level histograms of the encountered surfaces of both volumes. These confidence histograms for the two surfaces help making the decision on which volume **FV** or $\mathbf{CV}_m$ will transfer its labels to the final solution in this 6-connected area. We propose to choose the volume with the highest mean confidence level, but other competing scores could easily be proposed and tested from confidence histograms.

When the decision is made, a second pass is run. The same walk-through in the area is performed in order to resolve the inconsistency by copying labels of the chosen volume into the other. One could have thought that when the chosen volume is **FV** nothing needs be done, but the first pass and the decision making would then be repeated for every voxel of the area which is far from efficient. Therefore, during this second pass, when a voxel labelled $surf_i$ and $surf_j$ is encountered, its best confidence level $(max(i, j)$ is kept in both volumes.

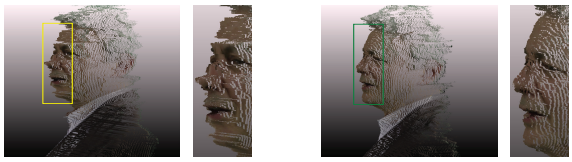
This process clearly relies on a consistent topology in both volumes. This point is ensured by the VH carving step described in section 3.4.2. This topological consistency further permits to keep our 6-connected area walk-through topologically consistent: it starts from an inside position (in or $surf_i$) in one of the volumes $\mathbf{V}_i$ and an outside position (out or $surf_i$) in the other volume $\mathbf{V}_o$. This per-volume topological position has to be ensured over the whole traversed area. No shift from *in* label to *out* label should occur in each volume across a 6-connection. Thus, ensuring topological consistency consists in avoiding 6-connections transgressing initial inside/outside position in any volume. This could occur in $\mathbf{V}_i$ for voxels on the surface connected to *out* voxels as in $\mathbf{V}_o$ for voxels on the surface connected to *in* voxels.

3.5.3 *Refinements*

A rough application of the process described in section 3.5.2 is not satisfactory because the walk-through areas sometimes appear as several,rather broad and distant, *blobs* of non surface voxels connected by thin lines or surfaces. The decision is made once for the whole area, while it should be differentiated for each blob and connection line or surface. This yields inconvenient decisions which need to be corrected. In order to do so, we apply several times the merging process of section 3.5.2 (three times in the present implementation) with less and less restrictive conditions on inconsistent voxels:

1. Considered voxels have to be labelled in/out or out/in. Furthermore a sufficient part of their 6-neighbours has to be labelled in the same way (at least 40% in our implementation). This step treats broad in/out blobs.
2. Considered voxels are the remaining in/out or out/in ones. This step treats rather thin areas.
3. Considered voxels are any other inconsistent ones. This steps finalizes the resolution and treats very thin areas with no (in, out) or (out, in) voxel.

Results from this refinement are illustrated in figure 7.



(a) (b)

Figure 5: Resulted point cloud of a real actor "Jacques". (a) point cloud obtained with integer disparity values without VH guidance and zoom in its yellow area. (b) point cloud obtained with integer disparity values with VH guided stereovision and zoom in its green area.

4 RESULTS AND DISCUSSION

To evaluate our framework described in figure 1, we used the studio layout scheme presented in section 3.2.1 both for real and virtual shooting and applied our framework to the views they produced. These experimental conditions apply to each result discussed in this section.

Figure 5 illustrates that the VH guided stereovision method described in section 3.3 improves the materiality map derived from a previous multi-baseline stereovision method [7] by ridding it of outliers outside the visual hull. Moreover, in non specular textured or concave areas, the materiality map solution proves to be more accurate than the visual hull as illustrated in first rows of figure 6 which clearly show that concavities, such as eye cavities, are carved out by our stereovision method both for virtual and actual shootings.

Figure 6 shows the results of the carving process described in section 3.4 on two view sets: the first one, of a virtual actor "Simon", shot under ideal calibration conditions by computer graphics software and the second one, of a real actor "Philippe", captured in the RECOVER 3D dedicated studio. Comparing the carved volume to the point cloud on each row of these figures, qualitatively validates our carving method. The evolutions obtained on both figures from each row to the

ISBN 978-80-86943-41-1

next, demonstrate the relevance of the disparity interpolation and smoothing steps.

The fusion of every multiscopic unit outcomes (see section 3.5) provides robust reconstruction, especially in the areas where two or more multiscopic units compete. Figure 7 demonstrates this with results obtained from a virtual and a real data set. One should notice the results' quality despite the low number of implied multiscopic units: three for the actual shooting and four for the virtual one.

To compare our results to state of the art, we apply our data (masks, RGB images, and camera parameters) to the PMVS method proposed by Yasu Furukawa [1]. We also apply chosen steps to all the results derived from multiscopic units in order to get one robust object modelling using CGAL library [2]. It includes the following steps: outlier removal, simplification to reduce the number of input points, smoothing to reduce noise in the input data, normal estimation and orientation, and Poisson surface reconstruction method.

We compare the results on the virtual data set "Simons". The first column of figure 8 shows the reconstructed visual hulls. The reconstruction using CGAL lacks overall precision, especially in the ear areas. The reconstruction using PMVS shows better results near the ear areas, but strong surface deformations, specifically at the salient parts. Our reconstruction is visually better, with smoother surface reconstruction, and specifically good results in difficult, concave regions such as the ears.

5 CONCLUSION

This paper describes a new way of combining visual hull and multi-baseline stereovision in a fully automatic process. In section 3.3, we explained how to exploit information from the VH to guide the materiality map process in order to increase its reconstruction accuracy and robustness.It was demonstrated that the materiality map framework can integrate the VH guidance in a powerful way thanks to its scene-based structure.

Our contributions are a new algorithm for VH carving from stereovision surface coded as central disparity map, and a novel framework to merge multiple carved VH obtained from different multiscopic units. This process yields a topologically consistent volume, crucial for many applications. We demonstrated on experimental examples the algorithm results, the relevance of our disparity interpolation and smoothing methods, and the efficiency of the proposed inconsistency handling on both virtual and actual shootings.

Altogether, these contributions yield a qualitative and robust omnidirectional 3D reconstruction tool. The proposed solution proves the advantages of using both multiscopic and monoscopic cameras in a studio system as well as combining multi-baseline stereovision with visual hull approaches.

[1] http://www.di.ens.fr/pmvs/

[2] https://www.cgal.org/

ACKNOWLEDGMENTS

This work was funded by the RECOVER 3D project supported by the French National Fund (FSN) for a Digital Society, and the ANR ReVeRY national fund (ANR-17-CE23-0020). We would like to thank our partners XD Productions for providing the models captured using their camera system.

6 REFERENCES

[1] L. Lucas, P. Souchet, M. Ismael, O. Nocent, C. Loscos, L. Blache, S. Prévost, and Y. Remion. Recover3d: A hybrid multi-view system for 4d reconstruction of moving actors. In *4th international conference on 3D Body Scanning Technologies, Long Beach, United States*, pages 219–230, 11 2013.

[2] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proceedings of the 7th European Conference on Computer Vision-Part III*, ECCV '02, pages 82–96, London, UK, UK, May 2002. Springer-Verlag.

[3] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, April 2002.

[4] C.H. Chien and J.K. Aggarwal. Volume/surface octrees for the representation of three-dimensional objects. *CVGIP*, 36:100–113, October 1986.

[5] E.Steinbach, B.Girod, P.Eisert, and A.Betz. 3-d reconstruction of real-world objects using extended voxels. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 1, pages 569–572, September 2000.

[6] G. K. M. Cheung, T. Kanade, J. Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 2, pages 714–720, 2000.

[7] M. Ismael, S. Prévost, C. Loscos, and Y. Remion. Materiality maps: A novel scene-based framework for direct multi-view stereovision reconstruction. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 5467–5471, October 2014.

[8] S.M. Seitz and D.R. Charles. Photorealistic scene reconstruction by voxel coloring. *Int. J. Comput. Vision*, 35(2):151–173, November 1999.

[9] K.N. Kutulakos and S.M. Seitz. A theory of shape by space carving. *Int. J. Comput. Vision*, 38(3):199–218, July 2000.

[10] K. Matsuda and N. Ukita. Direct shape carving: Smooth 3D points and normals for surface reconstruction. *IEICE TRANSACTIONS on Information and Systems*, 2011.

[11] Ming Li, H. Schirmacher, M. Magnor, and H.-P. Siedel. Combining stereo and visual hull information for on-line reconstruction and rendering of dynamic scenes. In *Multimedia Signal Processing, 2002 IEEE Workshop on*, pages 9–12, December 2002.

[12] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Comput. Vis. Image Underst.*, 96(3):367–392, December 2004.

[13] A. Hilton and J. Starck. Multiple view reconstruction of people. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 357–364, September 2004.

 ISBN 978-80-86943-41-1

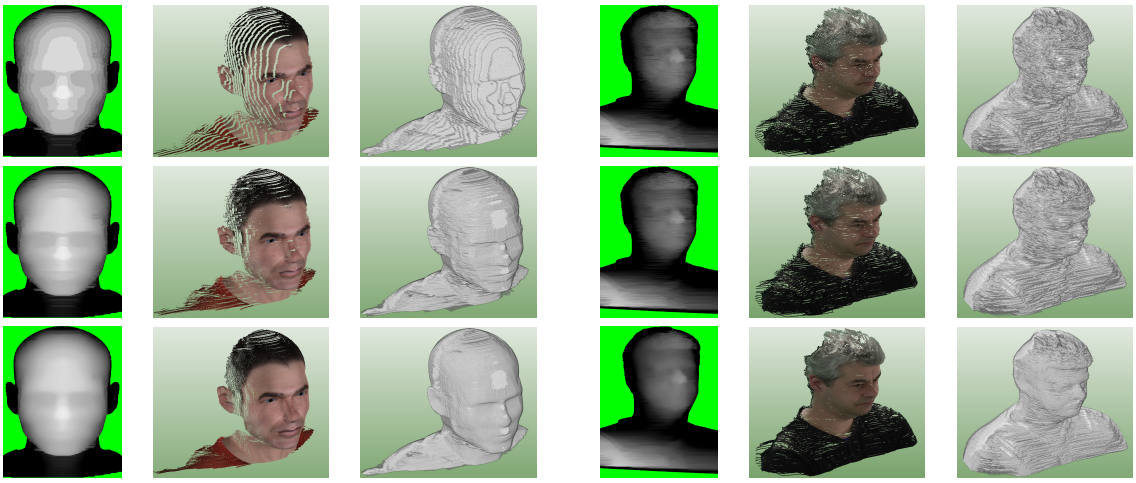


Figure 6: Results from one multiscopic unit (Left) for virtual Simon dataset, (Right) for real actor "Philippe". From top to bottom, results with: initial integer valued disparity; interpolated disparities according to 3.4.3; disparities smoothed by bilateral filtering described in 3.4.4. On each row, from left to right: disparity map, point cloud, and carved volume

Figure 7: Results of the entire pipeline. First row: several views of the point cloud and carved volume obtained from VH and four multiscopic units for virtual actor "Simon". Second row: several views of the global point cloud obtained for real actor "Jacques" from final volume resulting from VH and three multiscopic units. It corresponds to the union of the projection, per multiscopic unit, of the initial point cloud on the final volume.

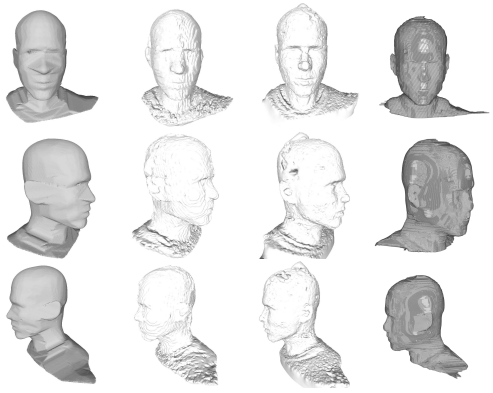


Figure 8: First column: visual hull of the ground truth virtual model "Simons". Second column: results from CGAL. Third column: results from PMVS. Fourth column: results from our framework.

[14] P. Song, X. Wu, and M. Wang. Volumetric stereo and silhouette fusion for image-based modeling. *The Visual Computer*, 26(12):1435–1450, December 2010.

[15] K.S. Narayan, J. Sha, A. Singh, and P. Abbeel. Range sensor and silhouette fusion for high-quality 3D scanning. *IEEE International Conference on Robotics and Automation, ICRA*, pages 3617–3624, 2015.

[16] R. A. Newcombe., S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. *Proceedings of 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.

[17] Y. Yang, A. Yuille, and J. Lu. Local, global, and multilevel stereo matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '93, pages 274–279. IEEE, June 1993.

[18] R. Szeliski and P. Golland. Stereo matching with transparency and matting. *Int. J. Comput. Vision*, 32(1):45–61, August 1999.

[19] C. Niquin, S. Prévost, and Y. Remion. An occlusion approach with consistency constraint for multiscopic depth extraction. *Int. J. Digital Multimedia Broadcasting*, 2010.

[20] C. Niquin. *Reconstruction du relief et mixage réel virtuel par caméras relief multi-points de vues*. Doctorate thesis, University of Reims Champagne-Ardenne, March 2011.

[21] T. Svoboda, D. Martinec, and T. Pajdla. A convenient multicamera self-calibration for virtual environments. *Presence*, 14(4):407–422, Aug 2005.

ISBN 978-80-86943-41-1