Face Reconstruction with structured light

John Congote[1,2], Iñigo Barandiaran[1], Javier Barandiaran[1], Marcos Nieto[1]
[1] *Vicomtech Research Center, Donostia - San Sebastian, Spain*
{*jcongote, ibarandiaran, jbarandiaran, mnieto*}*@vicomtech.org*
Oscar Ruiz[2]
[2] *CAD CAM CAE Laboratory, EAFIT University, Medellín, Colombia*
oruiz@eafit.edu.co

December 12, 2010

Abstract

This article presents a methodology for reconstruction of 3D faces which is based on stereoscopic images of the scene using active and passive surface reconstruction. A sequence of gray patterns is generated, which are projected onto the scene and their projection recorded by a pair of stereo cameras. The images are rectified to make coincident their epipolar planes and so to generate a stereo map of the scene. An algorithm for stereo matching is applied, whose result is a bijective mapping between subsets of the pixels of the images. A particular connected subset of the images (e.g. the face) is selected by a segmentation algorithm. The stereo mapping is applied to such a subset and enables the triangulation of the two image readings therefore rendering the (x,y,z) points of the face, which in turn allow the reconstruction of the triangular mesh of the face. Since the surface might have holes, bilateral filters are applied to have the holes filled. The algorithms are tested in real conditions and we evaluate their performance with virtual datasets. Our results show a good reconstruction of the faces and an improvement of the results of passive systems.

Keywords

3D reconstruction, structured light, Gray codes, depthmap

1 INTRODUCTION

1.1 Mathematical Context

In general, surface reconstruction from optical samples requires a function G relating pixels in an image of the scene $A \times B$ ($A, B \subset \mathbb{N}$) with points $p \in \mathbb{R}^3$. This function,

$G : A \times B \to \mathbb{R}^3$, is an injection since the image only records the visible part of the scene. G is not an onto function, as there are many points $p \in R^3$ for which there is no pixel $(i, j) \in A \times B$ in the image that records them.

Once this geometry function G is known, it is relatively simple to build a triangular mesh of the portion of the object visible in the image. Under a threshold of geometrical proximity, $G(i, j), G(i+i, j), G(i+1, j+1)$ may be considered the vertices of a triangular facet of the sought surface M. Moreover, the triangles being natural neighbors to triangle $t = [G(i, j), G(i+i, j), G(i+1, j+1)]$ are the ones involving pixels $(i, j+1), (i+2, j+1), (i, j-1)$, again, under thresholds of geometrical proximity. Stitching the different M triangular meshes originated in different views of the scene is known as zippering, and is not in the scope of our article. Literature on the topic might be found in [GT94], [MGC$^+$10] and [SOS04].

The discussion in this article involves two images, which may be labeled, without losing generality, as *right* and *left*, I_R and I_L. Simplifying the discussion, a color image is a mapping $I : A \times B \to [0, 255]^3$. For example, $I(i, j) = (143, 23, 112)$ means that the color registered in the pixel (i, j) of I corresponds to Red=143, Green=23 and Blue=112. A grey scale image has the form $I(i, j) = (k, k, k)$ due to the fact that in it the Red, Green and Blue graduations are identical ($k \in [0, 255]$).

Let S_L and S_R be the coordinate systems associated to images Left and Right, respectively. In the general configuration of the set-up, S_L and S_R such that (1) the Z axis of the coordinate system is normal to the capture plane of the image, and (2) the two cameras point to a common point $p \in \mathbb{R}^3$. In this article we assume that the images are *rectified*. This means, both of them have been rotated inside their own $X - Y$ plane (i.e. rotation around the Z axis of the image) in such a manner that the epipolar plane of the set-up is seen as the plane $y = E_e$ in both images. That means, the epipolar plane is seen as the same horizontal line in both images. We call the rectified images I_R and I_L and their rectified and their rectified coordinate systems S_L and S_R, respectively.

Let us consider a point $p \in \mathbb{R}^3$ recorded in both images I_R and I_L. Because the previous assumptions we have that $G_L(i, j) = p$ and $G_R(i, k) = p$. This means, the point p appears *in the same row i* of pixels in both images. The value $|k - j|$ is an offset that only occurs in the same pixel row of both images. Since we know that pixels (i, j) in image I_L and $(i, k) = p$ in image I_R record the same point $p \in \mathbb{R}^3$, the point p can be recovered by a usual triangulation procedure.

1.2 Informal Context

Human face reconstruction is a common problem in computer vision and computer graphics [SL09], where one possible objective is the generation and animation of a virtual model of it. The face is one of the most important identification regions of the human body, presenting commensurate technical challenges [ZCPR03]. A correct reconstruction of human faces is a precondition to both augmented reality and face recognition.

OJO: HERE, ONE MUST SAY WHY IS FACE RECONSTRUCION DIFFERENT FROM GENERAL SURFACE RECONSTRUCION.

3D surface reconstruction may be achieved by both passive and active methods. Passive ones do not change the environment in the process of reconstruction. Even

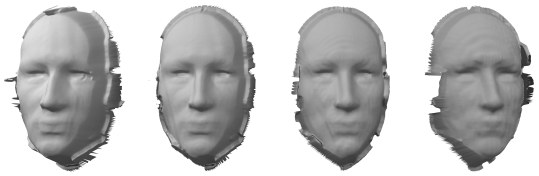

Figure 1: Results of the algorithm with the virtual dataset. Smooth surfaces are obtained with wider baselines.

thought passive methods obtain very good results, their setups are very expensive because they required a very high resolution required for obtaining reasonable results[BBB$^+$10].

Active systems obtain higher accuracy using off-the shelf components (OJO: WHO SAYS SO?). They modify or adapt the environment during the capture process, for example by (OJO: SAY WHICH ACTIONS TYPICALLY MAKE A SYSTEM TO BE 'ACTIVE'). Our active system uses the projection of a light pattern (i.e *structured light*), which is widely used for face surface reconstruction. In structured light systems any change on the setup requires new algorithms for face (surface) reconstruction.

The 3D surface reconstruction system implemented and informed in this article is part of a system used for full body reconstruction with visual hull algorithm [HP10]. Our setup applied to a face-body model produces a (OJO: IS IT A TRIANGULAR MESH ?) triangular mesh with high detail in the face region and low detail in the rest of the body. The reason for this differential resolution is that, while for the face region one requires high frequency details (e.g. texture of the skin), for the rest of the body such details are not required in our applications.

This article presents a system for face reconstruction which articulates non-proprietary hardware and our own software to obtain geometrical information from two images (possibly originated in 3D video - conference set ups). Our system also recovers the 3D geometry form the body region, although intentionally using lower resolution for neighborhoods other than the face.

This paper, Section 2 reviews previous works in face reconstruction. Section 3 presents the methodology implemented, including generation of the light patterns, capture, segmentation and reconstruction. Section 4 discusses the hardware set-up for the experiment and its configuration. Section 5 presents the results of the 3D surface reconstruction set-up and algorithms, and evaluates the reconstructed models against with real data. Section 6 concludes the work and proposes the future actions in this domain.

2 RELATED WORK

Face reconstruction is a widely studied topic. [PL05] presents a tutorial on face reconstruction, describing different problems and approaches from an artistic point of view, looking for a correct representation of the face and its expressions in multi- media. [SL09] presents a survey of 3D face reconstruction methods, classifying them in three different categories: *single image*, *stereo images* and *videos*.

Passive systems are commonly used for face reconstruction. One of the advantages of these systems is their non interaction with the environment, allowing to capture the geometry without interfering with other systems (OJO: what does this mean ?). [OTRT05] uses a system with four calibrated cameras applying a multi - view algorithm. A stochastic model is generated for the identification of the geometry, by minimizing a cost function. [LLWD05] compares different stereo algorithms for face reconstruction, and proposes an appropiate geometrical configuration of cameras to obtain accurate results. [ARL+09] presents a complex setup to a high resolution face reconstruction system. The methodology is based on an iterative reconstruction of the face by incrementing the size of the image and the number of stereo pairs used in each step. [BBB+10] extends the approach proposed in [ARL+09] by adding a postprocessing step that modifies the geometry of the face by using a texture, assuming that small dark regions of the face represent small hollows. This approach obtaining a more rich geometry.

Structured light for 3D reconstruction have been study for several years. The information obtained with this kind of systems is already being used as ground truth data for the evaluation of pasive reconstruction systems, such as stereo algorithms [SS03]. An extensive survey of structured light 3D reconstruction approaches can be found in [SFPL10], where a classification of different coding techniques are presented and evaluated. They identify the best coding approach for each one of the possible scenario configurations.

Real time capture of facial expresions is also an important feature in some systems. Several problems have to be addresed to accomplish this objective. One of these problems is the difficulty of projecting several patterns for a reconstruction , so other patterns have been employed which uses color models as [TFMS05] which allows a denser codification of the pattern, also single frame reconstruction with 2D coding is possible[CLZ08]. Another problem is hardware calibration to obtain several frames per second with a correct synchronization process between the projector and the cameras. An accepted synchronization approach can be found in [ZRY06]. Finaly, for a correct pattern codification of time variant patterns, a motion compensation should be implemented. This issue is especially critical for face reconstruction systems, where the person being reconstructed could move in a involuntary way,during adquisition [WLG07]

3 METHODOLOGY

Our algorithm of face reconstruction uses a set of stereo images captured at the same moment when a pattern is projected into the face. The images are captured in a setup previously calibrated. We assume that the object does not move between the different captures and the face is assumed to be a smooth surface without hair or beard and without highlight reflections. The result is a mesh of triangles correctly positionated in the space which represent the face region.

3.1 Stereo Calibration

Stereo calibration refers to the task of finding the relative pose between the cameras of a stereo pair. The objective is to feed subsequent stereo rectification processes that align the images such that the epipolar lines are horizontal and thus matching algorithms for 3D reconstruction can be implemented as one-dimensional searches.

Typically, stereo calibration is carried out by means of finding a number of point-correspondences between the images of the pair and retrieving the fundamental matrix. Let $\mathbf{x}$ be the image of a 3D point in the left image, and $\mathbf{x}'$ the image of the same point in the right image. The fundamental matrix restricts the position of $\mathbf{x}'$ to the epipolar line associated to $\mathbf{x}$, such that $\mathbf{x}'^{\top} F\mathbf{x} = 0$. It has been shown [HZ04], that the knowledge of the fundamental matrix can be used to retrieve the projection matrices of the two cameras of the pair, up to a projective ambiguity that can be solved with known restrictions of the camera.

Besides, images captured by real cameras show some tangential and radial distortion, which can be corrected applying the following functions:

$$u = p_x + (u - px)(1 + k_1 r + k_2 r^2 + k_3 r^3 + \ldots)$$
$$v = p_y + (v - py)(1 + k_1 r + k_2 r^2 + k_3 r^3 + \ldots)$$

where $r^2 = (u - p_x)^2 + (v - p_y)^2$ and $k_1, k_2, k_3, \ldots$ are the coefficients of the Taylor expansion of an arbitrary radial displacement function $L(r)$.

Parameter identification of the camera stereo pair is extracted from the calibration information of the full body reconstruction setup; which is further explained in [RVG08]. For our purposes we select the camera pair which are focused to the face region of the body, and we follow a 3D stereo reconstruction process with them.

3.2 PATTERN GENERATION

Pattern generation refers to the task of creating of a set of synthetic binary images to be projected as structured ligth in the scene. The objetive is to identify the coordinates of the projected pattern in the image scene and thus allowing a poing matching algorithm to became independient of the color in the captured scene.

The used patterns are represented as a matrix of boolean values. Let P be a matrix of M columns and N rows thus $P = \{P_{m,n} \in \{0,1\}\}$ with $0 < m < M$ and $0 < n < N$. Let C be a matrix of the same dimensions of P thus $C = \{C_{m,n} \in (0,M) \subseteq \mathbb{N}\}$. The restriction of the number of values in the matrix C is the same that the number of columns allows the correct identification of the column in the images. Let g be a function such as $g : \mathbb{N} \to \mathbb{N}$ which is bijective and transforms the numbers from binary representation to Gray representation as described in algorithm 1, the inverse Gray function g^{-1} is described in algorithm 2. The number of images to be projected depends of the number of columns of the matrix C, so $nPat = \lceil \log_2 M \rceil$

The $nPat$ patterns represented by the matrix P are generated as follows:

$$P^i_{j,k} = g(j) \bullet 2^i$$

Input: bin
Output: $gray$
return $bin^{(bin/2)}$

Algorithm 1: Gray function to convert from binary to gray code

Input: $gray$
Output: $bin, nPat$
$ish, ans, idiv \in \mathbb{N}$
$ish \leftarrow 1$
$ans \leftarrow gray$
while 1 **do**
  $idiv \leftarrow \frac{ans}{ish}$
  $ans \leftarrow ans \oplus idiv$
  if $idiv \leq 1 \vee ish = 32$ **then**
    return ans
  end
  $ish \leftarrow ish \times 2$
end

Algorithm 2: Gray function to convert from Gray code to binary

where $0 < i < nPat$ represent the number of the pattern, j,k the coordinates in the matrix P. The pattern structure can be depicted as a sequence of columns as can be visualized in the figure 2. The nature of this kind of patterns is 1D because the calibration setup already give us an epipolar constrain of the images. Therefore it is not neccesary to use of 2D patterns in this case.

3.3 PATTERN RECOGNITION

Pattern recognition refers to the task of creating a pair of images which maps the position of the projected patterns P in the set of stereo pair of images. The objetive is the identification of the projected pattern in the set of images, and calculate the value of the matrix C for each pixel. This matrix allow the point matching algorithm to became unambiguos since each point in the maps is labeled uniquely in each epipolar line.

Let $s = L, R$ be matrices of W columns and H rows, thus $L = \{L_{w,h} \in (0,255)\}$ with $0 < w < W$ and $0 < h < H$. The matrices L and R represent the information of the stereo pair cameras in grayscale. Let $O = \{O_{w,h} \in (0,M)\}$ be the decode maps OL, OR. Let $t : (0,255) \rightarrow \{0,1\}$ be a threshold function which binarize the images L and R. The threshold value could be calculated with the Otsu algorithm as explained in [KBW+09] or by means of calculating the alvedo of the images with the process described in [SS03] where each pattern is projected two times, each one with the original version and their negative.

$$O^s_{m,n} = g^{-1}\left(\bigvee_i t\left(s^i_{m,n}\right) \cdot 2^i\right)$$

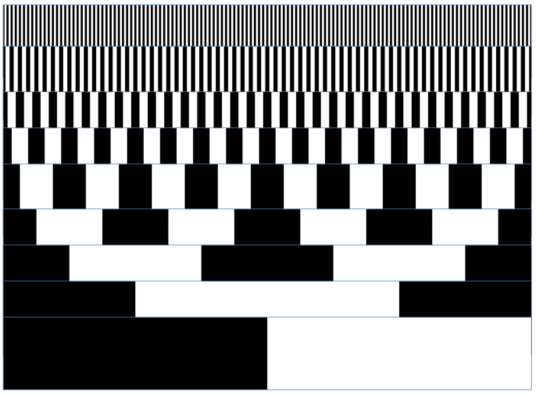

Figure 2: Gray code

As shown in figure 3 the set of *L* and *R* images are binarized. Stereo reconstruction images are combined, and a unique map *OL* and *OR* is processed. The maps should be rectified as shown in figure 4: this rectification process is possible because the camera information is already known as explained in section 3.1.

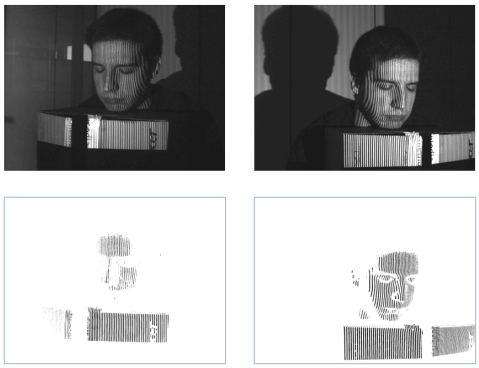

Figure 3: Stereo images captured from the cameras and their result of the threshold function

3.4 STEREO RECONSTRUCTION

Stereo reconstruction task calculates the point correspondence between two images. The objective is to calculate the 3D point coordinates of each pixel in the stereo images. The previous steps of the algorithm such as the stereo calibration assures the epipolar constrain. Pattern generation process gives the color independence of the images. Pattern recognition maps the set of images into a map without ambiguities. For dense depthmap calculation we used a simple box filter algorithm, based on sum of

Figure 4: *O* maps of the images, the gray value in each pixel represents the value of the *C* matrix in the stereo images and the rectified version

square diferences *SSD* and a Winner Takes All *WTA* for pixel selection [SS02].

Let D be a matrix of W columns and H rows, thus $D = \{D_{w,h} \in \mathbb{N}\}$. where D describes the pixel difference between the same point in the image L and R. The identification of the correct disparity the following function is applied.

$$D_{m,n} = minarg_l \left(SSD(OL, OR, m, n, l) \right)$$

$$SSD(L,R,m,n,l) = \sum_{e=m-b}^{m+b} \sum_{f=n-b}^{n+b} \left(L_{e,f} - R_{e,f+l} \right)^2$$

The figure 5 shows the result of the disparity maps *DL* and *DR*. The identification of wrong matched points is carried out by applying proccess such as *cross cheking* and *joint bilateral filter*. It is assumed that the biggest connected region represents the face, then a max blob algorithm is applied to filter regions outside the face. The mesh is generated by joining adjacent pixels in the image with their 3D coordinates. The topology of the mesh is correct since the difference between the coordinates of adjacent pixels are small.

4 SETUP CONFIGURATION

As our algorithm is part of a bigger chain of process where a full body reconstruction process is done. We tried to mantain the same setup for our algorithm, even we tried to use a passive system for face reconstruction, the resolution obtained where insuficient to fulfill our needs. Then, we put a stereo camera setup with a projector in the middle of the cameras. We identified that an small distance between the cameras does not give enough information for recovering 3D positions accurately. In opposite, a wide baseline between cameras generates occlusion regions, althought projection information is used for hole filling in post processing step.

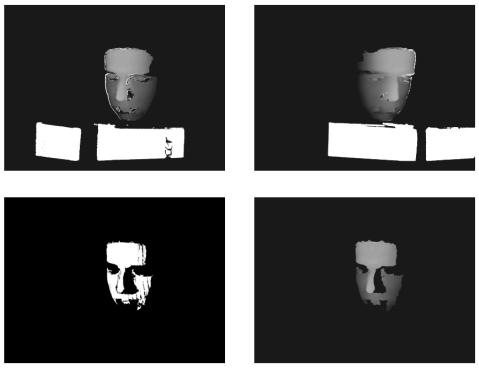

Figure 5: Stereo reconstruction of the face, the disparity images for the L and R, the maximum connected region and final depthmap result.

Figure 6: 3D face reconstruction, dark regions and occluded ones in the original images shows problems in the generated mesh

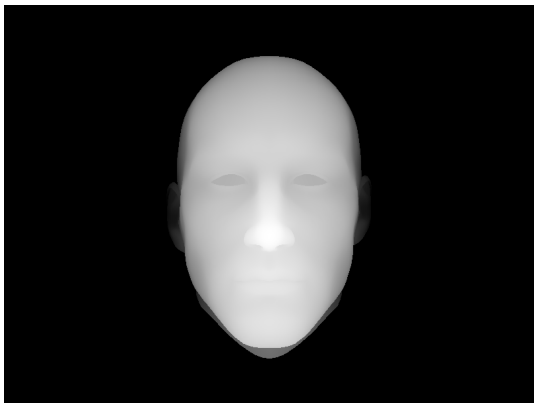

Figure 7: Ground truth depthmap image of a face

The cameras used where tested with different resolutions, 1024x768, 1240x960 and 1600x1200. Finally, the resolution was set to 1240x960. Also, the projector were defined at a resolution of 800x600 because an increase of resolution generates very high frecuency patterns, that are very difficutl to identify accurately at that resolution. We found that a minimum width of 4 pixels for each column of the pattern is necesary for a correct identification.

Different kind of binary patterns were used. Gray pattern generates the best results. Using binary or golay patterns shows in some aspects imposible to generate a workable results. In this way, we didn't consider these methods for the final version, and used only Gray codes. The binarization of the images present one of the biggest problems of the structured ligth setup. We used the Otsu method but it exhibits some problems, such as high sensitivity to areas of specular reflection. We finally choose the projection of the negative images with good results and a threshold t with a value of the half of the range of the gray image.

5 RESULTS

The evaluation of the algorithm presents several problems since the groundtruth information it is not available. However, we implemented a virtual environment which resembles a real setup. This approach allowed us to test our method and validate our results. We use Blender software for the generation of the setup and the identification of the groundtruth data, as shown in figure 8. The groundtruth was defined as a normalized depthmap with values between 0 and 255 using all the posible values in the image format as shown in the figure 7.

Different camera positions were tested in our virtual setup, for the identification of the best baseline distance. The results obtained with the algorithm are shown in the figure 8. and the position of the cameras are shown in the figure 9.

The groundtruth information and the results of the algorithm show a difference in scale, but not in position. We measured the difference of the results and the groundtruth

Figure 8: Result from the different cameras in the virtual setup

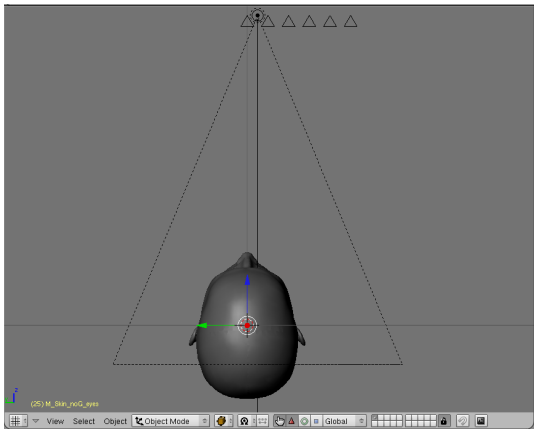

Figure 9: Camera positions for virtual framework

with a image correlation algorithm. The correlation gives us a value between the range of 0 and 1 where 0 is a bad results and 1 is the groundtruth. The table 1 presents the correlation values obtained for different camera position, and the figure 1 shows the 3D mesh generated.

6 CONCLUSIONS AND FUTURE WORK

We present a methodology for face reconstruction in a mixed enviroment of active-pasive setup. Structured light shows a quality improvement against the results obtained with pasive setups. Time multiplexing codification has the problem of motion between the captured images generating a waving efect in the reconstructions. Even robust algorithms of point matching for dense depthmaps were tested there were no real improvement in the results. We will try with color or 2D patterns which only requieres

Baseline distance	Value
1	0.904372
2	0.938449
3	0.958089
4	0.974051

Table 1: Correlation Results

one exposition that present a better aproach for the reconstruction of faces since the motion problem is not present.

7 ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish administration agency CDTI, under project CENIT-VISION 2007-1007. CAD/CAM/CAE Laboratory at EAFIT University and the Colombian Council for Science and Technology – COLCIENCIAS –.

References

[ARL+09] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*, SIGGRAPH '09, pages 12:1–12:15, New York, NY, USA, 2009. ACM.

[BBB+10] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 29(3), 2010.

[CLZ08] SY Chen, YF Li, and J. Zhang. Vision processing for realtime 3-D data acquisition based on coded structured light. *Image Processing, IEEE Transactions on*, 17(2):167–176, 2008.

[GT94] Marc Levoy Greg Turk. Zippered polygon meshes from range images. In *ACM SIGGRAPH. Computer Graphics Proceedings, Annual Conference Series*, pages 311–318, July 1994.

[HP10] Gloria Haro and Montse Pardís. Shape from incomplete silhouettes based on the reprojection error. *Image Vision Comput.*, 28:1354–1368, September 2010.

[HZ04] R. I. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2004.

[KBW+09] Petra Kramer, Fernando Boto, Diana Wald, Fabien Bessy, Celine Paloc, Carles Callol, A. Letamendia, Izaskun Ibarbia, O. Holgado, and J.M. Virto. Comparison of segmentation algorithms for the zebrafish heart in fluorescent microscopy images. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Yoshinori Kuno, Junxian Wang, Renato Pajarola, Peter Lindstrom, Andre Hinkenjann, Miguel L. Encarnacao, Claudio T. Silva, and Daniel Coming, editors, *Advances in Visual Computing*, Lecture Notes in Computer Science (LNCS), pages 1041–1050, Las Vegas, Nevada, USA, December 2009. Springer.

[LLWD05] Ph. Leclercq, J. Liu, A. Woodward, and P. Delmas. Which stereo matching algorithm for accurate 3d face creation. In Reinhard Klette and Jovisa Zunic, editors, *Combinatorial Image Analysis*, volume 3322 of *Lecture Notes in Computer Science*, pages 690–704. Springer Berlin - Heidelberg, 2005. 10.1007/978-3-540-30503-3_53.

[MGC+10] Stefano Marras, Fabio Ganovelli, Paolo Cignoni, Riccardo Scateni, and Roberto Scopigno. Controlled and adaptive mesh zippering. In *GRAPP - International Conference in Computer Graphics Theory and Applications*, 2010.

[OTRT05] D. Onofrio, S. Tubaro, A. Rama, and F. Tarres. 3D Face Reconstruction with a four camera acquisition system. In *Int'l Workshop on Very Low Bit-Rate Video Coding*, 2005.

[PL05] Frédéric Pighin and J. P. Lewis. Introduction. In *ACM SIGGRAPH 2005 Courses*, SIGGRAPH '05, New York, NY, USA, 2005. ACM.

[RVG08] José I. Ronda, Antonio Valdés, and Guillermo Gallego. Line geometry and camera autocalibration. *J. Math. Imaging Vis.*, 32:193–214, October 2008.

[SFPL10] Joaquim Salvi, Sergio Fernandez, Tomislav Pribanic, and Xavier Llado. A state of the art in structured light patterns for surface profilometry. *Pattern Recognition*, 43(8):2666 – 2680, 2010.

[SL09] G. Stylianou and A. Lanitis. Image based 3d face reconstruction: A survey. *International Journal of Image and Graphics*, 9(2):217–250, 2009.

[SOS04] Chen Shen, James O'brien, and Jonathan Shewchuk. Interpolating and approximating implicit surfaces from polygon soup. In *ACM Transactions on Graphics*, pages 896–904. ACM Press, 2004.

[SS02] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, 2002.

[SS03] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–195 – I–202 vol.1, june 2003.

[TFMS05] Filareti Tsalakanidou, Frank Forster, Sotiris Malassiotis, and Michael G. Strintzis. Real-time acquisition of depth and color images using structured light and its application to 3d face recognition. *Real-Time Imaging*, 11(5-6):358 – 369, 2005. Special Issue on Multi-Dimensional Image Processing.

[WLG07] T. Weise, B. Leibe, and L. Van Gool. Fast 3d scanning with automatic motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, June 2007.

[ZCPR03] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35:399–458, December 2003.

[ZRY06] Song Zhang, Dale Royer, and Shing-Tung Yau. High-resolution, real-time-geometry video acquisition. In *ACM SIGGRAPH 2006 Sketches*, SIGGRAPH '06, New York, NY, USA, 2006. ACM.