Proceedings of the IASTED International Conference
Robotics (Robo 2011)
November 7 - 9, 2011 Pittsburgh, USA

INCREMENTAL 3D BODY RECONSTRUCTION FRAMEWORK FOR ROBOTIC TELEPRESENCE APPLICATIONS

Luis Almeida
Institute of Systems and Robotics
University of Coimbra
IPT, Tomar, Portugal
email: laa@ipt.pt

Paulo Menezes
Institute of Systems and Robotics
University of Coimbra
Coimbra, Portugal
email: paulo@isr.uc.pt

Lakmal D. Seneviratne (*)
Division of Engineering
King's College
London, UK
email:
lakmal.seneviratne@kcl.ac.uk

Jorge Dias (*)
Institute of Systems and Robotics
University of Coimbra
Coimbra, Portugal
email: jorge@deec.uc.pt

ABSTRACT
This research proposes an on-line incremental 3D reconstruction framework that can be used on telepresence robots or human robot interaction (HRI). The aim is a low cost solution that enables users to communicate and interact remotely experiencing the benefits of a face-to-face meeting. By exploring computers graphics techniques and spatial audio we intend to induce sensations of being physical in the presence of other people. There is a wide variety of research opportunities including high performance imaging, multi-view video, virtual view synthesis, etc. One fundamental challenge in geometry reconstruction from traditional cameras array is the lack of accuracy in low-texture or repeated pattern region. Our approach explores virtual view synthesis through motion body estimation and hybrid sensors composed by video cameras and a depth camera based on structured-light or time-of-flight. We present a full 3D body reconstruction system that combines visual features and shape-based alignment. Modeling is based on meshes computed from dense depth maps in order lower the data to be processed and create a 3D mesh representation that is independent of view-point.

KEY WORDS
3D reconstruction, augmented reality, human robot interaction, tele-presence; virtual view synthesis.

1 Introduction

This work presents an on-line incremental 3D reconstruction framework that can be used on telepresence robots, human robot interaction (HRI) or augmented reality (AR) applications. The project, based on recent low cost depth sensors, intends to create a domestic easy to install 3D acquisition and display system that enables socialization, collaboration and entertainment. Exploring computers graphics techniques, spatial audio and artificial vision enables us to induce sensations of being physical in the presence of other people useful on several domains like elderly loneness minimization problem[20], tele-rehabilitation[18][32], companion robot, education, socialization, 3DTV, entertainment, etc.

Phones and internet chat/audio/video conferencing programs (ex: VOIP, NetMeeting, Skype) have been used for socialization, nevertheless they are not able to create the remote person presence feeling. Means of communications that enable eye contact, gestures reconnaissance, body language and facial expressions are required.

Figure 1. The concept goal: An mobile telepresence robot (Hilario) carrying an auto stereoscopic 3D display, video cameras, depth sensor, microphones and speakers enables users to communicate and interact remotely experiencing the benefits of a face-to-face meeting in full size.

The concept goal is depicted on figure 1. An mobile telepresence robot carrying an auto stereoscopic display, video cameras, depth sensor, microphones and speakers enables users to communicate and interact remotely experiencing the benefits of a face-to-face meeting in full size. On recent years, there has been significant advances on 3D displays not requiring special glasses, mostly due the game console industry and 3DTV. Flat-panel auto stereoscopic

[1]The authors gratefully acknowledge support from Institute of Systems and Robotics at University of Coimbra (ISR-UC). (*) on sabbatical leave at Khalifa University of Science, Technology and Research (KUSTAR), Abu Dhabi, UAE

solutions employing lenticular lenses or parallax barriers are common technologies nowadays, although they still constrain the user to certain point of views. By moving the robot in front of a person we can easily compensate wrong user point of views avoiding loosing the stereoscopic perception. Even with an accurately location track of viewer's head and rendering view dependent images on common screens (ex: TV's, LCD's) is possible to create the illusion of a real window. Our incremental on-line 3D human reconstruction solution should provide models easily rendered on any of those referred display technologies.

Augmented reality and particularly tele-immersion [16][5][21] can provide the technology means that enables users interact remotely and experience the benefits of a face-to-face meeting. The tele-immersive technology "combines virtual reality for rendering and display purpose, computer vision for image capturing and 3D reconstruction, and various networking techniques for transmitting data between remote sites in real-time with minimal delay" [22].

In order to aim an incremental on-line 3D human reconstruction solution useful for shared mixed reality workspace [19][30][18], we estimate the 3D world information using 2D image sequences and depth information using a depth camera, e.g. a time of flight camera (ToF) or structured light camera. This hybrid approach addresses the geometry reconstruction challenge from traditional cameras array, that is the lack of accuracy in low-texture or repeated pattern regions. The proposed real-time 3D full reconstruction system combines visual features and shape-based alignment. By detecting image point features for which tri-dimensional coordinates can be measured, a correspondence between 3D and 2D is established. Using those annotated 3D points, between consecutive point clouds, it is possible to estimate the motion transformation through a linear, closed form or iterative method, register them on one same referential and create a global model. Correspondence between consecutive image features in images is performed using SURF method [6]. Virtual view synthesis and modeling is based on 3D mesh from dense depth maps in order lower the data to be processed and to create a 3D mesh representation that is independent of view-point [9].

Mesh simplification are conducted reducing the number of vertices's and facets while keeping important object features or interest points in the model. The aim is to continuously generate a realistic body model, transfer the model and reconstruct on a remote common display or virtual environment according each users viewpoint by a tracking process. Figure 2 presents an overview of the algorithm.

The reminder of this paper is organized as follows. First a related work is presented on the following subsection. Section 2 describes the suggested methodology and section 3 present implementation and experimental results. Finally, section 4 presents the future work and conclusions.

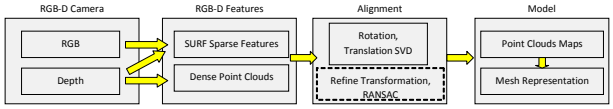


Figure 2. Algorithm overview. The proposed real-time 3D full reconstruction system combines visual features and shape-based alignment between consecutive point clouds. The model representation is updated incrementally

1.1 Background

Presently the field of robotic telepresence platforms is very active. Recently major groups on area directed by Professor Nadia Magnenat-Thalmann from Nanyang Technological University, Professor Markus Gross from Swiss Federal Institute of Technology (ETH) Zurich and Professor Henry Fuchs from University of North Carolina at Chapel Hill have created the *BeingThere Centre: International Research Centre for Telepresence* to pursue technological and systems-level advances leading to a credible, 3D experience of telepresence in a fitted room using a mobile platform. Efforts in achieving a semi-transparent autostereoscopic 3D displays that bring the illusion of the other person being present into a room, a laboratory or a hospital will soon be a reality. Requirements for future telepresence systems includes:

Seamless integration into everyday environments;

Autostereoscopic display and multiview holography;

Full gaze contact and perspective corrections;

Scalable and mobile;

Combine display, capture & robotics.

Market robotic telepresence platforms includes bot names like QB from AnyBots, TiLR from RoboDynamics, Jazz Connect from Gostai, Mantaro bot from Mantaro and VGo from VGo. In recent years, there has been a significant effort focusing immersive video conferencing and immersive television challenging research areas and consumers product industries. 3D cinema, 3D console games, 3D contends, 3D broadcast or 3DTV LCD displays are common technologies nowadays. The display technology, as a key component, is now able to recreate the stereoscopic perception of 3D depth for the viewer either using light active shutter glasses, passive polarized glasses or even without glasses, using flat-panel autostereoscopic solutions employing lenticular lenses or parallax barriers.

Notable works that realistically represent the user's appearance at tele-immersion lab at UC Berkeley [19] or at GrImage lab at Inria [31] are using traditional video cameras arrays to perform real-time full body 3D reconstructions. F. Isgro, Emanuele Trucco, Peter Kauff and Oliver Schreer present a good survey paper, titling *Three-Dimensional Image Processing in the Future of Immersive Media* [15], where they discuss "the three-dimensional image processing challenges posed by present and future immersive telecommunications, especially immersive video

conferencing and television". European funded projects [11] like VIRTUE, 3DTV, 3D4YOU, 2020-3D-MEDIA, MOBILE 3DTV, 3D PHONE and 3D Presence demonstrate the interest on the area.

Virtual view synthesis and modeling are the potential graphic tools to create the eye to eye contact illusion on tele-presence communications[15] [8]. Usually the body surface is reconstructed by merging sensors data from different views. Two types of information are required: depth data and sensor pose data. When there is no prior information about depth and pose, the reconstruction techniques bases on structure from motion. On such cases, the sensor ego-motion estimation is based on corresponding features found in consecutive images. The depth information, without absolute scale, is then computed using the obtained ego-motion information. When depth information is available a priori, but sensor pose is still unknown, using data resulting from a ToF or structured light depth camera, a laser scanner or a stereo camera without inertial sensors, the reconstruction techniques usually bases on the Iterative Closest Point (ICP) algorithm [7]. 3D point clouds acquired from different views are registered onto one same referential by iteratively matching overlap surfaces. This method is computationally heavy for real time applications. When depth data and sensor pose data are known a priori, no registration procedure is required to merge the data onto a global referential. The precision of depth measurements and sensor pose estimation influences the final surface reconstruction quality. Recent depth sensor devices provide 3D measurements and also RGB data, enabling the use of 2D image algorithms. It is possible to improve the 2D feature mapping between consecutive RGB images, associating the respective depth data and creating a 3D feature tracking. 2D image features mapping approaches are generally based on Kanade-Lucas-Tomasi (KLT) method [34][25][35], Scale-Invariant Feature Transform (SIFT) method [24] or Speed Up Robust Features (SURF) method [6]. Several works use these techniques to track 3D pose sensor changes either for object detection, path planning, for gesture recognition or for reconstruction purposes [14][29][1][26][27]. Our work intends to perform a real-time incremental body modeling.

2 Methodology

Building 3D body models is an important task for robotics with applications in grasping, manipulations, semantic mapping and tele-presence. We propose a real-time full 3D reconstruction system that combines visual features and shape-based alignment using Xbox Kinect device. Alignment between successive frames is computed by jointly optimizing over both appearance and shape matching. Appearance-based alignment is done over 2D SURF features annotated with 3D position. Although SIFT descriptor present better accuracy, we have choosen SURF method in order to achieve the real-time characteristic. Shape-based alignment is performed using the motion transformation estimation between consecutive annotated 3D point clouds through a linear method. There are several possible closed form solutions for rigid body transformation [12]: SVD [3][10][12] or iterative methods like Random Sample Consensus (RANSAC) [13][1][17]. Once obtained a 3D point model a mesh is generated through Delaunay triangulation.

2.1 Registration

Considerer the motion of a rigid body in front of a scanner and the estimation of the rigid transformation (rotation and translation). This information is important to register the body points on one same referential and create a global model.

Suppose the existence of two corresponding 3D points sets $\{\mathbf{x_i^t}\}$ and $\{\mathbf{x_i^{t+1}}\}$, $i = 1..N$, from consecutive t and $t+1$ scans, related through the following equation:

$$\mathbf{x_i^{t+1}} = \mathbf{R}\mathbf{x_i^t} + \mathbf{T} + \mathbf{V_i} \tag{1}$$

$\mathbf{R}$ represents a standard 3x3 rotation matrix, $\mathbf{T}$ stands for a 3D translation vector and $\mathbf{V_i}$ is a noise vector. The optimal transformation [R, T] that maps the set $\{\mathbf{x_i^t}\}$ on to $\{\mathbf{x_i^{t+1}}\}$ can be obtained through the minimization of the following equation using a least square criterion:

$$\varepsilon^2 = \sum_{i=1}^{N} \left\| \mathbf{x_i^{t+1}} - \mathbf{R}\mathbf{x_i^t} - \mathbf{T} \right\|^2 \tag{2}$$

The least square solution is the optimal transformation only if a correct correspondence between 3D point sets is guaranteed. Otherwise other methods should be selected (e.g. RANSAC). The singular value decomposition (SVD) of a matrix can be used to minimize Eq. 2 and obtain the rotation (standard orthonormal 3x3 matrix) and the translation (3D vector) [3][10][12]. In order to calculate rotation first, the least square solution requires that $\{\mathbf{x_i^t}\}$ and $\{\mathbf{x_i^{t+1}}\}$ point sets share a common centroid. With this constraint a new of equation can be written using the following definitions:

$$\overline{\mathbf{x_i^t}} = \frac{1}{N}\sum_{i=0}^{n} \mathbf{x_i^t} \qquad \overline{\mathbf{x_i^{t+1}}} = \frac{1}{N}\sum_{i=0}^{n} \mathbf{x_i^{t+1}} \tag{3}$$

$$\mathbf{x_{ci}^t} = \mathbf{x_i^t} - \overline{\mathbf{x_i^t}} \qquad \mathbf{x_{ci}^{t+1}} = \mathbf{x_i^{t+1}} - \overline{\mathbf{x_i^{t+1}}} \tag{4}$$

$$\varepsilon^2 = \sum_{i=1}^{N} \left\| \mathbf{x_{ci}^{t+1}} - \mathbf{R}\mathbf{x_{ci}^t} \right\|^2 \tag{5}$$

Maximizing $Trace(\mathbf{R}\,\mathbf{H})$ enable us to minimize the generated equation 5, with $\mathbf{H}$ being a 3x3 correlation matrix defined by $\mathbf{H} = \mathbf{x_{ci}^{t+1}}(\mathbf{x_{ci}^t})^{\mathbf{T}}$. Considering that the singular value decomposition of $\mathbf{H}$ results on $\mathbf{H=UDV}^T$, then

the optimal rotation matrix, $\mathbf{R}$, that maximizes the referred trace is $\mathbf{R}$= $\mathbf{U}$ diag(1; 1; det($\mathbf{UV}^T$)) $\mathbf{V}^T$ [3][10][12]:

$$\mathbf{R} = \mathbf{U}\mathbf{V}^{\mathbf{T}} \tag{6}$$

The optimal translation that aligns $\{\mathbf{x_i^{t+1}}\}$ centroid with the rotated $\{\mathbf{x_i^t}\}$ centroid is

$$\mathbf{T} = \overline{\mathbf{x_i^{t+1}}} - \mathbf{R}\overline{\mathbf{x_i^t}} \tag{7}$$

2.2 Model Mapping

Suppose that the mapping from the world coordinates to one of the scans of the sequence, is known (ex: to scan 0) and it is represented by the transformation ${}^{\mathbf{0}}\mathbf{H_w}$. As described before, for any consecutive pair of scans (t, t+1) from tracked points it is possible to measure rotation and translation and combine them into a single homogeneous matrix 4x4, ${}^{\mathbf{t+1}}\mathbf{H_t}$, $\mathbf{H} = [\mathbf{R}, \mathbf{T}]$. Therefore it is possible to compute Eq. 8:

$${}^{\mathbf{i}}\mathbf{H_0} = {}^{\mathbf{i}}\mathbf{H_{i-1}}{}^{\mathbf{i-1}}\mathbf{H_{i-2}} \ldots .. {}^{\mathbf{1}}\mathbf{H_0} \quad and \quad {}^{\mathbf{i}}\mathbf{H_w} = {}^{\mathbf{i}}\mathbf{H_0}{}^{\mathbf{0}}\mathbf{H_w} \tag{8}$$

To update the reconstructed model, each acquired 3D point set is transformed to the world coordinate system using ${}^{\mathbf{i}}\mathbf{H_w}$. This alignment step adds a new scan to the dense 3D model. Alignment between successive frames is a good method for tracking the body position over moderate distances. However, errors in alignment between a particular pair of frames, and noise and quantization in depth values, cause the estimation of body pose to drift over time, leading to inaccuracies in the map. This is most noticeable when the body follows a long path, eventually returning to a location previously visited. The cumulative error in frame alignment results in a map that has two representations of the same region in different locations.

2.3 Tracking

The system first undistorts the images, and then the SURF features are detected and matched. These features are invariant to affine transformations, so they allow detection of the feature points from different angles and range. Although SURF provides good distinctive descriptors, undesirable matches can occur related with background static areas and image body boundaries. To overcome this situation it possible to define a working reconstruction space for the body and a mask for the SURF algorithm.

After finding the set of matched image features, a correspondence between 2D and 3D is set up. These annotated 3D points pairs are then used to estimate the motion between two time consecutive point clouds. Assuming that the identification problem has been solved, we must compute the rigid transformation (rotation and translation) that align the two consecutive 3D scans. The solution should take in account that the data are typically affected by noise: correspondences may be false, and some key data patches may be partially occluded.

Virtual View Synthesis: On a 3D video conference, the real eye contact is preserved while each participant observes the others from their current perspective. Each user viewpoint changes according his movements around the shared meeting environment. Therefore new perspectives views have to be presented at each time instant depending on the viewers pose in front of the display. This requires a precise estimation of the viewers pose in 3D space, which can be accomplish by and head/body tracking module [37][33][4]. The selected approach is based on a facial feature tracker using eye feature [36][23]. The purpose of used Haar-like features is to meet the real-time requirement. The resulting 2D position of the eyes can then be associated to 3D points for the calculation of the 3D position of the head.

Algorithm: The global model reconstruction algorithm can be described as follow on Algorithm 1:

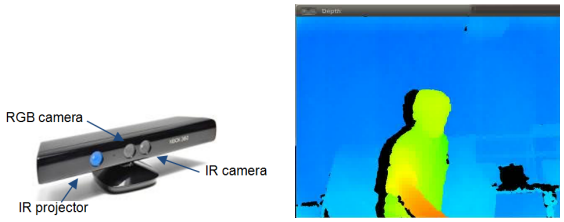
Algorithm 1 Model reconstruction algorithm: estimate the 3D world information using 2D image sequences and depth information using a depth camera. The proposed real-time 3D full reconstruction system combines visual features and shape-based alignment. By detecting image point features for which tri-dimensional coordinates can be measured, a correspondence between 3D and 2D is established. Using those annotated 3D points, between consecutive point clouds, we estimate the motion transformation through a closed form method, register them on one same referential and create a global model.

$R_g \leftarrow R_{init}; t_g \leftarrow t_{init}$
$f_1 \leftarrow undistort(adquire\_rgb\_image())$
$f_{1d} \leftarrow undistort(adquire\_depth\_image())$
$f_{1xyz} \leftarrow convert\_depth\_image\_to\_xyz\_data(f_{1d})$
$f_{1r} \leftarrow map\_rgbcolor\_to\_depth\_image(f_{1xyz}, f_1)$
for $(;;)$ **do**
  $f_2 \leftarrow undistort(adquire\_rgb\_image())$
  $f_{2d} \leftarrow undistort(adquire\_depth\_image())$
  $f_{2xyz} \leftarrow convert\_depth\_image\_to\_xyz\_data(f_{2d})$
  $f_{2r} \leftarrow map\_rgbcolor\_to\_depth\_image(f_{2xyz}, f_2)$

  $(surf_1, surf_2) \leftarrow detect\_SURF\_features(f_{1r}, f_{2r})$
  $matches2D \leftarrow SURF\_match(surf_1, surf_2)$
  $matches3D \leftarrow correspond2D3D(matches3D)$
  $(R, t) \leftarrow motion\_estimator(matches3D)$
  $(R_g, t_g) \leftarrow update\_global\_transformation(R, t)$
  $f_{1r} \leftarrow f_{2r}; f_{1xyz} \leftarrow f_{2xyz}$ {update_past_data}
  $model \leftarrow proj\_points\_to\_world\_coord(f_{2xyz}, R_g, t_g)$
  $mesh\_model\_generation$
end for

3 Implementation and Results

Novel depth sensors like PrimeSense camera or Xbox Kinect [28] can capture video images along with per-pixel depth information.



a) b)

Figure 3. a) Kinect Sensor b) Depth map with color representing the distance to sensor

To experimentally test the algorithm we register several 3D point clouds in order to create person model while he is rotating in front of Kinect device.

3.1 Calibrations

The Kinect device combines a regular RGB camera and a 3D scanner, consisting of an infrared (IR) projector and an IR camera as shown in figure 3a). The projector sends several thousand structured IR rays into the scene which are reflected by objects and recaptured by the IR camera. The distortion between the emitted and the received pattern is used to reconstruct the depth values for each reflected ray using triangulation. The driver interpolates the depth values between the rays and outputs a 640x480 depth grid with a precision of 11 bits @ 30 Hz. Microsoft officially specifies a depth range of 1.2–3.5m. The RBG image is provided in the same resolution and framerate as the depth data, however, the two signals do not naturally match due to different extrinsic and intrinsic camera parameters. The exact parameters may even vary among different Kinect devices which makes an individual calibration unavoidable. These parameters can be estimated using camera calibration methods. The cameras can be individually calibrated using chessboard patterns images and OpenCV's calibration routines. The aim is to undistort the RGB and IR images and map depth pixels with color pixels (see figure 4). The maximal range of the kinect raw depth is 2^{11}, and it is possible to convert the raw depth to metric depth using a linear approximation after a previous depth calibration $d_m(x_{ir}, y_{ir}) = f(rawdepth(x_{ir}, y_{ir}))$.

From the metric depth, the 3D metric position (X_{ir}, Y_{ir}, Z_{ir}) of the pixel, with the respect to the IR camera, can be computed using the following equation (9):

$$\begin{pmatrix} X_{ir} \\ Y_{ir} \\ Z_{ir} \end{pmatrix} = \begin{pmatrix} \frac{(x_{ir}-c_{xir})*d_m(x_{ir},y_{ir})}{f_{xir}} \\ \frac{(y_{ir}-c_{yir})*d_m(x_{ir},y_{ir})}{f_{yir}} \\ d_m(x_{ir}, y_{ir}) \end{pmatrix} \quad (9)$$

where x_{ir}, y_{ir} are the coordinates of the depth pixel in image, f_{xir},f_{yir} are the lengths in effective horizontal and vertical pixel size units (IR camera focal length), c_{xir}, c_{yir} are the coordinates of the image center of IR camera, and d_m is depth in meters.

The IR and RGB cameras are separated by a small baseline and using chessboard target data and stereo calibration algorithms, it is possible to determine the 6 DOF transform between them. Knowing the rotation $\mathbf{R}$ and translation $\mathbf{T}$ between the RGB and IR camera, we can then re-project each 3D point on the color image and get its color. The mapping between color image and depth image can be expressed by following equations (10):

$$\begin{pmatrix} X_{rgb} \\ Y_{rgb} \\ Z_{rgb} \end{pmatrix} = \mathbf{R} \begin{pmatrix} X_{ir} \\ Y_{ir} \\ Z_{ir} \end{pmatrix} + \mathbf{T} \qquad \begin{aligned} x_{rgb} &= \frac{(X_{rgb}*f_{xrgb})}{Z_{rgb}} + c_{xrgb} \\ y_{rgb} &= \frac{(Y_{rgb}*f_{yrgb})}{Z_{rgb}} + c_{yrgb} \end{aligned} \quad (10)$$

where x_{rgb}, y_{rgb} are the coordinates of the rgb pixel in image, f_{xir}, f_{yir} are the lengths in effective horizontal and vertical pixel size units (RGB camera focal length), c_{xrgb}, c_{yrgb} are the coordinates of the image center of RGB camera, and d_m is depth in meters.

a) b) c)

Figure 4. a) undistorted RGB image b) undistorted depth Image, the white pixels have unknown depth value, due occlusions or reflective surface material c) Map between undistorted RGB image and depth image.

On figure 5 we present an example of correspondence between consecutive image features in using SURF method (white lines indicate correspondent point). Some matches are undesirable and are related with background static areas. Our solution is to confine the reconstruction space with better limits or develop a movement segmentation filter. The contribution of erroneous matches is minimized by the number of good matches while using the described minimization method to obtain the transformation.

An example of off-line mesh generation, using unorganized kinect 3d points is provided on figure 6. Delaunay triangulation computation results on 99334 vertices and 1223930 faces. A further filtering is required to clean noisily points that increase the number of vertices.

Figure 7 depicts a sequence of scans that creates a 3D person model. They result from several 3D point clouds fused after applying successive 3D rigid body transformations.

Implementation: The system was developed using the C++ language, OpenCV library, OpenKinect library, OpenAR

Figure 5. SURF features matched on consecutive time frames

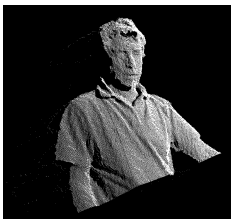


Figure 6. Mesh model with 99334 vertices and 1223930 faces

framework (an augmented reality framework under development on ISR-Coimbra). The processing unit, running Ubuntu Linux v10.10, is composed by a PC with an Intel Core 2 Duo CPU E8200 @2.66GHz, 2GB of RAM and an NVIDIA GPU 8600GT with 512MB. Typically the system has a performance of 2 HZ. The time consuming stage is related with the surf feature extraction and it takes an average of 300 ms. It depends on the number of detected good feature of the image, although we expect to speed up significantly this step by making use of GPU [2]. The involved number of points also influences the transformation time calculus. On table 1 we present some typically time measure involving some algorithm steps.

4 Conclusion

There is still a potential for algorithm speedup involving code optimization, GPU CUDA programming and stereo display graphics. The future work also includes studies conducing to a technological testbed that allow us to measure the sense of presence. Our approach explores virtual view synthesis through motion body estimation and hybrid sensors composed by video cameras and a low cost depth camera based on structured-light. The solution addresses the geometry reconstruction challenge from traditional video cameras array, that is, the lack of accuracy in low-texture or repeated pattern region. We present a full 3D body reconstruction system that combines visual features and shape-based alignment. Modeling is based on meshes computed from dense depth maps in order lower the data to be processed and create a 3D mesh representation that is independent of view-point. This work presents an online incremental 3D reconstruction framework that can be used on low cost telepresence robots or HRI applications applications to enable socialization and entertainment.

Table 1. Processing time measurements

Algorithm Steps	**(ms)**
Acquisition	1.55
Undistort Images	10.61
DepthRGB Map and last frame update	36.13
SURF feature extraction	314.853
Matching and transformation calculus	78.0282
Alignment, display and interaction	30.377
Total (framerate)	471.56 (f=2.12 Hz)

References

[1] A. Akbarzadeh, J.-M. Frahm, P. Mordohai, B. Clipp, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. N. Sinha, B. Talton, L. W. 0002, Q. Yang, H. Stewénius, R. Yang, G. Welch, H. Towles, D. Nistér, and M. Pollefeys. Towards urban 3d reconstruction from video. In *3DPVT*, pages 1–8. IEEE Computer Society, 2006.

[2] L. Almeida, P. Menezes, and J. Dias. Stereo vision head vergence using gpu cepstral filtering. In *VISAPP 2011 - Fifth International Conference on Computer Vision Theory and Applications*, Vilamoura, Algarve, Portugal, March 2011.

[3] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9:698–700, September 1987.

[4] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 1034 –1040, jun 1997.

[5] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. Recent advances in augmented reality. *IEEE Comput. Graph. Appl.*, 21:34–47, November 2001.

[6] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *In ECCV*, pages 404–417, 2006.

[7] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:239–256, February 1992.

[8] C. Bohil, C. Owen, E. Jeong, B. Alicea, and F. Biocca. *Virtual Reality and presence, 21st Century Communication: A reference handbook*. SAGE Publications, Inc, 2009.

[9] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577, July 2003.

[10] J. Challis. A procedure for determining rigid body transformation parameters. *Journal of Biomechanics*, 28(6):733–737, jun 1995.

[11] E. Commission. Cordis, the community research and development information service for science, research and development. 2011.

[12] D. W. Eggert, A. Lorusso, and R. B. Fisher. Estimating 3D rigid body transformations: a comparison of four major algorithms. *MAchine Vision and Applications*, 9:272–290, 1997.

[13] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, June 1981.

[14] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. RSS Workshop on Advanced Reasoning with Depth Cameras, 2010.

[15] F. Isgro, E. Trucco, P. Kauff, and O. Schreer. Three-dimensional image processing in the future of immersive media. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(3):288 – 303, march 2004.

[16] S.-H. Jung and R. Bajcsy. A framework for constructing real-time immersive environments for training physical activities. *Journal of Multimedia*, 1(7):9–17, 2006.

[17] K. Konolige and M. Agrawal. Frameslam: From bundle adjustment to real-time visual mapping. *Robotics, IEEE Transactions on*, 24(5):1066 –1077, oct. 2008.

[18] G. Kurillo, T. Koritnik, T. Bajd, and R. Bajcsy. Real-time 3d avatars for tele-rehabilitation in virtual reality. *Stud Health Technol Inform*, 163:290–6, 2011.

[19] G. Kurillo, R. Vasudevan, E. Lobaton, and R. Bajcsy. A framework for collaborative real-time 3d teleimmersion in a geographically distributed environment. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 111 –118, dec. 2008.

[20] B. Lange, P. Requejo, S. Flynn, A. Rizzo, F. Valero-Cuevas, L. Baker, and C. Winstein. The potential of virtual reality and gaming to assist successful aging with disability. *Physical Medicine and Rehabilitation Clinics of North America*, 21(2):339 – 356, 2010.

[21] J. Lanier. Virtually there. *j-SCI-AMER*, 284(4):66–75, apr 2001.

[22] J.-M. Lien, G. Kurillo, and R. Bajcsy. Skeleton-based data compression for multi-camera tele-immersion system. In *ISVC (1)*, pages 714–723, 2007.

[23] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *IEEE ICIP 2002*, pages 900–903, 2002.

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[25] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[26] S. May, D. Droeschel, D. Holz, S. Fuchs, E. Malis, A. Nüchter, and J. Hertzberg. Three-dimensional mapping with time-of-flight cameras. *J. Field Robot.*, 26:934–965, November 2009.

[27] P. Menezes, F. Lerasle, and J. Dias. Towards human motion capture from a camera mounted on a mobile robot. *IVC*, 29(6):382–393, May 2011.

[28] Microsoft. *Kinect for Xbox 360*. Microsoft Corporation Redmond WA, 2010.

[29] L. G. B. Mirisola, J. Lobo, and J. Dias. 3d map registration using vision/laser and inertial sensing. In *EMCR*, 2007.

[30] B. Petit, J.-D. Lesage, J.-S. Franco, E. Boyer, and B. Raffin. Grimage: 3d modeling for remote collaboration and telepresence. In *ACM Symposium on Virtual Reality Software and Technology*, October 2008.

[31] B. Petit, J.-D. Lesage, C. Menier, J. Allard, J.-S. Franco, B. Raffin, E. Boyer, and F. Faure. Multi-camera real-time 3d modeling for telepresence and remote collaboration. *INTERNATIONAL JOURNAL OF DIGITAL MULTIMEDIA BROADCASTING*, 2010:247108–12, 2009.

[32] A. A. Rizzo and G. J. Kim. A swot analysis of the field of virtual rehabilitation and therapy. *Presence*, 14(2):119–146, 2005.

[33] D. Scharstein. Stereo vision for view synthesis. In *In Proc. Computer Vision and Pattern Recognition Conf*, pages 852–858, 1996.

[34] J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 593 –600, jun 1994.

[35] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, International Journal of Computer Vision, 1991.

[36] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511 – I–518 vol.1, 2001.

[37] O. J. Woodford, I. D. Reid, P. H. S. Torr, and A. W. Fitzgibbon. On new view synthesis using multiview stereo. In *BMVC*. British Machine Vision Association, 2007.

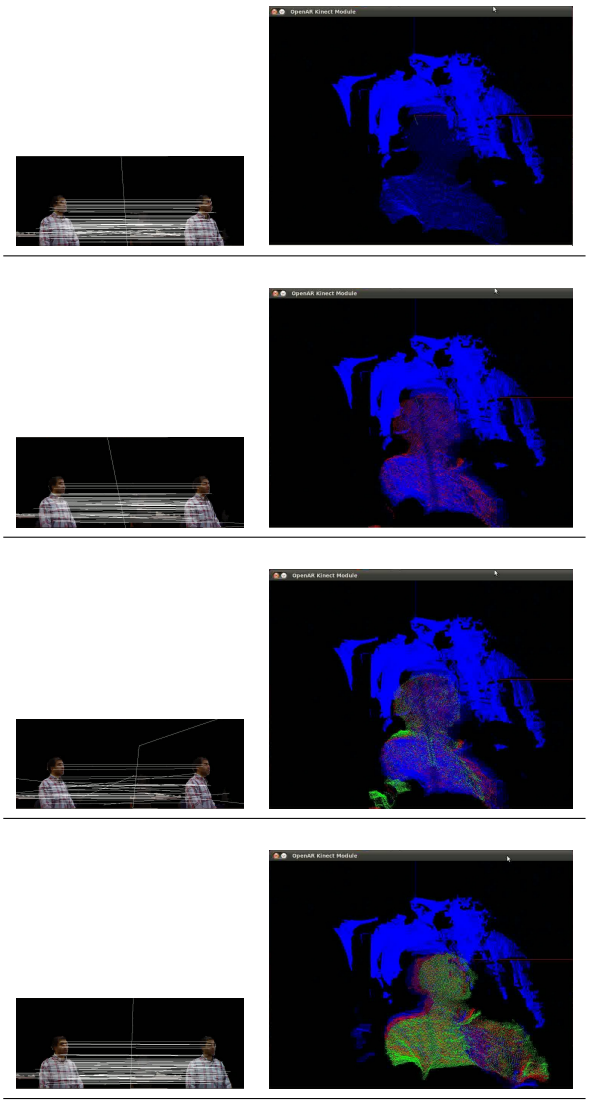


Figure 7. 3D Model, sequence of point clouds being registered on the same referential, each color represent time sequential scans