ON-LINE 3D BODY MODELLING FOR AUGMENTED REALITY

Luis Almeida[1,2], Paulo Menezes[1] and Jorge Dias[1]
[1] *Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal*
[2] *Institute Polytechnic of Tomar, Tomar, Portugal*
laa@ipt.pt, paulo@isr.uc.pt, jorge@deec.uc.pt

Keywords: Augmented Reality, 3D Reconstruction, Tele-presence, Virtual View Synthesis.

Abstract: Building 3D body models is an important task for virtual and augmented reality applications in tele-rehabilitation, education, 3DTV, entertainment and tele-presence. We propose a real-time full 3D reconstruction system that combines visual features and shape-based alignment using low cost depth sensor and video cameras targeting three-dimensional conferencing applications. With this approach we overcome the classic video based reconstruction problem in low-texture or repeated pattern regions. Alignment between successive frames is computed by jointly optimizing over both appearances and shape matching. Appearance-based alignment is done over 2D SURF features annotated with 3D position. Shape-based alignment is performed using the motion transformation estimation between consecutive annotated 3D point clouds through a linear method. A solution to avoid wrong annotated 3D matched points is proposed. 3D mesh model representation is used to lower the processed data and create a 3D representation that is independent of view-point.

1 INTRODUCTION

Immersive virtual applications are common technologies nowadays demanding new human machine interactions approaches. This paper presents an on-line incremental 3D reconstruction framework that can be used on mixed or augmented reality (AR) applications based on tele-presence. The project intends to create an affordable 3D acquisition and display system useful for socialization and entertainment using low cost depth sensors and video cameras. Exploring computers graphics, spatial audio and artificial vision, techniques enable us to induce sensations of being physical in the presence of other people useful on several domains like elderly loneness minimization problem(Lange et al., 2010), tele-rehabilitation(Kurillo et al., 2011)(Rizzo and Kim, 2005), education, socialization, 3DTV, entertainment, tele-presence (Nahrstedt et al., 2011), etc.

Internet chat/audio/video conferencing programs like Skype, VOIP, NetMeeting and phones have been used for socialization, nevertheless they are not able to create the remote person presence feeling. Means of communications that enable eye contact, facial expressions, body language, gestures reconnaissance are required to avoid this sense of disconnectedness. The concept goal is depicted on Figure 1. An example scenario consist on a system providing immersive tele-presence and natural representation of two remote checker players in a friendly and shared mixed reality space to enhance the quality of human-centered communication. The example, based on the principle of a shared virtual checkers board, tries to describe the correct eye contact and gestures reproduction.

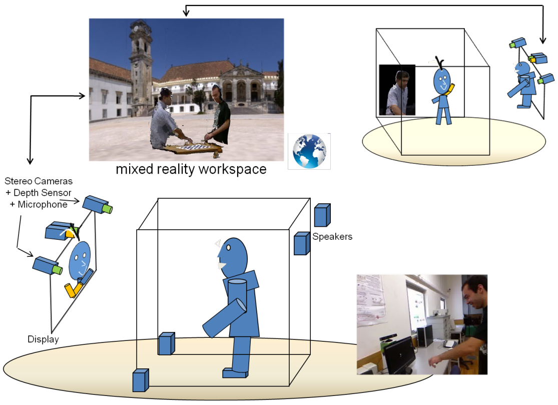

Figure 1: The concept explores computers graphics, spatial audio and artificial vision techniques to induce sensations of being physical in the presence. Checkers mixed reality space example.

Augmented reality and particularly tele-immersion (Jung and Bajcsy, 2006)(Azuma et al., 2001)(Lanier, 2001) provides the technology means that enable users to interact remotely while experiencing the benefits of a real face-to-face meeting. The tele-immersive technology integrates virtual reality for rendering and display purpose, artificial vision for image acquisition and 3D reconstruction, and various networking technologies for transmitting data between distant sites in real-time without significant delays. Virtual meeting spaces allows the possibility of socialization, collaborative work on 3D data, 3DTV, remote training and monitoring, and remote teaching of physical activities (e.g., rehabilitation, dance).

Aiming an incremental on-line 3D human body reconstruction solution useful for shared mixed reality workspace (Aliakbarpour et al., 2011)(Kurillo et al., 2008)(Petit et al., 2008)(Kurillo et al., 2011), we estimate the 3D world information using 2D image sequences and depth information using a depth camera, e.g. a structured light camera or time of flight camera (ToF). With this approach we overcome the classic video based reconstruction problem in low-texture or repeated pattern regions. The presented 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 (Bay et al., 2006). 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 viewpoint.

Mesh simplification is performed 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 existence of 3D human model that is incrementally updated according the user movements lowers the computational scanning resources and stands as an ideal data input solution for the emergent 3D display technology. New display devices are now able to provide a stereoscopic perception of 3-D depth to the viewer either using head mounted displays, light active shutter glasses, passive polarized glasses or without glasses, using flat-panel auto stereoscopic solutions employing lenticular lenses or parallax barriers. Even with an accurate viewer's head tracking and images view dependent rendering 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. The reminder of this paper is organized as follows. First a related work is presented on section 1.1 concerning the psychology nature of the sense of presence followed by the technological approaches to accomplish that. Section 2 describes the suggested methodology and section 3 present some experimental results and discussion. Finally, section 4 presents the future work and conclusions.

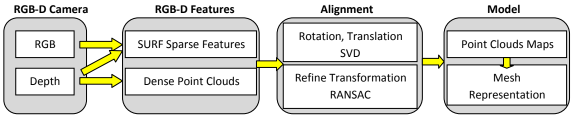

Figure 2: Algorithm overview. The proposed real-time 3D full reconstruction system combines visual features and shape-based alignment between consecutive point clouds. The model representation is updated incrementally.

1.1 Background

Virtual reality (VR) and Augmented reality (AR) creates a sensory and psychological experience for users as an alternative to reality (Bohil et al., 2009). The more one can provide the system with sensory inputs that simulate and effectively mimic those encountered in nature, the more convincing the resulting perceptual and cognitive experience will be for the user (Bohil et al., 2009). Immersive VR and AR perceptually surrounds the user, increasing his or her *sense of presence* (Steuer, 1992) or actually *being within* it. In immersive VR, sensory information is more psychologically prominent and engaging than the sensory information gleaned from other types of media (Lanier, 2001)(Bailenson et al., 2008).

Virtual view synthesis and modeling are the potential graphic tools to create the eye to eye contact illusion on tele-presence communications(Isgro et al., 2004) (Bohil et al., 2009). Real time 3D reconstruction approaches can be divided in three categories: silhouette-based reconstruction, voxel-based methods with space sampling and image-based reconstruction with dense stereo depth-maps. 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 (Besl and McKay, 1992). 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 act on the final surface reconstruction quality. Recent depth sensor devices provide precise 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 (Shi and Tomasi, 1994), Scale-Invariant Feature Transform (SIFT) method (Lowe, 2004) or Speed Up Robust Features (SURF) method (Bay et al., 2006). Several works use these techniques to track 3D pose sensor changes either for object detection, path planning, for gesture recognition or for reconstruction purposes (Henry et al., 2010)(Mirisola et al., 2007)(Akbarzadeh et al., 2006)(May et al., 2009)(Menezes et al., 2011). Our work intends to perform a real-time incremental body modeling.

2 METHODOLOGY

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 (Eggert et al., 1997): SVD (Arun et al., 1987)(Challis, 1995)(Eggert et al., 1997) or iterative methods like Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981)(Akbarzadeh et al., 2006)(Konolige and Agrawal, 2008). Once obtained a 3D point model a mesh is generated through Delaunay triangulation.

2.1 Registration

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 (1):

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

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

$\mathbf{R}$ 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 $[\mathrm{R},\mathrm{T}]$ that maps the set $\{\mathbf{x_i^t}\}$ on to $\{\mathbf{x_i^{t+1}}\}$ can be obtained through the minimization of the equation (2) using a least square criterion. The least square solution is the optimal transformation only if a correct correspondence between 3D point sets is guaranteed. Complementary methods are used to robust the correspondence (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) (Arun et al., 1987)(Challis, 1995)(Eggert et al., 1997). 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:

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

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

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

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,

R, that maximizes the referred trace is **R**= **U** diag(1; 1; det($\mathbf{UV}^T$)) $\mathbf{V}^T$:

$$\mathbf{R} = \mathbf{UV^T} \tag{6}$$

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 $\mathbf{{}^0H_w}$. 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, $\mathbf{{}^{t+1}H_t}$, $\mathbf{H} = [\mathbf{R}, \mathbf{T}]$. Therefore it is possible to compute equations:

$\mathbf{{}^iH_0} = \mathbf{{}^iH_{i-1}}{}^{\mathrm{i-1}}\mathbf{H_{i-2}}\ldots\ldots\mathbf{{}^1H_0}$ and $\mathbf{{}^iH_w} = \mathbf{{}^iH_0{}^0H_w}$

To update the reconstructed model, each acquired 3D point set is transformed to the world coordinate system using $\mathbf{{}^iH_w}$. 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.

2.3 Tracking and Registration Refining

SURF features are detected and matched over consecutive undistorted images. 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.

Registration Refining using RANSAC: False correspondent point pairs that wrongly biases the rigid body transformation estimation are removed using the RANSAC method. The approach randomly samples three 3D points correspondent pairs from consecutive scans and iteratively estimates the rigid body transformation (Arun et al., 1987) until find enough consensus or reach a maximum number of iteration based on the probability of outliers. The procedure starts to use a small initial data set and enlarges the number of samples consistent with the model. K iterations are performed while the eligible solution with highest number of inliers, based on sum of the distances between pair of correspondent point, is selected as the best transformation model. The K iterations number follows equation (8):

$$K = \frac{log(1-p)}{log(1-(n_{inliers}/N_{pts})^S)} \tag{8}$$

p stands for the desired probability of finding at least one model transformation without outliers within K iteration, $n_{inliers}$ is the number of eligibles pairs of points that fit the current estimation, N_{pts} represents the total number of pairs of points and S is the minimum number of eligible samples to fit the transformation model. Registration refining method is described in algorithm 1.

Algorithm 1: Registration refining algorithm - Outliers removal.

```
Input :$X_p, X_q$
{assumed correspondent 3D point pairs}
Output :$[R,t]$
{rigid body transformation estimation}
while ($i < MAXITER$) do
    randomly select 3 pairs of points
    $[R_i, t_i] \leftarrow$ estimate 6DOF rigid body transformation for these 3 pairs
    $X'_q = R_i * X_q + t_i$
    {apply the transformation to $X_q$ scan to map it into $X_p$ reference frame}
    $inliers_i = |(X'_q - X_p) < \tau|, number\_of\_inliers_i$
    {determine the set of data points which are within a Euclidean distance threshold $\tau$}
    if ($sizeof(inliers_i) > T_{threshold}$) then
        $[R,t] \leftarrow$ re-estimate the transformation model using all $inliers_i$
        $EXIT$
    end if
    if ($number\_of\_inliers_i > bestscore$) then
        $bestscore \leftarrow number\_of\_nliers_i$
        $best\_inliers \leftarrow inliers_i$
        {store cardinality of $inliers_i$ and $inliers_i$}
        update $MAXITER$ {using eq. 8}
    end if
    $i = i + 1$
end while
$[\mathbf{R}, \mathbf{t}] \leftarrow$ re-estimate the transformation model using all points from $best\_inliers$
```

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. The selected approach is based on a facial feature tracker using eye feature (Viola and Jones, 2001).The purpose of use Haar-like features is to meet the real-time requirement. The resulting eyes 2D position can then be associated to 3D points for the calculation of head 3D pose.

Algorithm 2: Model reconstruction algorithm.

```
**Input** :$rgb\_images, depth\_images$
**Output** :$3D\_mesh\_model$
initialize $R_g$, $t_g$, $f_1$, $f_{1d}$, $f_{1xyz}$, $f_{1r}$
**for** (;;) **do**
    $f_2 \leftarrow undistort(adquire\_rgb\_image())$
    $f_{2d} \leftarrow undistort(adquire\_depth\_image())$
    $f_{2xyz} \leftarrow convert\_depth\_image\_to\_xyz\_data(f_{2d})$
    $f_{2r} \leftarrow map\_rgbcolor\_to\_depth\_image(f_{2xyz}, f_2)$
    $(surf_1, surf_2) \leftarrow$
    $detect\_SURF\_features(f_{1r}, f_{2r})$
    $matches2D \leftarrow SURF\_match(surf_1, surf_2)$
    $matches3D \leftarrow correspond2D3D(matches2D)$
    $[R,t] \leftarrow motion\_estimator(matches3D)$
    $[R_g, t_g] \leftarrow update\_global\_transformation(R,t)$
    $f_{1r} \leftarrow f_{2r}, f_{1xyz} \leftarrow f_{2xyz}$ {update_past_data}
    $model \leftarrow$
    project_points_to_world_coordinates($f_{2xyz}, R_g, t_g$)
    $mesh\_model\_generation$
**end for**
```

3 IMPLEMENTATION AND RESULTS

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

Calibrations: The Kinect device combines a regular RGB camera and a 3D scanner, consisting of an infrared (IR) projector and an IR camera (figure 7a). A initial calibration step is required to undistort the RGB and IR images, and to map depth pixels with color pixels (6 DOF transform between RGB and IR cameras) (Almeida et al., 2011).

Implementation: The system was developed using the C++ language, OpenCV library, OpenKinect library, OpenAR framework (an augmented reality framework under development on ISR-Coimbra) and Ubuntu Linux v10.10.

Matching: On figure 3 we present an example of correspondence between consecutive image features 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 with outliers removal to obtain the transformation.

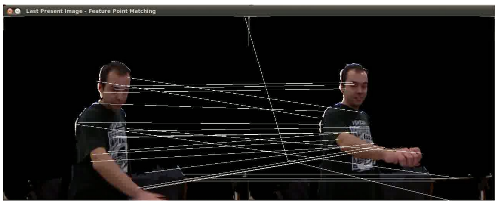

Figure 3: SURF features matched on consecutive time frames.

Outliers Removal: In order to analyze the registration refining improvement described on algorithm 1, we have measured the mean euclidean distance between several consecutive registrations with and with outliers removed after applying the transformation to X_q scan that maps it into X_p reference frame ($X'_q = R_i * X_q + t_i$, $|(X'_q - X_p)|$) (see figure 4). The red balls line (without outliers) presents a much lower error than considering all SURF matched point into rigid body transformation. Figure 5, presents for each consecutive rigid body transformation estimation the total number of SURF matched points (blue bars) and the number of inliers for that take (red bars).

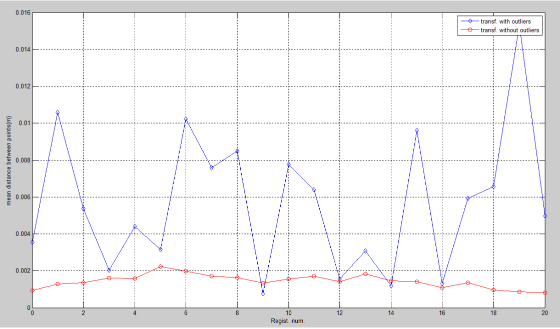

Figure 4: Mean euclidean distance between several consecutive registrations with and without outliers removed.

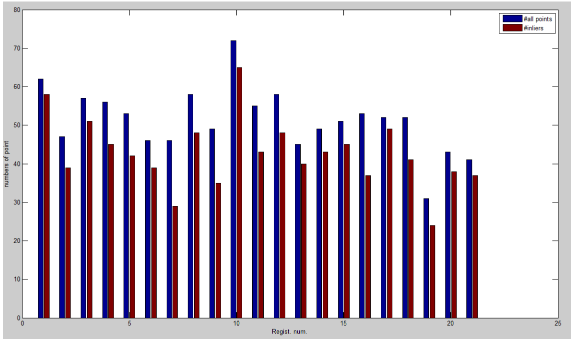

Figure 5: Number of points number (blue bars) vs Number of inlier's (red bars) on each registration.

Experimental results shows that considering a high number of inliers (not all SURF point features) makes the transformation estimation more robust and increases the alignment accuracy. Figure 6 depicts two correspondent 3D points sets, result from SURF algorithm that should be aligned. After applying the transformation to X_q scan to map it into X_p reference frame we obtain a new set of points $X_q' = R_i * X_q + t_i$ (green ball points). Applying the transformation to inlier's points only, we obtain magenta balls point.

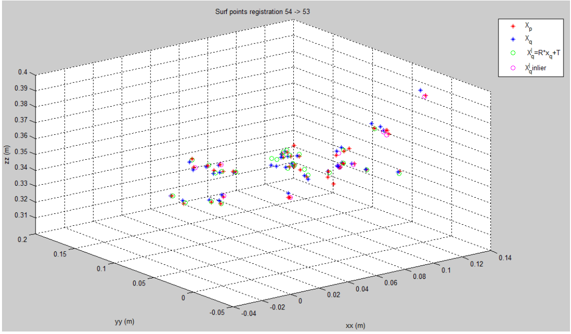

Figure 6: Applying the transformation to X_q scan to map it into X_p reference frame ($X_q' = R_i * X_q + t_i$), result into green ball for all points and magenta balls just for inliers.

3D Modeling: An example of off-line mesh generation, using unorganized kinect 3d points, is provided on figure 7b. Delaunay triangulation computation results on 99334 vertices and 1223930 faces.

Figure 8 depicts a sequence of scans that creates a 3D person model. They result from several 3D point clouds fused in real time after applying successive 3D rigid body transformations.

Processing Time Measurements: 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

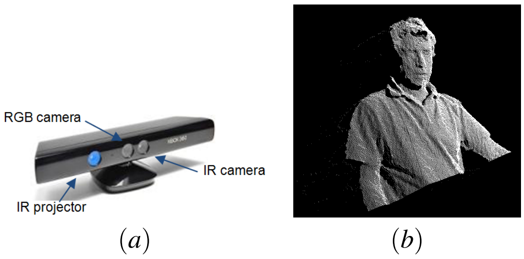

Figure 7: a) Kinect Sensor b) Mesh model with 99334 vertices and 1223930 faces.

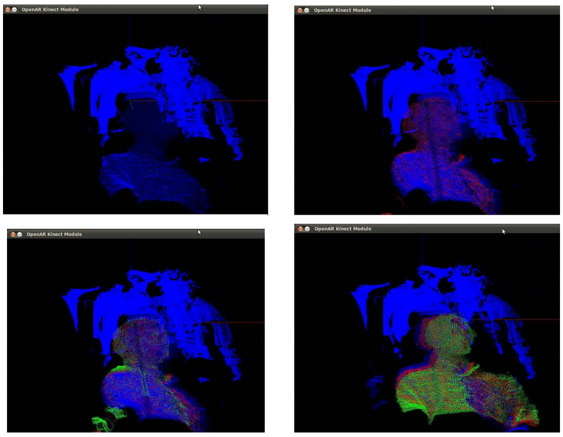

Figure 8: 3D Model, real time sequence of point clouds being registered on the same referential, each color represent time sequential scans.

we expect to speed up significantly this step by making use of GPU. 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 CONCLUSIONS

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

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	471.56 (f=2.12 Hz)

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. Experimental results shows that considering a high number of inliers (not all SURF point features) increases the alignment accuracy. 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 on-line incremental 3D reconstruction framework that can be used on low cost telepresence applications to enable socialization and entertainment.

REFERENCES

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

Aliakbarpour, H., Almeida, L., Menezes, P., and Dias, J. (2011). Multi-sensor 3d volumetric reconstruction using cuda. *3D Research*, 2:1–14. 10.1007/3DRes.04(2011)6.

Almeida, L., Menezes, P., Seneviratne, L., and Dias, J. (2011). Incremental 3d body reconstruction framework for robotic telepresence applications. In *Robo 2011: The 2nd IASTED International Conference on Robotics*, Pittsburgh, USA.

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

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

Bailenson, J., Patel, K., Nielsen, A., Bajscy, R., Jung, S.-H., and Kurillo, G. (2008). The Effect of Interactivity on Learning Physical Actions in Virtual Reality. *Media Psychology*, 11(3):354–376.

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

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

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

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

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

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

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

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

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

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

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

Kurillo, G., Vasudevan, R., Lobaton, E., and Bajcsy, R. (2008). 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.

Lange, B., Requejo, P., Flynn, S., Rizzo, A., Valero-Cuevas, F., Baker, L., and Winstein, C. (2010). 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.

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

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

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

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

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

Nahrstedt, K., Yang, Z., Wu, W., Arefin, M. A., and Rivas, R. (2011). Next generation session management for

3d teleimmersive interactive environments. *Multimedia Tools Appl.*, 51(2):593–623.

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

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

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

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *JOURNAL OF COMMUNICATION*, 42:73–93.

Viola, P. and Jones, M. (2001). 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.