COMPARATIVE ANALYSIS OF REAL-TIME MULTI-VIEW RECONSTRUCTION OF A SIGN LANGUAGE INTERPRETER

Andrej Satnik[1], Robin Ribback[2],
Krishna Chandramouli[1], Giacomo Inches[3], Mark Wheatley[4], Ebroul Izquierdo[1]

[1]Multimedia & Vision Research Group, Queen Mary University of London, United Kingdom
[2]SWISS TXT, Switzerland
[3]Fincons Group AG, Switzerland
[4]European Union of the Deaf, Belgium

ABSTRACT

Following the proliferation of smart handheld devices coupled with the launch of HbbTV 2.0 specification there is an emerging trend among broadcaster organisations to enable a higher degree of content accessibility across disabled communities. In this respect, the sign language interpreters play a crucial role in facilitating accessibility to main stream broadcaster content for the deaf community. Addressing the increasing demand placed on sign language interpreters and current limitations on delivering sign language content complementary to main stream media, this paper proposes the use of a low-cost 3D studio environment which enables photorealistic reconstruction of human avatars that captures the interpreter while eliminating the background. The core novelty of the proposed approach relies on the information fusion framework for correlating multi-sensor RGB-Depth sensors. The performance of the proposed approach for the reconstruction of a sign language interpreter has been evaluated using two low-cost sensors namely Kinect V2 and RealSense D435.

INTRODUCTION

In recent years, audio-visual equipment has become pervasive offering everyone accessibility to consume media services through a range of smart devices including mobiles, tablets and large-television sets. Among the global population, approximately 5% suffer from disabled hearing[1]. In 2006 the United Nations adopted the United Nations Convention on the Rights of Persons with Disabilities to explicitly state the need of providing inclusive services and products to persons with disabilities. In particular the UN convention states that the right of persons with disabilities to take part on an equal basis with others in cultural life, and shall take all appropriate measures to ensure this for persons with disabilities.

Following such recommendations, the Swiss public broadcaster SRG committed themselves via a contract with the national disabilities organisations to provide increased Signing of TV programs by more than 200% to 1000 hours first time signed programs, not counting any reused content. The same trend has been observed in other European countries like Germany (ZDF), Belgium (VRT) and the UK (BBC) while a shared agreement between

[1]http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2]https://www.ebu.ch/news/2017/02/broadcasters-and-disability-organisation-draw-up-common-recommendation-on-future-eu-rules-for-audiovisual-access-services

disability organisations and the European Broadcasting Union (EBU) was released in 2017[2]. In addition, with the expected launch of new technologies such as HbbTV 2.0 CSS[3], there is an emerging trend towards delivering accessible content across handheld devices, thus facilitating new services that are aimed to enhance the experience of members within the deaf community. While the delivery of existing services relies on the traditional use of high-end studio environments and capture setups that allow post-production tools to synchronise the main stream content with the sign language interpretation of the content.

Among the several challenges to be considered are the topics of multi-view sensing for robust aggregation of depth maps from multi-view sensing hardware. The disadvantage of single-view 3-D reconstruction techniques that allow to obtain digital objects only from single viewpoint leads to the usage of multiple RGB-Depth (RGB-D) sensors such as the Microsoft Kinect[4] and Intel's Realsense[5]. At the current state, for the signing of these programs additional studio resources need to be organized: studio, lighting, mixer operations, streaming setup, as well as the organisational efforts in term of resources and manpower. This infrastructure and process are cost intensive and can reach up to a thousand euros per hour of signing. The current way of producing signed content for a broadcaster is via a dedicated studio created at the broadcaster premise, which includes professional lighting, a professional camera, a special background for the signer (to later create overlay effects via Chroma Key) and screens that reproduce the original content to be signed and the subtitles or scene description.

In contrast, the emergence of 3D representation techniques facilitates the development of a user-centric reconstruction framework based on low-cost sensors. Therefore, in this paper addressing the need of photorealistic reconstruction of a sign language interpreter that remains independent of the environment (thus eliminating the need of high-cost studio setup), a multi-view reconstruction system is presented that includes an information fusion framework to synchronise the data captured from multiple RGB-D sensors. In order to facilitate an objective assessment of the overall algorithm, three low-cost sensors are interfaced with the algorithm, which differ only in the aggregation stage of the data capture.

RELATED WORK

When providing the sign-interpreted versions of content produced for the hearing, however, the experience of existing users should not be affected. In the literature there has been several reports on the topic of processing sign language [1, 2, 3, 4] spanning across more than five years.

RGB-Depth (RGB-D) sensors are used to identify colour and depth simultaneously in real time. With the development of low-cost commercial RGB-D sensors such as Kinect or RealSense, the availability of point-clouds and powerful computing devices has inspired researchers in several areas to develop many vision applications such as pose estimation gesture recognition [3, 5] or scene reconstruction [6]. However, depth sensors suffer from missing or inaccurate depth information. These problems are caused by the incorrect matching of infrared patterns causing numerous errors, such as optical noise, loss of depth values and flickering. Most of the works use bilateral filters and a variation of joint bilateral filters [7]. However, many of them were inaccurate or the processing time was too long to

[3]https://www.scribd.com/document/372066111/HbbTV-v202-Specification-2018-02-16

[4]https://developer.microsoft.com/en-us/windows/Kinect

[5]https://software.intel.com/en-us/realsense/d400

consider it as real-time filtering. In our prior work [8], several filtering approaches to improve depth map accuracy using low-cost hardware with three Time-of-Flight (ToF) sensors were used. Therefore, accuracy of point cloud can be improved in real- time with sufficient frame rate.

Systems with multiple ToF cameras have been previously proposed [9, 10, 11]. A multi-Kinect system to seamlessly render a scene from a wide range of angles was described in [10]. A challenge with multi-view sensing lays in filtration, correction, and reconstruction of sensed data. For instance, in approach [11] they introduced real-time surface refinement that requires no pre-processing stage, such as the definition of an intermediate surface, and that does not generate oversampling. It is based on the idea of adding a new point for every pair of neighbours in the cloud. The presented algorithm relied on the availability of minimal information such as per point normal and radius for each element of the point set. However, this approach is working directly with the point cloud and building up a new geometry. In [9], the authors stress the advantage of multi-view sensing using separate PCs since the processing time is reduced by distributing computational load among multiple devices. For computationally demanding algorithms, the frame rate becomes too low to achieve real-time processing as described in [9]

CONCEPT AND APPROACH

The main novelty of the holistic approach presented in this paper resides in the low-cost hardware and software suite able to deliver a high-quality 3D point cloud of a sign language interpreter. The overall concept and approach adopted for the development relies on three fundamental steps: hardware connection of multiple RGB-Depth (RGB-D) sensors using a single machine comprising low-cost hardware, calibration of the studio during its installation, which requires the positioning of the multiple depth sensors around capturing area, and point cloud enhancement using convenient methods capable of process large amount of data in real-time. The performance of the introduced system has already been assessed in a real-time application and it is presented in our prior work [8]. This system comprises three RGB-D sensors connected to a single computer producing reconstructed mesh and rendered in real-time. The overall functional architecture of the envisaged framework is presented in Figure 1.

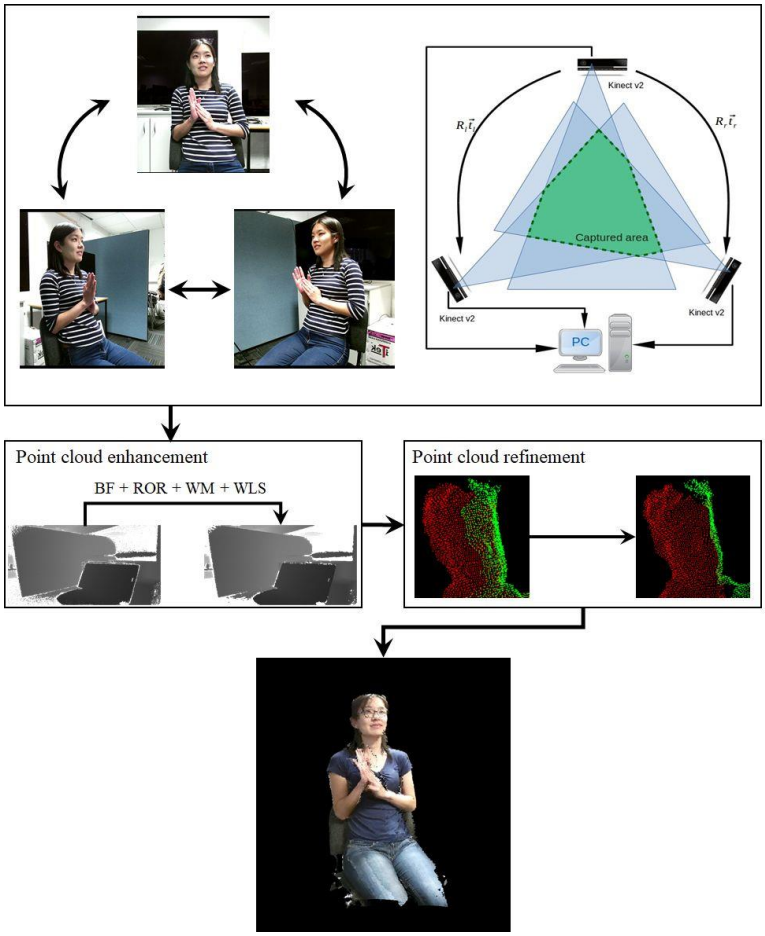

Figure 1 – Multi-view Sensing System architecture

Calibration

The reported challenge can be attributed to the calibration synchronisation of the multi-view sensing environment. In the development of the Multi-view Sensing System (MSS), the calibration tool suite was designed to be device agnostic and employs extrinsic parameters based on planar pattern detection and offline Iterative Closest Point (ICP). In addition to supporting the widely-used Microsoft Kinect V2 devices, the implementation carried out within MSS includes Intel RealSense RGB-D devices as well. The calibration stage is performed by finding extrinsic and intrinsic parameters of the sensor using a camera pinhole model presented in [8, 9].

[1]http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2]https://www.ebu.ch/news/2017/02/broadcasters-and-disability-organisation-draw-up-common-recommendation-on-future-eu-rules-for-audiovisual-access-services

$$p = K(RP + \vec{t}) \quad (1)$$

where *K* is the camera intrinsic matrix, *R* represent 3×3 camera orientation matrix and $\vec{t}$ represent translation vector.

The calibration has two components, calibration between depth and RBG camera and calibration of multiple RGB-D sensors. The depth and colour images come from two separate, slightly offset, cameras, so they do not perfectly overlap. However, for each pixel in the depth image, we can calculate its position in 3D space and reproject it into the image plane of the RGB camera and acquire colour for each depth pixel. Calibration between multiple RGB-D sensors is described and evaluated in [8].

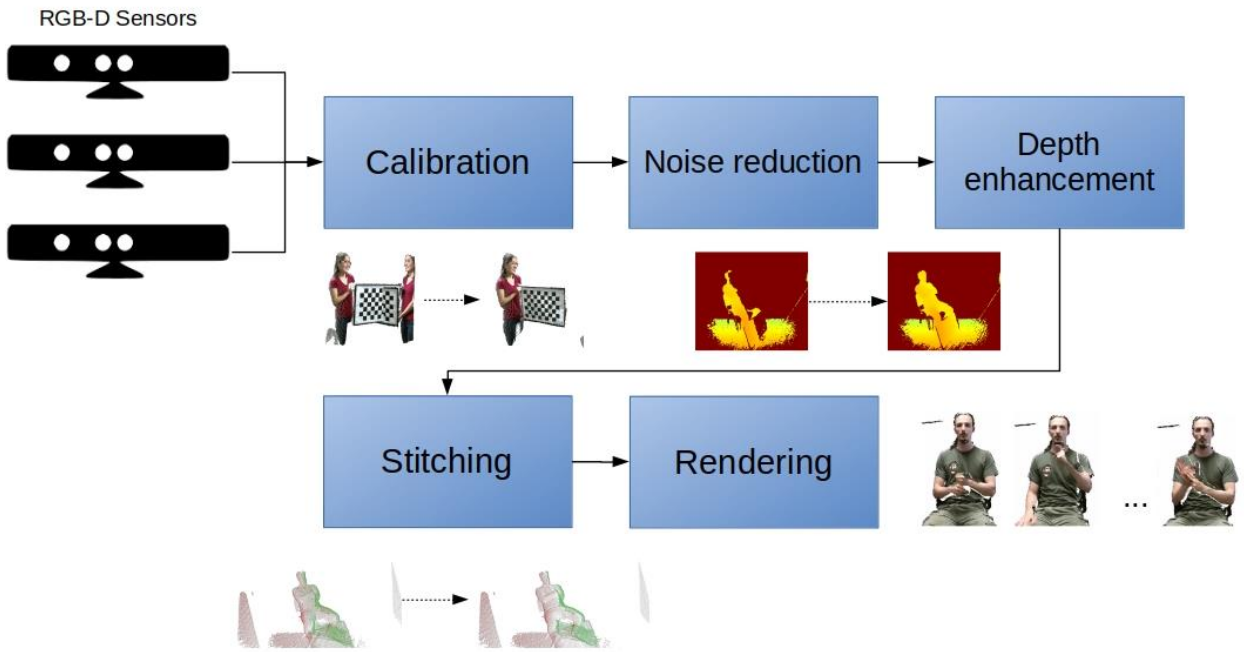

Figure 2 – Block diagram of Multi-view Sensing System for point cloud enhancement

Photorealistic reconstruction

Photorealistic reconstruction of 3D avatar models from multi-view sensing requires addressing the various challenges related to time-series synchronisation of sensed information from multiple sensors. The main drawback of related solutions for low-cost capture and reconstruction of photorealistic 3D avatar relates to depth inconsistencies between the measurements acquired with different RGB-D devices. The reported challenge can be attributed to the calibration synchronisation of the multi-view sensing environment.

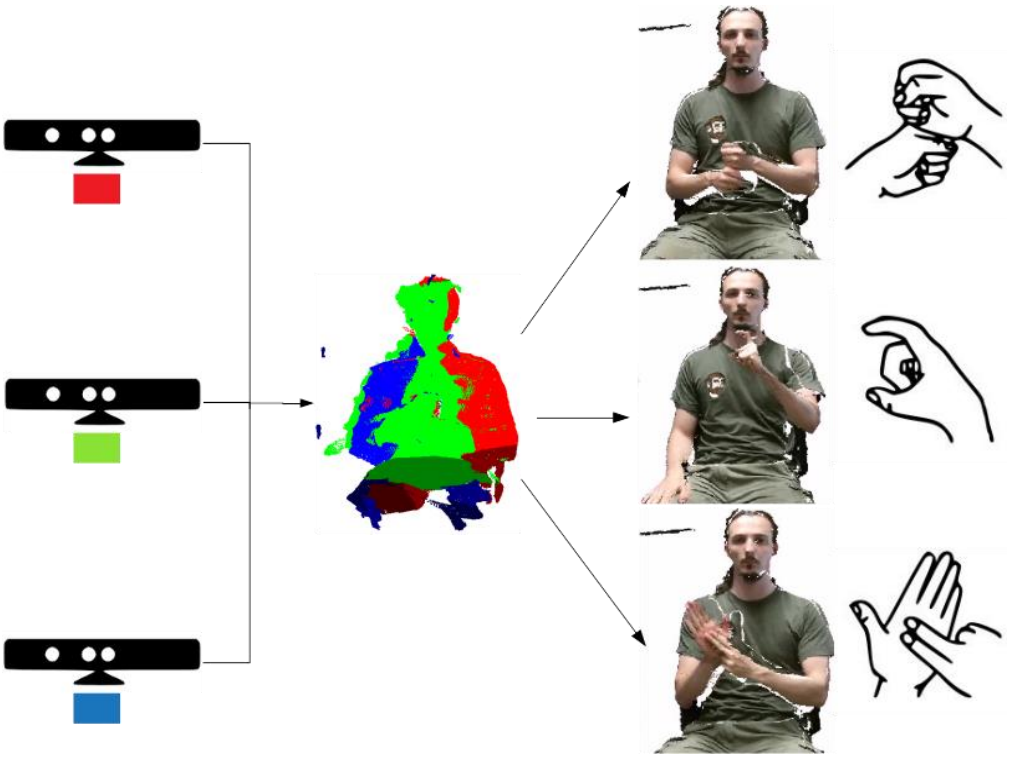

Figure 3 – Results of signs rendered in point cloud. Each colour represents separated point cloud captured by adjected sensor.

The quality of the depth image is a key factor for model generation and reconstruction. For instance, the infrared projector of Time-of-Flight cameras (ToF) emits the pseudo random pattern light by a diffractive mask [4], which can cause a distortion in the depth channel which generates wrong depth data. On the other hand, structural sensors such as RealSense D435 suffers from overexposing or underexposing the IR sensor which causes inaccuracies in the depth channel. This leads to inaccurate presentation of the point cloud and subsequently to deformations in reconstruction. Therefore, pre-processing is required to remove pixels on the edges because ToF cameras produce noisy edges as is stressed in [12]. This task can be finished by implementing several filtering approaches as was described and evaluated in our work [8]. Moreover, absence of topological information point cloud gives us a surface representation for complex geometries where the accuracy mainly depends on the number of points. Our framework integrates mechanisms to enable stitching of the acquired point cloud together to build a globally consistent representation using

adjected RGB-D sensors. Assuming that image coordinates of depth pixel *p'* of the adjected sensor can be estimated using reverse transformation:

$$p'.x = \frac{P'.x}{P'.z} f.x - 0.5 + c.x \qquad \text{and} \qquad z_{rs1} \begin{cases} z_{rs1}, & \text{if} \quad l_r < |z_{rs1} - z_{rs2}| \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

where $P' = R^T P - \vec{t}$ is reverse transformation of 3D point *P* from the first sensor, *f* and *c* are intrinsic parameters of second sensor then we can determine final depth pixel's value. The l_r represent minimal distance between occluded points, z_{rs1} is depth pixel on position *p* of the first RealSense sensor and z_{rs2} is depth pixel on position *p'* of the second RealSense sensor. This method must be applied for each sensor pair to remove overlapped points. The final building block of the Multi-view Sensing System (MSS) is the rendering framework that is able to display the sign-language gestures streamed from the broadcast studio.

IMPLEMENTATION

The technical development of the 3D reconstruction of a sign language presenter is achieved using CUDA parallel computing platform with GeForce 780GTX graphic card and the connection of three RGB-D sensors to a single machine as is depicted in Figure 1. The low-cost capture setup is further developed to include additional low-cost sensors to improve the quality of the initial point cloud capture system. The rendering of the 3D avatars across a range of heterogeneous display devices is achieved through system integration.

CONCLUSION

We believe, that we have proven the possibility of the multiple connection of RGB-D sensors and reconstruction point cloud of gestures for a sign language interpreter. The results of the preliminary experiments showed the effectiveness of combining a multi-view sensing system and the Intel RealSense technology using a reconstruction algorithm. Our work indicates that the performance of gesture reconstruction can be significantly improved by exploring and utilizing the more robust algorithms.

In addition, the work plan also dedicates effort for the development of a deep-learning framework for carrying out post-processing of the captured point cloud for improved quality of the photorealistic 3D avatars.

References

[1] C. Dong, M. C. Leu and Z. Yin, "American sign language alphabet recognition using microsoft kinect," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[2] H.-D. Yang, "Sign language recognition with the kinect sensor based on conditional random fields," *Sensors,* vol. 15, no. 1, pp. 135-147, 2015.

[3] S. Saha, S. Bhattacharya and A. Konar, “A novel approach to gesture recognition in sign language applications using avl tree and svm,” *Progress in Intelligent Computing Techniques: Theory, Practice,and Applications,* pp. 271-277, 2018.

[4] D. Konstantinidis, K. Dimitropoulos and P. Daras, “Sign language recognition based on hand and body skeletal data,” in *3DTV Conference*, Stockholm - Helsinki, 2018.

[5] P. Syrora, P. Kamencay, M. Zachariasova and R. Hudec, “Hand gesture recognition based on depth map,” in *ELEKTRO 2014*, Strbske pleso, 2014.

[6] R. A. Newcombe, D. Fox and S. M. Seitz, “Dynamicfusion: Reconstruction and tracking of nonrigid scenes in real-time,” in *Conference on Computer Vision and Pattern Recognition*, 2015.

[7] A. V. Le, S.-W. Jung and C. S. Won, “Directional joint bilateral filter for depth images,” *Sensors,* vol. 14, no. 7, p. 11362–11378, 2014.

[8] A. Satnik, E. Izquierdo and R. Orjesek, “Multiview 3d sensing and analysis for high quality point cloud reconstruction,” in *International Conference in Machine Vision*, Vienna, 2018.

[9] D. S. Alexiadis, D. Zarpalas and P. Daras, “Real-time full 3-d reconstruction of moving foreground objects from multiple consumer depth cameras,” *IEEE Transactions on Multimedia,* vol. 15, no. 2, pp. 339-358, 2013.

[10] M. Kowalski, J. Naruniec and M. Daniluk, “Livescan: A fast and inexpensive 3d data acquisition system for multiple kinect v2 sensors,” in *International Conference on 3D Vision*, 2015.

[11] G. Guennebaud, L. Barthe and a. M. Paulin, “Real-time point cloud refinement,” in *Proceedings of the First Eurographics Conference on Point-Based Graphics*, Switzerland, 2004.

[12] O. Wasenmüller and D. Stricker, “Comparison of kinect v1 and v2 depth images in terms of accuracy and precision,” in *Computer Vision Workshop* , 2017.