VISION-ASSISTED REMOTE ROBOTIC ASSEMBLY GUIDED BY SENSOR-DRIVEN 3D MODELS

Abdullah Mohammed, Lihui Wang, Mauro Onori

Department of Production Engineering, KTH Royal Institute of technology, Sweden

abdullah.mohammed@itm.kth.se

Abstract: This paper proposes an approach to developing a cyber-physical model-driven system that performs robotic distant assembly operations in decentralised manufacturing environment. The system uses an industrial robot to assemble components unknown in advance. The system captures and analyses the components' silhouettes, and constructs the corresponding 3D models. By the help of the models, the system is able to guide a distant human operator to assemble the real components in the actual robot cell. The results show that the developed system can construct the 3D models and assemble them within a suitable total time, imitating the human behaviour in a similar situation.

Keywords: robot, remote assembly, 3D modelling, cyber-physical system

1. INTRODUCTION

In the last few decades, growing needs for globalisation have been witnessed to leverage production cost vs. market share; therefore, it is highly expected that distributed manufacturing will become more common and play a highly important role in factories of the future. In a distributed environment where human operators and production equipment are not collocated, mutual collaboration towards remote real-time manufacturing is in demand (Krüger *et al.*, 2009). One example is human-guided robotic assembly.

The objective of this research is to bridge the gap between humans and robots while fully employing the capabilities of each in remote real-time assembly. This paper proposes a novel approach for real-time assembly achieved by remote human-robot collaboration using cyber-physical architecture. In the developed system, 3D models are used to guide an off-site human operator during remote assembly. 3D models of real physical components to be assembled by a robot are generated on the fly based on the objects' silhouettes captured by a robot-mounted camera. The camera is turned off during assembly to achieve a high communication performance. In this context, the robot is treated as a manipulator which mimics the human's operations remotely.

2. RELATED WORK

In recent years, researchers have developed various tools to program, monitor and control industrial robots. The aim is to reduce possible robot downtime and avoid collisions caused by inaccurate programming, through simulation and/or augmented reality (Nee *et al.,*2012). However, these tools require pre-knowledge about a robotic system. Introducing unknown objects to the robotic system may produce unrealistic solutions that cause a breakdown to the physical robotic system due to no-longer valid robot programs.

Both laser scanners and vision cameras are common techniques to convert unknown objects to virtual 3D models. Modelling objects using stereo vision cameras was a main focus for several researchers (Aristos and Tzafestas, 2009; Tian *et al.*, 2007; Samak *et al.*, 2007), whereas others including (Sumi *et al.,* 2002) adopted a

pre-defined library of 3D models to match the real desired objects. However, the stereo vision camera-based approach suffers from two drawbacks: (1) it requires expensive and less compact equipment, and (2) it lacks the ability to capture and model complex shapes from fixed single viewpoint due to limited visibility.

2D vision systems can also be applied to model unknown objects. By taking a number of snapshots of an object from different viewports, the object can be modelled based on analysing the captured silhouette in each snapshot. For example, (Petit *et al.,*2010; Atsushia *et al.,*2011) focused on modelling the object in high accuracy and with details.

Despite the fact that these approaches were successful in their reported applications, they are unable to model multiple objects in a single run. Besides, they lack the ability to model objects remotely.In this paper, we propose a new approach for constructing 3D models of multiple arbitrary objects, simultaneously, based on a set of snapshots taken for the objects from different angles. This approach is implemented through a system that analyses the captured silhouettes of the objects and constructs 3D representations for the objects in a web based environment. This allows an operator to perform assembly operations from a distance.

3. SYSTEM OVERVIEW

The proposed system demonstrates the ability to identify and model arbitrary unknown objects to be assembled using an industrial robot. The objects are then integrated with the existing 3D model of the robotic cell in a structured virtual environment – Wise-ShopFloor (Wang *et al.*, 2011), for 3D model-based distant assembly. The system consists of four modules (see Fig. 1): (1) an application server, responsible for image processing and 3D modelling; (2) a real industrial robot, for performing assembly operations; (3) a 2D network camera, mounted at the end effector of the robot for capturing silhouettes of unknown/new objects; and (4) a user interface for remote operator to monitor/control the entire assembly operation.

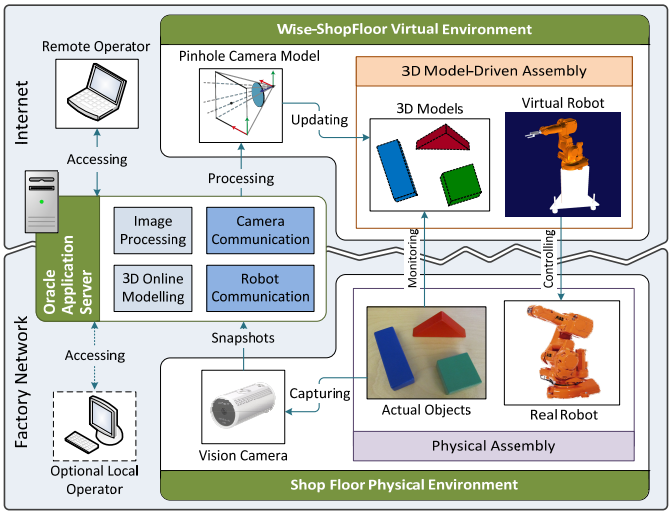

Fig. 1. System configuration.

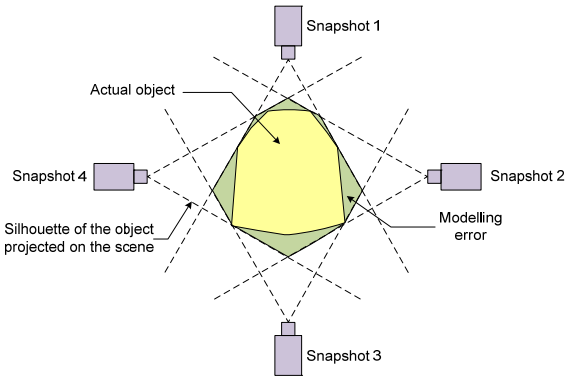

Fig. 2. Shape approximation by trimming of a 2D example.

The system has the ability to identify and model incoming parts of unknown objects and assemble them remotely. First, the robot moves to a position where the camera is facing the objects from above to capture the top-view snapshot. The system then construct the primary models of the objects by converting their silhouettes in the top-view snapshot to a set of vertical pillars with a default initial height. After that, the camera is used to take a sequence of new snapshots of the objects from other angles. Projecting the silhouettes of each snapshot back to the 3D space generates a number of trimmed pillars. The intersections of these pillars identify the final 3D models of the objects. Fig. 2 shows a simplified 2D trimming process of one object after the top-view snapshot, where the bounding polygon including errors is used to approximate the actual object.

4. SYSTEM IMPLEMENTATION

4.1 Image processing

Image processing steps are performed to recognise the geometry details of the captured objects through their extracted silhouettes. The details of those steps are explained below.

Converting to grayscale. To reduce the computational complexity, the captured colour images are converted to grayscale by taking the weighted average value of RGB values of each pixel in the images.

Adjusting brightness and contrast. Finding the right pixel intensity highly relies on the lighting conditions of the working environment and the settings of the camera. Therefore, the brightness and contrast are adjusted based on the lighting conditions of the developed system.

Gaussian smoothing. A zero-mean Gaussian filter given in eq. (1) is applied to each pixel at (i, j) in the image matrix of $(m \times n)$ to remove the noise and to smooth the image. The output image $H(i, j)$ is the convolution of an input image $f(i, j)$ and Gaussian mask $g(\alpha, \beta)$.

$$H(i,j) = f(i,j)\,g(\alpha,\beta) = \sum_{\alpha=-[(n-1)/2]}^{(n-1)/2} \sum_{\beta=-[(m-1)/2]}^{(m-1)/2} f(i-\alpha, j-\beta)\,g(\alpha,\beta) \tag{1}$$

The discrete form of convolution is performed which goes through each element in the convolution mask and multiply it with the value of the corresponding pixel of the input image; the sum of these multiplications is assigned to the pixel in the output image. The process is repeated for each pixel to produce the output image.

Image thresholding. This process identifies the silhouette pixels in the image by assigning a certain intensity values to them. It is started by scanning the image pixel by pixel while comparing its intensity value with a threshold value. Each pixel in the image will have either white or black intensity value based on whether it is higher or lower than the threshold value.

Silhouettes labelling. This process identifies each silhouette in the image by assigning a specific label to it. The connected component labelling algorithm (Jankowski and Kuska, 2004) is chosen due to its efficiency. The process starts by scanning the image pixel by pixel to find one that belongs to one of the silhouettes, followed by examining its neighbouring pixels. If one or more neighbouring pixels already have a label, the algorithm assigns the lowest label to the pixel. Otherwise, a new label is assigned. The outcome of labelling operation is a two-dimensional array where each element represents a pixel, and each silhouette is represented by a unique label. The pixels that do not belong to any silhouette have zero values.

4.2 3D modelling

Based on the silhouettes retrieved from the captured images, constructing 3D models of the captured objects is performed as follows:

Calibration of camera. Constructing 3D models precisely requires calibrating the camera to determine its parameters and identify its physical location, e.g. the camera's optical centre, image centre (see Fig. 3Ⓐ), and focal coefficients f_x and f_y, radial and tangential distortion coefficients (not shown). Therefore, a pinhole camera

model (Tsai, 1987) is adopted to calibrate the physical camera. A 2D coordinate system U-V is defined on the image plane, specifying pixel locations in a captured image. Moreover, the camera with respect to the robot's end-effector is specified by a transformation matrix, and its relationship to the robot base is defined similarly as shown in Fig. 3Ⓑ. The full calibration needs to be performed only once when the camera is mounted on the robot for the first time, with minor adjustments at regular service intervals to minimise any deviations.

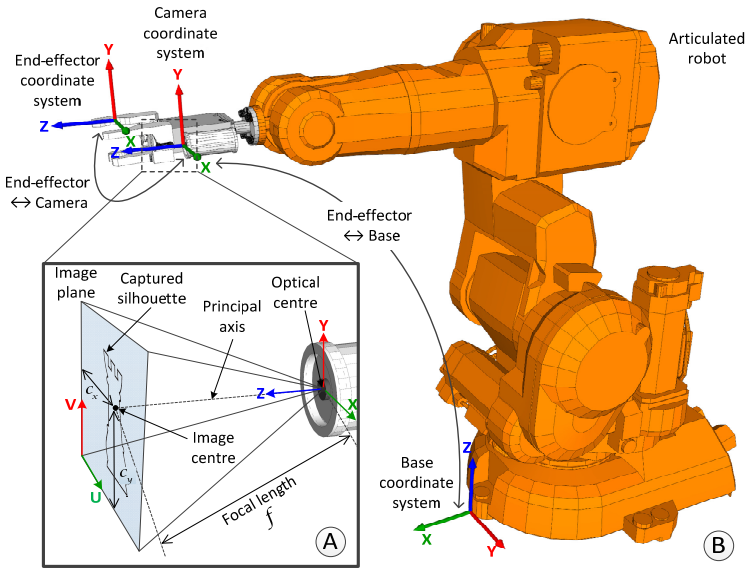

Fig. 3. Parameters and coordinate systems for camera calibration.

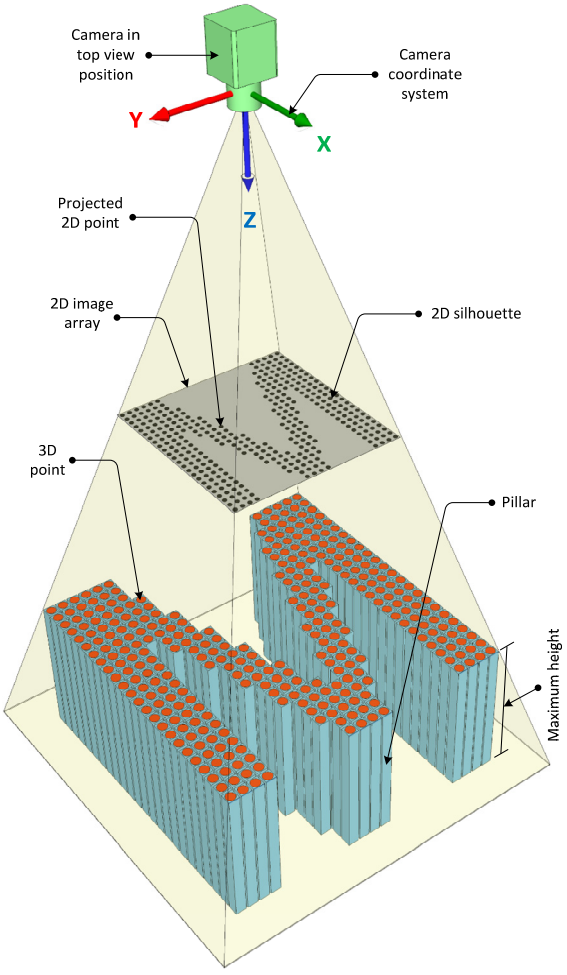

Fig. 4. Construction of initial pillars.

Construction of pillars. The first snapshot provides the silhouettes of the objects from the top view perspective. It is used to construct initial representation of the 3D models. These models are represented by a set of pillars of pixel diameter in 3D space. Fig. 4 depicts the construction of the initial pillars. A temporary value is assigned as the height to all pillars. This value is the maximum possible height of the objects. The construction of the initial pillars is accomplished by applying Tsai's pinhole camera model (Tsai, 1987) that is used in the calibration step, as shown in Fig. 3Ⓐ. Given a 3D point (x, y, z), its projected 2D point (u, v) on the U-V plane is described as

$$u = f_x \times x'' + c_x, \quad v = f_y \times y'' + c_y \tag{2}$$

$$x'' = x'(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + 2p_1 x'y' + p_2(r^2 + 2x'^2) \tag{3}$$

$$y'' = y'(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + 2p_2 x'y' + p_1(r^2 + 2y'^2) \tag{4}$$

$$x' = \frac{x}{z}, \quad y' = \frac{y}{z}, \quad r^2 = x'^2 + y'^2 \tag{5}$$

$$f_x = f \times s_x, \quad f_y = f \times s_y \tag{6}$$

where k_1, k_2, k_3 and p_1, p_2 are the radial distortion coefficients and tangential distortion coefficients, respectively. The dimensions of a pixel are defined by s_x and s_y. Eq. (2) and eq. (6) introduce two different focal coefficients: f_x and f_y. This is due to the fact that the individual pixels on a typical CCD image sensor are rectangles in shape.

Trimming of pillars. The trimming operation takes place after the second snapshot has been processed (silhouette extracted), and continues until the trimming by the last silhouette is done. Fig. 5 shows the trimming of two sample pillars with reference to one silhouette.

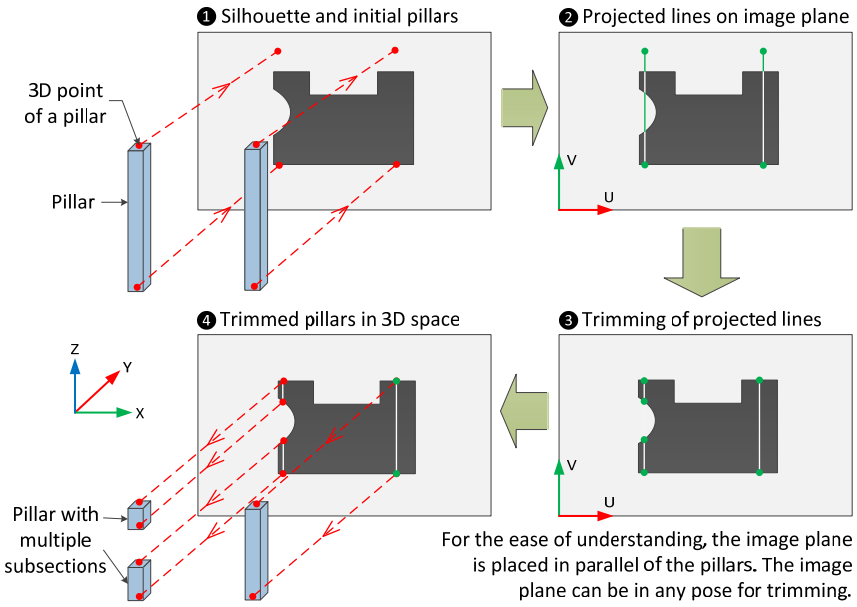

Fig. 5. Example of pillar-trimming process.

In other words, the trimming process starts by projecting the pillars one by one from the 3D space to the image plane. Since each pillar is represented by two 3D end points, the projection is only for the points. The projection of a pillar creates two 2D points, calculated by eq. (2), and a projected line on the image plane. Extracting the pixels shared by the projected line and the captured silhouette reveals a trimmed line. Finally, the trimmed 2D line is projected back to the 3D space, resulting in a trimmed pillar. The trimming process is repeated for all pillars and for all snapshots. The concavity of the object results in pillars with multiple subsections.

Solid prism representation. Despite the fact that the aforementioned processes can trim the pillars as closely as possible to mimic the real objects, the pillars alone are neither intuitive nor computationally efficient for 3D visualisation due to the fact of non-solid geometry and high redundant representation. Moreover, the modelled shapes need to be compatible with the robot 3D model in Wise-ShopFloor (Wang *et al.*, 2011). This, however, can be achieved by creating a solid prism representation for the trimmed pillars.

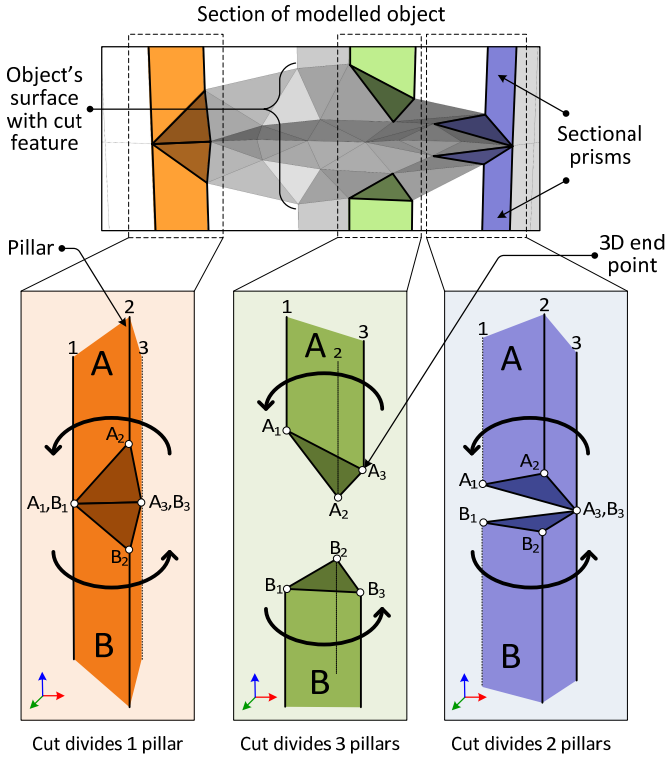

Fig. 6. Prism creation with different cut cases.

Two objectives are taken into consideration during the process: (1) to localise the process, and (2) to create a uniform representation of a given object. The pillars are first divided into groups of three according to their immediate neighbouring connectivity. The prism creation is then divided into three steps to construct: (1) the top surface, (2) the bottom surface, and (3) the three sides of each prism. Selecting the correct order of the end points is crucial when building a surface patch of each prism as its surface normal affects its visibility. As shown in Fig. 6, three end points in counter-clockwise order are used to create a surface patch with an outer visibility. Moreover, three cases of pillar division (cut) caused by the subsections of pillars are also considered during prism creation, as illustrated in Fig. 6.

5. CASE STUDY

As illustrated in Fig. 7, three simple parts are chosen for a proof-of-concept case study to evaluate the functionality of the developed 3D model-driven remote assembly system.

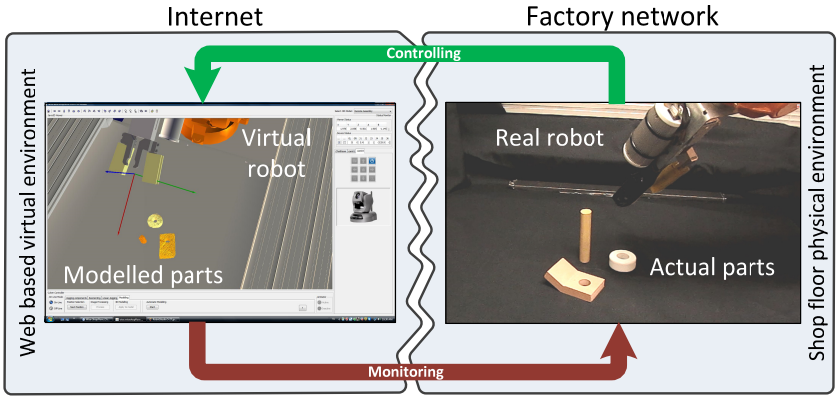

Fig. 7. 3D model-driven remote assembly.

The camera is switched off to save network bandwidth after modelling and integrating the 3D models of the assembly objects with the robotic assembly cell , leaving only low-volume data communication with the robot

controller alive. A remote operator assembles the 'parts' (3D models) using the 3D robot model in the virtual world. At the same time, the real robot mimics the virtual robot and assembles the actual parts simultaneously in the real world. During remote assembly, only the robot control commands are transmitted from the virtual robot to the real one instantly and automatically, without extra robot programming. It is worth mentioning that image processing can also identify the geometric centres and orientations of the parts, leading to semi-automatic pick-and-place and grasping operations during remote assembly. Fig. 8 depicts the results of 3D model creation of the parts, as well as those of 3D model-driven remote assembly.

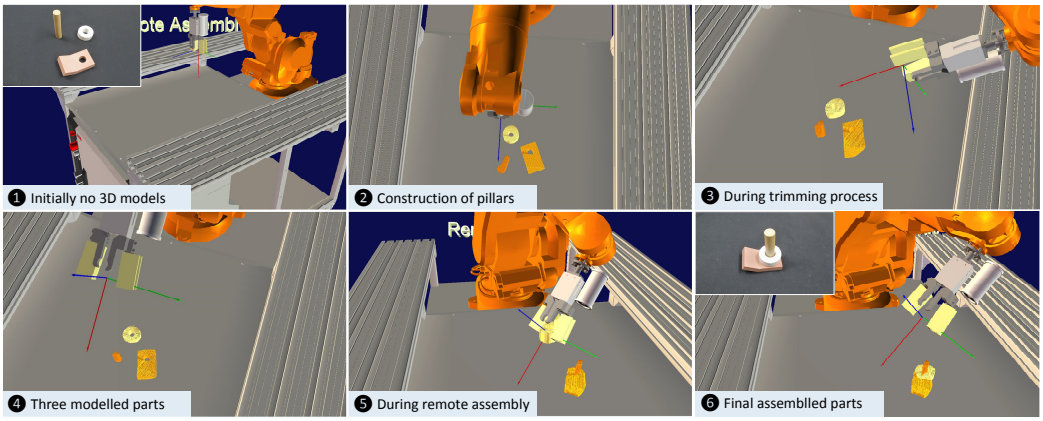

Fig. 8. Results of case study for 3D modelling and assembly.

In this case study, seven snapshots for the objects taken from different angles are used to model the parts. Improving the quality of the 3D models requires adding additional snapshots which leads to increase in the processing time. A performance analysis is therefore conducted under the following specifications to understand its relationship: Intel Core i5 processor of 2.6 GHz, graphics card of GeForce GT 240, a 4 GB RAM, and running under the operating system of Windows Vista.

The image-based 3D model generation has been tested for ten times and the average computation time for each processing step was calculated and recorded as shown in Fig. 9, with error bars indicating deviations in the recorded processing times. It is found that the silhouette labelling process consumes in average 1.75 sec in processing time, with the highest deviation. The labelling algorithm employed examines all the neighbouring pixels when spotting a non-zero pixel in an image during the pixel-by-pixel scanning. This explains why it consumes a high percentage of processing time and it varies from one test to another. Despite this fact, the system can process one image in about 4.8 sec.

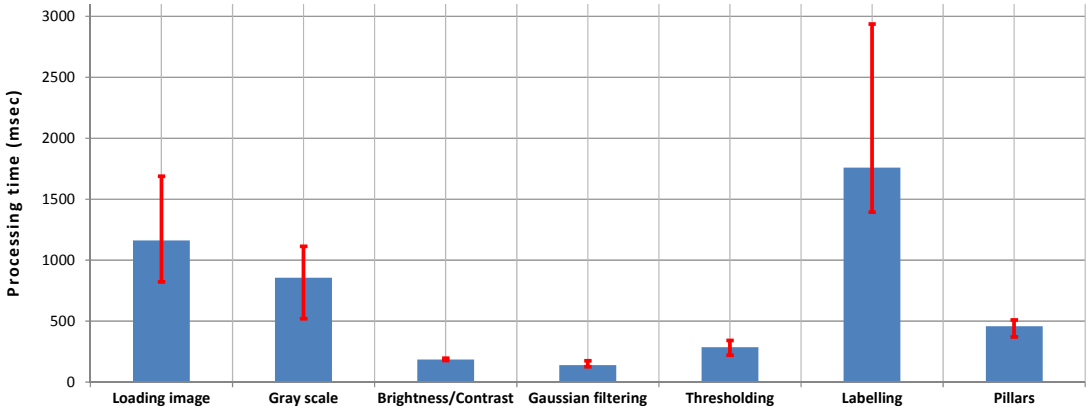

Fig. 9. Comparison of computation time of different processing steps.

The results of the pillar trimming process for each snapshot are also recorded. Fig. 10 manifests the accuracy by comparing the actual height of a real object and the trimmed pillar height of its 3D model after processing each snapshot excluding the first top-view image. As can be seen in the figure, the accuracy of pillar trimming is reasonably high after processing the 7th snapshot as the error has converged quickly to a small value in 22 sec.
By analysing the efficiency of remote assembly, the real robot lags behind the virtual robot by 30 msec over the Internet.

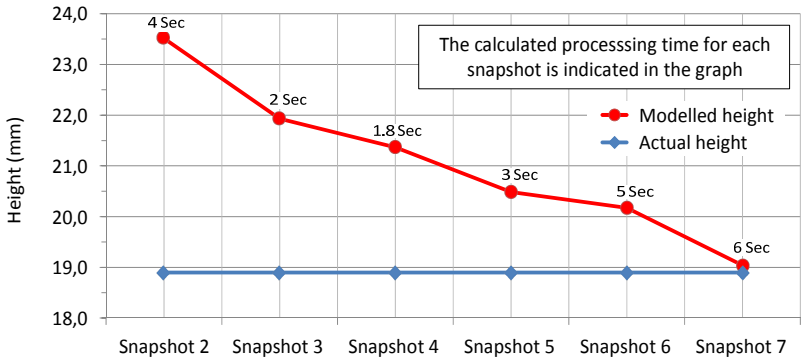

Fig. 10. Modelling error vs. number of snapshot processed.

6. CONCLUSIONS

A 3D model-driven remote robotic assembly approach is presented in this paper, where an off-site operator can manipulate a real robot instantly via virtual robot control. The 3D models of the parts to be assembled are generated based on the snapshots of the parts captured by a robot-mounted camera at the beginning. The generated 3D models are then integrated with the 3D model of a real robotic cell. The novelty of this research includes: (1) elimination of video-image streaming during remote assembly in real-time which reduced the communication bandwidth utilisation, and (2) user friendly robot programming-free environment. The needed robot control commands are generated automatically and transmitted from the virtual robot to the real one for physical assembly. The developed approach can be suitable for remote assembly operations in dangerous or rural locations. The results of the case study show that it can generate a set of 3D models in 22 sec by analysing seven snapshots from different points of view. The efficiency can be further improved by performing image processing tasks in parallel when moving the camera to the next position. For complex shapes, more snapshots could be used to improve the modelling accuracy. Our future work includes more comprehensive feature identification e.g. centre of hole etc. during 3D modelling and more tests of realistic parts assembly, the result of which will be reported separately in the future.

REFERENCES

Aristos, D. and S. Tzafestas(2009). Simultaneous Object Recognition and Position Tracking for Robotic Applications. *IEEE International Conference on Mechatronics*.

Atsushia K, H. Sueyasua, Y. Funayamab, T. Maekawa (2011). System for reconstruction of three-dimensional micro objects from multiple photographic images. *Computer-Aided Design*, **vol. 43, no. 8**, pp. 1045-1055.

Jankowski, M. and J. Kuska (2004), Connected Components Labelling – Algorithms in Mathematica, Java, C++ and C#. *International Mathematica Symposium*.

Krüger, J., TK. Lien, A. Verl (2009). Cooperation of Human and Machines in Assembly Lines. *CIRP Annals – Manufacturing Technology*, **vol. 58, no. 2**, pp. 628-646.

Nee, AYC., SK. Ong, G. Chryssolouris, D. Mourtzis (2012). Augmented Reality Applications in Design and Manufacturing. *CIRP Annals – Manufacturing Technology*, **vol. 61, no. 2**, pp. 657-679.

Petit, B., J. Lesage, C. Menier, J. Allard, J. Franco, B. Raffin, E. Boyer, and F. Faure (2010). Multicamera Real-Time 3D Modeling for Telepresence and Remote Collaboration. *International Journal of Digital Multimedia Broadcasting*, **vol. 2010, no. 247108**, pp. 1-12.

Samak, D., A. Fischer, D. Rittel (2007). 3D Reconstruction and Visualization of Microstructure Surfaces from 2D Images. *CIRP Annals – Manufacturing Technology*, **vol. 56, no. 1**, pp. 149-152.

Sumi , Y., Y. Kawai, T. Yoshimi and F. Tomita (2002). 3D Object Recognition in Cluttered Environments by Segment-Based Stereo Vision. *International Journal of Computer Vision*, **vol. 46. no. 1**, pp. 5-23.

Tian, X., H. Deng, M. Fujishima, K. Yamazaki (2007). Quick 3D Modeling of Machining Environment by Means of On-machine Stereo Vision with Digital Decomposition. *CIRP Annals – Manufacturing Technology*, **vol. 56, no. 1**, pp. 411-414.

Tsai, R. (1987). A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. *IEEE Journal of Robotics and Automation*, **vol. 3, no. 4**, pp. 323-344.

Wang L, M. Givehchi, G. Adamson, M. Holm (2011). A Sensor-Driven 3D Model-Based Approach to Remote Real-Time Monitoring. *CIRP Annals – Manufacturing Technology*, **vol. 60, no. 1**, pp. 493-496.