connected everything.

industrial systems in the digital age

Digitalisation of Collaborative Human-Robot Workspaces

Dr Niels Lohse
Loughborough University

EPSRC
Engineering and Physical Sciences Research Council

Executive Summary

This project has investigated the feasibility of achieving real-time digitalisation of complex human-robot workplaces by enhancing a network of 2D cameras with learned image segmentation and 3D reconstruction at the edge and centrally to the system respectively. The ambition was to demonstrate that using deep-learned models can achieve fast enough 3D reconstructions to track people and objects. The results show that people can be tracked with 10Hz refresh rates already without significant optimisation of the algorithms used. This clearly indicates that the use of networked intelligent 2D cameras has a very high potential. Challenges remain to create more effective approaches to train and verify the trained deep networks to establish their baseline accuracy as well as improve the processing time.

1. Research challenge

Close human-robot collaborative working will be essential to improve the competitiveness of high-wage manufacturing economies by increasing productivity without losing agility. For this to be achieved new ways for humans and robots to work together must be defined. Current approaches, to create inherently safe robots by limiting their size and capability, are not suitable for many manufacturing tasks. For example, in HVM sectors such as the Aerospace and Automotive industries, industrial robots with high payload capabilities that can move very fast and carry dangerous tools are essential. However, despite significant advances in robotics and autonomous systems, one of the most critical barriers for the successful introduction of these technologies in manufacturing is the lack of robust real-time high-fidelity awareness of the workspaces with all its actors. For collaborative and mixed autonomous human-robot systems to become safe, the whole workspace needs to be digitised in high detail.

New paradigms such as the German Industrie 4.0 and the American Industrial Internet of Things make it increasingly realistic to create Cyber-Physical Production Systems. Machines and robots are becoming able to share their process data and create a digitised presence of themselves in local or global cloud systems. However, the digitalisation of people and inert objects such as products, fixtures, tools, etc. has been largely unexplored so far. Yet, it is precisely the ability of a machine/robot to perceive the people and objects around it which will support safety assurances and traditional trust issues, thus enabling a breakthrough in collaborative and autonomous manufacturing systems. Currently, safety is achieved at the expense of productivity; either humans and robots are separated with physical barriers or the robot reduces its speed when a human gets closer. What is required are robust ubiquitous perception systems that can capture an industrial workspace to a level of detail, underpinning the prediction of safety and process critical events.

Figure 1. Complex manufacturing environments in which digitalisation, monitoring and machine awareness could help improve efficiency and productivity.

Hypothesis
A network of standard 2D smart cameras, combined with a purely data driven deep learning approach, can be used to recognise, localise and track multiple objects and people within a workspace; with robust and accurate real-time performance that rivals marker-based tracking systems (such as Vicon)

Objectives:

- Create an integrated state-of-the-art human-robot collaboration test cell, combining the best industrially available tracking technology (e.g. laser tracker and Vicon) with a network of high- resolution smart cameras.
- Investigate deep learning methods for robust object recognition and tracking using the smart camera network and learned ground truth observations from the highly accurate tracking systems.
- Determine the capabilities of trained multi sensor optical systems for real-time object tracking in complex, noisy industrial workspaces with people, objects and machines in close proximity to each other.
- Establish the necessary evidence to support more substantial research funding applications.

2. Approach

This section presents the methodology grouped under three areas: a) hardware infrastructure, b) software infrastructure, and c) reconstruction and tracking.

a) Hardware Infrastructure
In order to implement marker-less tracking and workspace monitoring, 8 Basler Ace acA2440-75um cameras (5 MPixel resolution, up to 75frames/s acquisition speed) with high resolution and wide field-of- view (each covering the entire cell) lenses have been mounted above the test workspace, in key locations (workspace corners and half way along each wall). Each of these cameras was connected to a dedicated PC (Intel i5 based architecture, including an NVIDIA GeForce GTX 1070 Ti graphics car, Gigabit Ethernet, running Ubuntu 16.4 and Basler Pylon camera interface package) via USB3. This configuration allowed us to easily control the image acquisition process and

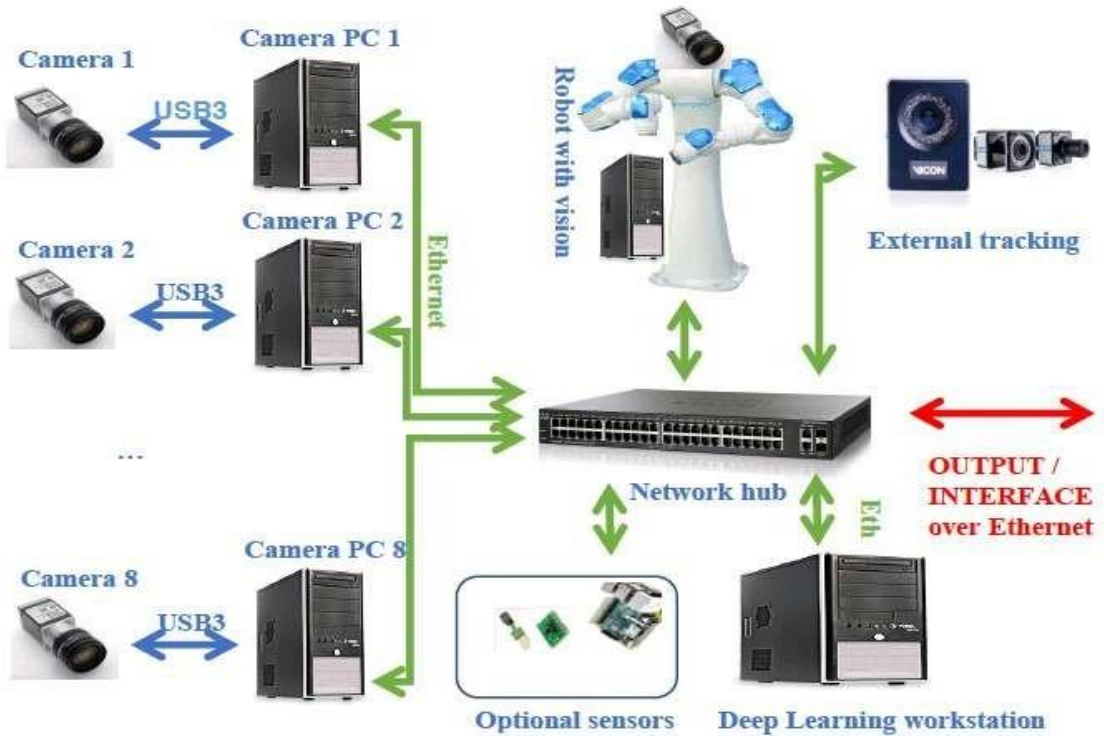

Figure 2. System hardware architecture

perform image pre-processing, providing a flexible and cost-effective development environment, in contrast to using FPGA-based smart cameras, which provide more performance but are cost intensive and require difficult re-programming (usually only available through the camera manufacturer).

A gigabit Ethernet network was established to connect the 8 camera PCs with a deep learning workstation (cards used: two NVIDIA Tesla K80) tasked with higher level, computationally intensive data processing, using the pre-processed image information from the individual camera-PCs transmitted over the network. This architecture is easily extendable, enabling easy integration of Ethernet- compatible sensors and computers for robot control. In order to further enhance compatibility and applicability to robotic applications (enabling human-robot interaction/collaboration), all network communication and control was implemented using ROS packages: Pylon-based packages for image acquisition, native ROS communication packages for data transfer, and custom written ROS nodes, using Python scripts for image pre-processing (on camera PCs).

In order to verify spatial resolution (tracking accuracy), a Vicon (Vantage) based external tracking system has been used. In a first approach, the two systems (camera based tracking and Vicon) have been used separately. However, in future work, these can be integrated in the overall architecture, using a custom written ROS node.

b) Software Infrastructure

The implemented pipeline is illustrated overleaf in Figure 3. Starting from the image acquisition from each camera, to the 3D reconstruction representation of the work cell environment, the processing pipeline was divided into several key necessary steps: pre-processing and processing of images from each camera independently, followed by a reconstruction that combines the outputs of the processed images.

Image Pre-Processing

After acquiring the images from the ROS nodes, the images are first pre-processed with OpenCV in Python. Pre-processing is beneficial as pre-trained deep learning models operate most effectively on images with similar properties as those on which the model was trained. For example, if the model was trained on images of size: 656 x 368 pixels then the images from the cameras are resized from 2448 × 2048 pixels to 656 x 368 pixels before running inference on those images.

Image Processing

A number of computer vision projects have been made available as open source on GitHub. For this feasibility study, we focused on implementing the latest state-of-the-art algorithms in semantic segmentation: DeepLab [1], instance segmentation: Mask R-CNN [2], and human keypoint detection: OpenPose [3] and DensePose [4].

c) Reconstruction and Tracking

By synthesising the processed images from each camera, we implemented 2 different types of reconstruction based on calibrated cameras: an occupancy-based volume carving reconstruction of people and objects that provides an estimate of the shape, size and location of objects in the 3D space; and a 3D skeletal-based reconstruction that computes the 3D locations of people and their key features such as eyes, neck, elbows, and knees in the 3D space. The 3D skeletal approach was then further extended to track individual instances of people over time (see Figure 4 overleaf).

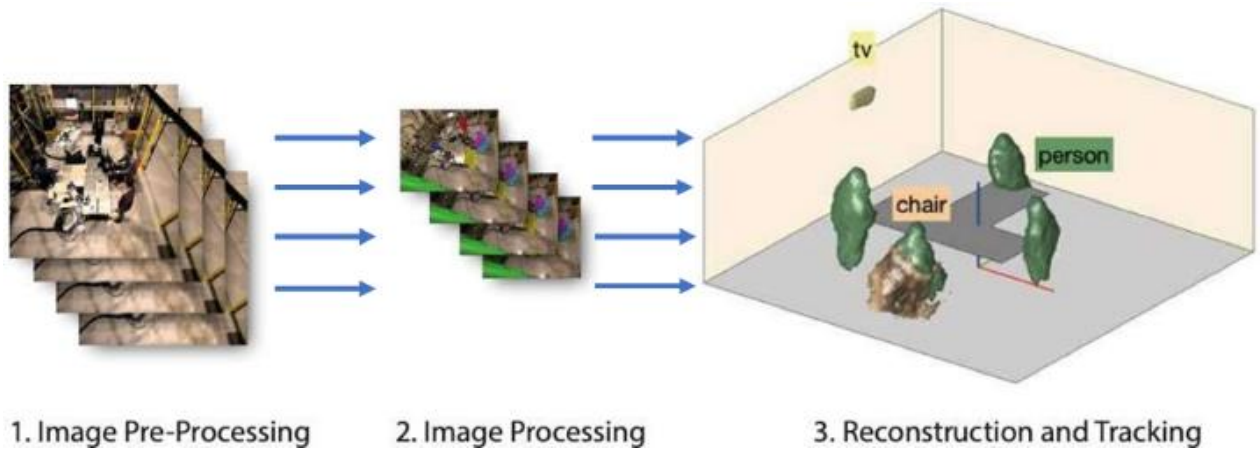

Figure 3. Data pipeline: from multiple images to 3D reconstruction.

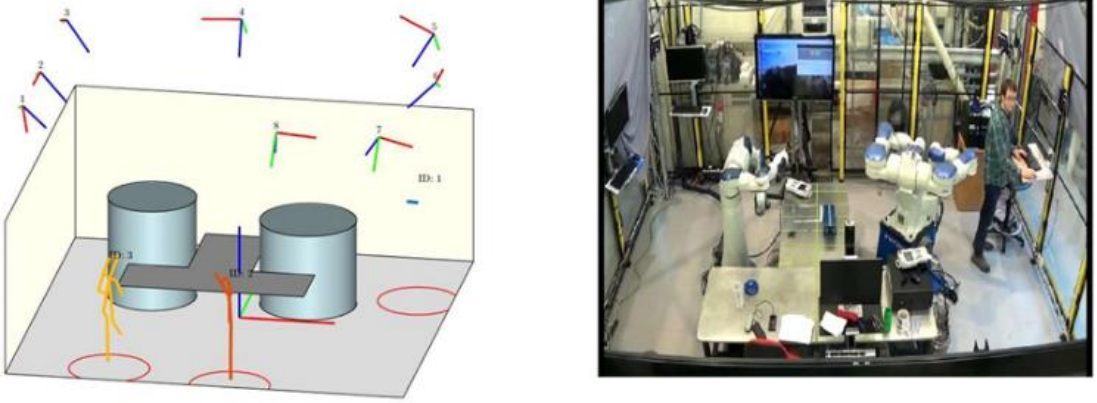

Figure 4. 3D skeletal tracking

3. Results

In this section, results are described for the different stages of the software pipeline (a) image processing, (b) 3D reconstruction, and (c) 3D skeletal reconstruction and tracking.

a) Image Processing

Following the implementation of the off-the-shelf image processing algorithms into our system, we analysed their performances qualitatively, and in terms of computational speed. In the following sections we summarise the results of these analyses.

Figure 5 shows example outputs of the segmentation algorithms in a manufacturing setting. **Mask R- CNN** was shown to perform well in detecting and segmenting different object instances. However, the algorithm misclassified some objects such as the detection of a robot as a fire hydrant. **DensePose** appeared to perform well on the task of detecting and segmenting humans. Occasionally the algorithm misclassified robots as human. **DeepLab** was generally able to classify humans, but the segmentations were noisy and robots were also misclassified.

Figure 6 shows example outputs of **OpenPose** in the manufacturing setting. As opposed to the previous algorithms, this algorithm identifies only human body parts, and connects them with lines. As it can be seen from Figure 6, not all images have a correct body structure. In some cases, e.g. 2nd top panel in Figure 6, non-human objects are mistakenly identified as humans.

Figure 5. Segmentation algorithms applied to a manufacturing setting. Left: Mask R-CNN [2]; Center: DensePose [4]; Right: DeepLab [1]

Qualitative assessment of distributed deep learning processing

Figure 6 OpenPose tested in the cell. The algorithm detects body parts and connects them creating a skeleton-like structure.

Computational speed of distributed deep learning processing

In order to assess the potential of using the off-the-shelf algorithms for real-time processing, we ran the algorithms on different hardware CPU and GPUs. The results are shown in Figure 7 as frames per second (fps). Timings for DensePose were only captured on the K80. The results showed that the most recent GPU, the 2080 ti, DeepLab and OpenPose were able to process images at a speed of 77 fps and 50 fps, respectively, when the image resolution was reduced to 512x512. The results also illustrate the trend of GPUs allowing for increasing processing speeds as the technology

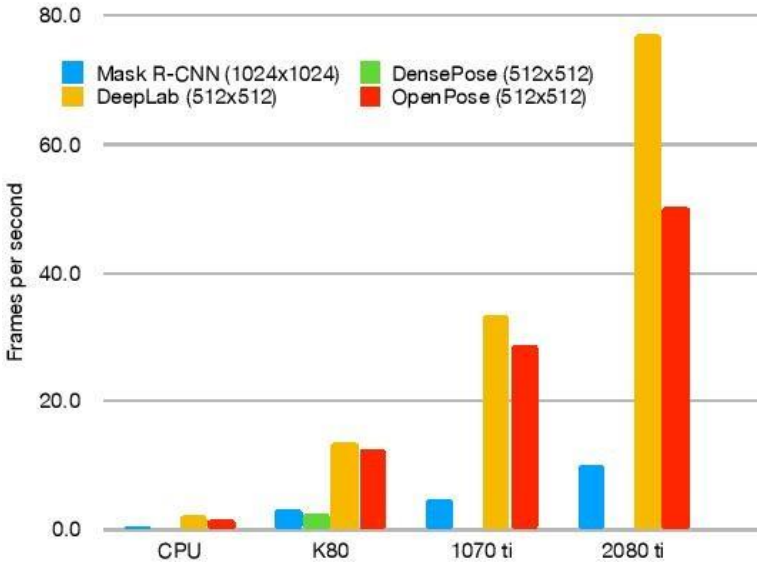

Figure 7. Processing speeds of algorithms using different hardware advances.

b) 3D reconstruction

Figure 8 shows the results of the 3D reconstruction using segmentations from Mask R-CNN, DeepLab, and a manual segmentation, respectively. Using a space carving algorithm in Matlab, the 3D voxel grid can be generated in under 0.1 seconds. The results illustrate the potential of our pipeline to provide a near real-time localization of objects in 3D space with an estimate of their size and shape.

c) 3D Skeletal reconstruction and tracking

OpenPose detects human body parts in a 2D image and attempts to reconstruct a structure of a person by joining related body parts with lines as previously shown in Figure 6. In our setting, we combined such body structures from each camera to reconstruct their 3D positions. Thus, people in the cell can be located in 3D based on the locations of their key body features. An example is shown in Figure 9.

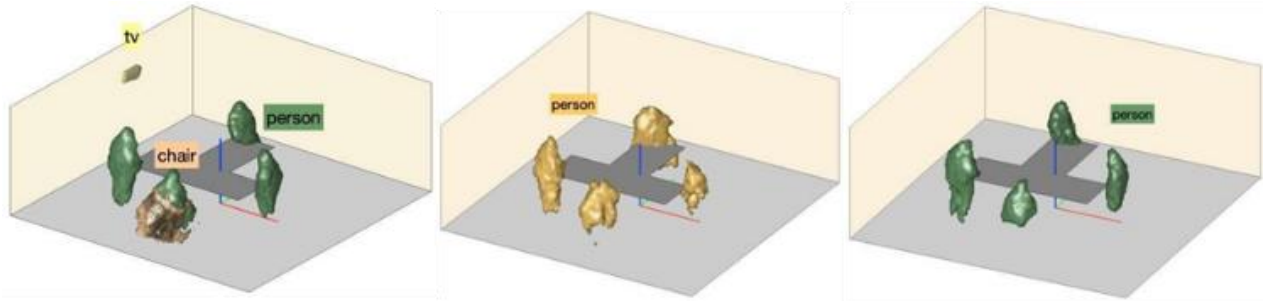

Figure 8. 3D reconstructions using segmentations as input. Left: 3D reconstruction using Mask R-CNN segmentations. Center: 3D reconstruction using DeepLab segmentations. Right: 3D reconstruction using manual segmentations.

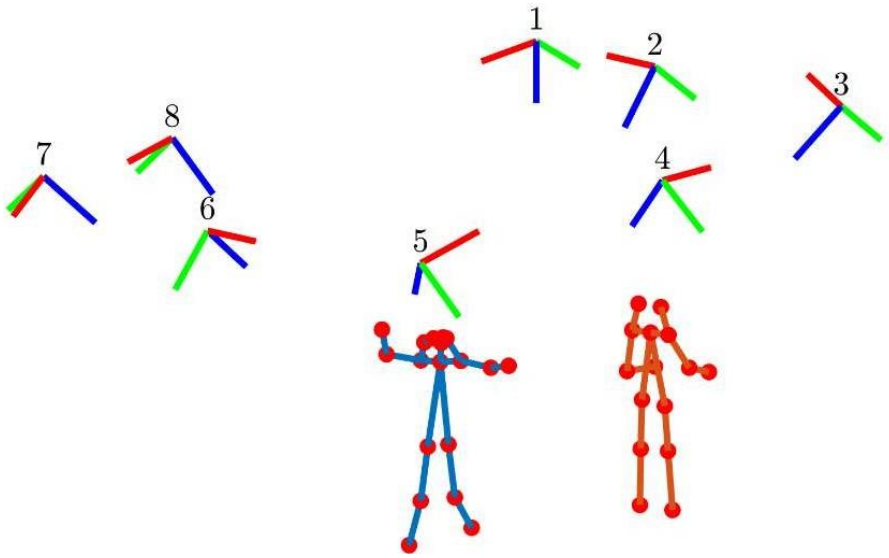

Figure 9. 3D reconstruction of body structures from OpenPose 2D outputs.

6. Conclusions

In this feasibility study, we investigated the possibility of real-time digitalisation of 3D manufacturing environments from a set of 2D standard RGB cameras.

The work presented in this report involved (1) the requirement analysis and purchase of the required equipment, (2) the construction of the hardware infrastructure at the Intelligent Automation Centre (Loughborough University), (3) the preparation of the communication network and software environment, (4) the preparation, testing and deployment of the deep learning algorithms and reconstruction as presented in this document, (5) the demonstration of the working prototype to industry.

The study allowed us to assess the potential of this approach experimentally for digitalisation of manufacturing environments. In particular, we validated the hypothesis that current technology allows for the localisation and tracking of a variety of objects and people in a 3D space from a set of 2D cameras. The hardware and software developments also indicate a fast increasing potential for higher precision, fast frame rate and reduced costs in the near future. One important finding is that off-the- shelf deep learning algorithms do not detect uncommon objects that may be of interest in specialised manufacturing environments.

Another important result of this feasibility study is the established collaboration and cross-pollination of ideas between the IA centre and Computer Science Department. The utility of this collaboration is becoming increasingly more important with the convergence of technology, from digital and data-driven artificial intelligence to engineering, manufacturing and metrology processes.

7. References

[1] Chen L.C, Papandreou G, Kokkinos I, Murphy K and Yuille AL. 2018. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". IEEE Trans Pattern Anal Mach Intell. **40**, (4), 834–848.

[2] He K,Gkioxari G, Dollar P and Girshick R. 2017. "Mask R-CNN," Proc. IEEE Int. Conf. Comput. Vis., vol. 2017 Oct, 2980–2988.

[3] Cao Z, Hidalgo G, Simon T, Wei SE and Sheikh Y. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," Dec. 2018.

[4] Güler RA, Neverova N and Kokkinos I. "DensePose: Dense Human Pose Estimation In The Wild," Feb. 2018.

8. Feasibility study team members

The study was conducted by a team of researchers from **Loughborough University.**

Dr Niels Lohse, Intelligent Automation
Dr Peter Kinnell, Metrology
Dr Andrea Soltoggio, Lifelong Learning Machines
Dr Ella-Mae Hubbard, Systems Ergonomics
Dr Joanna Turner, Deep Learning Algorithms
Dr Istvan Biro, System Architecture
Dr John Hudgson, Camera network algorithms and triangulation algorithms.

connected everything.

industrial systems in the digital age

Connected Everything
Faculty of Engineering
University of Nottingham
University Park
Nottingham
NG7 7RD
UK.

www.connectedeverything.ac.uk