Event Analysis and Interpretation of Human Activity for Augmented Reality-based Assistant Systems

Alexander Bannat, Jürgen Gast, Gerhard Rigoll, Frank Wallhoff
Technische Universitaet Muenchen
Department of Electrical Engineering
and Information Technology
Institute for Human-Machine Communication
Theresienstr. 90, 80333 Munich, Germany

Abstract

In this paper a concept and its implementation of an ergonomic cognitive assistant system for supporting human workers at complex assembly tasks in industrial environments is introduced. Depending on the level of the user's product knowledge this mixed-initiative system follows and gains knowledge from the human worker's construction steps while it is also able to automatically give hints and instruct the worker whenever needed. The presented agent bases on a closed human-machine interaction loop consisting of the multimodal perception of the worker's action, the comparison with the system's knowledge about the production task, and the displaying of the adequate next assembly instruction step. First experimental results of the assistant system are demonstrated on a simplified use case with the construction of a small toy car using augmented reality display techniques.

1 Introduction

Material goods on todays production lines are usually directly manufactured according to their customers' demands, i. e. a car's interior with a certain radio model etc. Thus, a high demand for flexible systems and reliable workers immediately arises. Furthermore, the provision and storage of a huge number of parts and devices represents an additional and challenging aspect and increases the complexity of production processes [12], which are all embedded into the so-called Cognitive Factory. The Cognitive Factory [10] is a demonstration scenario within The Cluster of Excellence CoTeSys – "Cognition for Technical Systems" [2]. Besides a central control unit and an automated stock for parts, the factory consists of three different assembly stations [5]:

- fully automated assembly by machines only
- hybrid assembly with joint actions between humans and machines
- pure manual assembly with humans only

The main aspect of this factory is that all deployed systems and units work in a cognitive manner, which means that they are self aware, they can adapt to new situations, and they can learn from observed knowledge. "More specifically, a technical system becomes a cognitive system, if it can reason substantial amounts of appropriately represented knowledge, learn from its experience so that it performs better tomorrow than it did today, explain itself and be told what to do, be aware of its own capabilities and reflect on its own behavior, and respond robustly to surprise." [7]

In the following we concentrate on the design and implementation of an ergonomic assistant system for the human worker in one of the earlier mentioned hybrid or manual assembly cells, using his actions and gestures as inputs.

A well equipped smart working environment will optimize a worker's productiveness and reduce his acquainting time for novel products with small batches but highly customized and varying models, which is an important economic aspect especially in high loan countries.

Such a smart environment is capable of giving assistance based on the actual work step, the human worker is currently fulfilling. This will further allow a flexible work plan where both – the machine or the worker – can take the initiative to identify the next best assembly or construction step. In order to not patronize or disturb the human worker, such a system has to work in the background and analyze the observed scene. Furthermore, manufacturing errors can directly be recognized and corrected, thus pushing down the margin for errors. The main components of the proposed cognitive system and its human-machine interaction loop therefore are:

- Action tracking and recognition of the human worker

From: Proc. of ICCP 2008, Cluj-Napoca, Romania, IEEE

(perception and self awareness)

- Prediction of next assembly step based on observed events; prepare assistance if worker halts (learning, reasoning and knowledge about the production task)
- Display instructions to help the worker in fulfilling his task (action)

The paper is organized as follows: in section 2 the setup of the assembly workbench is summarized. The introduction of the deployed input modalities follows in section 3. The used techniques for generating events and representing knowledge are shown in section 4. The output modalities are presented in section 5, followed by the description of a concrete use case in section 6. The paper closes with conclusions and an outlook.

2 Assembly Workbench

The assembly station mainly consists of a workbench for manual assembly. The centered area on the workbench is mainly foreseen as space for mounting the parts. Parts for the assembly process are stored in several storage boxes, which may be moved along the surface of the table. Required components can be fetched from these storage boxes. Visual feedback is given over a monitor and artificial reality techniques, such as a table projection and a head mounted display (HMD). Direct interaction with the system can happen over the mouse or a touchscreen. Figure 1 depicts the actual setup of the workbench in our lab. The major image shows the workbench. It is equipped with a touchscreen, a standard PC mouse, storage boxes, a P5 data glove (described in subsection 3.2) and a touchpad (left to right). The upper left image shows a PMD-Camera (depicted in subsection 3.3) and a video projector. The upper right image shows a top-down-view camera.

Due to their importance for the subsequent processing steps these multimodal inputs will be introduced in more detail.

3 Input Modalities

To be able to monitor human actions, the workbench features passive as well as active input modalities. The vision-based and body-mounted sensors are used to track events in the background compared to the input devices, which allow direct manipulation e.g. of presented information. Observations by all inputs and sensors generate events that have an impact on the computation of the actual work-piece status.

3.1 Top-down view Camera

A global top-down view camera is mounted above the workbench. This device has the overview over the entire work-area and makes it possible to watch the actions on the workbench and locate objects on the surface.

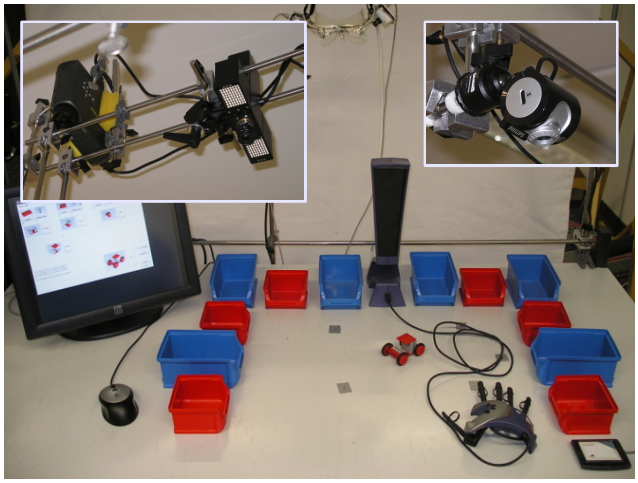

Figure 1. Workbench-Setup: ceiling with PMD-camera, table projector (upper left) and webcam (upper right); work-area with touchscreen, PC mouse, storage boxes, P5 data glove and touchpad (main picture, left to right).

3.1.1 Box detection

To give the right assistance at the right time, the system has to know, what parts have been taken out of the boxes on the workbench. This requires to find these boxes first, which is done by analyzing the acquired image from the top-down view camera.

In our setup different boxes are used, varying in color (red and blue) and size (nearly square and rectangular). To detect their position in pixel coordinates, a color-based image segmentation is performed on the camera picture. Thresholding filters are used in the HSV-color-space to extract the relevant areas and purge background information [1].

After the filter operation, the image is converted into binary masks for each box color. To determine whether the areas are boxes or not, the size of the found regions have to fit into an allowed height-to-width value of 0.5 to 2. The binary region extracted from the filter mask using connected-components analysis is further on called ”blob”. The blob of a box has also certain features, for example the number of non-zero pixels within the blob-borders. The system identifies blobs as boxes, if their size is in the interval between 1% up to 5% of the image-area. By these constraits it can be determined, if an area is a box or not.

Figure 2 shows a typical result of the box detection module on the workbench. The center of gravity for each box

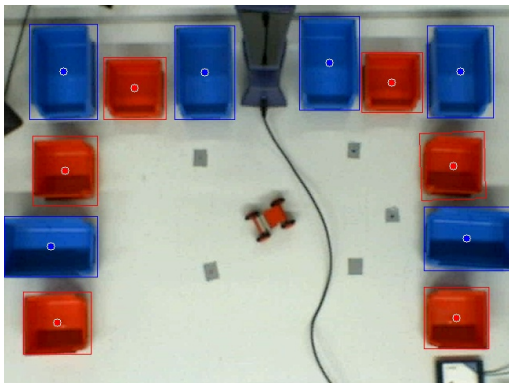

Figure 2. Image-based workbench observation: box detection

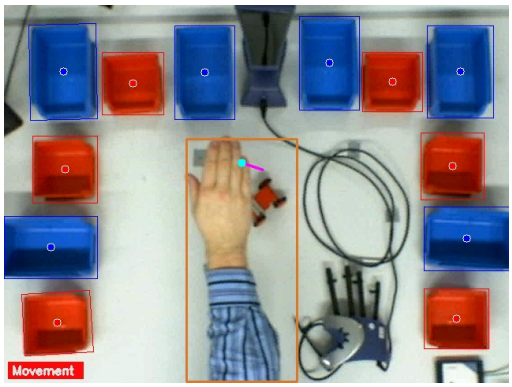

Figure 3. Image-based motion analysis for hand movement detection: area with motion (brown rectangle), hand motion-history (magenta line) and the approximated hand position (cyan dot)

is calculated and illustrated as a dot with the corresponding box-color.
By enabling the automatic recognition of the box coordinates, the free positioning becomes possible, which is an important factor for the ergonomic layout of the table; the worker can decide where he wants to place the boxes. One could assume, that a frequently used part is stationed closer to the work-area, as a rather rarely required one.

3.1.2 Hand detection

The system is capable of detecting where the storage boxes with the parts are located on the workbench. In a next step it can be detected, which parts are currently used for assembly. Therefore it is necessary to know, where the hand of the human worker is located.

Motion-Detector Module In the context of image-based hand-tracking, not every image-area has to be analyzed. Hands can only be detected in regions, where the image has changed during an analysis-period. To crop the relevant area, a motion-detector has been implemented. If the value of the motion energy in the image increases over a threshold of 0.5% pixel units, the detector sets the corresponding areas in a binary mask to logical `true`. Bringing the mask and the real image together, only the changed regions in the image-space are visible and will further be analyzed. The unchanged areas are purged. Figure 3 shows the result of the motion analysis. The brown rectangle marks the areas, which have been changed during the last analysis-period. The magenta-colored line indicates the motion history of the hand. An additional condition defining possible hand movements in image coordinates allows to improve the motion-detector and obtain more robust results against distortions. The current hand position is finally approximated by the most up and centered position in the rectangle, drawn in Figure 3 as a cyan-colored dot.

Modality skin color based hand-tracking Another modality for detecting the worker's hands is by applying a skin color filter operation to the image. A similar color-segmentation filter, as for the boxes is used. It is possible to find areas in the image, containing only human skin-color. The parameters of this filter operation are adjusted according to the skin color model of [6] in the rg-Chroma color space. Due to our setup, only hands are visible to the camera. Because hands have a certain height-to-width value of about 0.75 to 1.3, regions with other values can be purged. Thus, the resulting mask has only hand related blobs in it. Again, the center of gravity of these hand-blobs are calculated, representing the actual position of the hand. Combined with the motion-detector, the value for the extracted location of the hand in image coordinates can be improved.

3.2 Data Glove

Another modality for tracking the position of the worker's hand is the usage of a P5 data glove from Essential Reality [3]. The glove is equipped with active infrared LEDs. These emit pulsed infrared flashes, which are tracked by IR-sensors installed in the base unit of the device, see Figure 1. The detected flashes are used to calculate the position of the glove in 3D-coordinates.
In addition to the position, it is possible to get data about the bending of each finger. The glove has resistance strain gauges for every finger. Bending the hinges results in a changed value of the gauges, corresponding to the bending angle of the fingers. This sensor information is used to detect if the worker has grasped a part or laid one down.

3.3 Photonic Mixer Device (PMD)

The Photonic Mixer Devide (PMD) emits infrared light and acquires depth information based on the time-of-flight principle. This depth information can be used to make the observation task faster and also more reliable [4] compared to most vision algorithms which use only planar data for segmentation of an arbitrary scenery. In [9] we point out, how such a segmentation is done. The result of the segmentation of the worker's arm is depicted in figure 4. The

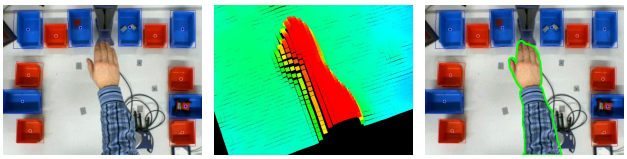

Figure 4. Segmentation of the worker's arm: raw color image (left), color coded depth map (middle) and in green highlighted contour of the worker's arm (right).

distance between the camera and an object is directly accessible through every pixel value. The "image" of a PMD-Camera has to be understood as a depth map. This sensor can therefore be applied to track a person's hand or recognize three-dimensional object parameters, e.g. height information.

3.4 Standard Human Interface Devices (HID)

As an active input modality of the table, a finger touchpad (comparable with notebook-pads) is used. This device has a fixed position on the table and is accessibly in every situation. For more precise operations, like zooming into details on the screen, a standard PC mouse is also connected to the system.

3.5 Touchscreen

A second active interaction unit is a touchscreen. This device makes it possible to directly navigate through menus. Simple menu structures allow an intuitive way of getting special information. For example, the last step can be reviewed for quality check reasons and so on.

4 Knowledge Representation and Event Analysis

The focus of this work lies on the surveillance of human worker actions without restraining his flexibility in completing his assembly task. The overall assembly task is to build up a toy car, depicted in Figure 5.

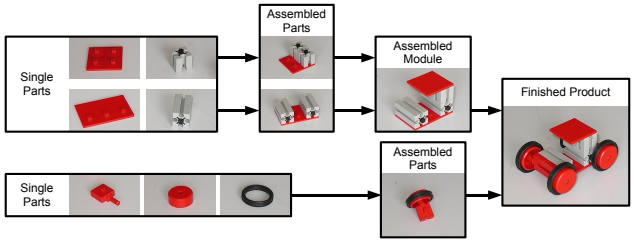

Figure 5. Work-piece: toy car. To get the finished car, three modules have to be built first. These intermediate modules can consist either of single parts or also pre-built modules.

This toy car has several different parts which have to be mounted together. It is also possible to assemble it in more than one finite way – meaning the assistant system has to be capable of reacting on different construction strategies.

4.1 Task knowledge representation

The first important module is a representation of the actual production task. To decide how the parts have to be mounted together, an expert analyzes the work-piece and defines the necessary assembly steps for task completition. Required intermediate steps have been carried out, labeled as "Assembled Parts" and "Assembled Module" in figure 5. These stages have to be passed to complete the task. This task knowledge is compiled into a "knowledge-system". Two possible implementation variants have been included in our setup.

4.1.1 Tree-based dependency representation

The first dependency representation is a tree-based form. Every interim step only depends on the assembly steps done before. By checking, if the previous parts have been fetched or the previous modules have been completely built, the system is able to distinguish, whether the actual module can also be finished or not. If some parts are missing, the assistant system can advise the worker, what other parts one needed.
The single parts are the base units for the assembly task. The worker is able to start with what ever part he likes to. The system adapts to the taken parts and is capable of tracking the production stage the worker currently is in.

4.1.2 Prolog-Tasker

Another way of implementing this task knowledge is by representing the part-dependencies with a logical programming language, like Prolog. This system has the advantage,

that it is able to do reasoning. The output of such a system can generate an event, telling if two parts can be mounted together, or not. An example of such a representation looks like the following pseudo-code:

```
join_able(A, B) :-
partPresent(A), partHasMountport(A),
partPresent(B), partHasMountport(B),
mountPortConnection(A, B).
```

Evaluating this command results in a boolean value, defining whether part `A` and `B` can be mounted together or not. Furthermore, the system can find impossible paths in the production cycle and avoid falling into them. Based on this information, the assistant system could display a warning message for the worker, saying something like: "Please check your current assembly step! You might run into problems, if you continue in this way."This prevents the worker from doing false construction step.

4.2 Scene representation

As introduced, several tracking systems and modalities are available. Each of them works with its own local coordinate system. For comparing the different data, these coordinate systems have to be transfered into a common world system. The worker is handling the items on the table. This is why the table surface is used as the base coordinate system.
Taking into account that the P5 data glove delivers 3D-coordinates, whereas the camera only has 2D-coordinates, a rather complex mathematical dependency between the devices arises.
To bring all these systems together, a general method has been conceived to calculate the individual dependencies. The basis for this computation is the selection of several predefined points on the table surface. The underlying condition is, that every tracker has to be able to detect these points without occlusions. Using mathematical descriptions among each tracker- and table-coordinate system allows e.g. the coordinate comparison of the same points between different coordinate systems. Figure 6 shows an example of the 3D P5 data glove coordinates mapped into 2D camera coordinates. The green-colored dots represent 2D-webcam coordinates. The corresponding 3D coordinates of the P5 data glove are plotted in red. The mapping accuracy is about 3 to 5 Pixels into each direction, which is precise enough for the desired recognition of actions.

4.2.1 Sensor fusion for hand-over-box-event

The presented methods for hand detection in section 3.1.2 allow the system to get accurate data of the location of the worker's hand. The positions of the boxes on the surface of the workbench can be gathered from the box-tracker, presented in section 3.1.1. The next step is to generate a *hand-over-box-event* based on the hand and the box position. The distance between their centers of gravity is calculated with the city block distance for 2D-points:

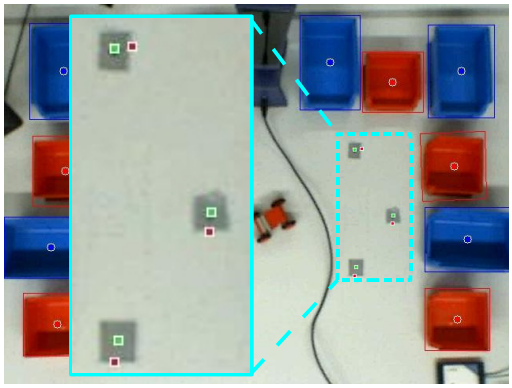

Figure 6. Mapping of two different coordinate systems into one common base system: camera coordinates (green, light) and P5 data glove coordinates (red, dark). The area in the cyan rectangle was digitally enhanced.

$$d_{\text{box2hand}} = (x_{\text{box}} - x_{\text{hand}}) + (y_{\text{box}} - y_{\text{hand}}) \qquad (1)$$

The indicator telling if a *hand-over-box-event* has occurred follows the condition:

$$\text{hand\_over\_box} = \begin{cases} true, & \text{if } d_{\text{box2hand}} \leq 50 \text{ pixels} \\ false, & \text{else} \end{cases} \qquad (2)$$

The value 50 pixels results of the approximated diameter of a box in pixels. In figure 7 the allocation of a box to the hand position is depicted. The accordant box is highlighted with green color. At this point, it is possible to recognize over which box the worker's hand is hovering. This information is used to generate the grasp-event, described in the following paragraph.

4.2.2 Sensor fusion for grasp-event

To be able to detect, if the worker has grasped a part, it is necessary to analyze the bending of his fingers. In our experiments the P5-handler was set to send the *grasp-event*, when all fingers were bent by 45°. The second condition is, that the hand has to be in a storage box to grasp a part. For that purpose, the hand has to be beneath a certain threshold relatively to the table surface. The elevation $\text{height}_{\text{hand}}$ is detected with the PMD-camera. The heights

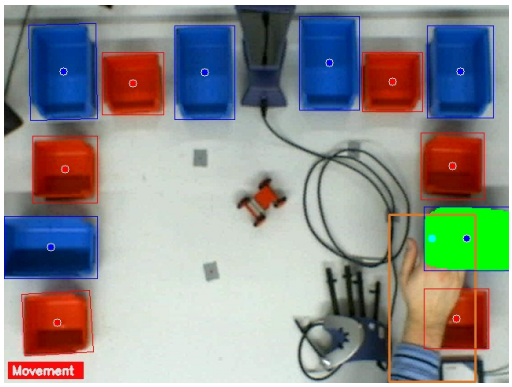

Figure 7. Visualized position of the worker's hand, hovering the accordant box (highlighted in green).

of the boxes ($\text{height}_{\text{box}}$) are stored in an initial scan of the surface. The subtraction of the relevant box-height and the hand-elevation is an other measure:

$$d_{\text{heights}} = \text{height}_{\text{box}} - \text{height}_{\text{hand}} \quad (3)$$

If the difference d_{heights} between the two values is less than $\frac{1}{2} \cdot height_{\text{box}}$, then the hand is in the selected box. This status combined with the value of the finger bending concludes to the *grasp-event*.

4.3 Knowledge base with box-part association

The system has to have knowledge about which storage box contains which part of the toy car. In an initial step, the worker has to load the boxes manually with all available and needed parts for the assembly task. The assistant system is monitoring this process in a calibration phase. That step enables it to generate a knowledge base containing the needed associations of boxes to assembly parts. The detection of the *lay-down-event* is the inverse operation to the *grasping-event*, described in section 4.2.2. If a hand is over or in a box and the system is in the calibration phase, then the deposed part will be assigned to the box. This is done for every single part. This operation also has to be performed on the boxes which are reserved for intermediate modules.

5 Output Modalities

Giving assistance requires facilities for presenting helping instructions or details on the task. Thus, three different kinds of outputs are integrated in the setup.

5.1 Flatscreen

Gathering information over the visual sense is very efficient. The worker can access detailed information about the assembly steps over a standard TFT-monitor. Furthermore, it is possible to display the progress of the built-up work piece. This device can be regarded to as a central display unit. It could display status flags and so on. The worker can decide whether he takes care of the presented information or not.

5.2 Head-Mounted Display (HMD)

One important fact is that the worker should be able to get important information instantly, with no respect to place he is currently looking at. A HMD is independent from the view angle because it is mounted in front of the eye and projects visual information directly in the worker's retina, as shown in Figure 8. The HMD delivers critical information directly to the work-space in an easy to use head-up system, i.e. the contact analog highlighting of destinations or important objects [8].

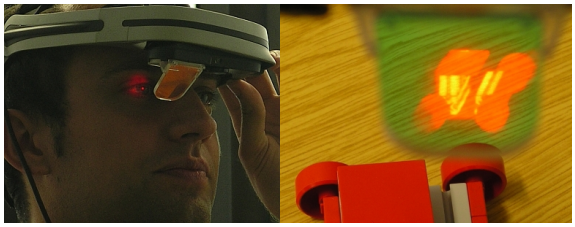

Figure 8. Microvision ND2000 HMD (left) and displayed construction information in the worker's field of view (right).

Due to the reason of discomfort during the experiments, this modality was discarded and replaced by a more ergonomic display strategy.

5.3 The table projection

Another promising way of bringing information into the worker's field of view is by directly projecting information onto the surface of the workbench. This is done with a video projector mounted above the work-area. With this modality, it is possible to show contact analog instructions, e.g. project a "taxiway" toward a storage case or even highlight needed parts.

6 Use Case and First Results

As a demonstration scenario, we chose the construction of a small toy car. This car has already been introduced

in section 4. At each assembly step the system works in the background. It only becomes active, when the worker needs assistance. This assistance can either be requested by the human or determined by the system, e.g. when grasping the next part takes more time than expected. In both cases, the box, containing the next required part, is highlighted with the table projector. With this method, the worker can instantly see from where he has to fetch the next part. An example for this technique is shown in Figure 9.

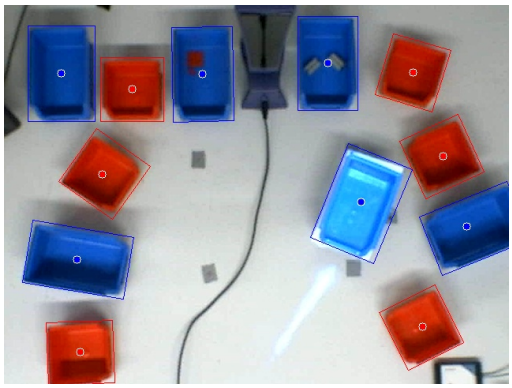

Figure 9. Photo of the running system, showing one highlighted blue box.

It becomes obvious, that the boxes have been freely rearranged. The highlighted area has been automatically adapted to the new position of the box. With the implemented table projection system the assistant system is able to use contact analog display techniques bringing objects into the visual focus of the worker.
After the highlighting of the box with the required part the system waits a predefined period of time, here five seconds. If the worker wont grasp the highlighted part within this period of time, an additional artificial pointer is also projected onto the workbench. Pointing at the location of the highlighted box shifts the worker's attention towards it. Figure 10 depicts a typical situation of the working demonstrator. In the small image, the touchscreen is depicted. One can see the highlighting of the blue box and the artificial pointer directing to it.
Furthermore, the user has the opportunity to highlight a box manually. As it is further shown in the last figure, the touchscreen is provided with a Graphical User Interface (GUI). For each part or intermediate module, an icon is visible. If the worker wants to see where a specific part is currently located, he simply has to press the according icon on the touchscreen. By the projection the storage box will then be highlighted.

To make such an on-line assistance work, all agents have to deliver their measurements in real-time. Therefore, the

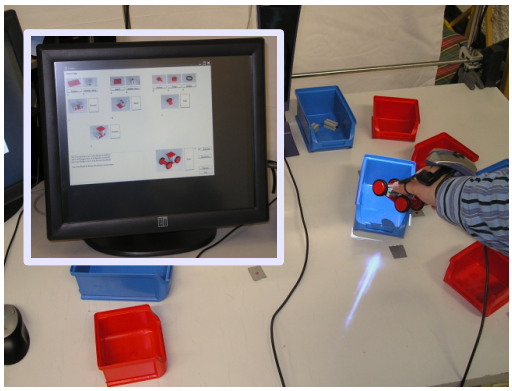

Figure 10. Photo of the running system: touchscreen showing the helper-GUI (overlayed picture), artificial pointer directing to a highlighted box while the worker is disposing the toy car.

above shown components are kept simple in their complexity in order to guarantee robust measurements and a computation in real-time.

7 Conclusions

In this paper an assistant system for the detection and interpretation of human activity has been presented. The observed activities are transformed into high-level events, which again are analyzed in the context of a manual assembly scenario. In the presented use case a human worker has to build a simple toy car. While performing this task, he is monitored by the assistant system. As soon as he stops working, the system is able to react on his behavior. Based on the activity analysis, it is possible to give help instructions. To enable the worker in fulfilling his work, a tree structure-like representation of the assembly steps has been implemented. This form of knowledge representation allows to track every stage of the production plan. Therefore, an adaptable guidance is guaranteed without a static predefined plan. With an augmented reality table projection system, contact analog display techniques make it possible to visually highlight storage boxes on the workbench and attract the workers visual focus.

8 Outlook

In parallel to the expansion of the tracking systems and the implementation of non-invasive hand gesture recognition, the production task will be extended to the assembly of complex robot parts.

The output data of our reasoning unit will be generalized, enabling human-robot collaboration scenarios. With the underlying concepts it is possible to let the human worker perform the assembly task in collaboration with an industrial robot in one common work space.
Furthermore, the knowledge representation unit will be enhanced by learning capabilities. Thus, it is possible to adapt the system to the human worker's preferences and his permanently evolving working-skills.

9 Acknowledgments

This work is supported in part within the DFG excellence initiative research cluster *Cognition for Technical Systems — CoTeSys*, see also `www.cotesys.org`. Furthermore, the authors would like to thank their research partners Anna Schubö, Sonja Stork, Florian Friesdorf and Mathey Wiesbeck from the ACIPE project [11].

References

[1] A. Bannat. Development of a Videobased Traffic Sign Recognition System. Master's thesis, Faculty of Electrical Engineering and Information Technology, Technische Universität München, Germany, Nov. 2006. in German.

[2] M. Buss, M. Beetz, and D. Wollherr. CoTeSys - Cognition for Technical Systems. In *Proceedings of the 4th COE Workshop on Human Adaptive Mechatronics (HAM)*, 2007.

[3] Essential Reality. `http://www.vrealities.com/P5.html`, 2008.

[4] T. Möller, H. Kraft, J. Frey, M. Albrecht, and R. Lange. Robust 3D Measurement with PMD Sensors. In *Proceedings of the 1st Range Imaging Research Day at ETH Zurich*, page "Supplement to the Proceedings", Zurich, Switzerland, 2005.

[5] S. M. Saad and M. D. Byrne. Comprehensive Simulation Analysis of a Flexible Hybrid Assembly System. In *Integrated Manufacturing Systems*, volume 9 of *Emerald Group Publishing Limited*, pages 156–167, 1998.

[6] M. Soriano, S. Huovinen, B. Martinkauppi, and M. Laaksonen. Skin Detection in Video under Changing Illumination Conditions. In *Proc. 15th International Conference on Pattern Recognition*, pages 839–842, Barcelona, Spain, 2000.

[7] Technische Universität München. Homepage of the Excellence Cluster Cognitive Technical Systems. Internet Publication: `http://www.cotesys.org/`, 12 2006.

[8] F. Wallhoff, M. Ablaßmeier, A. Bannat, S. Buchta, A. Rauschert, G. Rigoll, and M. Wiesbeck. Adaptive Human-Machine Interfaces in Cognitive Production Environments. In *IEEE Proceeding ICME 2007*, pages 2246–2249, Beijing, China, July 2-5 2007.

[9] F. Wallhoff, M. Ruß, G. Rigoll, J. Göbel, and H. Diehl. Improved Image Segmentation Using Photonic Mixer Devices. In *IEEE Proceeding ICIP 2007*, volume VI, pages 53–56, San Antonio, Texas, USA, 16.-19.09. 2007.

[10] M. F. Zäh, C. Lau, M. Wiesbeck, M. Ostgathe, and W. Vogl. Towards the Cognitive Factory. In *Proceedings of the 2nd International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV)*, Toronto, Canada, July 2007.

[11] M. F. Zäh, M. Wiesbeck, F. Engstler, F. Friesdorf, A. Schubö, S. Stork, A. Bannat, and F. Wallhoff. Kognitive Assistenzsysteme in der Manuellen Montage. In *wt Werkstattstechnik online*, volume 97, 9, pages 644–650. Springer-VDI-Verlag, 2007.

[12] M. F. Zäh, M. Wiesbeck, H. Rudolf, and W. Vogl. Virtual and Augmented Reality. In *Proceedings of Virtual Concept*, 2006.