

Research Report RR-13-02



Cognitive Workflow Capturing and Rendering with On-Body Sensor Networks (COGNITO)

Final Report of the European project FP7-ICT-2009.2.1 – 248290

02/2013

Research Report RR-13-02

German Research Center for Artificial Intelligence (DFKI) GmbH

Editorial Board: Prof. Dr. Frank Kirchner, Prof. Dr. Prof. h.c. Andreas Dengel, Prof. Dr. Hans Uszkoreit, Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster

Bibliographic information published by the German National Library

The German National Library lists this publication in the German National Biography; detailed bibliographic data are available in the internet at <http://dnb.ddb.de>.

Editorial Board:

Prof. Dr. Frank Kirchner

Prof. Dr. Prof. h.c. Andreas Dengel

Prof. Dr. Hans Uszkoreit

Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster

©German Research Center for Artificial Intelligence (DFKI) GmbH, 2013

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of the German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern, Federal Republic of Germany; an acknowledgement of the authors and individual contributors to the work; all applicable portions of this copyright notice. Copying, reproducing, or republishing for any other purpose shall require a licence with payment of fee to German Research Center for Artificial Intelligence (DFKI) GmbH.

Issue RR-13-02 (2013)

ISSN 0946-008x

German Research Center for Artificial Intelligence
Deutsches Forschungszentrum für Künstliche Intelligenz
DFKI GmbH

Founded in 1988, DFKI today is one of the largest nonprofit contract research institutes in the field of innovative software technology based on Artificial Intelligence (AI) methods. DFKI is focusing on the complete cycle of innovation – from world-class basic research and technology development through leading-edge demonstrators and prototypes to product functions and commercialization.

Based in Kaiserslautern, Saarbrücken and Bremen, the German Research Center for Artificial Intelligence ranks among the important ‘Centers of Excellence’ worldwide. An important element of DFKI’s mission is to move innovations as quickly as possible from the lab into the marketplace. Only by maintaining research projects at the forefront of science DFKI has the strength to meet its technology transfer goals.

The key directors of DFKI are Prof. Wolfgang Wahlster (CEO) and Dr. Walter Olthoff (CFO). DFKI’s research departments are directed by internationally recognized research scientists:

- Knowledge Management (Prof. A. Dengel)
- Cyber-Physical Systems (Prof. R. Drechsler)
- Robotics Innovation Center (Prof. F. Kirchner)
- Innovative Retail Laboratory (Prof. A. Krüger)
- Institute for Information Systems (Prof. P. Loos)
- Embedded Intelligence (Prof. P. Lukowicz)
- Agents and Simulated Reality (Prof. P. Slusallek)
- Augmented Vision (Prof. D. Stricker)
- Language Technology (Prof. H. Uszkoreit)
- Intelligent User Interfaces (Prof. W. Wahlster)
- Innovative Factory Systems (Prof. D. Zühlke)

In this series, DFKI publishes research reports, technical memos, documents (eg. workshop proceedings), and final project reports. The aim is to make new results, ideas, and software available as quickly as possible.

Prof. Wolfgang Wahlster
Director

Cognitive Workflow Capturing and Rendering with On-Body Sensor Networks (COGNITO)

Final Report of the European project FP7-ICT-2009.2.1-248290

Project duration: Jan 1, 2010 – Dec 31, 2012

Co-ordinator: Prof. Dr. Didier Stricker, German Research Center for Artificial Intelligence (DFKI)

Project website address: www.ict-cognito.org



Table of Contents

AUTHORS	3
EXECUTIVE SUMMARY	5
PROJECT CONTEXT AND OBJECTIVES	6
PROJECT ACHIEVEMENTS AND RESULTS	7
1. SYSTEM ARCHITECTURE	8
2. USE CASE SCENARIOS AND TEST DATASETS	8
3. ON-BODY SENSOR NETWORK AND HEAD MOUNTED DISPLAY	10
4. LOW-LEVEL SENSOR PROCESSING.....	13
4.1. SCENE CHARACTERISATION AND MONITORING	13
4.2. USER ACTIVITY MONITORING.....	17
5. HIGHER-LEVEL MODELLING	20
5.1. WORKFLOW RECOVERY AND MONITORING.....	21
5.2. BIOMECHANICAL ANALYSIS	24
5.3. MONOCULAR WORKFLOW SEGMENTATION, AUTHORING AND MONITORING	26
6. USER INTERFACE	27
6.1. WORKFLOW EDITOR AND AR PLAYER.....	27
6.2. EVALUATION AND USER TESTING.....	30
7. SUMMARY AND CONCLUSION	32
POTENTIAL IMPACT.....	33
DISSEMINATION	34
EXPLOITATION	35
ACKNOWLEDGEMENTS	36
THE COGNITO CONSORTIUM.....	37
REFERENCES	38

Editor

Gabriele Bleser, German Research Center for Artificial Intelligence, Germany

Authors

Gabriele Bleser, German Research Center for Artificial Intelligence, Germany

Luis Almeida, Center for Computer Graphics, Portugal

Ardhendu Behera, University of Leeds, UK

Andrew Calway, University of Bristol, UK

Anthony Cohn, University of Leeds, UK

Dima Damen, University of Bristol, UK

Hugo Domingues, Center for Computer Graphics, Portugal

Andrew Gee, University of Bristol, UK

Dominic Gorecky, Technologie-Initiative SmartFactory KL e.V., Germany

David Hogg, University of Leeds, UK

Michael Kraly, Trivisio Prototyping GmbH, Germany

Gustavo Mações, Center for Computer Graphics, Portugal

Frederic Marin, University of Compiègne, France

Walterio Mayol-Cuevas, University of Bristol, UK

Markus Miezal, German Research Center for Artificial Intelligence, Germany

Katharina Mura, Technologie-Initiative SmartFactory KL e.V., Germany

Nils Petersen, German Research Center for Artificial Intelligence, Germany

Nicolas Vignais, University of Compiègne, France

Luis Paulo Santos, Center for Computer Graphics, Portugal

Gerrit Spaas, Trivisio Prototyping GmbH, Germany

Didier Stricker, German Research Center for Artificial Intelligence, Germany

Executive summary

The major goal of COGNITO was to develop enabling technologies for intelligent user assistance systems with focus on industrial manual tasks. This comprises technology for capturing user activity in relation to industrial workspaces through on-body sensors, algorithms for linking the captured information to underlying workflow patterns and adaptive user interfaces which use this higher-level information for providing tailored feedback and support through adaptive augmented reality (AR) techniques. The work was organised in four layers:

- Hardware layer: development of an on-body sensor network consisting of highly integrated visual and inertial units and a head-mounted-display (HMD) with integrated tracking capabilities and good ergonomics for industrial use.
- Low-level processing layer: research and development of computer vision and sensor fusion algorithms for capturing user activity and workspace structures including the position and handling of key objects, such as tools and parts.
- Higher-level processing layer: research and development of workflow recovery and monitoring algorithms as well as methods for biomechanical analysis of workers and assessment of stress levels associated to tasks.
- User interface layer: iterative design, development and evaluation of multimodal, user- and context-adaptive interaction and feedback mechanisms supporting and exploiting the available low- and higher-level information while following the general principles of usability.

The major contributions of COGNITO are in line with its execution objectives and can be summarised as follows:

- A new generation of precise wireless inertial measurement units (IMUs) and an HMD with integrated visual-inertial unit and gaze tracking apparatus;
- A novel method for dense 3D workspace reconstruction and robust camera tracking including fast relocalisation for agile movements based on colour and depth (RGBD) information; a novel learning-based method for detection and recognition of handled textureless objects; a complete multi-object detection and tracking framework built upon the previous components;
- A novel visual-inertial upper body tracking method using egocentric vision measurements for increased robustness against magnetic disturbances and, built upon this, an innovative system for online global biomechanical analysis and feedback according to the RULA (rapid upper limb assessment) standard;
- A detailed musculoskeletal model of the hand and forearm and a database of muscle forces and articular loads for typical industrial tasks enabling online biomechanical analysis;
- A novel method for hand detection and tracking based on a monocular RGB camera, which can be linked to the aforementioned hand model;
- Novel, domain-independent methods for workflow recovery and monitoring based on spatiotemporal pairwise relations deduced from scene features, objects and user motions, which handle workspace appearance changes and are robust against broken tracks and missed detections;
- Workflow authoring tools and user interfaces exploiting the low- and higher-level information for context-sensitive user feedback.

Besides the major advances in various key technologies as reported above, we have also developed an integrated system, which makes use of these key technologies and addresses the complete action-perception-feedback loop, linking data capture with workflow and ergonomic understanding and feedback. Moreover, we have developed and tested this system and its components on three increasingly complex test datasets based on typical industrial manual tasks and have thoroughly documented the results. Finally, we have also developed lightweight monocular workflow segmentation, authoring and monitoring tools and a demonstrator system based on this.

Project context and objectives

As the complexity of workflows and manual tasks in industry and production increases, there is an increasing need for intelligent user assistance systems. The goal of COGNITO was to develop the enabling technologies for these systems, i.e. technology for capturing information about the user and the environment through sensors, reasoning about this and providing tailored feedback and support through adaptive augmented reality (AR) techniques. Below, the COGNITO concept is introduced through a demonstrative use case scenario, followed by the concrete objectives of the project.

The envisioned COGNITO system initially learns a complex manual task by “observing” an expert user during demonstrations. Afterwards, the system is able to assist an inexperienced user in executing the same task. During the initial learning phase, the system captures the activity of the expert based on an on-body sensor network comprising inertial sensors (IMUs) and cameras, derives a meaningful

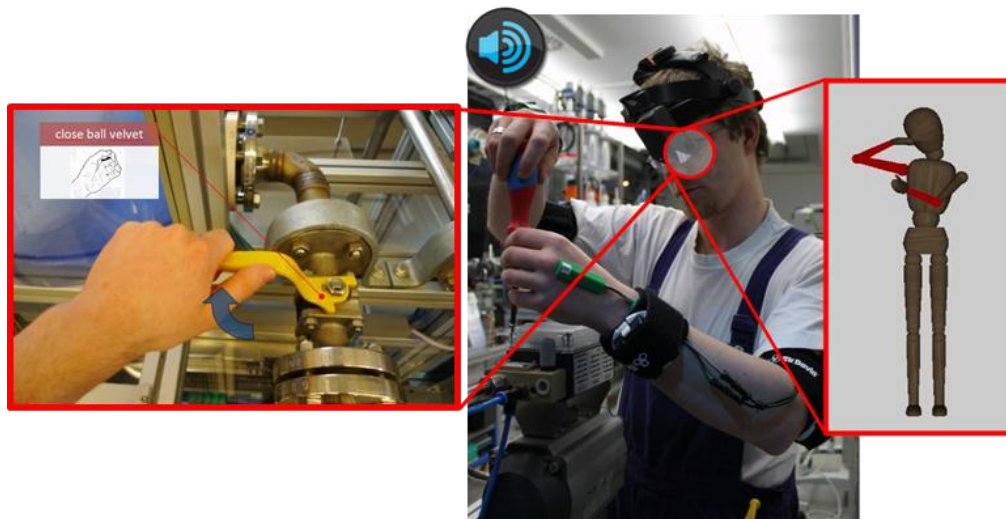


Figure 1 Support of an industrial maintenance task with audible and visual instructions and ergonomic feedback.

workflow model and extracts representative information for each action, such as video clips or prototypical images. Intuitive authoring tools then allow the expert to enrich the workflow model with further descriptive information, such as labels, text or graphics. During the subsequent assistance phase, the previously learnt statistical workflow model is applied to the data continuously captured from an inexperienced user when performing the same task. This allows monitoring the user, tracking the current workflow position, predicting the next steps and providing tailored support in a head-mounted AR display. At the same time, the system monitors the postures and movements of the worker and ensures an ergonomic behaviour by providing appropriate feedback (see Figure 1). Since the COGNITO system is completely mobile and building on generic and learning-based processing algorithms, it can be applied at any location and in any domain.

While advanced user assistance systems based on AR and VR technology became more and more common during the last years, in contrast to the concept described above, they were mostly based on hard wired content generated in a time consuming authoring process. Moreover, while such systems indeed may be adaptive to local conditions, they are in general not so intelligent as to be adaptive to the relations of both the current user activity and the environmental conditions. The goal of the COGNITO project was to address these shortcomings by *advancing the capabilities of human activity capture, learning and monitoring and by linking these cognitive capabilities to the user interface* layer in order to enable a genuinely intelligent assistance system. In detail, the following objectives were defined:

- To develop an on-body sensor network consisting of miniature inertial and vision sensors capable of simultaneous monitoring of user and task space activity in a mobile egocentric manner, avoiding the need for a static capture infrastructure;
- To utilise osteo-articular models of the human body to allow the capture of user pose and motions based on visual and inertial data from the sensor network;
- To develop computer vision algorithms for monitoring and recognising user activity relative to the task space, based on video data from the on-body cameras;
- To develop sensor fusion algorithms for integrating body pose and activity data with visual observations obtained from the sensor network;
- To develop a framework and algorithms for learning workflow patterns and task completion strategies based on the interpreted sensor data;
- To develop a framework, models and algorithms for global ergonomic and musculoskeletal assessment of a worker during task execution from the interpreted sensor data in order to prevent musculoskeletal disorders;
- To develop novel rendering mechanisms to allow effective visualisation and delivery of the captured workflow patterns in a user-adaptive manner;
- To develop and evaluate a demonstrator system which illustrates the capability of the technology in the context of a skilled industrial manual task.

Project achievements and results

The achievements of the COGNITO project are in line with the above stated objectives and concern novel developments on hardware, algorithm, software component and system level. The final integrated system architecture is shown in Figure 2.

The system is made up of several layers (sensing, low-level processing, higher-level processing, feedback) and components (user activity monitoring, workspace monitoring, biomechanical user activity analysis, workflow analysis), representing its different aspects and levels of information. On the lowest level, the on-body sensor network comprising visual and inertial units worn by the user provides information about both the user and the workspace (and interactions between both) in a completely mobile way. On the highest level, the user interface component provides – through a monocular HMD – tailored multimodal and hands-free user feedback (cf. Figure 1) driven by the workflow and biomechanical monitoring components in the higher-level modelling layer. The latter is decoupled from the raw sensor information (images, kinematics) and is fed by an intermediate low-level processing layer, which refines the raw sensor measurements into object-level information, such as features, workspace geometry, object positions and user’s egomotion, all fused in a global three-dimensional frame of reference. Decoupling higher-level analysis from raw sensory information increases the invariance of the system with respect to user and environmental conditions. Sections 3 through 6 describe the major contributions that the project has made in each of these components.

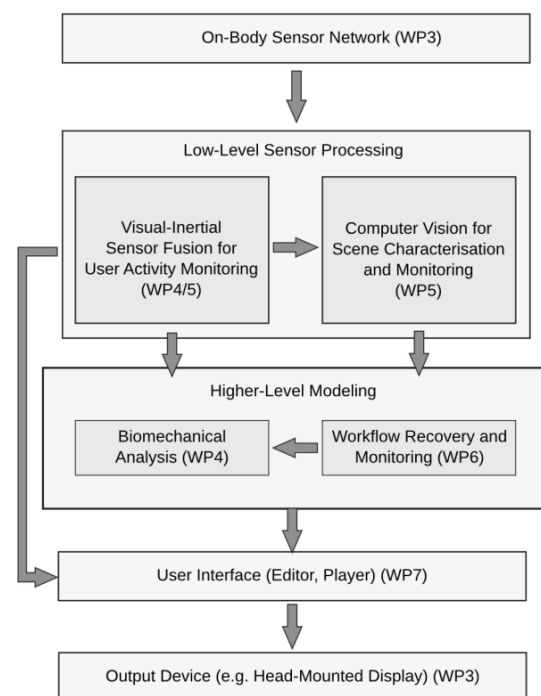


Figure 2 COGNITO system architecture in relation to the work-packages.

While the focus of COGNITO was on the enabling technologies, we have also developed an integrated system prototype according to the architecture in Figure 2, which covers the complete action-perception-feedback loop. Section 1 details the COGNITO system architecture including the information flow between the building blocks. Moreover, we have developed three increasingly complex use case scenarios based on typical industrial manual tasks and have captured datasets for integration and testing of the system and its components. These are outlined in Section 2. Finally, the development of the COGNITO system, in particular the user interaction and feedback components, were iterative and driven by continuous user testing. The evaluation activities and results are described in conjunction with the user interface in Section 6.

1. System architecture

As outlined in Figure 2, the COGNITO system consists of five major building blocks, which gradually refine the information obtained from the on-body sensor network (i.e. heterogeneous streamed sensor data) towards the user interaction layer, which offers rendering capabilities in an AR display. Moreover, two major analyses are performed in parallel, workflow analysis and biomechanical analysis. The building blocks are described in more detail below:

On-Body Sensor Network (BSN) and Output Device (cf. Section 3): IMUs, cameras, and RGBD sensors are combined in a sensor network worn on the body (cf. Figure 3). A monocular see-through HMD with integrated tracking capabilities provides the system feedback and user assistance information during workflow execution.

Low-Level Sensor Processing (cf. Section 4): processes the measurements from the BSN in order to provide information about the operator's activity (upper body/limb motion, hand postures/positions) and the workspace (features, positions of relevant objects) in a global workspace coordinate frame.

Workflow Recovery and Monitoring (cf. Section 5.1): processes the instantaneous information from the low-level sensor processing and provides an estimate of the current and a prediction for the next atomic action in the considered workflow model. During the learning phase, it builds a statistical workflow model from the received data.

Biomechanical Analysis (cf. Section 5.2): receives a sequence of postures and forces for the operator's upper body and hands and performs two biomechanical evaluations. An online global estimation is done based on the ergonomic tool Rapid Upper Limb Assessment (RULA) index. Assessing the RULA score online permits the worker to modify his posture in real-time when the movement leads to an important musculoskeletal disorders risk exposure. An offline local estimation of muscle forces and articular loads for hands and forearms is done based on a detailed musculoskeletal model.

User interface (cf. Section 6): provides the means for editing recovered workflows and enriching them with descriptive information as well as aiding the user during task execution with context sensitive user feedback using AR techniques.

2. Use case scenarios and test datasets

Table 1 provides an overview of the workflows and test datasets used by the consortium for integrating and evaluating the system and its components. In addition to these common datasets, each partner has also performed individual testing focused on their own research and system components. Figure 3 and Figure 4 show the on-body sensor network and snapshots from different data captures.

Table 1 COGNITO's workflows and common test datasets (increasing complexity from left to right).

Workflow	Nails & Screws	Labelling & Packaging	Ball valve installation
Workflow summary	Hammer 3 nails and fasten 3 screws onto a wooden piece.	Attach labels to two objects and package them within a box. Then seal the box and write the address.	Install a new ball valve and assemble the components of a pump system.
Remarks	<ul style="list-style-type: none"> • Simple operations • Magnetic disturbances 	<ul style="list-style-type: none"> • Bimanual manipulations • Complex N-wise relationships between objects 	<ul style="list-style-type: none"> • Bimanual manipulations • Complex operations • Many tools • Shiny and small objects • Magnetic disturbances
Objects (besides user's left and right wrist)	<ul style="list-style-type: none"> • Box • Wooden baton • Hammer • Screwdriver 	<ul style="list-style-type: none"> • Bottle • Box • Pen • Tape dispenser 	<ul style="list-style-type: none"> • Ball valve • Ball valve casing • Electrical positioner • Positioner covering • Connecting pipe • Screwdriver • Spanner
Atomic events	<ul style="list-style-type: none"> • Take/put box • Take baton • Pick hammer • Pick nail/screw • Hammer nail • Put down hammer • Pick screwdriver • Drive screw • Put down screwdriver 	<ul style="list-style-type: none"> • Pick/put bottle • Stick label • Pick/put box • Remove cover • Put bottle inside box • Take/put cover • Write address • Take/put sticky tape dispenser • Seal box 	<ul style="list-style-type: none"> • Pick and attach ball valve into base • Pick and attach bearing onto base • Fix casing with nuts and bolts (4) • Pick spanner • Tighten nuts (2) • Pick and put electrical positioner • Pick and fix positioner covering • Pick screwdriver • Fasten screws of electric cover (4) • Put down screwdriver • Attach electric positioner to actuator • Fix positioner with nuts (2) • Tighten nuts (2) • Pick and fix pipe • Remove cap • Fix cap to pipe • Attach pipe to the base
Captured data	<ul style="list-style-type: none"> • IMU data (3D acceleration, angular velocity, magnetic field) from 5 IMUs (chest, upper arms, forearms) at 100 Hz • RGB images from a chest-mounted fisheye camera at 20 Hz • RGB and D images from an overhead and a statically mounted sensor at 30 Hz 	<ul style="list-style-type: none"> • IMU data (3D acceleration, angular velocity, magnetic field) from 5 IMUs (chest, upper arms, forearms) at 100 Hz • RGB images from a chest-mounted fisheye camera at 20 Hz • RGB and D images from an overhead sensor at 30 Hz 	<ul style="list-style-type: none"> • IMU data (3D acceleration, angular velocity, magnetic field) from 7 IMUs (chest, pelvis, head, upper arms, forearms) at 100 Hz • RGB and D images from an overhead sensor at 30 Hz
Collected sequences	<ul style="list-style-type: none"> • 5 participants • 5-6 workflow executions each • 1 individual variation 	<ul style="list-style-type: none"> • 5 participants • 4-5 workflow executions each • Varying background clutter • Varying order of events (e.g. order of sealing and writing) 	<ul style="list-style-type: none"> • 6 participants • 4-5 workflow executions each

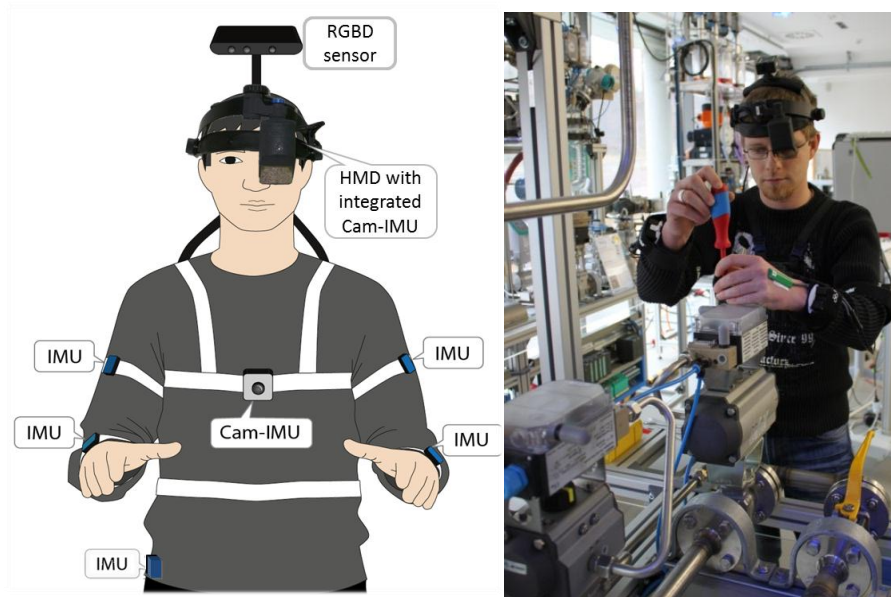


Figure 3 Final on-body sensor network as used for the last data capture (left) and snapshot from a user study evaluating the effectiveness of the online global ergonomic assessment in preventing hazardous postures (right).



Figure 4 Snapshots from the three common test datasets (left to right): Nails & Screws, Labelling & Packaging, Ball Valve installation.

3. On-body sensor network and head mounted display

The COGNITO hardware platform consists of two components, the on-body sensor network (BSN), which allows the COGNITO system to acquire information about the user in relation to the workspace for subsequent workspace monitoring and biomechanical analysis, and the HMD, which provides the means for hands-free user feedback in graphical and audible form. Hence, the BSN and the HMD represent the necessary input and output devices of the system. The complete setup is shown in Figure 3 (left). It consists of five IMUs (two on each arm and one on the pelvis), one visual-inertial unit on the chest, one visual-inertial unit attached to the HMD and one overhead RGBD sensor attached to a backpack.

While the RGBD sensor is based on a commercially available component (cf. Section 4.1), a new generation of wireless IMUs, a highly integrated visual-inertial unit (Cam-IMU), a monocular see-through HMD with integrated gaze tracking apparatus and a powerful processing unit enabling a completely mobile setup have been developed by Trivisio within the project, since no such components were available on the market.

ColibriW IMU

Each ColibriW IMU contains a tri-axial accelerometer, a tri-axial gyroscope, a tri-axial magnetic sensor and a temperature sensor. The IMUs feature an internal battery, which can be recharged via a USB port and provides the tracker with power. In addition, the USB port provides the possibility to flash the firmware and program the internal memory, e.g. by setting individual calibration parameters. For each built-in sensor a different calibration device is required. In order to ensure accurate data, a new semi-automatic calibration procedure including the compensation of temperature effects has been developed.



Figure 5 Inertial measurement unit (left) and USB transceiver (right)

The raw IMU data is transmitted over the 2.4 GHz band using a radio frequency protocol. Different channels can be selected in order to ensure an interference free transmission. The transferred data includes, for example, a time stamp, the measured accelerations, angular velocities and magnetic fields, all sampled at 100 Hz. A software development kit including an application programming interface for configuring and operating the IMUs as well as a basic orientation filter for fusing the raw sensor signals to an absolute rotation (given as roll, pitch and yaw angles with respect to a local magnetic north and gravity aligned coordinate system) has been developed by DFKI. Up to ten IMUs can be registered on one USB receiver dongle for synchronised wireless communication (see Figure 5). By using a second dongle, the network can be expanded even further. A synchronisation cable between two dongles ensures synchronous data. The ColibriW IMU generation has been commercialised by Trivisio early during the project and is now already used by several customers world-wide, among those the a big German car company.

Cam-IMU

A compact visual-inertial unit has been developed by integrating a small and lightweight camera with an IMU in the same housing (see Figure 6). The IMU is setup to trigger the camera. Hence, the camera images are hardware synchronised with the IMU measurements. The video data stream from the camera is transferred over a USB 2.0 interface. This port also provides the power for the IMU. The integrated IMU is registered on the ColibriW dongle for wireless communication and is thus synchronised with the rest of the IMU network.



Figure 6 Cam-IMU with strap fixation (left) and ball joint holding arm (right)

The Cam-IMU provides a rigid transformation between camera and IMU, which is calibrated once. At the same time two adjustable fixations have been developed. With the first one the Cam-IMU can be attached to the chest and the second fixation provides a way to attach the unit to the HMD, while keeping it adjustable using a ball joint on the holder. The final Cam-IMU version features a monochrome camera with SXGA (1280 x 1024) resolution and a fisheye lens providing visibility of the complete frontal hemisphere.

Head-mounted display (HMD)

According to an exploratory user study during the first part of the COGNITO project, a monocular see-through HMD was preferred as output device for hands-free user assistance. With such a device, as shown in Figure 7, the user can see a virtual image in front of his eyes above the real world. The image signal is transferred by DVI from a graphics card, e.g. as integrated into the wearable processing unit described below.



Figure 7 Monocular see-through HMD with integrated gaze tracking camera and attached Cam-IMU.

This monocular optical-see through HMD presents the system feedback and user assistance information in both graphical and audible form. The transparency/reflective ratio of the current beam splitter is 50%/50%. For the presentation of the virtual content on the display, a resolution of SVGA (800x600) with a frame rate of 60 Hz is used. The HMD has a diagonal field of view of 29° and the accommodation of the virtual image is 1580 mm from the user's eye. A microphone is built into the HMD, as well as a 3.5 mm stereo jack for connecting headphones.

As an additional feature, a mini eye-tracking camera has been integrated into a flexible gooseneck holder filming the eye from below. The eye region is illuminated by six infra-red LEDs. In order to reduce the disturbing influence of incident light, the lens of the eye tracking camera is covered with an IR-pass filter strip, which blocks the visible light facilitating recognition of the pupil.

Wearable Processing Unit

In order to enable a fully mobile setup, a wearable processing unit with a powerful Intel Core i7 2.67 GHz processor has been developed. The unit is mounted on the back of a weight lifting belt and offers good wearing comfort (see Figure 8). The computer is housed in a robust aluminum case and is therefore particularly suitable for everyday use in an industrial environment. The HMD can be directly connected to the computer via DVI. Six USB ports enable the inclusion of and power supply for all COGNITO input and output devices (e.g. HMD, ColibriW dongle, eye tracking camera). The unit is equipped with a 128 GB SSD (solid state drive) for software installation and can be operated as a normal computer. Also, the system features a wireless module with internal antenna.



Figure 8 Wearable processing unit on the back of a weight lifting belt.

4. Low-level sensor processing

The low-level sensor processing layer processes the raw measurements from the BSN (colour, depth images, IMU measurements) and deduces information about the operator's own activity in terms of upper body motion, hand positions and postures and the workspace configuration in terms of visual features and the identity and position of relevant objects, such as tools and parts, in a global workspace coordinate frame. The information is then provided to the next processing layer. The responsibility is shared between two components, both of which are described subsequently.

4.1. Scene characterisation and monitoring

Scene Characterisation and Monitoring was achieved by Bristol within the Computer Vision package (WP5). The purpose of WP5 is to develop the computer vision algorithms required to analyse video-data from on-body cameras. The key results can be summarised as follows:

1. The development of a real-time technique for fusing multi-frame depth measurements captured from a moving sensor. This allows the building of dense 3D structural models of large workspaces and segmentation of foreground objects.
2. The development of a fast localisation technique for locating the sensor with respect to the workspace. The method is novel and outperforms alternative methods.
3. The development of a novel scalable real-time object detection and recognition algorithm geared towards texture-minimal tools and components. The method outperforms existing techniques in terms of combined speed and accuracy.
4. The development of a technique for tracking objects within the workspace in 3D. Coupled with the above two techniques it allows robust recognition and tracking of tools and components within the workspace.
5. Initial tests on integrating gaze tracking for priming object learning and detection in real-time.
6. Extensive tests of the approach on multiple operators and three tasks of varying complexity levels.

Figure 9 shows the prototype for egocentric real-time workspace monitoring. A real-time structured light sensor is used to provide image and depth data (RGB-D), mounted above the user’s head on a backpack as shown in Figure 9a. The 3D pose of the sensor is tracked in real-time and foreground objects are recognised, positioned and tracked across time. Object appearance and handling characteristics are learnt independently in real-time for each user.

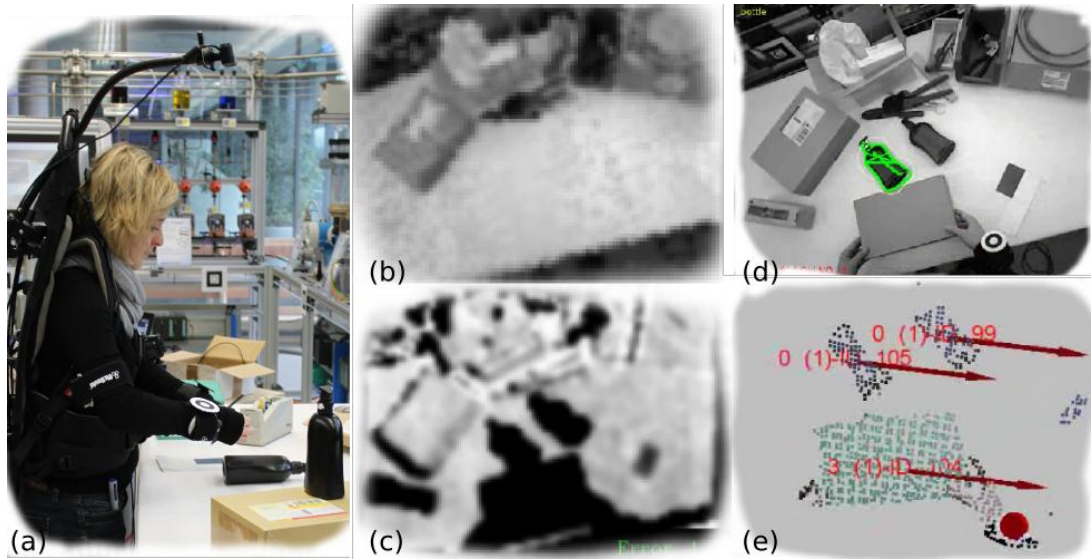


Figure 9 Prototype for egocentric mobile scene characterisation and monitoring. The setup (a) tracks an RGB-D sensor using dense depth and appearance modelling of the workspace (b), segments outliers indicating the presence of new objects (c), known objects are recognised (d) and tracked in 3D (e).

Users walk up to the workspace and monitoring is started, with minimal set up and pre-calibration of the environment. There are two key components in the method. First, an efficient optimisation strategy fuses depth maps and appearance frames from the RGB-D sensor to build a dense 3D reconstruction of the workspace. This then forms the basis of robust tracking and re-localisation of the sensor and the segmentation of foreground objects. Second, a fast object recognition method provides reliable identification of multiple previously seen objects. Recognition is based on characteristic edge features on the objects, which suits the minimal-textured tools and components usually encountered in industrial manufacturing. Details of each component are summarised below. The prototype has been evaluated extensively in a variety of scenarios, including the three COGNITO test tasks involving multiple operators completing the tasks multiple times. An evaluation is available in [1].

Background Mapping

Prior to task execution, a dense 3D map of the workspace is constructed by fusing data from the RGB-D sensor (Asus Xtion Pro Live). The method uses a framework akin to the depth fusion approach of Newcombe *et al* [2], although we incorporate appearance information and a robust relocalisation strategy. Successfully tracked frames are fused into a textured 3D occupancy grid map to build a representation of the static environment. The implementation is based on the open-source KinectFusion system, which is part of the PCL software library.

Sensor Tracking

Sensor tracking estimates the global sensor pose in each frame by aligning appearance and depth information with the stored map. This is achieved by a combination of motion prediction between frames and iterative closest point (ICP) alignment of the point clouds. To enable the system to run at 30 Hz, we run five ICP iterations per each frame on 80 x 60 pixels down sampled images. Additionally, the YUV colour information and surface normal directions are used to improve correspondence

estimation and generate a weighted inlier image that enables simple segmentation of foreground objects from the static map.

Relocalisation

One common failure case of the tracking system occurs during periods of rapid or erratic camera motion. To achieve fast relocalisation in such cases, knowledge of the map and prior information about the expected pose of the user is exploited. Using a view-based method, intensity and depth information are combined to estimate full 6D pose using regression over a database of synthetic views of the map. In contrast to standard methods, which usually take several frames to provide a solution (>100 ms), the approach enables relocalisation at frame rate (<10 ms). Comparative results are available at [3].

Foreground Segmentation

With the sensor's pose established, weighted appearance and depth differences are used to segment foreground pixels. Edge discontinuities produce noisy segmentations, which are cleared using image-based erosion. The foreground pixels are converted into a 3D point cloud, and are passed to the cluster-based tracker.

Real-time Learning and Detection of Texture-minimal Objects

In the industrial tasks, tools and components often have little texture and adopt a wide range of 3D poses. Thus a shape-based method is used for detection and recognition. This is occlusion-tolerant, scalable and fast. It is based on characterising shape by constellations of edge features. To give speed and scalability, pre-defined paths for constellations are used, enabling tractable search and real-time operation. Figure 10 illustrates the key components. An important element is that hand configuration is accounted for in recognition, reflecting the fact that user grip is highly dependent on tool or component shape. This gives significant improvements in performance, particularly for small tools and components. The method has been evaluated extensively on large data sets and shown to give superior performance to existing methods in terms of the trade-off between speed, scalability and accuracy (details in [4]). Figure 11 shows examples of recognised objects for two of the COGNITO tasks.

Cluster-based Tracking

For each frame, the foreground segmentation is clustered into connected components based on 3D spatial proximity, with small clusters being ignored. Clusters are then assigned to trajectories maintained from the previous frame based on spatial proximity and size similarity. New trajectories are created for unassigned clusters. The tracker operates at 30 fps. For each new trajectory, the clustered points are projected into the current frame to produce an image mask which is used to focus the object recognition algorithm.

Gaze Tracking

Tracking user gaze gives the potential to prime object recognition - when picking up an object, the person first gazes at the object and then approaches it. Initial investigations suggest that this can be utilised in future development of the prototype system. To illustrate, Figure 12 shows a sequence of images with accompanying gaze information indicating the advantage of gaze information in recognising objects in clutter using the developed scalable object detector. Gaze has an additional advantage over foreground segmentations in that it can recognise objects prior to their initial manipulation. This preliminary work suggests that combining gaze with background subtraction has the potential to significantly improve performance of the scene monitoring module.

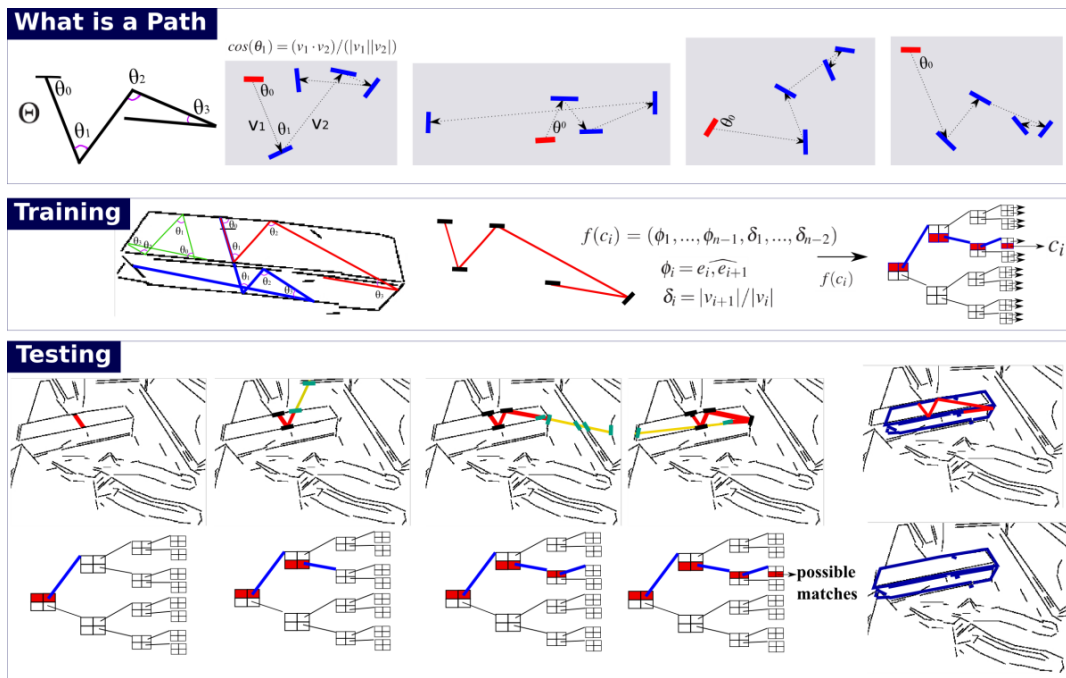


Figure 10 A path defines the relative direction between a constellation's constituent Edgelets. For a given path, edgelet constellations are traced out from training views. During training, a descriptor for each constellation is inserted into a hierarchical hash table. During testing, constellations are traced out using the same paths and descriptors enable identification of possible matches in the hash table. These are assessed using homography mapping of the view's edgelets to the test image.

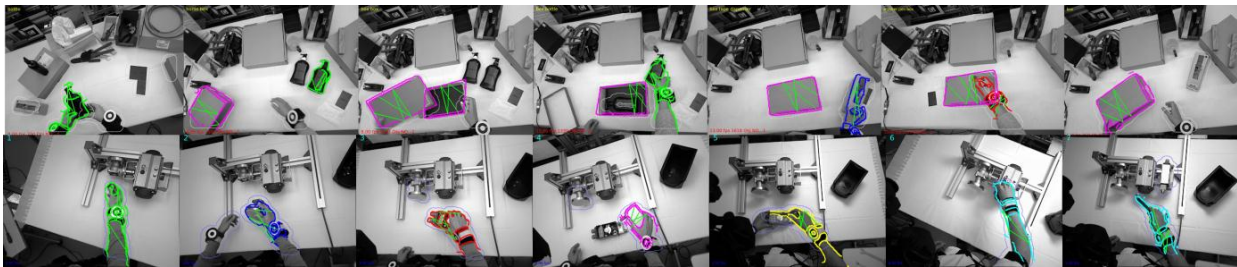


Figure 11 A collection of hand-held objects being detected using the scalable shape-based detector method, for both the bottle packaging scenario (top) and the ball valve scenario (bottom).

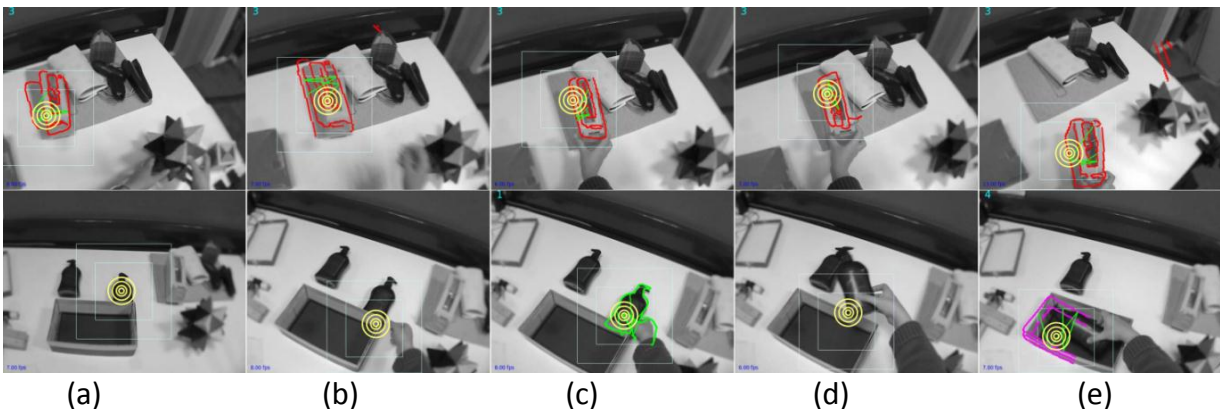


Figure 12 Two examples are shown where the direction of the person's gaze (indicated by yellow concentric circles) follows the pattern: the person looks at the object before approaching it (a), while approaching it (b) and while manipulating it (c,d). The person then directs his/her gaze to where the object will be placed (e). Objects are recognised within segmented regions around the estimated gaze position.

4.2. User activity monitoring

User activity monitoring was achieved by DFKI within a collaboration of both the Computer Vision package (WP5) and the Biomechanical Parameter Estimation package (WP4). The work was mainly performed by DFKI. The purpose of this component is to analyse the heterogeneous sensor data streams received from several on-body IMUs and cameras and to deduce relevant user activity information for workflow recovery and monitoring, biomechanical analysis and also user feedback. This information comprises the user's upper body motions and postures, her head poses and her hand positions and postures in 3D space. The focus was on the development of visual-inertial sensor fusion algorithms for robust and accurate real-time body motion tracking and the development of algorithms for real-time monocular hand detection and tracking. The key results can be summarised as follows:

1. The development of a novel inertial upper body tracking method using egocentric vision measurements for increased robustness against magnetic disturbances.
2. Built upon the above, a real-time estimation of the RULA (rapid upper limb assessment) scores during task execution and its integration into a complete system for online global biomechanical analysis and feedback as further described in Section 5.2.
3. The development, implementation and benchmarking of different visual-inertial sensor fusion strategies for 6 DoF tracking. This enables (1) global positioning of the user's upper body in the workspace and (2) robust, accurate and highly responsive head tracking suited for the rendering of 3D augmentations in an optical see-through HMD.
4. A novel learning-based method for real-time markerless hand detection and tracking in perspective RGB images.

The above key results are based on the final on-body sensor network configuration as shown in Figure 3.

Visual-inertial body motion tracking

In COGNITO, we have developed a novel method, which precisely and in real-time captures the motions of the upper body based on the miniature IMUs and cameras in the on-body sensor network (cf. Section 3). Each IMU provides synchronised 3D acceleration, angular velocity and magnetic field data. This information can be used for tracking relative body motions, given that one IMU is attached on each major body segment. The cameras are then used as aiding sensors as well as for global positioning.

The method is based on a biomechanical model, which represents the configuration and DoF of the user's upper body. This functional model is composed of 8 rigid segments (trunk, clavicles, upper arms, forearms and head) connected by anatomically motivated restricted articulations (pelvis, neck joint, sternoclavicular joints, shoulders, elbows). Figure 13 illustrates the upper body model with indicated IMU positions. In order to align the technical IMU frames with the anatomical reference frames, an easy-to-perform calibration procedure has been developed. It requires the user to stand upright with the arms straight down and the thumbs forward (N pose), and then bent over. The sought-after rotations are then deduced from the acceleration and magnetic field measurements captured in the two static poses. The relative positions of the IMUs with respect to adjoining segments are derived from the assumed segment lengths and placement protocol.

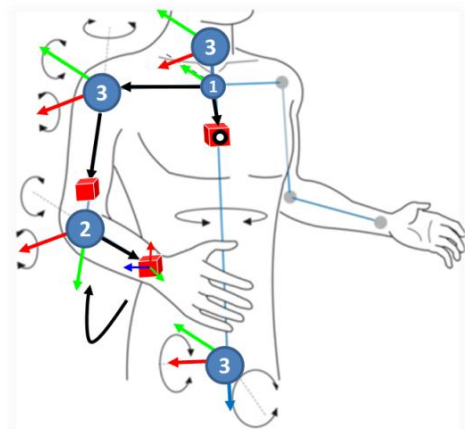


Figure 13: Upper body model with indicated sensor placement, calibration parameters and joint angles (and their DoF).

Once the IMU sensor network has been calibrated, joint angles and kinematics are estimated in real-time by fusing the IMU measurements and the model in a set of nested extended Kalman filters (EKFs). The measurement equations are based on forward kinematic equations derived from the biomechanical model. Given the joint angles from the filter and the model, the body pose is fully determined. The novelties of the proposed tracking solution are: (1) It is based on a body model, which directly builds in anatomical constraints, such as the universal elbow, resulting in higher robustness; (2) Instead of relying on each IMU to estimate its orientation separately, the raw IMU measurements are jointly fused in a statistical filter and effects due to linear accelerations are handled by the model. This results in higher flexibility and accuracy of the approach.

While the above setup offers already a good estimate of the worker's posture, the COGNITO use case required a further novel extension. First, the manipulation of ferromagnetic materials and tools during task execution results in strong magnetic disturbances (especially in the forearm IMUs), which rendered the magnetic field measurements used to compensate for drift useless. Second, pure inertial tracking can only provide relative positioning, while registration in the workspace is required for COGNITO.

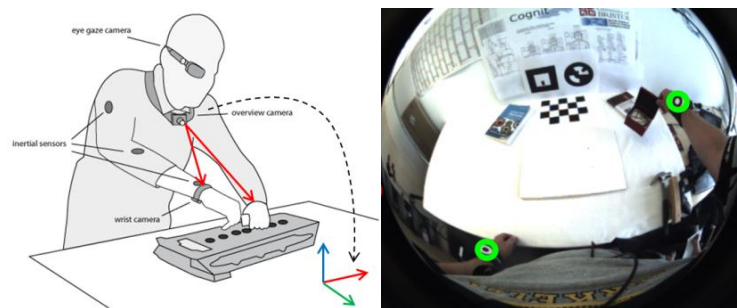


Figure 14: Illustration of the benefits of egocentric vision measurements (left); Egocentric view from the Cam-IMU attached to the user's chest with detected wrist positions (right).

Both of the above described issues have been solved by using visual information from the Cam-IMU attached to the user's chest. The idea is illustrated in Figure 14. The wide-angle camera provides visibility of the workspace at the same time as the user's own limbs are visible. Hence, we used the camera images for registering the torso in the workspace coordinate system as well as for obtaining information about the position of the upper limbs with respect to the torso. The latter information in terms of visual detections of the wrists (cf. Figure 14) is incorporated as an additional measurement into the above described filter setup. This relaxes the need for magnetometer measurements and results in a novel visual-inertial body tracking scheme, which has been shown to provide superior performance over the purely inertial approach in [5]. While a marker-based approach has been used for reliably detecting the hands in the fisheye images, a logical next step would be to benefit from the novel hand detection and tracking methods presented further below.

Visual-inertial 6 DoF tracking

In order to enable 6 DoF positioning of the Cam-IMUs in the on-body sensor network, different visual-inertial fusion strategies have been designed according to the following major requirements: stable tracking under agile motions (e.g. quick head motions); low latency and sufficient accuracy suited for the see-through AR display; continued rough positioning and orientation tracking during periods without 3D information. The focus of the work was on the development and evaluation of the fusion strategy rather than on the image processing algorithms. Hence, a circular marker was used for providing visual anchor points (in terms of up to four 2D/3D correspondences) in the workspace. Due to the fisheye view and the tracking method being designed to work with minimal 3D information, a single marker provided good tracking quality over a wide range of head motions, even during periods where as few as two anchor points could be detected.

Starting from our earlier work [6] [7], different visual-inertial fusion models have been designed, implemented and evaluated (using the EKF as estimation tool). These models differ in the fusion level (pose vs. measurement), assumptions about the translational dynamics (constant acceleration/ velocity/ position, decaying velocity), handling of the magnetometer information (bias estimation, omission), handling of the accelerometer measurements (gravity vs. linear acceleration), visual measurements (pose, 2D/3D correspondences, 2D/2D correspondences), among others. An extensive evaluation and comparison of the different approaches showed: Overall, fusion on measurement level provided more flexibility and higher accuracy than fusion on pose level. A constant acceleration model provided extremely responsive tracking during agile camera motions, however, turned out to be too fragile during extended periods without 3D information available, even if this effect could be considerably reduced by exploiting information from temporary 2D/2D tracks between frames. The final model fulfilling the requirements of COGNITO performs fusion on measurement level, assuming decaying velocity and constant angular velocity. The gyroscope measurements are modelled as control inputs and the accelerometers are modelled as inclinometers, while the magnetic field measurements are not used. Details on the different models and evaluation results can be found in the public evaluation reports¹ and in [8].

Monocular hand detection and tracking

The role of hand detection and tracking in COGNITO is threefold: First, the wrist positions are one piece of information used for workflow recovery and monitoring. Second, the wrist positions are used for aiding the upper body tracking as described above. Third, the full hand posture could potentially be connected to the musculoskeletal hand model developed by UTC in order to perform a detailed online biomechanical analysis of the hand (cf. Section 5.2).

In this context, DFKI has developed several methods for wrist and hand detection and tracking. Since the IMUs positioned on the forearms provide a good opportunity to integrate an easy-to-detect pattern, a marker-based solution has been developed for the purpose of providing reliable positioning of the wrists, even in a distorted fisheye view, serving the first two purposes above.

The major contribution in this context, however, consists in the development of novel markerless methods for hand detection and accurate hand posture estimation in perspective RGB images.

In early stages of COGNITO, we have developed a feature-based method, which fuses posture-invariants of the hands and fingers in order to obtain robust detection above cluttered background. The algorithm utilises parallel edges and typical shading of the fingers and deliberately avoids relying on skin-colour based segmentation as first processing step, since this tends to be highly dependent on the camera's photometric calibration and the lighting conditions. This approach is further detailed in [9].

The next step consisted in enhancing the feature-based method to actual 3D pose estimation. For this, a kinematic hand model with 26 DoF was developed in collaboration with UTC (see Figure 15). Although an extremely challenging task, the algorithms developed within COGNITO made it possible to estimate the parameters of this model at interactive framerates, given the images from a standard perspective colour camera. The approach is based on the combination of an efficient database search and a local kinematic model-fitting step. The algorithm consists of two phases. First, a rough hand pose is obtained from template matching within a big database of hand views under different postures. This step is highly efficient due to a beam search based on the posture history and adaptive search tree database indexing. In a second step, the rough pose is refined by a subsequent model-fitting step, where the kinematic hand model is fit into the current view using a particle swarm

¹ Deliverable D2.4 (Report on the trials I) and D2.7 (Report on the trials II) are publicly available on the website (www.ict-cognito.org).

optimiser. The model-fitting step is less efficient, however, it is real-time capable, given that the rough pose from the first step is close to the actual solution.

For both, the generation of a dense hand view database from very few labelled example images, and the model-fitting step, a novel billboard morphing technique has been developed. In combination with the kinematic hand model, this method allows synthesising in real-time previously unseen hand views from two prototypes through advanced image-based rendering (axis aligned pixel-wise warping). In conclusion, the developed approach and its above described key features enable robust hand tracking from a standard RGB camera in complex scenes (cf. Figure 15). After a training phase, the adaptive approach runs at interactive frame rates in pre-seen environments. More details can be found in [10] [11].

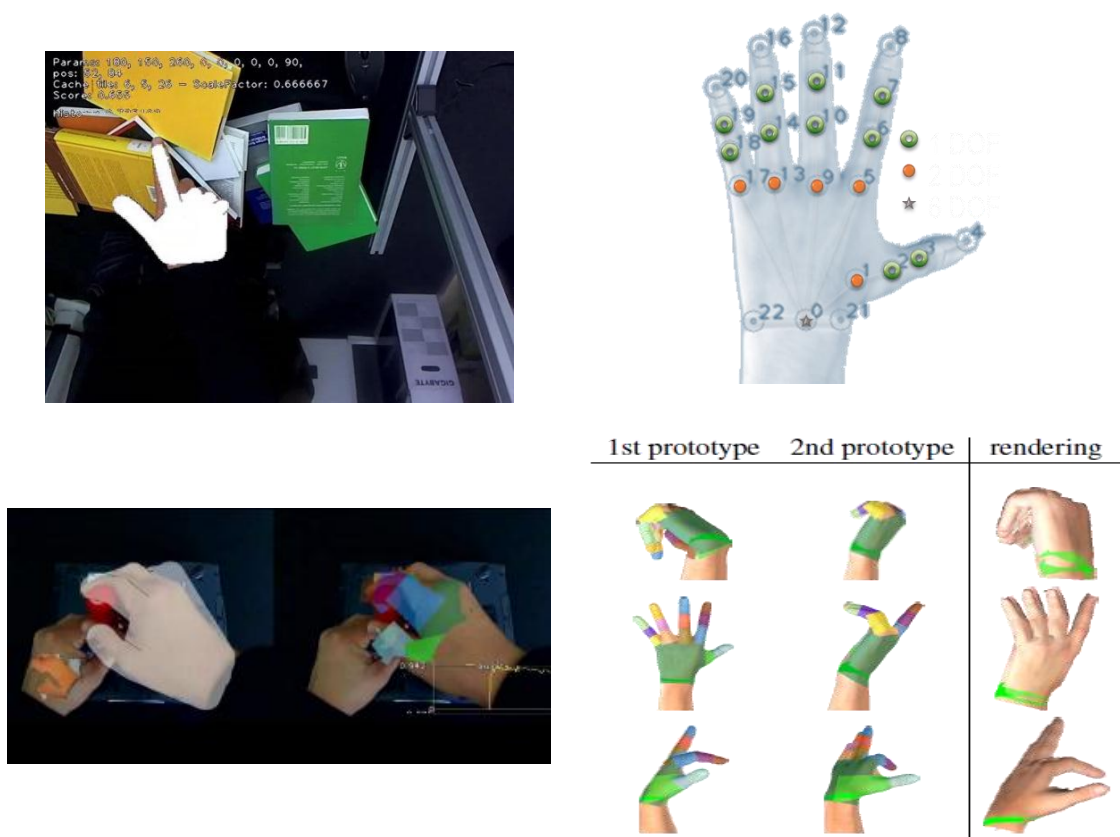


Figure 15 Examples of hand tracking above cluttered background and during task execution (left); Kinematic hand model and examples of image-based renderings (right).

Although this hasn't been done yet, a logical next step would be to connect this approach to the detailed biomechanical analysis of hand and forearm as described in Section 5.2 and by this achieve a more sophisticated online assessment.

5. Higher-level modelling

The higher-level modelling layer processes the instantaneous object-level information obtained from the low-level sensor processing components and performs two types of analysis: Workflow recovery and monitoring links the user activity and workspace information to the underlying workflow structure. The work has been performed by Leeds. Biomechanical analysis assesses the global postural discomfort and the hands' muscle forces and articular loads during task execution. This work has been performed by CNRS, in collaboration with DFKI. As an alternative approach to workflow

recovery and monitoring, DFKI has developed a lightweight system based on a single RGB camera along with intuitive authoring tools. The key results are detailed subsequently.

5.1. Workflow recovery and monitoring

Leeds explored a number of approaches to workflow modelling and monitoring. The challenge was to find a method that worked sufficiently well with small datasets for parameter estimation (learning) and errors in visual analysis arising from the challenging datasets used in our evaluation.

At an early stage, we focussed attention on a representation that makes explicit the spatiotemporal pairwise relationships between body parts, tools and machine parts in the workspace. The motivation was that these relationships are invariant to position and viewpoint, and correlate well with functional relationships between objects (e.g. picking up an object with the hand involves contact between the body part and the object; hitting a nail with a hammer involves a high speed of approach between the hammer and the nail). In our prior work on activity analysis, we have already found purely spatial relationships (e.g. contact, no-contact) to be a useful form of representation.

We have combined this representation for the objects in the workspace with an HMM to model workflow and observation uncertainty. Details of this model and associated evaluation can be found in [12]. We summarise the model here (see also Figure 16).

A workflow is assumed to be a temporally ordered set of procedural steps or *atomic events* for accomplishing a task. The atomic events are associated with the states of an HMM, with a conditional observation distribution over a summary of the expected spatiotemporal relationships between pairs of objects in the workspace within a sliding time window. These objects include parts of the human body such as the wrists, work tools such as hammers and screwdrivers, and machine parts. The observation distribution is represented by a probabilistic multi-class SVM. The key novelty in our model is the way in which we characterise the movement and interaction of a varying number of objects using a fixed observation space, and in such a way that recognition performance is invariant to broken tracks and missed detections.

The input is a time-series of the object detections appearing at each time-step obtained from visual analysis (WP5). Each detection consists of an object identifier (associating detections through time), a position in 3D, and the class of the object (from a set C, e.g. hammer, wrist). At each time-step, the detections are converted into a set of pairwise spatiotemporal relationships between objects, each consisting of an identifier for the relation (from a pre-determined set R), and the classes of the two objects involved.

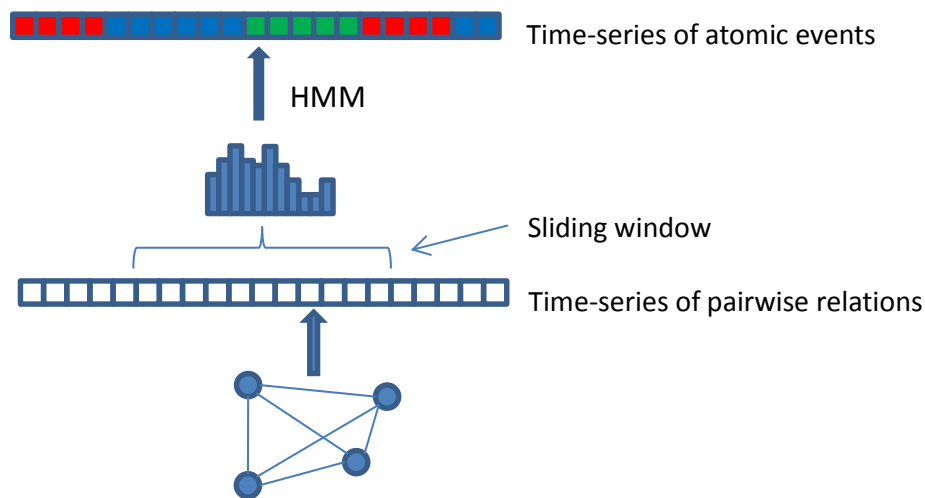


Figure 16 HMM over histograms of relations

The relationships over a fixed time window (typically 1-3 seconds) are pooled and summarised as a histogram, recording the number of times each relation in R occurs. Thus a time-series of histograms is obtained using a sliding window that moves through time. Finally, the counts are accumulated separately for every pairwise combination of object classes, and the histograms are combined into a single feature vector. For example, there is a separate histogram associated with the relationships between wrists and hammers. Thus, the feature vector has dimension $|C| \times |C| \times |R|$. In our experiments on two datasets, $|C|$ is 6, and $|R|$ is between 2 and 16. Thus the feature vector has between 72 and 576 dimensions. We call this representation a *Bag-of-Relations* by analogy with Bag-of-Features.

Similarly, body motion estimates (WP4) are converted into a second set of pairwise spatiotemporal relationships using codebooks which are generated separately from the relative joint angles, and their first two time-derivatives, between elbow-shoulder and shoulder-torso pairs. There are six different codebooks (three each for elbow-shoulder and shoulder-torso relation) and applying to both left and right arms.

To learn a new workflow model in a supervised fashion, we assume multiple examples of the workflow along with ground-truth of the desired atomic event label at each time-step. Each workflow example is processed to detect and track the objects it depicts (cf. Section 4.1), and we also acquire motion data from the body of the operator (cf. Section 4.2). We model the pairwise relationships between visually-detected objects by the distance between objects and the first derivative of this distance (Figure 17), a representation that is invariant to rotation and translation, but not scaling. A codebook is generated for these tuples using K-means clustering on a training dataset. We have found that performance improves when we use a separate codebook for each pair of object classes.



Figure 17 Quantisation of pairwise spatiotemporal relationships between key objects in the workspace.

Thus, for each position of the sliding window, we obtain a feature vector from the time-series of object detections and body motions. The probabilistic multi-class SVM is trained from these feature vectors and the corresponding ground-truth atomic events at the time-step in the centre of the sliding window. In all experiments we used a χ^2 -kernel for the SVM. Finally, the initial state and state transition distributions are estimated from the sequences of ground-truth atomic events.

Given a learnt model, workflow monitoring proceeds on-line from a stream of object detections and body motions. We infer the most likely sequence of atomic events leading up to the current time, by inference on the HMM.

We have evaluated the workflow monitoring off-line using the HMM model on two datasets: *Nails & Screws and Labelling & Packaging* (cf. Table 1). Each dataset is composed of sequences made up of 9 distinct atomic events. We evaluated our model using a ‘one-vs-all-subjects’ approach i.e. the workflow sequences from all subjects except one are used for training and the sequences belonging to the left-out subject are used for testing. In the initial experiments we used a 2 second sliding window. The performance is set out in the following table:

Recognition rates in [%]	Vision	IMU	STIP	Vision+ IMU	Vision+ STIP	IMU+ STIP	Vision+ IMU+ STIP
<i>Nails & Screws</i>	60.5	49.8	58.6	62.8	68.3	61.7	65.3
<i>Labelling & Packaging</i>	62.1	59.1	51.3	73.4	69.1	63.8	77.1

Several approaches are compared, visual analysis (Vision) alone, body motion data (IMU) alone, the addition of spatiotemporal descriptors at STIP points, and combinations of these. The incorporation of descriptors at STIP points results in a hybrid approach that exploits both visual features extracted from the egocentric video stream (cf. Figure 14) and spatial relations between objects that have been detected within the visual stream – two different conceptual levels of analysis.

From the confusion matrices, it is evident that the predicted atomic events are often confused with the previous and next atomic events. This is a typical synchronisation error for sequential data. It is partly due to the manual assignment of labels while preparing ground-truth, since it is difficult for humans to assign boundaries consistently between consecutive events. Details on the evaluation procedure and results can be found in [12].

In the course of development, we explored several extensions and variations of the approach. We experimented with partitioning the sliding window into equal sub-windows, computing a separate set of relation histograms from each, and then concatenating the histograms to form a single feature vector.

We have investigated two ways of using spatiotemporal relationships between visual features instead of between objects. The motivation was to exploit information from the video stream that does not depend on object detection and tracking, and promises robustness to occlusion and background variation that is especially prominent in an egocentric setup. In the first approach, we make class predictions from single video frames (i.e. no sliding window) by encoding the distances between pairs of visual features, and associated orientations in the image plane. In the second, we combine Bag-of-Features using SURF with additional relational features obtained from the distance and first derivative of distance between selected pairs of SURF features – we choose only pairs that are assigned the same descriptor codeword, although other selection mechanisms would probably have worked equally well. At present, features are accumulated over the duration of a ground-truth instance used in training or testing, rather than a fixed length sliding window. Although we have not yet done this, an obvious next step would be to combine the feature vectors obtained from pairwise object relationships with those obtained from pairwise feature relationships.

Most recently, we have undertaken unsupervised training of the HMM by clustering input histograms. The idea here is to allow annotation of novel datasets without having to predefine the atomic events in advance, and prepare ground-truth.

In the earlier part of the project, we experimented with a Petri Net sitting in place of the HMM. Formed from states, temporal markers and conditional transition arcs, the properties of a Petri-Net allowed us to capture and model a number of workflow restrictions such as the number of repetitions of a sub-action. Although in principle the Petri-Net structure could be induced from example workflow sequences, this required some supervision. This work is reported in [13].

We also explored a pLSA model feeding the HMM (details in [14]), but for performance reasons adopted the final approach reported above.

In the final year of the project, we explored an entirely different model motivated by our earlier work on the use of logical induction for inferring atomic events and prior work combining a classifier with rule induction to classify logical data. Specifically, we developed a novel method in which the HMM in our principal model is replaced by a rule-based predictor of atomic events, working from a binarised version of the previous histogram vector (all non-zero entries are set to one). The rules are expressed in logic and inferred from training data using an existing induction mechanism. On top of this sits a ‘deep learner’ composed of a layering of one or more Restricted Boltzmann Machines. This is trained to map between the binary vector of rules that fire with a given input (1 if rule fires and 0 otherwise) and the set of atomic events. The results are promising but not as good as those obtained recently with the HMM model. On the *Labelling & Packaging* dataset the average accuracy was 59.9% and on the *Nails & Screws* dataset it was 62%. This work has not yet been published.

5.2. Biomechanical analysis

Besides the structural workflow analysis, the biomechanical analysis of the worker during industrial tasks is another major component of COGNITO. This part had been supervised by CNRS. In order to achieve this ambitious goal, two levels of analysis were developed.

The first level of analysis is an ergonomics analysis of the subject during the industrial task based on the RULA table. With help of the inertial upper body motion capture as described in Section 4.2 (see also Figure 18), it was possible to compute real time scores, which quantified the postural discomfort. The implementation into the COGNITO system has been achieved with DFKI.

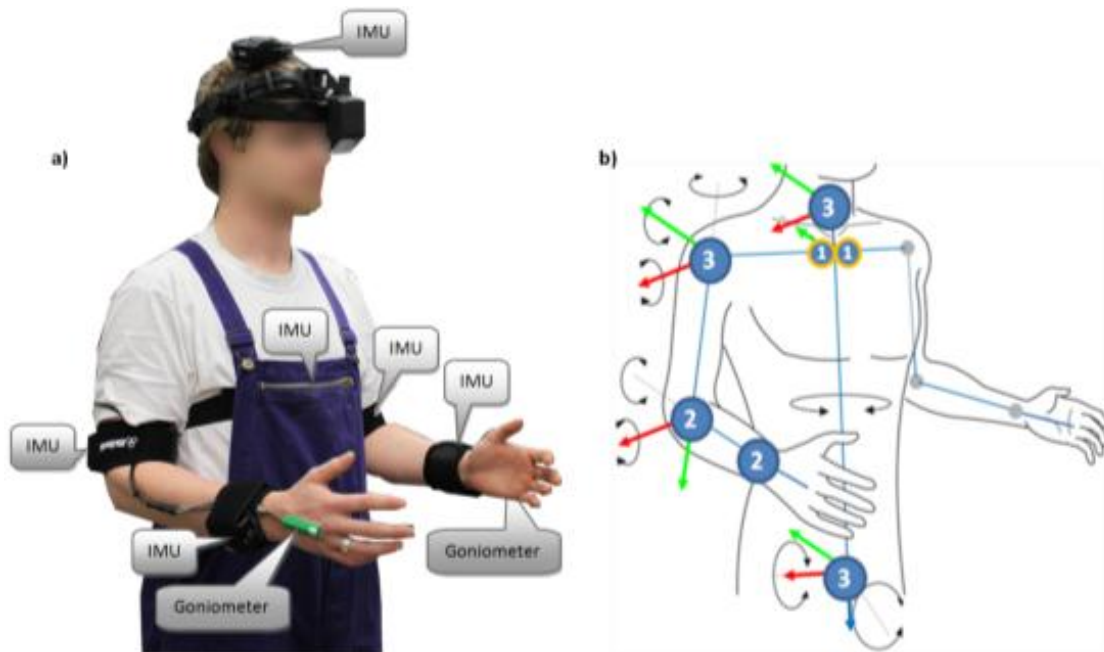


Figure 18: Hardware platform (on-body sensor network and HMD) (a) and biomechanical model of the upper limbs (b). Compared to the model in Figure 13, hand segments and universal wrist joints have been added and their configurations have been tracked by adding goniometers to the on-body sensor network as shown in (a).

The ergonomic tool RULA estimates the exposure to the upper limb musculoskeletal disorders (MSDs) by computing a global risk score. For a given posture, a global score ranging from 1 to 7 is computed. A score of 1 means that the posture is comfortable and safe for the subject, whereas a score of 7 means that the posture is not convenient for the subject and presents a risk of musculoskeletal troubles. This scoring is based on posture, muscle use, weight of loads, task

duration, and repetitiveness. In order to refine the analysis a local score is also computed. The local score is based on the joint angles.

For the current COGNITO system the two types of feedback for the worker are: (1) an acoustic signal linked with the global score and (2) a visual information related to local scores provided through the see-through HMD.

The acoustic signal is given as a warning, if the global score of 7 is reached for a period of at least 0.5 seconds. It reflects the recommendation of the RULA table to change the posture immediately. If the score is 5 or 6, the RULA table suggests modifying the posture soon. In order to communicate this recommendation, the acoustic warning is given after 5 seconds in this range. In the same way, each time a local score is greater than a predefined threshold, the concerned joint and segment are highlighted in red in a schematic representation of the operator's upper body on the see-through HMD (Figure 19). The visual warning is given, if the local score of the shoulder and upper arm is higher than 5, if the local score of the elbow and lower arm is > 3 , if the local score of the wrist and hand is > 5 , if the local score of the neck is > 5 , if the local score of the head is > 4 , and if the local score of the pelvis and trunk is > 4 . Besides the online feedback, a printable report summarising the RULA statistics is generated after task execution and serves as further documentation.

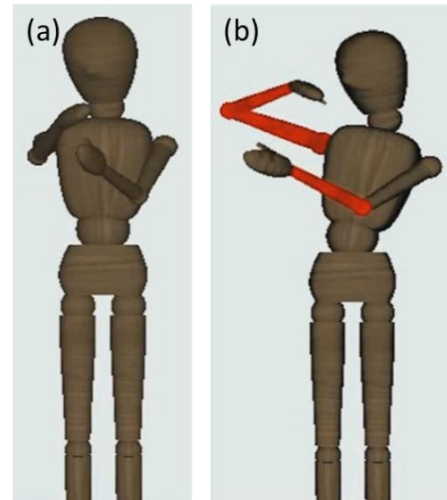


Figure 19: Visual feedback for a manual task performed in a healthy manner (a) and in a hazardous way (b). For case b segments are highlighted in red.

The whole procedure and the efficiency of this analysis were tested in user sessions with the help of the project partner SmartFactory. The work has been published in a scientific article [15] and provided very satisfying results.

The second level of analysis was a musculoskeletal modelling of the upper segment during an industrial task. Based on an inverse-to-forward dynamics simulation in order to estimate muscle force and joint load during an industrial task as previously defined, an accurate musculoskeletal model of the hand and forearm has been obtained including a motion capture protocol and inverse to-forward dynamics calculation. This part was a very challenging task as hand modelling is not generic. In addition, for the COGNITO project the highest level of description of musculoskeletal modelling of the hand had been expected.

In order to achieve this very ambitious part, hand and forearm segments have been modelled as kinematical chains of 21 rigid segments: the forearm composed of ulna and radius bones, the carpus which includes eight carpal bones, five metacarpals, five proximal phalanges, four middle phalanges for digits 2 to 4, and five distal phalanges. In addition to this, 46 musculo-tendon units as force generators on the musculoskeletal system have been included to generate segmental motion (Figure 20)².

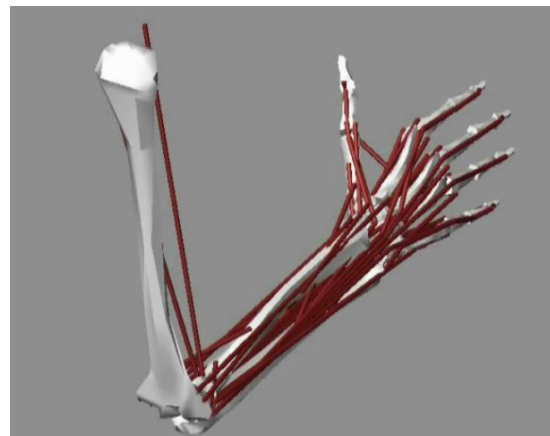


Figure 20: Musculoskeletal model of the upper segment

² Details on the modelling procedure are given in the deliverable D4.3 (Biomechanical analysis of the worker), which is publicly available on the project website www.ict-cognito.org

The musculoskeletal modelling is currently an offline procedure, due to the fact that all the calculations necessary for the estimation of muscle and joint forces could not be performed real time. Consequently, a musculoskeletal modelling database based on the COGNITO scenarios (cf. Table 1) has been generated for a 50th percentile subject (age 30, 1.78m height and 77kg weight). The integration of these results into the COGNITO system has then been performed as follows: Once the current manual task has been identified by the workflow monitoring component, the worker can request the corresponding video of the task performed by the musculoskeletal model of the hand and forearm. Maximal values of the muscle forces and joint loads are also displayed on the video. This information is valuable for the worker in order to control the stress on forearm and hand and in particular for experts in order to plan tasks appropriately. It also contributes to the documentation of the cumulative physical stress on the worker during a task. This is in accordance with prevention and health surveillance principles as explained by the Community strategy on health and safety at work from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions of the 21st February 2007.

The biomechanical analysis of the worker during industrial task was a very challenging part due to the fact that no commercial tools are currently available to do this. New tools and analysis were developed for the COGNITO system. It was an interdisciplinary task necessitating competences as present inside of the COGNITO consortium.

5.3. Monocular workflow segmentation, authoring and monitoring

In parallel to the sophisticated workflow recovery and monitoring methods developed by Leeds, DFKI has explored the potentials of a lightweight monocular approach to workflow segmentation and monitoring. In particular, the goal was to reduce the amount of manual labelling and authoring required for workflow recovery and augmentation. Starting from an approach for automatically segmenting a workflow into elementary actions based on a single video example taken from a head camera, an intuitive authoring tool and an online workflow monitoring component based on image feature similarities have been developed in the last period of the COGNITO project. The overall system is called the Augmented Reality Manual and has been published in [16]. The key features of the system are: (1) automatic segmentation of visually recorded workflows into meaningful chapters based on a robust image distance measure (cf. Figure 21); (2) automatic generation of augmentations for user assistance (cf. Figure 22) and (3) recognition of a segmented workflow (internally represented as a Markov chain) in a video stream for user assistance (cf. Figure 23).

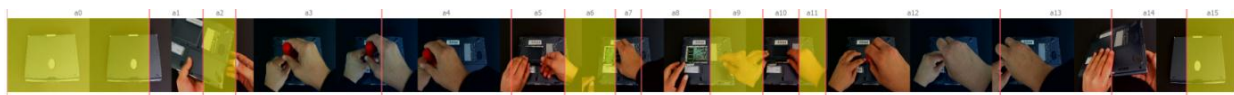


Figure 21 An image sequence is automatically segmented into single actions. A statistical representation in the form of an HMM is then learned for subsequent monitoring.

While the developed approach assumes video sequences without significant head motions and with strong focus on the scene and is, hence, inferior to the methods described in Section 5.1 in terms of invariance to changes in viewpoint, environment and user, it has the strong advantages of being more lightweight (based on a single RGB camera), easier to setup and use and reducing the amount of manual labelling by the expert user. The system has been successfully demonstrated at the CeBIT 2012 and will be shown in an extended version at the CeBIT 2013.

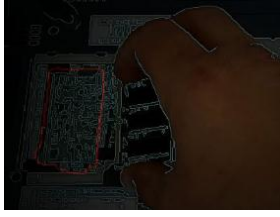


Figure 22 The AR manual system can automatically create augmentations, e.g. by generating animations from the video clips associated to each action in a segmented workflow and by performing before-after image analysis for each action and highlighting differences as areas of interest (left). Also, in order to emphasize more details the user can add visual overlays using an intuitive authoring tool (right).



Figure 23 Example of online workflow assistance. The user wears an HMD with integrated camera. The task consists in changing the memory of a laptop. The right image shows what the user sees in the HMD while performing the task.

6. User interface

The user interface building block provides the means for editing recovered workflow structures and enriching them with descriptive information (workflow editor) as well as assisting the user during task execution based on AR techniques (AR player). For the latter, different communication interfaces have been established to the underlying system components delivering the information necessary for providing context-sensitive feedback registered to the real world. This information comprises: the user's head pose, the position of key objects in the workspace, the current position in the workflow, the predicted next action and the postural discomfort. With this information it is possible to provide feedback that is tailored to the current workflow context and user activity. The user interface has been developed by CCG in close collaboration with SmartFactory, who supported its design and development through various exploratory studies, user tests and expert evaluations. Both contributions are described subsequently.

6.1. Workflow editor and AR player

As a first contribution, CCG has in collaboration with Leeds developed an XML format (Template of Actions (ToA)), which allows formalising and storing a recovered workflow as a sequence of elementary actions (primitive events (PE)) with associated elements for information presentation. This file format is exported by the workflow recovery and monitoring component (cf. Section 5.1) and serves as basis for the communication between this and the user interface.

The workflow editor, as the name says, provides the functionality of editing a workflow, i.e. modifying a given ToA. For instance, an expert can group or separate PEs and by this modify the workflow structure that has been recovered automatically in an unsupervised learning step. This provides the means for iterative semi-automatic workflow recovery, which can be seen as an extension of the supervised algorithms presented in Section 5.1. The second use case consists in labelling PEs and adding descriptive multimedia-based information, such as images, videos, textual explanations and 3D graphics explaining the respective PE. This information will be visualised in the AR player during task execution, as soon as it becomes relevant, i.e. when the associated event is detected or predicted. The overall structure of the editor interface is shown in Figure 24.

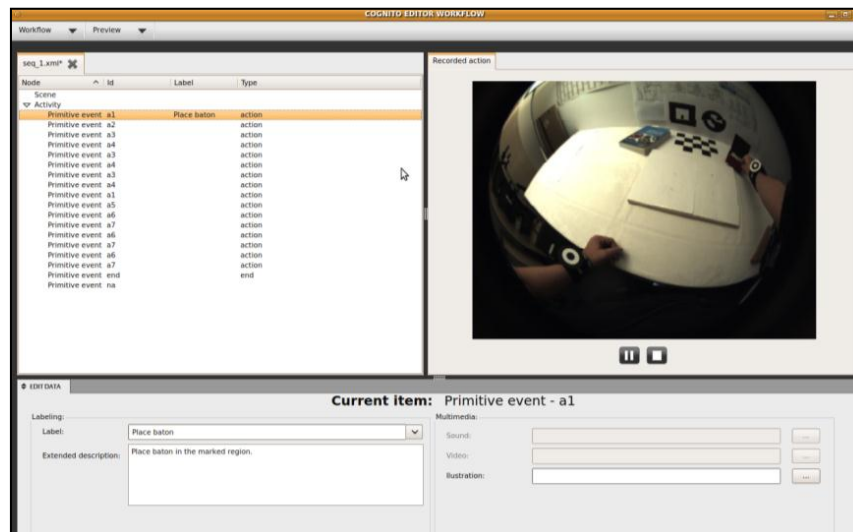


Figure 24 - Illustration of the COGNITO workflow editor

The AR player provides two different utilisation modes:

Controlled by the workflow monitoring module (WMM): In this mode, the AR player receives from the WMM the information about the current PE. Having this information, it can present context-sensitive information, i.e. the multimedia-based content associated to the given PE. In this case, the WMM and the underlying low-level processing components provide cognitive capabilities to the user interface and allow for context-sensitive information delivery.

Controlled by the user: In this interactive mode, the AR player is controlled by the user. Via speech commands (and a respective voice recognition system), the user can step through a workflow or ask for specific information. The interaction is hands-free in order to prevent interference with the manual task to be performed. This interactive mode allows inexperienced users to benefit from the assistance system without relying on workflow monitoring and the required capturing equipment. In this way, the user interface can be used as a stand-alone module for demonstration and user testing.

For both of the above modes, the types of content available are the same, depending on the information that is received from the low-level sensor processing components. While textual overlays can be displayed without further information, the user's head pose with respect to the workspace coordinate system is required for being able to render 3D graphics registered to the real world in the see-through HMD (see Figure 25). A connection to the biomechanical analysis enables feedback in the form of audible and graphical warnings concerning the worker's posture (cf. Figure 19). A connection to the computer vision module and its knowledge about the current location of known objects within the workspace enables highlighting key objects that are relevant for the current PE, such as tools or parts (see Figure 25). The AR player reacts flexible to the given information channels and can also be customised by the user concerning the amount and type of information to be shown.

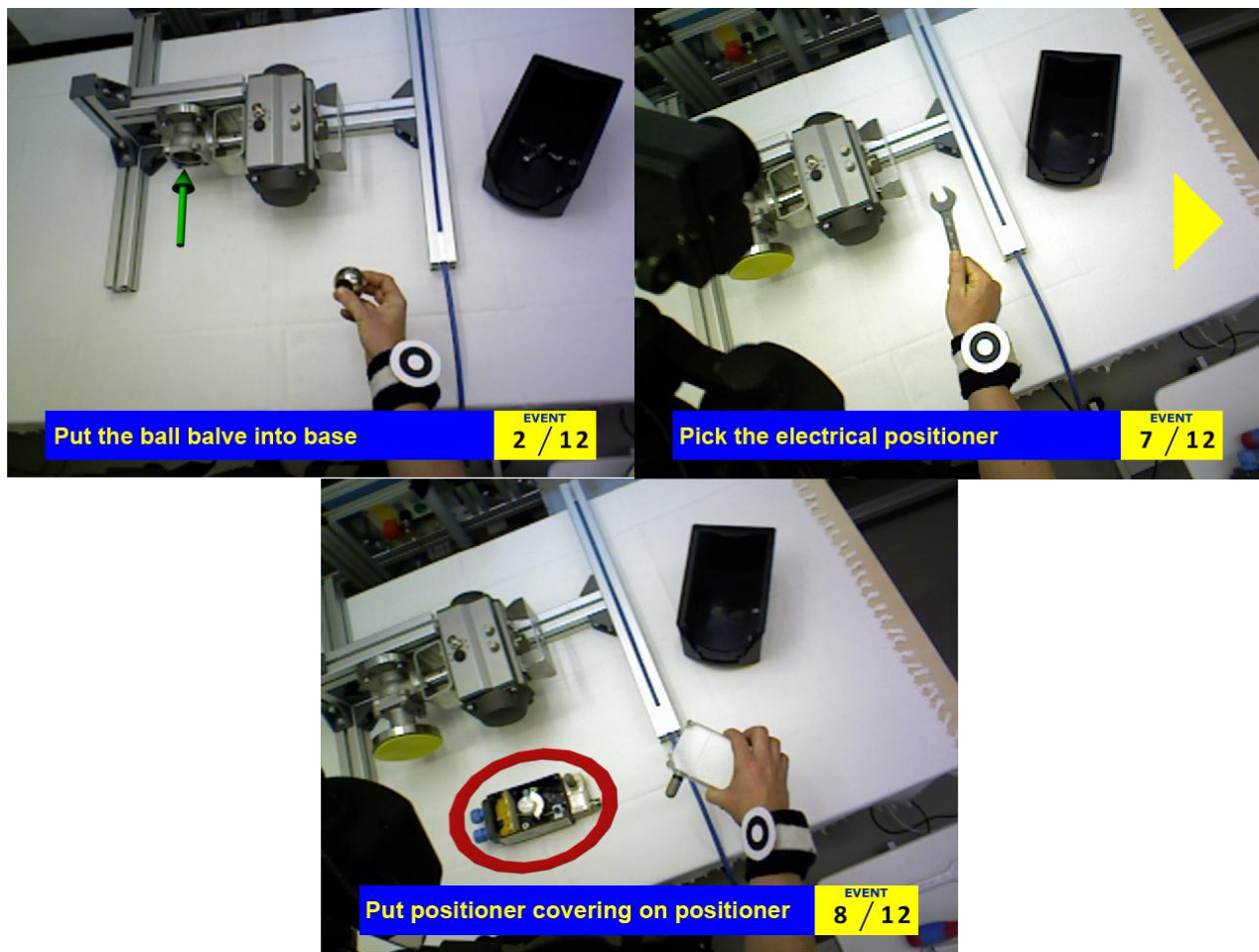


Figure 25 – Examples of information overlaid in a see-through HMD.

Other major features for the AR player application are:

- The process of visualising graphical overlay information, contextually to the background image as seen through the HMD, implied the implementation of a camera-eye calibration procedure. An interactive tool based on the visual-inertial 6 DoF tracking approach outlined in Section 4.2 and the SPAAM algorithm [17] has been developed by DFKI.
- Typically the HMD field of view (FOV) does not cover the entire working area. This means that sometimes the user is not able to see the augmented graphical instructions, as the HMD's FOV does not include the respective area. In such cases, additional cues are necessary for guiding the user's attention to the right location. Two variants have been developed: (1) a line in rubber band style connects the centre of the HMD screen to the point where the user should look at, or (2) flashing arrows indicate the direction to where the user should move the head and focus (cf. Figure 25). The type of visualisation is under the user's control. Therefore, some degree of intelligence is implemented in the AR player itself, as it determines, whether the HMD FOV encloses the part of the workspace to be augmented.
- The AR player is flexible and complete, permitting the visualisation of multi-modal feedback, such as animated 3D models for actions/movement instructions, the 3D representation of tools and parts to be manipulated and its animations, the display of textual current PE as a label and more extensive explanation as audio, the display of short videos illustrating the event to be performed, the visualisation of graphical body posture warnings and appropriated audio warnings, if an unhealthy posture was detected by the biomechanical analysis, the display of arm, forearm and hand musculoskeletal forces and joint loads as explained in Section 5.2, among other features.

- Extended explanations associated to the PE to be performed are provided as audio. The audio is automatically generated by means of real-time text-to-speech conversion. The audio feedback prevents overloading the display with graphical information and distracting the user from the workspace.

The above presented two components are the user interface. In particular the AR player is an important part of the integrated COGNITO system, where all information received from the low-level sensor processing and the higher-level modelling components are exploited for providing context-sensitive and timely, multimodal and hands-free user feedback.

6.2. Evaluation and user testing

The COGNITO concept and its user interfaces were continuously evaluated in the SmartFactory Living Lab. A major evaluation of the COGNITO concept after the first project phase convincingly demonstrated the potential of an AR-based assistance in comparison to conventional video and paper-based instruction methods and provided initial design decisions for the user interface, such as the preference of a monocular HMD over a binocular one. During the second project phase, where the focus moved from the technology to the user interface, short development-evaluation cycles enabled fruitful cooperation between the respective partners. While formative evaluations yielded important input for upcoming design decisions, summative evaluations enabled assessment of the achieved quality of user interaction. Different data assessment methods were chosen according to the evaluation objectives. Table 2 summarizes all user evaluations performed throughout the COGNITO project³.

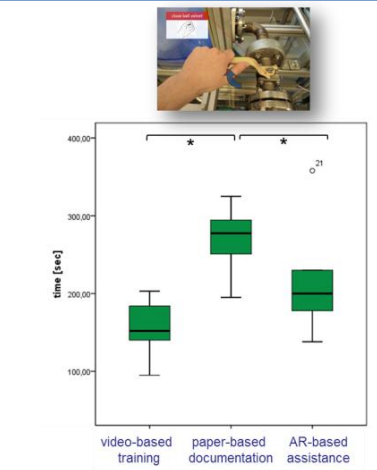

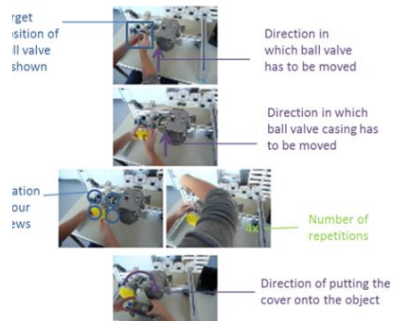
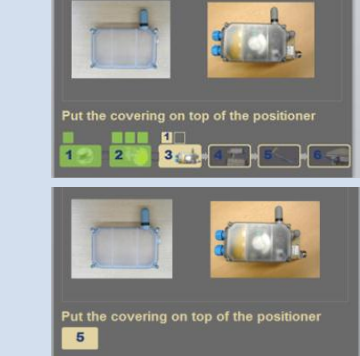
In conclusion, the evaluation activities enabled us to choose the right output device (i.e. a monocular see-through HMD), to quantify the benefit of an AR-based assistance system over traditional instruction methods, to demonstrate the value of an online ergonomic feedback for the avoidance of hazardous postures, to get insights into human understanding of the structure of an assembly task and to improve the usability of user interaction including both input and output devices. While these outcomes were already taken into account during the development of the COGNITO user interface as described above, we also explored ways of enhancing comprehensibility and memorability of information presentation, which provide routes for future improvements.

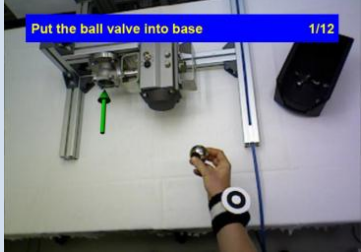
Table 2: Overview of all evaluation studies concerning user interaction of the COGNITO system over the whole project.

Topic of evaluation activity	Research question(s)	Method, procedure, and main result
See-through HMD	Is the monocular or the binocular HMD more suitable for assistive usage in industrial contexts?	Two-within-group randomized experiment (N = 22); Puzzle tasks with monocular versus binocular HMD; Time for HMD adjustment, subjective comfort and image perception were preferable in the monocular HMD.



³ Detailed documentations are provided in the deliverables D2.4 (Report on the trials I) and D2.7 (Report on the trials II), which are publicly available on the website (www.ict-cognito.org)

<p>COGNITO assistance concept</p>	<p>Is the COGNITO system effective, efficient, and accepted for supporting inexperienced users in manual industrial tasks and in comparison to traditional methods?</p>	<p>Three-group randomized experiment (N = 27);</p> <p>Each group performed two scenarios (Nails & Screws, Ball Valve) by help of either a paper manual, an instructional video, or the simulated COGNITO system;</p> <p>Execution time of COGNITO system was faster than with paper manual, subjective preference for COGNITO system.</p>	 <p>The box plot shows execution time in seconds for three methods: video-based training (median ~160s), paper-based documentation (median ~280s), and AR-based assistance (median ~200s). Significance markers (*) indicate differences between video and paper, and between paper and AR. An outlier is present for AR-based assistance at approximately 360s.</p>
<p>RULA feedback</p>	<p>Does visual and audible feedback on hazardous postures support users' ergonomic behaviour?</p>	<p>Two-group randomized experiment (N = 12);</p> <p>During Ball Valve scenario: RULA feedback with visual and audible warnings versus control group without feedback;</p> <p>Objective and subjective benefits of RULA feedback.</p>	
<p>Structure and information content of Ball Valve scenario</p>	<p>How do people structure the Ball Valve workflow? Which information content do people find important to highlight?</p>	<p>User survey (N = 10)</p> <p>Task: segmentation of ball valve workflow, choice of helpful overlays; Number of consensual steps of ball valve, sketches of "augmentations".</p>	 <p>Annotations include: 'Direction in which ball valve has to be moved', 'Direction in which ball valve casing has to be moved', 'Number of repetitions', and 'Direction of putting the cover onto the object'.</p>
<p>Suitability of assistance systems for understanding and learning</p>	<p>How could information be presented in a way that promotes understanding and memory of the task?</p>	<p>Two-group randomized experiment (N = 24);</p> <p>Structured versus unstructured user interface for instructions (during Ball Valve scenario);</p> <p>Structured interface more interesting and more thought-provoking.</p>	 <p>The structured interface shows a sequence of steps: 'Put the covering on top of the positioner' with numbered icons 1, 2, 3, 4, 5. The unstructured interface shows the same instruction with only icon 5 highlighted.</p>

AR player with choice of modes and speech control	Which usability problems can be identified and how can we fix them?	Heuristic evaluation; Exploration of user interaction by experts based on given usability principles; List of usability problems, appropriate suggestions for improvement, and their importance/feasibility	<table border="1"> <thead> <tr> <th>Usability Heuristic</th> </tr> </thead> <tbody> <tr> <td>Visibility of system status</td> </tr> <tr> <td>Match between system and the real world</td> </tr> <tr> <td>User control and freedom</td> </tr> <tr> <td>Consistency and standards</td> </tr> <tr> <td>Error prevention</td> </tr> </tbody> </table>	Usability Heuristic	Visibility of system status	Match between system and the real world	User control and freedom	Consistency and standards	Error prevention
Usability Heuristic									
Visibility of system status									
Match between system and the real world									
User control and freedom									
Consistency and standards									
Error prevention									
User-controlled AR player with animated 3D objects augmented in the HMD	Which conclusive statements can be made on the usability of the system?	Heuristic evaluation; Exploration of user interaction by experts based on given usability principles; List of usability problems, appropriate suggestions for improvement, and their importance/feasibility.							

7. Summary and conclusion

The major goal of COGNITO was to develop enabling technologies for intelligent user assistance systems specifically targeted to industrial manual tasks, i.e. technology for capturing information about the user activity in relation to the environment through sensors, reasoning about this and providing tailored feedback and support through adaptive augmented reality (AR) techniques. Moreover, the system should be mobile by being purely based on sensors attached to the user's body.

The achievements of COGNITO are in line with these objectives. As described in the above sections, we have made important contributions in all of the proposed areas. We have designed a system architecture, which reflects the above requirements, and, thanks to its layered structure (decoupling high-level processing from raw sensory information), provides a workflow learning and monitoring approach, which can tolerate user and workspace variations. The developed algorithms for workspace and workflow monitoring are learning-based and thus provide a generic and domain-independent approach. Also, the system considers two aspects in parallel, workflow analysis and support and biomechanical assessment of the worker. Hence, we address two major trends of today's industry and production section, increasing complexity of workflows and increasing need for ergonomic behaviour in the context of the demographic change. In parallel to the above system architecture, we have also developed a lightweight approach and intuitive tools for segmentation, authoring and monitoring of workflows based on a single RGB camera. The resulting system is easy to setup and use and reduces the amount of manual work required from an expert user to document a workflow. Moreover, the system has proven suitable for demonstration and technology transfer.

While our focus in COGNITO was on the enabling technologies, we have also developed an integrated system covering the complete action-perception-feedback loop, linking data capture with workflow and ergonomic understanding and feedback. Moreover, we have developed and tested this system and its components on three increasingly complex test datasets based on typical industrial manual tasks and thoroughly documented the results in publicly deliverables. At the same time as novel methods enabling new cognitive capabilities have been developed and promising results have been achieved, we have also discovered limitations of the current system, such as for instance the recognition and tracking of small objects and objects with reflective and transparent surfaces and the

recognition and monitoring of fine-grained manipulations, which require further investigation in follow-up projects. Moreover, the ergonomics of the sensor platform, the form factor of the IMUs and the capabilities of the vision sensors in terms of resolution and covered volume need to be improved, which is feasible thanks to latest developments in these areas.

Besides the technical evaluations, continuous formative and summative evaluations of the user interaction have helped us to design and iteratively improve the user interface. Moreover, exploration of ways for enhancing comprehensibility and memorability of information presentation provide routes for future improvements beyond the end of the COGNITO project.

Potential impact

In the course of rapidly changing technological, societal and economic developments industrial work is already today and will be in the future characterised by increasing needs concerning workers' flexibility, adaptability, and qualification. Demographic change involving more ageing and less young skilled workers indicates on the one hand that the expertise of the more experienced has to be captured and transferred efficiently to the less experienced and on the other hand that older workers have to be supported by appropriate cognitive assistance. Furthermore, a shortage of skilled workers requires making the most of every single employee investing in his/ her empowerment and demanding a broader range of tasks and responsibilities from each individual. Finally, accelerated product lifecycles, rapidly developing technologies in manufacturing industry, and increased product variability demand constant adaptability and innovation of the organisations and constant acquisition of new knowledge of the employees. In sum, the industrial work situation is characterised by an increasing creation of knowledge-based value.

Consequently, the demand for appropriate knowledge delivery, assistance and training will increase. Especially for the countries of the EU it is important to develop and invest in advanced assistive technologies in order to stay competitive compared to low-wage countries. Through intelligent systems and innovative assistance and training solutions which foster workers' qualifications, organisations' productivity, product quality and error prevention a competitive advantage may be established. The solutions have to focus on systems that enable an efficient knowledge management by capturing existing skills (that is often implicitly stored in the heads of the workers) and transferring them to co-workers, e.g. novices.

Moreover, in the age of demographic change, ergonomic working conditions will have to be assured, e.g. in order to reduce the risk of musculoskeletal disorders in labour intensive sectors and by this reduce absence due to illness and enable employees to work effectively until a high age. For this, it is necessary to have tools available, which can assess cumulative physical stress during task execution.

COGNITO jointly addresses all of the above challenges by ambitious technologies that enable an assistance system to (1) capture automatically the nature of an industrial workflow during execution by a worker and transferring the content in terms of instructions to an inexperienced or untrained user and (2) documenting postural discomfort during task execution according to the well-known RULA standard and assessing cumulative physical stress (muscle forces and joint loads) associated to manual tasks.

When it comes to user interaction Augmented Reality technology revealed to be the method of choice for user assistance in industrial settings. While in COGNITO the focus was on head-mounted displays, industrial exploitation showed growing interest also in tablet PCs as well as smartphones (cf. [18]). Disregarding the hardware in use, AR demonstrates its usefulness as supportive technology in several application scenarios for instance facility design, quality control, assembly, logistics, maintenance, and operator training (e.g. [19], [20], [21], [22], [23]). From a technological point of view, the COGNITO project may be understood as one of the pioneer projects in the area of AR-based support and cognitive systems engineering for industrial purposes, since it demonstrates the

variety of technologies that have to be integrated in order to provide adaptive, flexible, mobile, and intelligent support that requires minimal further editing.

Although the enabling technologies were the main scope, socio-economically, the COGNITO system addresses all of the before mentioned changes and needs for future education, for (ergonomic) assistance, and for knowledge exchange. COGNITO addresses an independent, worker-controlled on-the-job training and assistance where task-dependent information is displayed through an HMD while executing industrial work. The instructions are automatically presented where and when they are needed. The worker is enabled to cope with tasks he never performed before and train procedures and skills during execution itself without reading manuals, attending training sessions or asking others for help. Furthermore, COGNITO represents an approach to sensor integration and workflow monitoring that result in a sophisticated and challenging way how, for instance, tacit knowledge and motoric skills may be transferred from one person to another by help of cognitive systems. In addition, COGNITO incorporates an innovative way how the combination of sensor integration and biomechanical parameters enables an objective ergonomic assessment of workers' movements. In sum, the project results contribute to safer, more efficient, less error prone, higher-quality, educative and satisfying work conditions.

In a wider societal context the relationship between human workers and intelligent systems and robots respectively is of great interest and importance. Many people are sceptical if and to which extent robotic systems should replace or affect human work and human thinking. With growing sophistication of technological opportunities it will be to decide how to implement them, either in a way where humans remain only a helping hand of the machine or rather in a user-centred way fostering human skills, willingness to exchange their knowledge, and (decisional) control. With the COGNITO project we clearly followed the latter approach which is shown by the continuous testing and evaluation activities. Future research dealing with a wider social impact of such systems has to address new ways of how qualification may be implemented (cf. [24]), and to involve the organisational context by providing appropriate incentives (cf. [25]), cooperation concepts and work design means. The COGNITO project is supposed to encourage further development in this area both technologically and on a higher social level.

Dissemination

Throughout the project lifetime, the whole consortium actively participated in raising the external awareness of COGNITO and in disseminating its results through various channels, targeting the scientific, technical, user and professional community, as well as the public at large. Dissemination activities ranged from flyers and press releases over publications and demonstrations in conferences to exhibitions and dissemination is still ongoing.

For reaching the broader public, the COGNITO website (www.ict-cognito.org) was regularly updated with the latest results and news informing about the project progress.

All partners actively published and demonstrated their results in relevant conferences (e.g. ISMAR, BMVC, IROS, Ueware, ACCV), thus raising the awareness of the project in the scientific community. An article, jointly written by UTC, DFKI and SmartFactory was accepted for publication in *Applied Ergonomics*. Further articles are under review and a final joint publication is planned in *PLOS ONE*.

As a joint event, a workshop "Cognitive assistive systems: closing the action-perception loop" was organised by the consortium in collaboration with a team from KTH University (Sweden) at the major robotics conference IROS (International Conference on Intelligent Robots and Systems). This workshop was used as a platform for presenting the COGNITO results and for discussions with other experts working in the field of cognitive systems. Since the workshop was a successful event and had attracted well-known experts from the field, it is planned to continue this in the next year. Moreover,

based on the advice of some of the participants, a proposal for a special issue has been submitted to RAS (Robotics and Autonomous Systems).

Also, COGNITO results were demonstrated at major exhibitions, such as the Hannover Messe and CeBIT, gaining high interest and positive feedback.

Throughout the project the SmartFactory has acted as a mediator and actively communicated the COGNITO concept and results to visitors of the SmartFactory living lab and through the organisation of major events, such as their yearly Innovation Day, where in 2012, among other exhibits, three COGNITO demonstrators have been presented to around 70 invited visitors from industry as well as researchers and politicians.

Major dissemination and exploitation activities are also documented on the project website (www.ict-cognito.org) in the news, publications and press category.

Exploitation

Industrial exploitation has been realised throughout the project lifetime in several fairs (e.g. CeBIT, Hannover-Messe) and events, such as the SmartFactory Innovation Day, where industrial partners come together to experience and discuss new technologies once a year. Several presentations and demonstrations including the Augmented Reality Manual, the Online Ergonomic Assessment, the on-body sensor network and motion capturing capabilities, the object learning and tracking framework and the output device ensured public visibility of the project.

Furthermore, in the Living Lab of SmartFactory the idea of COGNITO was taken up in a less sophisticated (fully marker-based and stationary) but permanently visible demonstrator for industrial visitors and press in order to present and communicate the potential of Augmented Reality-based assistance and training for industrial purposes.

In the course of two industrial competitions, Audi Production Award 2012 and award of Software AG 2012, SmartFactory elaborated the impact of AR-based assistance and training systems for manual industrial tasks and won the first prize respectively. In both cases the prizes are starting points for further development and exploitation of the idea for real application in cooperation with the industrial partners (AUDI AG and Software AG).

Furthermore, the industrial partner Trivisio exploited the results of COGNITO by bringing new products to the market. The new monocular see-through HMD and the new wireless IMU generation, ColibriW, were commercialised at early stages of the project. The wireless units are sold in conjunction with the software development kit and application programming interface developed by DFKI (cf. Section 3) and are today, after continuous improvements in hardware, calibration and software, already used by customers world-wide, including a global player in the German car industry.

The monocular workflow segmentation, authoring and monitoring tools have been demonstrated at CeBIT 2012 and attracted great interest among industrial visitors. A small-scale study contracted by an international enterprise has been performed with the perspective of a larger follow-up project.

One of the most promising exploitations of motion tracking developed in the course of COGNITO lies within the ergonomic analysis during task execution. The Online Ergonomic Assessment demonstrator as published in [15] gained high interest of media and industry. The significance of such a system is clearly promoted by the need for objective assessment of ergonomic risk factors at every working place in the context of the prevention and health surveillance principles as explained by the Community strategy on health and safety at work from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions of the 21st February 2007. Also the markerless hand tracking technology, the detailed musculoskeletal hand model and the tools for assessing muscle forces and joint loads based on these, are necessary enabling technologies for being able to document the cumulative physical stress

on the worker during a task. We expect that the commercial need for both online and offline objective assessment tools will increase in the future.

The impact of the generic COGNITO technologies, such as motion tracking, object recognition and tracking, pattern recognition, etc. go beyond the industrial area of the COGNITO project. These enabling technologies are relevant in many other applications, such as interaction in virtual environments (e.g. training simulations), robotics (e.g. teleoperation, programming by demonstration, autonomous robots), medicine and rehabilitation (e.g. diagnosis of false postures, support of rehabilitation exercises) as well as video games and animated movies (e.g. Pixar movies), to name just a few examples. Accordingly, the technologies developed within COGNITO have already been the basis for acquiring industrial projects as well as follow-up research projects and for submitting proposals formulating continuative project ideas to the latest FP7 Call 10.

Another route of exploitation, particularly important for the universities and research institutes within the COGNITO consortium, is the academic exploitation, which has been continuously realised throughout the project lifetime. The COGNITO topics and results have been used for lectures, seminars, courses and training of students at the participating universities' sites, thus transferring the newly gained insights to the next generation of researchers. Also, the consortium has followed an active policy of publishing the project results in refereed journals and conferences and plans to continue these activities after the end of the project in order to make the results available to the scientific community.

Acknowledgements

The COGNITO project was funded under the seventh framework program (ICT-2009.2.1, grant number 248290). The consortium would like to thank the EU for the financial support. For more information, please visit the website www.ictcognito.org.

The COGNITO consortium

Number	Partner name / main contact and e-mail	Partner short name	Country
1	German Research Center for Artificial Intelligence Didier Stricker, Didier.Stricker@dfki.de	DFKI	DE
2	University of Bristol Andrew Calway, andrew@cs.bris.ac.uk	Bristol	UK
3	University of Leeds David Hogg, d.c.hogg@leeds.ac.uk	LEEDS	UK
4	CNRS / University of Compiegne Frederic Marin, frederic.marin@utc.fr	CNRS	FR
5	Trivisio Prototyping GmbH Gerrit Spaas, spaas@trivisio.com	Trivisio	DE
6	Center for Computer Graphics Luis Almeida, Luis.Almeida@ccg.pt	CCG	PT
7	Technologie-Initiative SmartFactory KL e.V. Katharina Mura, mura@smartfactory.de	SmartFactory	DE

References

- [1] D. Damen, A. Gee, W. Mayol-Cuevas and A. Calway, "Egocentric Real-time Workspace Monitoring using an RGB-D Camera," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, 2012.
- [2] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohli, J. Shotton, S. Hodges and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.
- [3] A. Gee and W. Mayol-Cuevas, "6D Relocalisation for RGBD Cameras Using Synthetic View Regression," in *British Machine Vision Conference (BMVC)*, Surrey, UK, 2012.
- [4] D. Damen, P. Bunnun, A. Calway and W. Mayol-Cuevas, "Real-time Learning and Detection of 3D Texture-less Objects: A Scalable Approach," in *British Machine Vision Conference (BMVC)*, Surrey, UK, 2012.
- [5] G. Bleser, G. Hendeby and M. Miezal, "Using Egocentric Vision to Achieve Robust Inertial Body Tracking under Magnetic Disturbances," in *International Symposium on Mixed and Augmented Reality (ISMAR)*, Basel, Switzerland, 2011.
- [6] G. Bleser and D. Stricker, "Advanced tracking through efficient image processing and visual-inertial sensor fusion," *Computer & Graphics*, pp. 59-72, 2009.
- [7] G. Bleser and G. Hendeby, "Using optical flow for filling the gaps in visual-inertial tracking," *Proc. European Signal Processing Conference (EUSIPCO)*, 2010.
- [8] M. Miezal, B. Gabriele, S. Didier and J. Tümler, "Towards practical inside-out head tracking for mobile seating bucks," in *ISMAR 2012 Workshop on Tracking Methods and Applications*, Atlanta, Georgia, USA, 2012.
- [9] N. Petersen and D. Stricker, "Fast Hand Detection Using Posture Invariant Constraints," *KI*, pp. 106-113, 2009.
- [10] N. Petersen and D. Stricker, "Morphing billboards for accurate reproduction of shape and shading of articulated objects with an application to real-time hand tracking," in *Computational Modeling of Objects Presented in Images (ComplImage)*, Rome, Italy, 2012.
- [11] N. Petersen and D. Stricker, "Adaptive Search-Tree Database-Indexing for Hand Tracking," in *Computer Graphics, Visualization, Computer Vision and Image Processing 2012. IADIS International Conference Computer Graphics, Visualization, Computer Vision and Image Processing (CGVCVIP-2012)*, Lisbon, Portugal, 2012.
- [12] A. Behera, D. C. Hogg and A. G. Cohn, "Egocentric Activity Monitoring and Recovery," in *The 11th Asian Conference on Computer Vision (ACCV)*, Daejeon, Korea, 2012.

- [13] S. F. Worgan, A. Behera, A. G. Cohn and D. C. Hogg, "Exploiting petri-net structure for activity classification and user instruction within an industrial setting," in *International Conference on Multimodal Interfaces*, Alicante, 2011.
- [14] A. Behera, A. G. Cohn and D. Hogg, "Workflow Activity Monitoring using the Dynamics of Pairwise Qualitative Spatial Relations," in *International Conference on MultiMedia Modeling (MMM)*, Klagenfurt, Austria, 2012.
- [15] N. Vignais, M. Miezal, G. Bleser, K. Mura, D. Gorecky and F. Marin, "Innovative system for real-time ergonomic feedback in industrial manufacturing (in press)," *Applied Ergonomics*, 2013.
- [16] N. Petersen and D. Stricker, "Learning Task Structure from Video Examples for Workflow Tracking and Authoring," in *International Symposium on Mixed and Augmented Reality (ISMAR)*, Atlanta, GA, 2012.
- [17] M. Tuceryan and N. Navab, "Single point active alignment method (SPAAM) for optical see-through HMD calibration for AR," *Proceedings of the IEEE and ACM International Symposium on Augmented Reality*, pp. 149 - 158, 2000.
- [18] D. Gorecky, R. Campos and G. Meixner, "Seamless Augmented-Reality Support on the Shopfloor Based On Cyber-Physical-Systems," in *MobileHCI*, San Francisco, CA, USA, 2012.
- [19] T. Alt, *Augmented Reality in der Produktion*, München: Herbert Utz Verlag GmbH, 2003.
- [20] N. Gavish, T. Gutierrez, S. Webel, J. Rodriguez and F. Tecchia, "Design Guidelines for the Development of Virtual Reality and Augmented Reality Training Systems for Maintenance and Assembly Tasks," in *BIO Web of Conferences*, 2011.
- [21] W. Friedrich, D. Jahn and L. Schmidt, "ARVIKA - Augmented Reality for development, production and service," *Proceedings of the International Status Conference (HCI)*, vol. 3, p. 34, 2001.
- [22] S. Webel, U. Bockholt, T. Engelke, M. Peveri, M. Olbrich and C. Preusche, "Augmented Reality Training for Assembly and Maintenance Skills," in *BIO Web of Conference*, 2011.
- [23] O. Korn, "Industrial Playgrounds, How Gamification helps to enrich work of elderly or impaired persons in production," in *EICS 2012*, 2012.
- [24] E. Eiriksdottir and R. Catrambone, "Procedural Instruction, Principles, and Examples: How to Structure Instructions for Procedural Tasks to Enhance Performance, Learning and Transfer," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 6, pp. 749-770, 2011.
- [25] J. H. Bishop, "What We Know About Employer-Provided Training: A Review of Literature," in *CAHRS Working Paper Series*, 1996.

German Research Center for Artificial Intelligence (DFKI) GmbH

DFKI Bremen

Robert-Hooke-Straße 5

28359 Bremen

Germany

Phone: +49 421 178 45 4100

Fax: +49 421 178 45 4150

DFKI Saarbrücken

Stuhlsatzenhausweg 3

Campus D3 2

66123 Saarbrücken

Germany

Phone: +49 681 875 75 0

Fax: +49 681 857 75 5341

DFKI Kaiserslautern

Trippstadter Straße 122

67608 Kaiserslautern

Germany

Phone: +49 631 205 75 0

Fax: +49 631 205 75 5030

DFKI Projektbüro Berlin

Alt-Moabit 91c

10559 Berlin

Germany

Phone: +49 30 238 95 1800

E-mail:

reports@dfki.de

Further information:

<http://www.dfki.de>