Tools for semi-automatic monitoring of industrial workflows

Roland Mörzinger, Marcus Thaler, Schuster Rene, Hofmann Albert and Georg Thallinger
JOANNEUM RESEARCH Forschungsgesellschaft mbH
Graz, Austria
roland.moerzinger@joanneum.at

Manolis Sardis, Vassileios Anagnostopoulos, Dimitrios Kosmopoulos, Athanasios Voulodimos, Constantinos Lalos, Nikolaos Doulamis and Theodora Varvarigou
National Technical University of Athens
Athens, Greece
sardis@telecom.ntua.gr

Igor Rosenberg, Rolando Palma Zelada and Ignacio Soler Jubert
ATOS Origin
Barcelona, Spain
igor.rosenberg@atosresearch.eu

Helmut Grabner, Severin Stalder and Luc Van Gool
ETH, Vision Lab
Zurich, Switzerland
grabner@vision.ee.ethz.ch

Galina Veres, Lee Middleton and Zoheir Sabeur
University of Southampton, IT Innovation Centre, UK
zas@it-innovation.soton.ac.uk

Imed Bouchrika, Banafshe Arbab-Zavar, John Carter and Mark Nixon
Electronics & Computer Science Southampton University, UK
msn@ecs.soton.ac.uk

ABSTRACT

This paper describes a tool chain for monitoring complex workflows. Statistics obtained from automatic workflow monitoring in a car assembly environment assist in improving industrial safety and process quality. To this end, we propose automatic detection and tracking of humans and their activity in multiple networked cameras. The described tools offer human operators retrospective analysis of a huge amount of pre-recorded and analyzed footage from multiple cameras in order to get a comprehensive overview of the workflows. Furthermore, the tools help technical administrators in adjusting algorithms by letting the user correct detections (for relevance feedback) and ground truth for evaluation. Another important feature of the tool chain is the capability to inform the employees about potentially risky conditions using the tool for automatic detection of unusual scenes.

Categories and Subject Descriptors

I.5.5 [**Pattern Recognition**]: Applications – *Computer vision.*

General Terms

Algorithms, Security, Human Factors.

Keywords

Computer vision, industrial environments, applications, human detection and tracking, workflow recognition

1. INTRODUCTION

Large-scale enterprises like industrial plants or public infrastructure organizations have a clear need for supervision services to guarantee: (a) quality – adherence to predefined procedures for production or services and (b) security and safety – prevention of actions that may lead to hazardous situations. Such supervision services are frequently of vital importance for the enterprise/organization. However, they are performed manually and thus inefficiently and subjectively. The inefficiency stems from the fact that the videos from many cameras are displayed on monitors that switch between cameras, thus no 100% monitoring is possible, even if we assume that the operators are constantly concentrated on their task. Regarding subjectivity, recent studies [1] have proven that the attention of the operators of current surveillance systems is mainly attracted by the appearance of monitored individuals and not by their behavior. This practice of course raises several privacy issues.

Recently several supervision systems have been presented, which are able to detect a limited number of predefined security-

related events using a small number of cameras. Although all these systems could theoretically be applied for supervision purposes, their application is limited especially in large scale, which requires a high number of collaborating cameras. The system configuration in those cases is extremely laborious and the complexity of identifying activities through a network of collaborating cameras makes the system application questionable. Furthermore, the high diversity and complexity of the behaviors which need to be monitored makes offline scene modeling unrealistic. It is clear that the next generation of such systems will have to require minimal human intervention and will have to be scalable, so that they can be easily applied in large enterprises [2]. These systems must be able to operate in a largely unsupervised way, by (i) automatically learning salient objects and behaviours, (ii) focusing attention on really important events, (iii) self-configuring and coordinating the cameras, (iv) detecting novelty regarding salient objects and behaviours, (v) self-evolution and adaptation with respect to the changes and variations of the environmental conditions and the monitored objects and behaviours.

In this paper, we propose the main innovative tools that have been already developed and tested within the European Union initiative SCOVIS (Self-Configurable Cognitive Video Supervision) [2] in a real industrial environment. Specifically, the following tools have been developed within the framework of this project; the SCOVIS Data Browser (SDB), the Unusual Scene Detector (USD), the Workflow Recognition through Artificial Intelligence Tool (WRAIT) and the Annotation Feedback Tool (AFT). The USD and the WRAIT tool are designed for online application, whereas the SDB and AFT are for offline usage. Each tool answers distinct needs, but they are all operating on the same dataset and they share algorithmic outputs, as described in the following sections.

2. INDUSTRIAL DATASET

In the course of the project a real-world industrial dataset was recorded in the NISSAN Motor Iberica SA plant. In this environment an automotive assembly process with 3 operators is monitored where one operator provides pieces to be assembled and two other ones are handling pieces to put them over an assembly (robotic) machine. The dataset, depicted in Figure 8, is composed of 5 JPEG image sequences captured at a varying frame rate of 18-25fps with resolution 704x576 and 640x480. Specifically, the scene was captured by 4 AXIS 213 PTZ network cameras with partly overlapping views and one wide-angle AXIS 212 PTZ camera mounted on top of the working cell. The total duration of each sequence is approximately 15 hours (2 shifts on two consecutive days). Synchronicity is ensured by time-stamped filenames.

3. OVERVIEW OF ALGORITHMS

The proposed tools are applied in real scenarios where many employees are part of the scene. Therefore, static or moving object detection is applied in multiple areas that are monitored by a system of multiple cameras. Multiple camera systems are used in order to increase the efficiency of tracking algorithms and to prevent the system from losing the tracked target in case of object occlusions. First we are detecting objects and track them through the industrial workflows videos. Then by applying multi-camera calibration techniques and information fusion per camera, we are identifying and characterizing the workflows through the segmentation of internal subtasks that each workflow contains. The related techniques are outlined in the following sections. For detailed information on the used techniques, the reader is referred to the publications referenced in the text.

3.1 Basic Object Detection and Tracking

Person detection and tracking from visual observations are highly challenging especially within an industrial environment. In such conditions, it is difficult to discern persons due to sparks and vibrations, difficult structured background (e.g., upright racks, heavy occlusions of the workers in most parts of the image), and other moving objects (e.g., welding machines and forklifts). Existing methods like color background modeling and generic person detectors perform dismally on that dataset. Therefore we proposed to learn grid detectors [3] which are more adapted to the observed scene rather than relying on pre-trained detectors. Based on this line of work to improve person detectors locally and using scene specific context, we recently proposed a novel approach to increase the robustness of any generic person detection algorithm. The cascaded confidence filter [4] successively incorporates constraints on the size of the object, on the appearance of the background and on the smoothness of the trajectories. In fact, we model the continuous detection confidences similarly to traditional background pixel color modeling. The smoothness of trajectories is ensured through a process analogous to vessel filtering in medical imaging. The approach does not learn specific object models, incorporate scene specific constraints, reason about complete trajectories nor use multiple cameras. Therefore, it can serve as preprocessing step to robustify many tracking algorithms. The experimental validation shows significantly improved performance in case of difficult scenes while conserving the formidable performance of a person detector to detect fully visible persons. A simple example is shown in Figure 1 in which the cascaded confidence filters robustly detect persons in the NISSAN industrial environment.

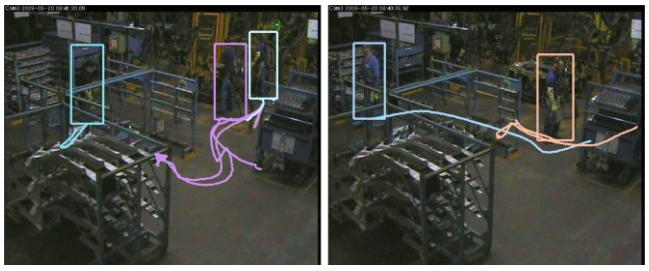

Figure 1 Illustration of typical tracking results using the cascaded confidence filter to increase the robustness of person detection. Such long trajectories become only possible using the intermediate detection filtering steps, developed during the project.

Conference'10, Month 1–2, 2010, City, State, Country.

3.2 Multiple Camera Calibration

The correspondence in distributed cameras with overlapping views is a prerequisite for consistent monitoring, e.g. for identification and resolving occlusions. In the SCOVIS project we developed an approach for automatically establishing the cross-camera correspondence for planar scenes. In many real world scenarios automatic calibration based on correspondence between image features (colour, SIFT) is not applicable. Therefore we propose to automatically estimate an inter-image homography from person detections only without using any prior correspondence information [5]. Detections are filtered by automatic scene scale estimation [6] and moreover temporal and geometric conditions are exploited to identify the true correspondences in the set of all possible detections pairs. The approach is self-configurable, adaptive, provides robustness over time and is also suited for wide angles between the two camera views. Experiments show that compared to a RANSAC-based baseline approach our method is able to produce adequate results on difficult datasets, which feature a small ratio of true to false correspondence. This calibration information is a prerequisite for multi-camera tracking. Figure 2 shows an example where single view tracking results from two cameras are fused on a ground plane using a constrained nearest neighbor clustering for obtaining occlusion free human detections.

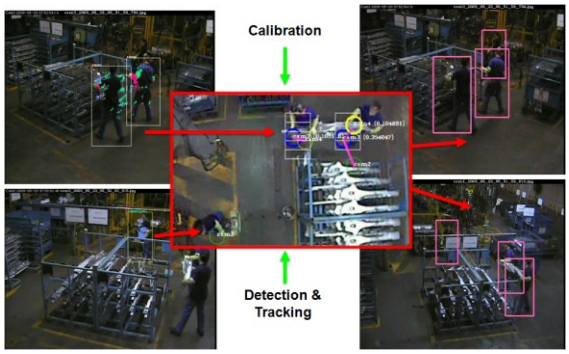

Figure 2 Illustration of multi-camera tracking. Single-view detection and tracking results (left) are fused on the ground plane using the automatically estimated calibration information (center) for occlusion-free human person detection and tracking (right).

3.3 Workflow Analysis

Robust automated workflow monitoring using visual sensors in industrial environments such as those described in earlier sections is a notoriously difficult problem in computer vision. Under those industrial conditions, it is very hard to acquire good quality data for achieving intelligent workflow analysis tools. Nevertheless, it is important to address such problems by building generic quasi-performing automated workflow monitoring tools for industrial operation management purposes. This view will contribute in the improvement of process quality standards and health and safety in industrial manufacturing environments of the future [7].

Eight hours of video camera scenes have been acquired, analysed and led to the description of specific operational workflows conducted in a car assembly environment. These workflows were split into specific tasks which were significantly distinct in their corresponding activities. The tasks can be of varying lengths and may occur in any sequential order. Furthermore, two tasks can happen concurrently and make their detection and recognition more complex.

For real-time workflow monitoring two approaches are proposed. The first one is based on the use of holistic scene descriptors (background subtraction, Pixel Change History, Zernike Moments) and Hidden Markov Models (HMM). HMMs are very popular as a classification means, mainly due to the fact that they can efficiently model stochastic time series at various time scales. We are experimenting using both Gaussian and Student-t distributions as the observation likelihood in order to achieve more robust modeling. The problems of limited visibility and occlusions are addressed by extending the framework for multiple cameras. We use fused HMMs in order to exploit complementarities of the different views. For this, we experiment with various fusion approaches, such as feature fusion, synchronous HMM, parallel HMM, and multi-stream fused HMM, which seems to provide the most significant enhancement in comparison to the single stream HMM. The model is being further enhanced to support online real-time task and workflow recognition based on a Bayesian filtering method.

The second methodology consists of robust and simple scene descriptor to detect features and an efficient time series classifier. Specifically, as an input we used local motion descriptors. In order to learn from these descriptors an efficient time series analysis technique called the Echo State Network (ESN) [8] was applied. The ESN consists of a large set of connected nodes with a fixed number of trainable outputs and a number of randomly selected fixed inputs and hidden states. ESN is easy to train and update online, i.e. it allows the inclusion of prior knowledge and adaptation to new workflows. Multiple activities which could occur simultaneously can be easily incorporated while the approach is still robust to irrelevant motions. Recorded data was equally partitioned for training and testing purposes.

Experimental results on the testing set show that our approach can successfully monitor complex workflows automatically with a recall of 75% and precision reaching a maximum of 80%. Illustrative results of our method in automated workflow monitoring are shown in Figure 3 below.

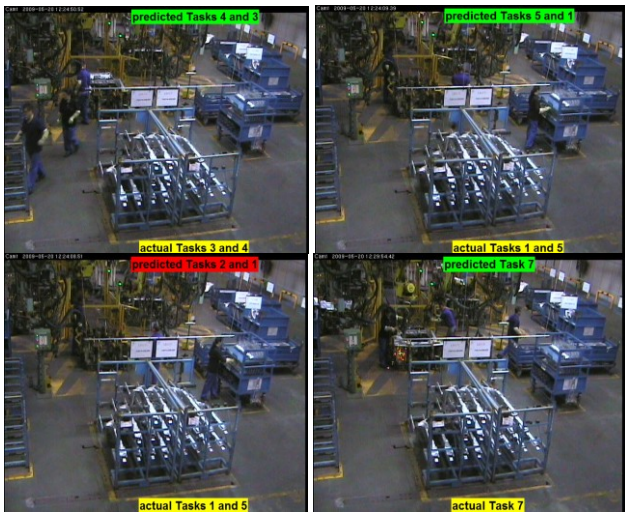

Figure 3 Successful recognitions and failures over one working cycle.

3.4 Human Activity Detection

The detection of human activity has been performed using an overhead view. This was chosen over more conventional views as it does not suffer from occlusion, but still retains powerful cues about the identity/activity [9] of individuals. A simple blob tracker has been used to track the most significant moving parts i.e. human beings. The output of the tracking stage was manually labeled into 4 distinct categories: walking; carrying; handling and standing still which taken together form the basic building blocks of a higher work flow description. These were used to train a decision tree using one subset of the data. On independent testing it was found that the activity was correctly identified in more than 80% of the video frames analyzed. Figure 4 shows some results of individual frames and how they might be combined together to form a work flow task.

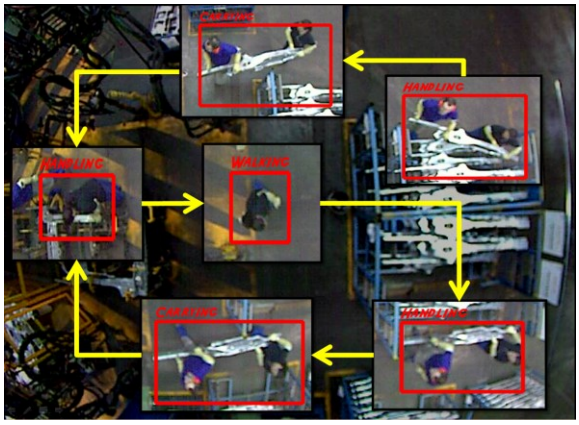

Figure 4 Output of the human activity detector.

4. SCOVIS DATA BROWSER

Managing a lean factory requires systems that enable quick access to information about process efficiency and quality. Vision systems using multiple cameras are often used for monitoring large-scale manufacturing areas. Human observers cannot observe many information streams simultaneously, as it is the case in surveillance control rooms where multiple streams converge.

The SCOVIS data browser (SDB) is a tool which offers retrospective analysis of huge amount of prerecorded footage from multiple cameras. The SDB performs the task of automatic camera selection based on recognition probability, thus reducing redundancy and creating a low-cost video documentary. Intelligent algorithms for multi-camera fusion and integration help the users to get a quick access and comprehensive overview about the work-in-process. The SDB also serves as platform for the integration of results from various algorithms such as object detection, tracking, activity, task, and workflow identification. Interactive temporal exploration and navigation is supported by different visualizations, such as key frame summaries, skims and timelines indicating object appearances and detected workflows. Case studies where information is extracted from videos obtained in a multi-camera manufacturing environment demonstrate the utility of the system. Figure 5 shows a screen shot of the SDB implemented in C++. On the right hand, the visual content is depicted while on the left hand some additional information is shown.

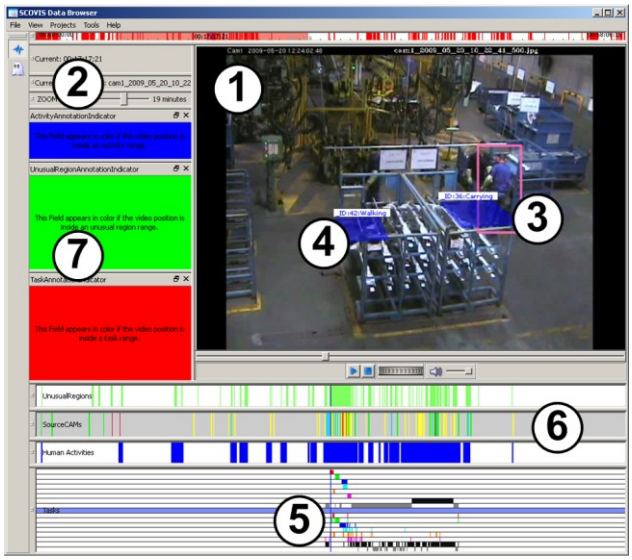

Figure 5 Screenshot of the SCOVIS Data Browser (SDB) with different views. Input video streams (1) can be inspected using large-scale data browsing facilities (2) and various timelines presenting different meta-data, such as the "best-view" camera selection (6) and automatic time-based task detection (5). Results from human detection and tracking (3) and human activity recognition (4) are presented. Colored widgets indicate the happening of various important events (7).

5. UNUSUAL SCENE DETECTOR

Automatic and online identification of unusual incidents is important for critical event detection, alarm systems and the implementation of health and safety measures. In today's camera surveillance solutions the video streams are displayed on-screen for human operators, e.g. in large multi-screen control centers. This in turn requires the attention of operators for unusual events and urgent response. We have developed an unusual scene detector (USD) that is able to automatically identify unusual visual content in video streams in real-time. A scene (image region) is identified as unusual if similar image content on the same place has not or rarely been observed in the past. In contrast to explicitly modelling specific unusual events, the USD incrementally learns from the visual source in order to build an adaptive model of the significant content of the past. At the same time, the current image is compared to the model to identify potential unusual regions in the scene. Both tasks can be performed in parallel and run in real-time (20 fps) on 640x480 video streams (Dual Core 2.66 Ghz/ 3 GB RAM).

Long-term experiments demonstrate that our method finds plausible unusual scenes in traffic monitoring and industrial manufacturing data, but the USD is not limited to any domain. Due to the high run-time performance and universal applicability, the USD can be demonstrated live, i.e. interaction with the audience via webcam. Figure 6 shows the flexible user interface of the unusual scene detector implemented in C++.

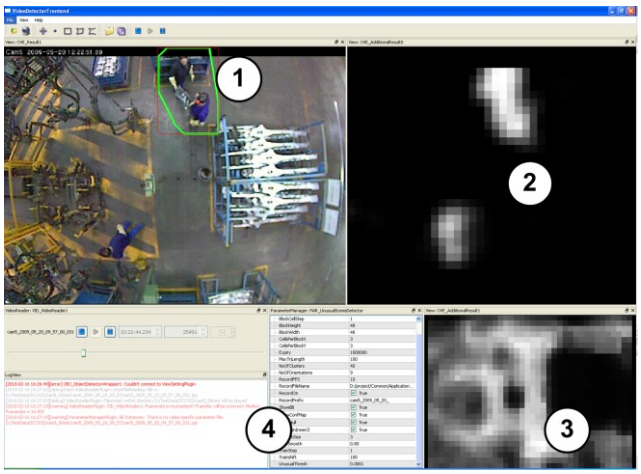

Figure 6 Screenshot of the Unusual Scene Detector (USD) with different views. Input video stream with indicated unusual areas (1), confidence map (2), visualization of the model of usualness (3) and widgets for administration and logging (4).

6. WORKFLOW RECOGNITION USING ARTIFICIAL INTELLIGENCE TOOL

The scope of the WRAIT tool is to identify and recognize workflows of single humans/objects based on the results from human detection and tracking algorithm [10]. The combined technologies used are artificial intelligence and computer vision.

The artificial intelligence part handles the transformation of the trajectories of the objects to semantic events and inference upon the normality of the sequence of events. This normality can be validated against a rule table. Sequences of semantic events are checked for conformance to a rule. The semantic events take the form of *enter/leave* events of points in the image space. These semantic events will be called henceforth micro-tasks. These points in space are called trap points since they are visual traps for reporting. The micro-tasks summarize a trajectory of the worker. The user combines these elementary tasks/events into meaningful sequences that should be tracked across frame sequences. Every rule is a sequence of micro-tasks. Every micro-task has an expiration time. In order to understand this requirement it is enough to realize that in everyday activities a human object accomplishes a job in a finite amount of time. In industrial environments this amount of time comes with some lower and upper bounds. Moreover even in walking tasks the small amount of time necessary to emit the enter/leave events with respect to a trap point play a crucial role to detect accidents. For this reason a micro-task with its expiration time are crucial in our approach.

Parallelizing the processing of these tasks is straightforward since trap points are autonomous entities doing the same thing. This autonomy can be translated directly to an agent based implementation. For this reason we achieved a proper decomposition of the mechanism used to monitor these points into an artificial intelligence and in specific using multi-agents infrastructure that executes a distributed algorithm [11]. Not only load is balanced but we also mimic the processes occurring in nature. Multi-processing happens in a higher level in contrast to the usual lower level that speeds up calculations. We developed a cross-platform tool for experimentation that loads "rules", in other words relevant sequences of micro tasks to identify single-object behaviors. In Figure 7 we present the layout of the WRAIT Tool.

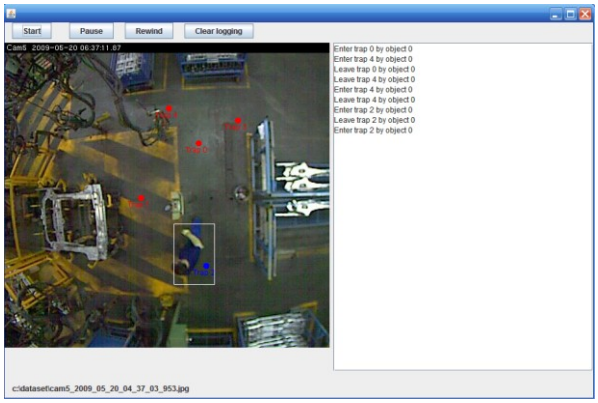

Figure 7 Check point events and related workflows recognitions.

7. ANNOTATION FEEDBACK TOOL

Although self-configured algorithms are the final objectives, many algorithms still require training. A visual interface is required to allow non-scientific user to define the ground truth or bring corrections. This is the role of the Annotation Feedback Tool AFT (see Figure 8). The user corrections (interpreted as feedback) can then be given back to the original tools which produced them.

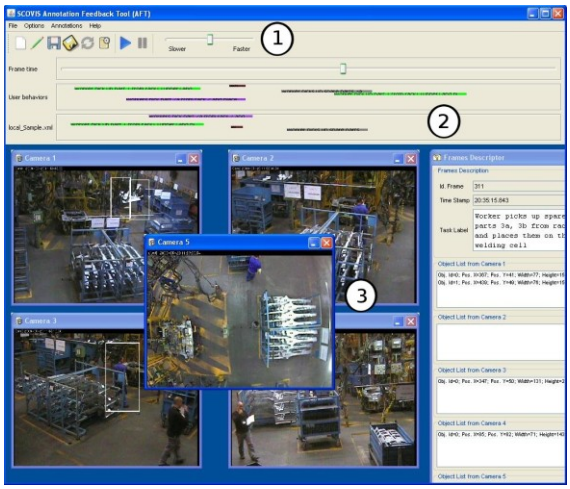

Figure 8 Screenshot of the AFT main window with (1) menu for interaction to load image data, annotations files, and control of which elements to display; (2) the time panel, which presents the frames of the image data as a horizontal timeline; (3) the annotations (which are created manually or loaded as output of tools) overlaid to the video stream.

A statistics window allows reviewing the annotations in a statistical view, by comparing occurrences. This offers the end-user a business view of the annotations, focusing on the real value the annotations are bringing, and possibly highlighting discrepancies. On a side note, as this tool offers the system technician the ability to create new annotations, thus it can be used to create a ground truth data set, which then is used to evaluate the algorithms. The annotation files accepted are plain text files, with each line referencing a given frame, followed by the relevant annotations. The annotations can be bounding boxes, or behaviors (for example "Welding part B to part A"). For highly

detailed annotations, the SAM metadata model [12] could be used. The data sets, once loaded, appear graphically in the main view, and also in the time line. The user can modify the annotations which define bounding boxes. These annotations will typically define the position of people (operators of the factory).

The AFT is ideal for a situation requiring the review of annotations. In an "annotation and feedback for adaptation phase" scenario, the end-user needs to review the results produced by annotation algorithms, in order to tweak their training. The statistics viewer helps to understand the global aspect of the whole sequence. The AFT is used as a user interface to provide the relevance feedback to the tools which expect to be trained. When a specific event requires an en even more detailed view of a specific sequence, and easy switch to the SDB tool is offered.

8. RESULTS AND CONCLUSIONS

This paper describes algorithms and end-user tools developed in the course of the SCOVIS project [2]. The project's goal is to investigate weakly supervised learning algorithms and self-adaptation strategies for analysis of visually observable procedures in an industrial environment. The used industrial dataset comprises extensive recordings from multiple cameras for many hours of the production plant under very difficult imaging conditions. This dataset triggered many interesting and innovative research questions and scientific challenges which include, but are not limited to, object detection in occluded low light scenes, and vague human activity identification. The dataset is a perfectly suited test bed for algorithms that are to be employed in challenging environments.

This has led to significantly more robust person detection algorithms which nevertheless remain very general. The recent improvements in the field of multi-camera coordination and large scale multi-camera monitoring allows for placing cameras in arbitrary situations with minimal manual calibration effort. In addition to that an approach that can successfully monitor complex workflows automatically has been proposed.

The technological approach adopted is able to react and comprehend the executed procedures and the employees' behaviour without requiring extensive training. Through its capabilities of on-line adaptation and self-learning, the developed approach is able to work in open-ended environments and react to unforeseen situations. The adaptation will be performed either autonomously through an evolutionary learning process, or through high-level feedback.

The developed end-user tools support the semi-automatic monitoring of industrial workflows and increase the efficiency and quality of the car production process. In the SDB tool interactive temporal exploration and navigation is supported by different visualizations, such as key frame summaries, skims and timelines indicating object appearances and detected workflows. These features enable quick access and manipulation to a huge amount of video data and vision based algorithmic results obtained from multiple cameras (hundreds of hours). The AFT tool is novel due to its ease of use, its support of multiple cameras and the integration of a variety of different algorithmic results. It allows reviewing and modifying the results of detection tools in a synthetic way. The WRAIT provides rule based workflow analysis in a distributed architecture. This cross-platform tool operates with sequences of micro tasks in order to identify single-object behaviors. To support large multi-screen control centers we developed a system for automatic and online identification of unusual incidents (critical event detection). This tool, called USD, is particularly useful to implement health and safety measures in industrial environments. All of the tools are designed and built in a way that can be easily applied in a variety of different environments where human or objects activities occur.

9. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 - Challenge 2 - Cognitive Systems, Interaction, Robotics - under grant agreement n. 216465 – SCOVIS project.

10. REFERENCES

[1] Smith G.J.D: Behind the screens: Examining constructions of deviance and informal practices among cctv control room operators in the UK. *In Surveillance and Society*, Vol 2, 2004.

[2] Doulamis, A., Kosmopoulos, D., Sardis, M., and Varvarigou, T.: An Architecture for a Self Configurable Video Supervision. *In ACM Workshop in Analysis and Retrieval of Events, Actions, Workflows (AREA) in conjunction with ACM Multimedia*, Vancouver, Canada, 2008.

[3] Stalder, S., Grabner, H. and Van Gool, L.: Exploring Context to Learn Scene Specific Object Detectors. *In Proceedings CVPR09 Workshop on PETS,* 2009

[4] Stalder, S., Grabner, H. and Van Gool, L.: Cascaded Confidence Filtering for Improved Tracking-by-Detection *In Proceedings ECCV,* 2010

[5] Thaler, M. and Mörzinger, R.: Automatic Inter-Image Homography Estimation from Person Detections. *(to appear) In Int. Conf. on Advanced Video and Signal-Based Surveillance*, AVSS 2010

[6] Mörzinger, R. and Thaler, M.: Improving Person Detection in Videos By Automatic Scene Adaptation. *In VISAPP* 2010.

[7] EU Commission 2010. A new European model of production systems for the factories of the future**.** http://ec.europa.eu/research/industrial_technologies/lists/factories-of-the-future_en.html, last accessed 2010-06-25

[8] Jaeger, H.: The "echo state" approach to analyzing and training recurrent neural networks. *Technical Report GMD Report 148, German National Research Institute for Computer Science*, 2001

[9] Seely, R. D., Samangooei, S., Middleton, L., Carter, J. and Nixon, M.: The University of Southampton Multi-Biometric Tunnel and introducing a novel 3D gait dataset. *In Biometrics: Theory, Applications and Systems*, 2008

[10] Sardis, M., Anagnostopoulos, V. and Varvarigou, T.: Workflows Recognition through Multi Agents in Surveillance systems, *In IFIP Conference on "Artificial Intelligence,* 2010.

[11] Sardis, E., Anagnostopoulos, V., Varvarigou, T.: Multi-agent Based Surveillance of Workflows. *The 3rd WS on Logics for Intelligent Agents and Multi-Agent Systems, held in Conjunction with IEEE/WIC/ACM (WI-IAT'10)*, 2010

[12] Schallauer, P., Bailer W., Hofmann, A., Mörzinger, R.: SAM - An Interoperable Metadata Model for Multimodal Surveillance Applications. *In SPIE Defense, Security, and Sensing*, 2009.