

# A Threefold Dataset for Activity and Workflow Recognition in Complex Industrial Environments

Athanasios Voulodimos  
*National Technical University of Athens, Greece*

Dimitrios Kosmopoulos  
*Rutgers University, The State University of New Jersey*

Georgios Vasileiou  
*TEI of Athens, Greece*

Emmanuel Sardis, Vasileios Anagnostopoulos, and  
Constantinos Lalos  
*National Technical University of Athens, Greece*

Anastasios Doulamis  
*Technical University of Crete, Greece*

Theodora Varvarigou  
*National Technical University of Athens, Greece*

Unlike any previous effort, the Workflow Recognition (WR) large-scale dataset is a collection of video sequences from the real industrial manufacturing environment of a major automobile manufacturer.

Behavior recognition in video is a focal point of research in the computer vision, image processing, and multimedia communities. Driven by applications such as assistive technologies, security, intelligent transportation, and human-computer interaction, a considerable body of work targets hierarchical event detection, workflow monitoring, and structured activity modeling in real-world scenarios.

More specifically, smart monitoring in industrial settings involves staff safety and

security, cost reduction, production scheduling, and production process quality. Such quality is guaranteed by enforcing strictly predefined procedures and activities for production or service provision. Production scheduling in particular—such as allocating available production resources (including raw materials, equipment, utilities, and manpower) to tasks over a scheduling horizon—is one of the most crucial managerial tasks within a large-scale industry because it can guarantee the sufficient execution of predefined operations (high-quality products construction and assembly), satisfy predefined task deadlines (production consistency), and improve the economic growth of the industry.<sup>1</sup> Smart video surveillance systems properly modified to satisfy industrial requirements and robust enough to work in harsh industrial environments can enrich modern factories with automatic decision-making tools and intelligent/smart capabilities increasing productivity and competitiveness.

Several challenging issues still remain, however, such as efficiently addressing occlusions and outliers, recognizing simultaneous actions, and tackling the execution of activities that can be carried out more than one way (high variability). The availability of appropriate datasets is essential for objective comparison and algorithm development. Although significant data collections comprising footage from public surveillance cameras and everyday activities exist, there is a dearth of significantly sized datasets depicting real-world activities, let alone with ground truth. (See the “Related Work in Behavior and Workflow Recognition” sidebar for previous research in this area.)

In this context, we introduce the Workflow Recognition (WR) dataset, which consists of multicamera video sequences taken from the production line of a major automobile manufacturer. The dataset depicts workers executing specific activities (workflow tasks such as assembling a car chassis) in a cluttered industrial environment. To our knowledge, no other dataset displays real workflows as they occur in a natural setting. We captured video during three different time periods using four cameras, obtaining more than 35 hours of data. We then separated the video footage into three parts: workflow 1, 2, and 3. Together with the dataset, we provide the ground truth in terms of activity labeling and a set of holistic features for scene representation.

To date, we have used the WR dataset as a benchmark for behavior- and workflow-recognition algorithms with promising results. The WR dataset was first presented in an earlier paper.<sup>2</sup> In this extended work, we present the WR dataset in its entirety, including the recent addition of workflow 3, which is 14 hours and 40 minutes long, takes place in a slightly different industrial setting, and adds more than 150 workflow instances to the collection. Furthermore, we thoroughly explain annotation and feature extraction and evaluate the collection based on new behavior- and workflow-recognition approaches.

### WR Dataset

The WR dataset consists of video sequences from the production line of a major automobile manufacturer. We obtained footage of three separate workflow scenarios, resulting in three dataset segments: workflows 1, 2, and 3. (The full dataset is publicly available for download at [www.scovis.eu](http://www.scovis.eu).) The environment is similar in all the datasets. The most prominent elements are workers (usually two or three in the foreground) dressed in blue or black uniforms, blue racks filled with metallic spare parts, a welding cell onto which workers transfer the spare parts, welding tools surrounding the cell, and a robot that picks up the assembled car chassis at the end of a workflow execution. Because this is a real environment, not a laboratory, other elements are present in the sequences as well (mainly in the background), such as small red and green lights, yellow and gray colored pipes, other employees wearing clothes of various colors, and forklifts that remove empty racks and replace them with full ones.

For data acquisition, we used AXIS 212-213 pan-tilt-zoom (PTZ) camera models. The frame-grabbing software we used in the PTZ setup is a free open source component that runs on a variety of operating systems. Most importantly, we deactivated the internal denoising algorithms in the camera setup. The denoising algorithms produced serious artifacts because the recording field of view was dark, and high gain values resulted in increased noise. Other side effects of the denoising were motion blur due to the nature of the algorithm and that the implementation made the camera overheat easily and slightly destabilized the frame rate. The average recording frame rate

was 25 frames per second (fps) with relative jitter bounded by 1.6 percent. Most of the jitter stemmed from the saving process. The resulting frame files were 50 to 60 Kbytes.

### Workflow 1

The first part of the dataset consists of four JPEG image sequences captured at 18 to 25 fps at  $704 \times 576$  resolution and 60 percent compression. Each image sequence corresponds to a different camera offering a different viewpoint of the scene. The goal was to capture the widest possible scene coverage, and the resulting overlapping views let us exploit redundancies in order to solve occlusions. The footage is approximately 5 hours and 10 minutes long.

Each of the four sequences includes repetitions of the same succession of executed tasks. These tasks involve workers picking up parts from racks, carrying them, and placing them on the welding cell, after welding them with the aid of the welding tools. This work cycle (or scenario) is repeated 20 times in workflow 1. However, the activity during each work cycle is not identical. For instance, sometimes the order of the executed tasks or the number of workers executing a task changes. Furthermore, there are some unpredicted, “abnormal” events, such as a bicycle passing right in front of the welding cell. There are also intervals of inactivity—for example, when employees are standing or chatting.

### Workflow 2

The second part of the dataset, workflow 2, is longer and richer than workflow 1 in many ways. This part captured two days of labor (as opposed to one day in workflow 1). During each day, 20 work cycles were executed for an additional 15 hours and 40 minutes of footage.

The content in workflow 2 differs significantly from workflow 1, making it more challenging for computer vision algorithms. Three employees perform work at the same time, simultaneously executing more than one task. This creates a much more complex foreground and makes it more difficult to recognize which task is being executed and to track a moving person. The order in which the tasks are executed is far less specific than in workflow 1, and the number of workers executing each task also varies. There are larger gaps of inactivity, both during a specific work cycle and

## Related Work in Behavior and Workflow Recognition

The computer vision community has attempted to systematically compare different algorithms by experimenting on a number of common datasets. To begin with, the Performance Evaluation of Tracking and Surveillance (PETS) workshop series provides interesting benchmarking datasets with a different focus each year. For instance, PETS 2002 covered indoor person tracking (<http://pets2002.visualsurveillance.org>), PETS 2006 focused on public space surveillance (<http://pets2006.net>), and PETS 2007 addressed attended luggage theft and detection (<http://pets2007.net>). The Caviar project recorded and released sequences including people walking, meeting others, entering and exiting shops, and so forth,<sup>1</sup> and the iLids dataset (<ftp://motinas.elec.qmul.ac.uk/pub/iLids>) focused on parked-vehicle detection, abandoned baggage detection, and doorway surveillance.

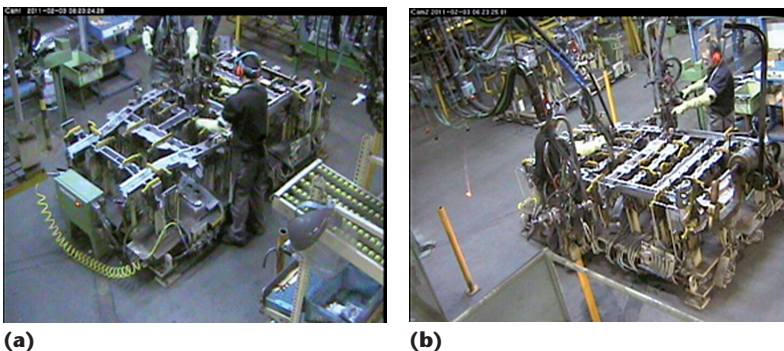
Several motion-capture-only datasets are also available, such as the Carnegie Mellon University (CMU) Motion Capture Database (<http://mocap.cs.cmu.edu>) and the MPI HDM05 Motion Capture Database ([www.mpi-inf.mpg.de/resources/HDM05](http://www.mpi-inf.mpg.de/resources/HDM05)), that provide large data collections. In these cases, the available motions are extremely articulated and well separated, and they bear little resemblance to natural, everyday activities. In addition, these datasets involve no manipulation or interaction tasks.

The CMU Kitchen Dataset contains multimodal observations of several cooking tasks, including calibrated cameras and motion-capture data (<http://kitchen.cs.cmu.edu>). This

dataset contains more realistic motions, but the numerous actions, sometimes unnatural movements, and uniforms worn detract from the dataset's naturalness. Another example is the Technical University Munich (TUM) Kitchen Dataset, which contains sequences of everyday manipulation activities in a natural kitchen environment with a focus on workflow execution, such as setting a dining table.<sup>2</sup> However, the monitored objects act far from naturally, actually simulating robot-like behavior that eliminates effects such as co-articulation of motion or unpredicted events.

Each of these datasets is more or less suitable for certain research goals or applications. Most of these datasets were actually recorded for security purposes and not for industrial workflow monitoring in complex environments. For instance, the Caviar dataset contains people walking in an almost open area, and the image content is captured from above. Thus, the footage of humans is distorted as they move closer and further away from the camera. The iLids dataset is more complicated because it focuses on more crowded conditions. However, it is unsuitable for behavior recognition because the majority of people recorded are simply walking in airport halls. The TUM kitchen dataset is more suitable for workflow recognition, but its footage is relatively unnatural and thus not particularly challenging.

On the contrary, our Workflow Recognition (WR) dataset depicts workers executing real industrial tasks in a real-world, visually complicated industrial environment.



**Figure 1.** Camera viewpoints in workflow 3. The two synchronized pan-tilt-zoom (PTZ) cameras, (a) cam 102 and (b) cam 103, captured 14 hours and 40 minutes of data.

between work cycles. Numerous irrelevant events and activities occur that are not linked to the work cycle (such as workers waiting, holding parts without carrying them, and walking around), which significantly hinders attempts at event and behavior recognition.

All these particularities led us to define an additional series of events and repeat the labeling process, which we describe in detail later on.

### Workflow 3

The third and most recent part of the dataset was shot in a different area of the industrial plant. We mounted the cameras so as to cover the full view of humans and the spare parts transfer paths to allow for detection and tracking of both the workers and spare parts, as well as event and workflow recognition. The environment is clearer with better lighting conditions. In this instance, we used two well-synchronized PTZ cameras. Figure 1 depicts the two cameras' viewpoints. The captured data collection is 14 hours and 40 minutes long.

The workflow itself follows the general idea of picking up and placing parts, welding, and collecting the assembled chassis, but this time

The multicamera view in the WR dataset (four cameras at different viewpoints) can help researchers address occlusions. Finally, despite the image complexity, the recorded processes are fairly structured, which is an important attribute for machine-learning algorithms.

There is a growing interest among the research community for structured-action recognition, workflow monitoring, and hierarchical event detection. For example, Nam Nguyen and his colleagues focused on recognizing a behavior and primitive hierarchy imposing temporal constraints,<sup>3</sup> whereas Nuria Oliver, Ashutosh Garg, and Eric Horvitz proposed layered hidden Markov models (HMMs) to capture different abstraction levels and corresponding time granularities for event recognition in meetings.<sup>4</sup> Another work modeled and recognized the workflow in a hospital operating room using a hierarchical HMM approach.<sup>5</sup> Yifan Shi, Aaron Bobick, and Irfan Essa proposed propagation networks (P-nets) to model and detect from video the primitive actions of a task performed by a tracked person.<sup>6</sup> Other works have addressed analysis and detection of hierarchical events using Rao Blackwell particle filters along with a dynamic Bayesian network.<sup>7</sup>

This increased interest in visual workflow analysis and recognition suffers from a lack of publicly available corpora that could serve as a basis for related algorithms comparison and testing. In this sense, the WR dataset is a unique, valuable data collection that stems from a real-world industrial environment, has a significant size, and is accompanied by the ground-truth annotations of the observed activities.

the workers place 12 small and two large spare parts to assemble a larger chassis and then weld them together. One worker then handles a yellow crane, which lifts and transfers the chassis away from the scene, leaving the cell empty and ready to host a new work cycle (assembly process).

The mean length of a work cycle is 6.5 minutes, which is significantly shorter than in the two previous cases, thus yielding more than 150 work cycles. Given this abundance of available data for training and testing, workflow 3 is a promising data collection for algorithm testing.

### Challenges

The WR dataset contains numerous challenging sequences that include serious visibility problems, severe occlusions, self occlusions, and outliers. In particular, vibrations, sparks, and cluttered backgrounds with upright racks, welding machines, and forklifts create

a cluttered environment that often occludes the workers. Additional exacerbating factors include frequent illumination changes and the fact that the workers' blue clothes closely resemble the color of the racks. Figure 2 shows some examples of the challenges faced in the dataset.

In certain cases, the high-intraclass and low-interclass variance among tasks makes task discerning difficult even for the human eye. Significant deviations in the workflow process can occur (especially in workflow 2). Several tasks within a workflow can have fluctuating durations and no clearly defined beginning or ending. Furthermore, the tasks might entail both human actions and machine movements.

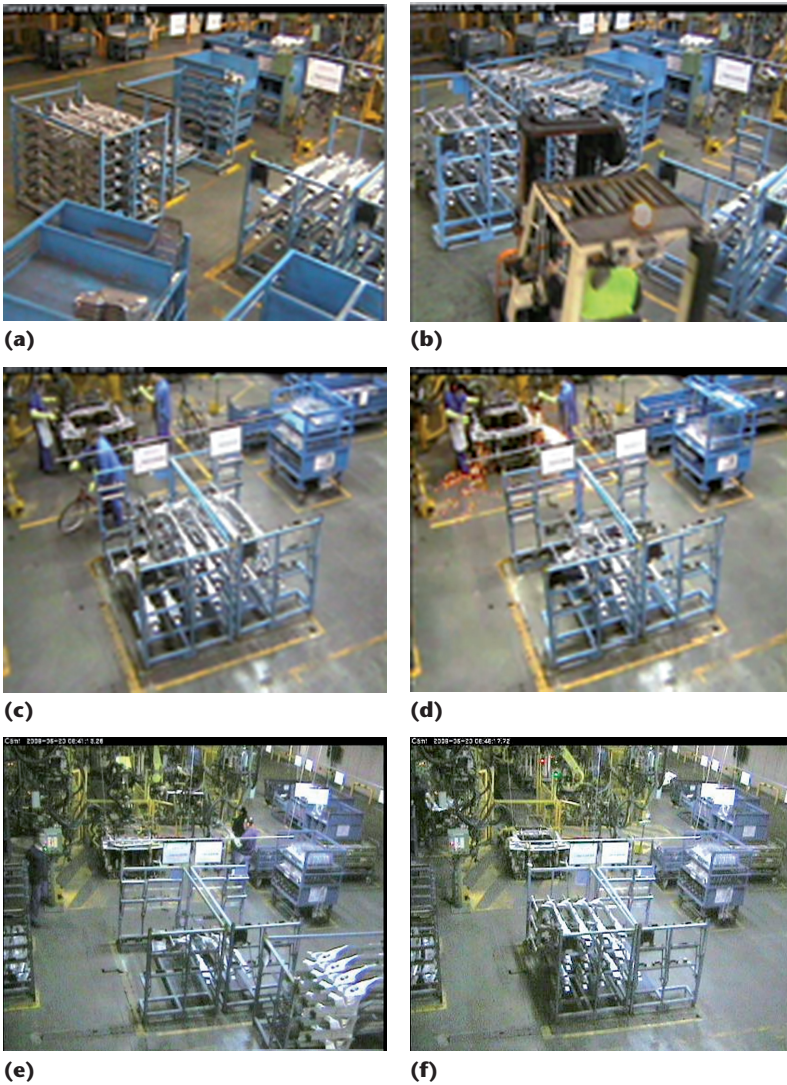
### Data Labeling: Creating Ground Truth

After acquiring the dataset and completing the activity-definition process, which we

## References

1. R. Fisher, "The PETS04 Surveillance Ground-Truth Data Sets," *Proc. 6th IEEE Int'l Workshop Performance Evaluation of Tracking and Surveillance (PETS)*, 2004, pp. 1–5; <http://homepages.inf.ed.ac.uk/rbf/PAPERS/pets04.pdf>.
2. M. Tenorth, J. Bandouch, and M. Beetz, "The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition," *Proc. 12th IEEE Int'l Conf. Computer Vision (ICCV) Workshops*, 2009, IEEE CS Press, pp. 1089–1096.
3. N. Nguyen et al., "Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Model," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 2, IEEE CS Press, 2005, pp. 955–960.
4. N. Oliver, A. Garg, and E. Horvitz, "Layered Representations for Learning and Inferring Office Activity from Multiple Sensory Channels," *Computer Vision and Image Understanding*, vol. 96, no. 2, 2004, pp. 163–180.
5. N. Padoy et al., "Workflow Monitoring Based on 3D Motion Features," *Proc. 12th IEEE Int'l Conf. Computer Vision Workshops (ICCV)*, IEEE CS Press, 2009, pp. 585–592.
6. Y. Shi, A. Bobick, and I. Essa, "Learning Temporal Sequence Model from Partially Labeled Data," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 2, IEEE CS Press, 2006, pp. 1631–1638.
7. X. Xiaoling and L. Layuan, "Real Time Analysis of Situation Events for Intelligent Surveillance," *Proc. Int'l Symp. Computational Intelligence and Design (ISCID)*, vol. 2, IEEE CS Press, 2008, pp. 122–125.





**Figure 2.** Example challenges in the Workflow Recognition (WR) dataset. The footage of real-world workflow scenarios includes (a, b) occlusions, (c) abnormal events, (d) sparks and visibility problems, (e) irrelevant activity, and (f) periods of inactivity.

performed with the assistance of industrial engineers, the next step was to create the ground truth for activity and workflow recognition. Two doctoral students annotated the ground truth for the observed activities with clear guidelines to ensure maximum consistency and agreement. When the students disagreed, a professor solved the ambiguity to guarantee consistent annotations.

#### Workflow 1

To enable activity and workflow recognition, we split each workflow into seven discrete tasks. We consulted expert industrial engineers to ensure that the defined tasks were

meaningful to the production process. With the help of the experts, we defined the following tasks:

1. One worker selects part 1 from rack 1 and places it on the welding cell.
2. Two workers select part 2a from rack 2 and place it on the welding cell.
3. Two workers get part 2b from rack 3 and place it on the welding cell.
4. A worker picks up parts 3a and 3b from rack 4 and places them on the welding cell.
5. A worker picks up part 4 from rack 1 and places it on the welding cell.
6. Two workers pick up part 5 from rack 5 and place it on the welding cell.
7. Two workers grab the welding tools and weld the parts together.

A single worker might perform a task that is normally executed by two workers. Each task is executed once in each work cycle, but the order is not strict and permutations are allowed. The duration of “pick up and place part” tasks ranges from 200 to 400 frames (approximately 8 to 16 seconds), whereas welding takes longer and lasts for 1,600 to 2,000 frames. To assist the community in exploiting the dataset for algorithm testing, we provide the task-labeling files (which we describe shortly) together with the dataset.

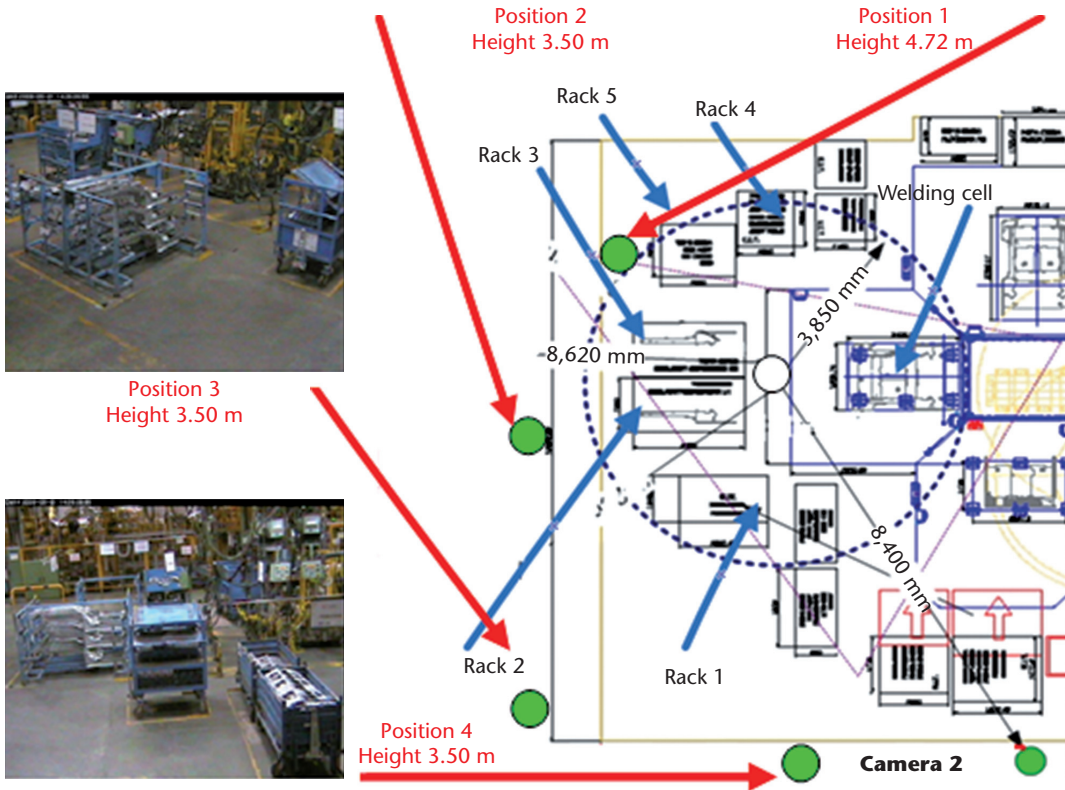
Figure 3 shows the workspace configuration and camera positions. To ensure we fully captured the observed workflows, we selected the camera settings and positions with the aid of industrial engineers who are experts in the operational details of the particular assembly line. This also helped us keep the number of cameras we used to a minimum while conforming to industrial safety regulations.

#### Workflow 2

Because the workflow 2 footage is more complex, making activity recognition more difficult, the data requires a supplementary alternative task definition. Therefore, in addition to labeling workflow 2 sequences as we did for workflow 1, we further split the seven



Figure 3. Work-cell depiction. This workspace configuration shows the position of the four cameras and racks 1 through 5 for workflows 1 and 2.



tasks to identify smaller, shorter events in an effort to enable event-driven recognition. We defined these “microtasks” to ensure the fulfillment of several assumptions:

- Workflow recognition is possible through microtask recognition.
- Micro-tasks should be as spatially confined as possible so that we can efficiently define and observe regions of interest (ROIs).
- Micro-tasks are as temporally short as possible so that overlapping instances are less frequent.

Micro-tasks are actually segments of the tasks that signify characteristic actions that might in a later step lead to task recognition.

We defined the following micro-tasks for workflow 2:

1. Worker selects part 1 from rack 1.
2. Worker places part 1 on the welding cell.
3. Two workers get part 2a from rack 2.
4. Two workers place part 2a on the welding cell.
5. Two workers select part 2b from rack 3.
6. Two workers place part 2b on the welding cell.
7. Worker picks up parts 3a and 3b from rack 4.
8. Worker places parts 3a and 3b on the welding cell.
9. Worker picks up part 4 from rack 1.

10. Worker places part 4 on the welding cell.
11. Worker(s) get part 5 from rack 5.
12. Worker(s) place part 5 on the welding cell.
13. Worker grabs the first welding tool, approaches the cell, and starts welding.
14. Robot collects the assembled chassis.

In this 14-microtask definition, each pair of sequential micro-tasks (up until micro-task 12) defines the beginning and end of a task. That is, micro-task 1 begins a task, and micro-task 2 completes it. Micro-task 13 denotes the start of the welding process, and microtask 14 signifies the end of the work cycle. The duration of each micro-task ranges from 50 to 150 frames (2 to 6 seconds) for micro-tasks 1 through 12, whereas micro-tasks 13 and 14 take a little longer. The definition and labeling of micro-tasks, which are spatially confined in ROIs, allows for scrutinizing alternative event-detection methods—for example, based on ROI motion history observation.

### Workflow 3

Because workflow 3 consists of a greater number of shorter tasks in each work cycle, we modified the rationale behind the annotation in this workflow. The workflow involves placing 14 parts on the cell, with many placements occurring simultaneously and under heavy occlusions, and of several welding actions. Thus, the human silhouettes often overlap while executing tasks simultaneously. Because of the major overlap among actions and the camera's zooming on the welding cell itself, we decided to annotate the moments when each part has been correctly placed on its final position by the worker(s), as well as the beginning and end of each series of welding actions performed using the same welding tool. All these positions correspond to particular ROIs on the image.

There are six welding tools, so the actions to be labeled include 14 part placements, six welding activities, and finally the lifting of the assembled chassis, which concludes the workflow. This part of the dataset consists of more than 150 instances. The mean duration of a work cycle is approximately 6.5 minutes (9,000 to 10,000 frames).

### File Labeling and Description

The task-labeling files currently available include the entire workflow 1 (cameras 32 and 34), the first day of workflow 2 (cameras 1 and 3), and part of workflow 3. The annotation effort, which is tedious, is still ongoing. We have provided a .txt file for each combination of camera and workflow. These text files contain information about the beginning and ending time (either in terms of frame number or timestamp) of every task for all 20 work cycles in the day. The same information is given for the 14-microtask definition for workflow 2. Researchers can use these labeling files to process the footage to train and test recognition algorithms. When a timestamp is provided, the frame's timestamp is in the JPEG File Interchange Format (JFIF) header.

An example record of the .txt file for camera 1 in workflow-2 would be

```
scenario_filename
1 06.41.25.36 06.41.36.36
2 06.41.45.44 06.41.56.76
...
```

This record means that task 1 of this particular scenario begins at 06.41.25.36 (timestamp on the frame) and ends at 06.41.36.36, and so forth.

### Scene Representation

Using features directly extracted from the video frames has the significant advantage of obviating the need to detect and track salient scene objects, which is exceptionally difficult in the dataset's complex industrial environment. Initially, we tested the efficiency of tracking<sup>3</sup> and pure detection histogram of oriented gradients (HOG), but both methods failed. We employed both local and holistic features; some typical examples include cascaded confidence filtering,<sup>4</sup> object flow,<sup>5</sup> and local motion classifiers.<sup>6</sup> Here, we present a set of holistic features, which provide a more robust representation, leading to more successful and long-term activity recognition.

First, we performed background subtraction. We used the foreground regions to represent the multiscale spatiotemporal changes at the pixel level. For this purpose, we used a concept similar to motion history images,<sup>7</sup> but one that has better representation capabilities.



The pixel change history (PCH) of a pixel is defined as

$$P_{\zeta,\tau}(x,y,t) = \begin{cases} \min(P_{\zeta,\tau}(x,y,t-1) + \frac{255}{\zeta}, 255) \\ \text{if } D(x,y,t) = 1 \\ \max(P_{\zeta,\tau}(x,y,t-1) - \frac{255}{\tau}, 0) \\ \text{otherwise} \end{cases} \quad (1)$$

where  $P_{\zeta,\tau}(x,y,t)$  is the PCH for a pixel at  $(x,y)$ ,  $D(x,y,t)$  is a binary value corresponding to a pixel indicating the foreground region at time  $t$ ,  $\zeta$  is an accumulation factor, and  $\tau$  is a decay factor. By setting appropriate values for  $\zeta$  and  $\tau$ , we can capture pixel-level changes over time.

To represent the resulting PCH images, we used Zernike moments because of their noise resiliency, reduced information redundancy, and reconstruction capability. The complex Zernike moments of order  $p$  are defined as

$$A_{pq} = \frac{p+1}{\pi} \int_0^1 \int_{-\pi}^{\pi} R_{pq}(r) e^{-iq\theta} f(r,\theta) r dr d\theta \quad (2)$$

where  $r = \sqrt{x^2 + y^2}$ ,  $\theta = \tan^{-1}(y/x)$ ,  $-1 < x, y < 1$ , and

$$R_{pq}(r) = \sum_{s=0}^{\frac{p-q}{2}} (-1)^s \frac{(p-s)!}{s! \left(\frac{p+q}{2} - s\right)! \left(\frac{p-q}{2} - s\right)!} r^{p-2s} \quad (3)$$

where  $p - q$  is even and  $0 \leq q \leq p$ . The higher the order of moments used, the more detailed the region reconstruction will be, but also the more processing power will be required.

To capture the spatiotemporal variations, we set the parameters at the empirically defined values  $\zeta = 10$  and  $\tau = 70$ . As feature vectors, we used the Zernike moments up to the sixth order. Specifically, these are the complex moments  $A_{00}, A_{11}, A_{20}, A_{22}, A_{31}, A_{33}, A_{40}, A_{42}, A_{44}, A_{51}, A_{53}, A_{55}, A_{60}, A_{62}, A_{64}$ , and  $A_{66}$  for which we used the norm and the angle, except for  $A_{00}, A_{20}, A_{40}$ , and  $A_{60}$ , for which the angle was always constant. Additionally, we used the center of gravity and area for a total of 31 parameters. Thus, we obtained good scene reconstruction without having a vector dimension that is too high.

Of course, this is by no means the only feasible approach for feature definition and

extraction. Other holistic features such as Chebyshev moments or local descriptors (SIFT, SURF, and so on) could have been used as well.

The features we extracted are publicly available along with the dataset. The feature vectors currently available concern the seven-task definition. Regarding workflow 1, the features cover all 20 work cycles (scenarios) for cameras 32 and 34. For workflow 2, the features cover the 20 work cycles of the first day for cameras 1 and 3. The features are in a .mat Matlab file. There are four cell arrays, one for each combination of camera and dataset. The cell arrays have the following format: `camXXf1, taskgf1, scenariog <31_num of frames >`, where XX is the camera ID and takes the values {32, 34, 1, 3}. The first two values correspond to dataset 1, the other two correspond to dataset 2. The task number (as we defined earlier) takes values 1 through 7, and the scenario number takes values 1 through 20. The number of frames describes the specific task of the specific scenario, and 31 is the dimension of the feature vector. In addition, we provide features based on the 14-micro-task definition for the first day of workflow 2 for camera 1 extracted through the same process. The features for workflow 3 will be provided as continuous cell arrays containing one 31D feature vector for each frame. Similar to annotation, feature extraction for the remaining cameras and parts of the dataset is an ongoing effort.

In the near future we intend to additionally provide features based on local descriptors, although because of the many occlusions and the relatively low resolution, the results probably will not be as good as with the region descriptors.

## Evaluation

The primary goal of the WR dataset is to serve as a testbed for activity and workflow recognition, although it can also be used for object detection and tracking. An initial criterion for grouping behavior recognition approaches pertains to whether the data used as input are segmented or continuous. Segmented sequences, where each segment corresponds to a task, serve as input to the classifiers so they can be recognized as one of the available tasks. Earlier work proposed a behavior-recognition method using multiple cameras based on the proposed PCH-Zernike moments features and fused hidden Markov models (HMMs).<sup>3</sup> Experimenting



Table 1. Comparison of activity classification approaches on presegmented sequences and continuous stream data.

Method	Workflow	Single camera		Multicamera	
		Precision (%)	Recall (%)	Precision (%)	Recall (%)
<i>Presegmented sequences</i>					
Hidden Markov model (HMM) <sup>3</sup>	1	84.1	88.6	89.8	92.1
HMM and evaluative rectification (HMM+ER) <sup>9</sup>	1	90.7	93.5	91.8	94.2
Rectification-driven fused HMM (RDFHMM) <sup>9</sup>	1	–	–	93.2	95.0
HMM <sup>3</sup>	2	53.4	56.3	57.8	61.2
HMM+ER <sup>9</sup>	2	62.3	73.2	72.3	78.9
RDFHMM <sup>9</sup>	2	–	–	77.3	79.8
<i>Continuous stream data</i>					
Genetic algorithms and HMM (GAHMM) <sup>6</sup>	1	89.8	90.0	90.3	90.4
Echo state network (ESN) <sup>6</sup>	1	73.5	72.8	76.2	73.6
Bayesian filters and recursive neural-network-based approach (BF+rNN) <sup>10</sup>	1	87.9	86.8	–	–
Region of interest and string matching (ROI+SM) <sup>11</sup>	2	81.3	78.6	–	–

on workflow 1, we obtained a maximum precision of 89.8 percent and a recall of 92.1 percent when using the multistream fused HMMs<sup>8</sup> and employing the multivariate student's *t*-distribution as an observation model of the HMMs, whose heavier tails (compared to a Gaussian distribution model) permit efficient handling of outliers.<sup>3</sup> The evaluative rectification (ER) approach<sup>9</sup> aims to dynamically correct erroneous task classifications by exploiting an expert user's feedback through a neural-network-based framework that readjusts the HMM likelihoods, significantly decreasing the overall misclassification rates.

As the upper part of Table 1 shows, the ER approach further improves the already high-performance rates in both single-camera and multicamera configurations. The biggest enhancement is attained by the rectification-driven fused HMM (RDFHMM) approach,<sup>9</sup> which fuses the rectified single streams (instead of rectifying the fused results), leading to a recall of 95.0 percent. When applied to the more challenging workflow 2, the same methods yield recall rates ranging from 56.3 percent (single HMM only) to 79.8 percent (RDFHMM), highlighting the improvement induced by the neural-network-based rectification method.

The most realistic, challenging, and high-impact problem, however, is online behavior recognition of continuous data streams. A mixed Gaussian and HMM based approach<sup>6</sup> addresses online segmentation of the sequences.

We classified the outcomes using an HMM framework that incorporates a priori knowledge via genetic algorithms (GAHMM). Yielding a 90.3 percent precision in the multicamera scheme, this framework outperforms an echo state network (ESN) based approach, which achieves only 76.2 percent precision (see the lower part of Table 1).

A different approach for behavior recognition assumes particles and uses Bayesian filters to complement HMMs, together with a recursive implementation of the aforementioned neural-network-based readjustment (BF+rNN).<sup>10</sup> Although we cannot directly compare these methods because they use different sets of features, this approach attains similar, if slightly inferior precision and recall.

Finally, in a previous work,<sup>11</sup> we addressed the problem of concurrent activity recognition that arises in workflow 2. There we used an approach based on ROIs and a modified string matching (SM) technique that regards workflows as strings and micro-tasks as characters, attaining promising results. Nonetheless, research in this area is currently ongoing.

### Future Enhancements

To the best of our knowledge, with its multicamera video sequences from a real-world production line, the WR dataset is a novel contribution. The heavy occlusions, outliers, human-machine interaction, and visually complicated environment make this dataset a challenging testbed for computer vision and

multimedia algorithms. We hope it is a useful addition for benchmarking of workflow recognition, which is rapidly gaining momentum. Initial evaluation of the dataset has inspired new behavior- and workflow-recognition approaches. The dataset can also be used for unsupervised-learning experiments. We expect this research effort to trigger the interest of the scientific community, and we welcome contributions in feature definition and extraction as well as annotation efforts, which are still ongoing.

**MM**

## Acknowledgments

This work has been funded by the European Union's Seventh Framework Program ([FP7/2007-2013]) under grant agreement 216465. We thank everyone involved in the SCOVIS project for their contribution.

## References

1. J. Palma et al., "Scheduling of Maintenance Work: A Constraint-Based Approach," *Expert Systems with Applications*, vol. 37, 2010, pp. 2963–2973.
2. A. Voulodimos et al., "A Dataset for Workflow Recognition in Industrial Scenes," *Proc. 18th IEEE Int'l Conf. Image Processing (ICIP)*, IEEE Press, 2011, pp. 3310–3313.
3. D. Kosmopoulos and S. Chatzis, "Robust Visual Behavior Recognition," *IEEE Signal Processing Magazine*, vol. 27, no. 5, 2010, pp. 34–45.
4. S. Stalder, H. Grabner, and L. Van Gool, "Cascaded Confidence Filtering for Improved Tracking-By-Detection," *Proc. European Conf. Computer Vision (ECCV)*, Springer-Verlag, 2010, pp. 369–382.
5. C. Lalos et al., "Efficient Tracking Using a Robust Motion Estimation Technique," to be published in *Multimedia Tools and Applications*, doi:10.1007/s11042-012-0994-3.
6. A. Voulodimos et al., "Online Classification of Visual Tasks for Industrial Workflow Monitoring," *Neural Networks*, vol. 24, no. 8, 2011, pp. 852–860.
7. T. Xiang and S. Gong, "Beyond Tracking: Modeling Activity and Understanding Behaviour," *Int'l J. Computer Vision*, vol. 67, no. 1, 2006, pp. 21–51.
8. Z. Zeng et al., "Audiovisual Affective Expression Recognition through Multistream Fused HMM," *IEEE Trans. Multimedia*, vol. 10, no. 4, 2008, pp. 570–577.
9. A.S. Voulodimos et al., "Improving Multi-camera Activity Recognition by Employing Neural Network Based Readjustment," *Applied Artificial Intelligence*, vol. 26, nos. 1–2, 2012, pp. 97–118.
10. D.I. Kosmopoulos, N.D. Doulamis, and A.S. Voulodimos, "Bayesian Filter Based Behavior Recognition in Workflows Allowing for User Feedback," *Computer Vision and Image Understanding*, vol. 116, no. 3, 2012, pp. 422–434.
11. A. Voulodimos et al., "A Top-Down Event-Driven Approach for Concurrent Activity Recognition," to be published in *Multimedia Tools and Applications*, 2012, doi:10.1007/s11042-012-0993-4.

**Athanasios Voulodimos** is a researcher in the School of Electrical and Computer Engineering at the National Technical University of Athens. His research interests include computer vision, machine learning, and ubiquitous and cloud computing. Voulodimos has a PhD in computer vision and machine learning from the National Technical University of Athens. He is a member of IEEE. Contact him at thanosv@mail.ntua.gr.

**Dimitrios Kosmopoulos** is a research professor in the Department of Computer Science at Rutgers University. His research interests include computer vision, machine learning, and robotics. Kosmopoulos has a PhD in computer vision and robotics from the National Technical University of Athens. He is a member of IEEE. Contact him at dkosmo@ieee.org.

**Georgios Vasileiou** is a senior at the National Technical University of Athens completing a degree in informatics. His research interests include various aspects of artificial intelligence. Contact him at grgs.vsl@gmail.com.

**Emmanuel Sardis** is a visiting assistant professor at the Technical University of Crete and a senior researcher coordinating numerous national and European projects in the areas of Internet technologies, embedded systems, and distributed architectures in the Distributed, Knowledge and Media Systems Laboratory at the National Technical University of Athens. His research interests include agents, embedded systems, and distributed architectures. Sardis has a PhD in electrical and computer engineering from the National Technical University of Athens. He is a member of IEEE. Contact him at sardism@mail.ntua.gr.

**Vasileios Anagnostopoulos** is a researcher in the School of Electrical and Computer Engineering at the National Technical University of Athens.


His research interests include discrete event simulation methodologies, Banach space geometry, and Symmetric group combinatorics. Anagnostopoulos has a PhD in simulation methodology and online routing from the National Technical University of Athens. Contact him at vanag@mail.ntua.gr.

**Constantinos Lalos** is a doctoral student and a research associate in the Distributed Knowledge and Media Systems Laboratory at the National Technical University of Athens. His research interests include computer vision and content-based retrieval. Lalos has an electronics engineering diploma from the University of Bristol, UK. Contact him at lalos@mail.ntua.gr.

**Anastasios Doulamis** is an assistant professor at the Technical University of Crete, Greece. His research interests include nonlinear analysis, neural networks, multimedia content description, and intelligent

techniques for video processing. Doulamis has a PhD in electrical and computer engineering from the National Technical University of Athens. He is a member of IEEE. Contact him at adoulam@dpem.tuc.gr.

**Theodora Varvarigou** is a professor at the National Technical University of Athens. Her research interests include parallel algorithms and architectures, fault-tolerant computation, optimization algorithms, and content management. Varvarigou has a PhD in computer science from Stanford University. Contact her at dora@telecom.ntua.gr.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

## IEEE computer society

**PURPOSE:** The IEEE Computer Society is the world's largest association of computing professionals and is the leading provider of technical information in the field.

**MEMBERSHIP:** Members receive the monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.

**COMPUTER SOCIETY WEBSITE:** [www.computer.org](http://www.computer.org)

**Next Board Meeting:** 5–6 Nov., New Brunswick, NJ, USA

### EXECUTIVE COMMITTEE

**President:** John W. Walz\*

**President-Elect:** David Alan Grier;\* **Past President:** Sorel Reisman;\* **VP, Standards Activities:** Charlene (Chuck) Walrad;† **Secretary:** Andre Ivanov (2nd VP);\* **VP, Educational Activities:** Elizabeth L. Burd;\* **VP, Member & Geographic Activities:** Sattupathuv Sankaran;† **VP, Publications:** Tom M. Conte (1st VP);\* **VP, Professional Activities:** Paul K. Joannou;\* **VP, Technical & Conference Activities:** Paul R. Croll;† **Treasurer:** James W. Moore, CSDP;\* **2011–2012 IEEE Division VIII Director:** Susan K. (Kathy) Land, CSDP;† **2012–2013 IEEE Division V Director:** James W. Moore, CSDP;† **2012 IEEE Division Director VIII Director-Elect:** Roger U. Fujii†

\*voting member of the Board of Governors †nonvoting member of the Board of Governors

### BOARD OF GOVERNORS

**Term Expiring 2012:** Elizabeth L. Burd, Thomas M. Conte, Frank E. Ferrante, Jean-Luc Gaudiot, Paul K. Joannou, Luis Kun, James W. Moore, William (Bill) Pitts

**Term Expiring 2013:** Pierre Bourque, Dennis J. Frailey, Atsuhiko Goto, André Ivanov, Dejan S. Milojicic, Paolo Montuschi, Jane Chu Prey, Charlene (Chuck) Walrad

### EXECUTIVE STAFF

**Executive Director:** Angela R. Burgess; **Associate Executive Director, Director, Governance:** Anne Marie Kelly; **Director, Finance & Accounting:** John Miller; **Director, Information Technology & Services:** Ray Kahn; **Director, Membership Development:** Violet S. Doan; **Director, Products & Services:** Evan Butterfield; **Director, Sales & Marketing:** Chris Jensen

### COMPUTER SOCIETY OFFICES

**Washington, D.C.:** 2001 L St., Ste. 700, Washington, D.C. 20036-4928

**Phone:** +1 202 371 0101 • **Fax:** +1 202 728 9614

**Email:** [hq.ofc@computer.org](mailto:hq.ofc@computer.org)

**Los Alamitos:** 10662 Los Vaqueros Circle, Los Alamitos, CA 90720-1314

**Phone:** +1 714 821 8380 • **Email:** [help@computer.org](mailto:help@computer.org)

### MEMBERSHIP & PUBLICATION ORDERS

**Phone:** +1 800 272 6657 • **Fax:** +1 714 821 4641 • **Email:** [help@computer.org](mailto:help@computer.org)

**Asia/Pacific:** Watanabe Building, 1-4-2 Minami-Aoyama, Minato-ku, Tokyo 107-0062, Japan

**Phone:** +81 3 3408 3118 • **Fax:** +81 3 3408 3553

**Email:** [tokyo.ofc@computer.org](mailto:tokyo.ofc@computer.org)

### IEEE OFFICERS

**President:** Gordon W. Day; **President-Elect:** Peter W. Staecker; **Past President:** Moshe Kam; **Secretary:** Celia L. Desmond; **Treasurer:** Harold L. Flescher; **President, Standards Association Board of Governors:** Steven M. Mills; **VP, Educational Activities:** Michael R. Lightner; **VP, Membership & Geographic Activities:** Howard E. Michel; **VP, Publication Services & Products:** David A. Hodges; **VP, Technical Activities:** Frederick C. Mintzer; **IEEE Division V Director:** James W. Moore, CSDP; **IEEE Division VIII Director:** Susan K. (Kathy) Land, CSDP; **IEEE Division VIII Director-Elect:** Roger U. Fujii; **President, IEEE-USA:** James M. Howard

revised 22 May 2012

