A DATASET FOR WORKFLOW RECOGNITION IN INDUSTRIAL SCENES

Athanasios Voulodimos[1], *Dimitrios Kosmopoulos*[1], *Georgios Vassileiou*[2], *Emmanuel Sardis*[1],
Anastasios Doulamis[3], *Vassileios Anagnostopoulos*[1], *Constantinos Lalos*[1], *Theodora Varvarigou*[1]

[1]National Technical University of Athens [2]TEI of Athens [3]Technical University of Crete

ABSTRACT

In this paper we introduce the WR (Workflow Recognition) dataset. Recorded in the production line of a major automobile manufacturer, this dataset consists of sequences that depict workers executing industrial workflows. The heavy occlusions, outliers, the visually complicated background and the human-machinery interaction are among the factors that make this dataset a very challenging testbed for computer vision and image processing algorithms. We provide the original video sequences together with event labeling, as well as feature vectors extracted through our proposed scene representation methodology, and we refer to our results so far in workflow recognition using this dataset.

Index Terms— Video surveillance, industrial environments, tracking, activity recognition, workflow recognition.

1. INTRODUCTION

Computer vision is a research area that has been vastly researched upon, since it entails a significant number of important and challenging open issues, such as object detection and tracking, multi-camera calibration and activity recognition. The availability of appropriate datasets is essential in computer vision, so that the community can objectively compare the performance of algorithms. Given the vast range of different applications and the large number of competing algorithms, having the widest possible assortment of testbed datasets is desirable. In this context, we introduce a new dataset comprising very challenging sequences from the production line of a major automobile manufacturer. These sequences can serve as a testbed for studying the performance of algorithms from all the aforementioned research fields in a complex industrial environment. Two different footages at two different time periods were shot, therefore the two resulting sets of video sequences will be separately described.

The remainder of this paper is organized as follows: In Section 2 we describe the video sequences and their challenges. Section 3 explains the event labeling and feature extraction processes. Section 4 mentions approaches on workflow recognition that have already generated promising results on the WR dataset, while Section 5 concludes the paper.

1.1. Related work

In recent years the computer vision community has attempted to systematically compare different algorithms by experimenting on a number of common datasets. To begin with, the PETS workshop series provide interesting datasets for benchmarking, with a different focus each year. For instance, PETS 2002[1] was about indoor person tracking, PETS 2006[2] focused on surveillance of public spaces, while PETS 2007[3] particularized the topics to attended luggage theft and detection (among others). What's more, a number of sequences including people walking, meeting others, entering and exiting shops, etc. were recorded and released as part of the CAVIAR project [1], while the i-Lids[4] dataset focuses on parked vehicle detection, adandoned baggage detection and doorway surveillance.

Moreover, there are several motion-capture-only datasets available, such as the CMU Motion Capture Database[5] or the MPI HDM05 Motion Capture Database[6] providing large collections of data. In these cases the available motions are extremely articulated, well separated, and they bear little resemblance to natural everyday activities. In addition, there is no manipulation or interaction tasks involved in these datasets. On the contrary, the CMU Kitchen Dataset[7] contains multimodal observations of several cooking tasks, including calibrated cameras and motion capture data. This dataset contains more natural motions, but the large number of actions and the high variation between the actors pose serious challenges for action recognition. Another example is the TUM Kitchen Dataset, whose sequences consist of everyday manipulation activities in a natural kitchen environment with a focus on realistic motions [2]. Other testbed examples include the USF dataset for gait recognition algorithms [3].

The research leading to these results has received funding from the European Community's FP7/2007-2013 under grant agreement n. 216465. The authors would like to thank all partners involved in the SCOVIS project for their help and cooperation.

[1] http://pets2002.visualsurveillance.org/

[2] http://pets 2006.net/

[3] http://pets2007.net/

[4] ftp://motinas.elec.qmul.ac.uk/pub/iLids/

[5] http://mocap.cs.cmu.edu

[6] http://www.mpi-inf.mpg.de/resources/HDM05/

[7] http://kitchen.cs.cmu.edu

Each of the aforementioned datasets is more or less suitable for some research goals or applications and presents a series of advantages but also drawbacks compared to the WR dataset. The main advantage of the WR dataset over these datasets is that the latter have been actually recorded for security purposes and not for industrial workflow monitoring in complex situations. As can be seen by comparing the content of all these datasets with the initial recording content in the automobile manufacturer infrastructure, direct application of already developed computer vision algorithms to the WR dataset is not straightforward. For instance, CAVIAR dataset mostly contains people walking in an almost open area, while the image content is captured from above. Thus, humans are presented distortedly as they get close or far away from the camera. iLids dataset is more complicated, as it focuses on more crowded conditions (lots of people walking in front of cameras). However, even in this case, this dataset is not suitable for behavior detection since the majority of persons recorded are just walking in the airport halls. Another significant difference lies in the particular nature of the occlusions in our industrial environment. These occlusions do not stem from rigid bodies or humans. They mostly come from slim periodic structures, non-rigid bodies and grids. To this end, the multi-camera view in the WR dataset (4-5 cameras at different viewpoints) can be a useful tool for occlusion solving. Finally, what makes the WR dataset extremely interesting as a testbed is that despite the image complexity the recorded processes are fairly structured, which is important and meaningful for machine learning algorithms.

2. THE WR DATASET

The Workflow Recognition (WR) dataset[8] consists of video sequences from the production line of the automobile manufacturer. Two different shots took place: *dataset-1* and *dataset-2*. In both datsets the environment is the same. Its most prominent constituent elements are: a number of workers (usually two to three in the foreground) dressed in blue uniforms; blue racks filled with metallic spare parts; a welding cell, onto which these spare parts are transferred by workers; surrounding the cell, six welding tools; and a robot that picks up the assembled car chassis in the end of a workflow execution. Of course other elements are present as well in the sequences (mainly in the background), such as small red and green lights, pipes, other workers, forklifts, etc.

2.1. Dataset-1

Dataset-1 consists of four jpeg image sequences captured at 18-25fps at a resolution of 704×576, and compression 60%. Each image sequence corresponds to a different camera offering a different viewpoint of the scene. The goal was to have the widest possible scene coverage and the resulting overlapping views provide the possibility to exploit redundancies in order to solve occlusions. The overall duration of the footage is approximately 5 hours and 10 minutes. Each of the four sequences displays repetitions of the same succession of tasks executed. These tasks involve picking up parts from racks, carrying them, and placing them on the welding cell, after welding them with the aid of the welding tools (for a detailed description of the task definition see subsection 3.1). The aforementioned workcycle (or scenario) is repeated 20 times in dataset-1. However, this does not imply that the activity during each workcycle is identical. For instance, sometimes the order of the executed tasks or the number of workers executing a task changes. Furthermore, there are some unpredicted events, which are considered as ”abnormal”, e.g. a bicycle passing right in front of the welding cell, whereas there are also intervals of inactivity but are captured from different cameras and contain abnormal behavior, accidents and intervals of inactivity.

2.2. Dataset-2

Dataset-2 is longer and richer than dataset-1 in many ways. To begin with, a ”fish-eye” camera was deployed in addition to the four side cameras, providing an overall panoramic view of the scene. Furthermore, two days of labour were captured (as opposed to one day in dataset-1). During each day 20 workcycles were executed, thus augmenting the duration of dataset-2 to 15 hours and 30 minutes.

The content of dataset-2 has some significant differences in comparison to that of dataset-1, thus making the former more challenging for computer vision algorithms testing. The number of the workers working at the same time is increased to three, thus allowing for the simultaneous execution of more than one task. This creates a much more complex foreground and makes it more difficult to recognize which task is being executed, or to track a moving person. Figure 1 shows an example of a sequence where two tasks are performed in parallel, thus increasing the complexity. The order in which the tasks are executed is far less specific than in dataset-1 and the number of workers executing each task varies too. There are larger gaps of inactivity. Numerous irrelevant events and activities occur that are not linked to the workcycle (e.g. workers waiting, holding parts whithout carrying them, walking around, etc), thus hindering significantly event and behavior recognition attempts. All these particularities led us to define a new series of events and repeat the labeling process, as can be read in subsection 3.1.

2.3. Challenges

As can be inferred by a brief look (let alone by attempting to perform object tracking and behavior recognition) the dataset

[8] The dataset is publicly available for download on http://www.scovis.eu/. Please cite the paper at hand when publishing results obtained using this dataset.

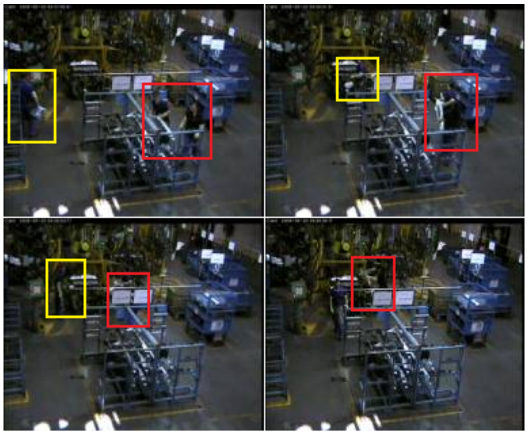

Fig. 1. An exexution of task 2 (red). In the first three frames a third worker is performing task 4 (yellow) in parallel.

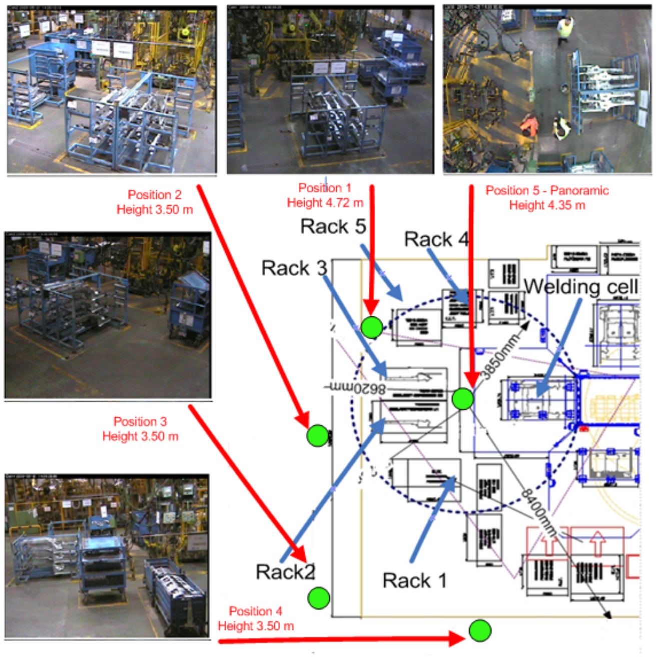

Fig. 2. Depiction of workcell along with the position of the five cameras and racks #1-5.

contains some very challenging sequences. There are serious visibility problems, heavy occlusions, self occlusions, and outliers. In particular, vibrations, sparks, difficult structured background, such as upright racks, welding machines and forklifts creating heavy occlusions of the workers compose a very challenging environment. The frequent illumination changes, along with the fact that the blue color of the workers' clothes bears a great resemblance to the blue color of the racks are additional exacerbating factors.

Regarding action/activity/behavior recognition, the high intraclass and low interclass variance among the tasks makes task discerning hard even for the human eye in certain cases. Significant deviations in the workflow process can occur (especially in dataset-2). Several tasks within a workflow can have fluctuating durations and no clear definition of beginning/ending. Furthermore, the tasks may entail both human actions and motions of machinery in the observed process. Taking these factors into consideration, the introduced WR dataset appears as a challenging testbed for computer vision, machine learning, and multimedia related algorithms.

3. DATA ANNOTATION - SCENE REPRESENTATION

The camera models used for data acquisition are AXIS 212-213 PTZ. We recorded at 25fps with relative jitter bounded by 1.6% on frame rate.

3.1. Activity labeling

Dataset-1. For purposes of enabling behavior and workflow recognition we split each workflow into seven discrete tasks (of meaning to the production process), which include picking different parts from different racks and placing them on a designated cell some meters away, where welding took place. The workspace configuration and the positioning of the cameras is given in Figure 2. More specifically we define the following tasks: (1) One worker picks part #1 from rack #1 and places it on the welding cell. (2) Two workers pick part #2a from rack #2 and place it on the welding cell. (3) Two workers pick 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) Welding: two workers grab the welding tools and weld the parts together. It is possible that sometimes a single worker performs a task normally executed by two workers.

Dataset-2. As mentioned, dataset-2 is more complex, and the specific problems analyzed in Section 2.2 further encumber behavior recognition, thus dictating the need for a new task definition. We therefore further split the seven tasks, thus introducing smaller, shorter ”micro-tasks”, whose appropriate execution indicates the execution of tasks, in an endeavour to improve the recognition process performance. The ”micro-tasks” definition was done is such, as to ensure the fulfillment of the following assumptions: (i) Workflow recognition as described above is possible through micro-tasks recognition. (ii) Micro-tasks are as spatially confined as possible, so that we can define and observe Regions Of Interest (ROIs) efficiently. (iii) Micro-tasks are as temporally short as possible so that simultaneity cases are less often.

The ”micro tasks” are actually parts of the ”tasks” which signify some characteristic actions that may in a later step lead to task recognition. The micro-tasks we defined are listed as follows: (1) Worker picks part #1 from rack #1. (2) Worker places part #1 on the welding cell. (3) Two workers pick part #2a from rack #2. (4) Two workers place part #2a on the welding cell. (5) Two workers pick 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) pick up part #5 from rack #5. (12) Worker(s) place part #5 on the welding cell. (13) Worker grabs first welding tool, approaches cell and starts welding. (14) Robot collects assembled chassis.

3.2. Scene representation

Along with the dataset we provide a set of features that represent the sequences. Initially we tested the efficiency of tracking (e.g. [4]) and pure detection (HOG [5]), but both methods failed in our complex environment. We therefore resorted to another kind of features, the extraction process of which briefly involves background subtraction, Pixel Change History calculation and, finally, representation of the resulting images through sixth order Zernike moments. For details regarding feature extraction see [6]. Zernike moments are known for their noise resiliency, reduced information redundancy and reconstruction capability. Hence, as can be corroborated by the research results mentioned in Section 4, our features lead to a good scene representation of the complex industrial environment. Detailed explanations on the format and structure of the annotation file and the accompanying feature sets are provided online together with the dataset.

4. APPLICATIONS - RESEARCH RESULTS

The WR dataset can serve as a testbed for motion tracking and behavior and workflow recognition. There has been a number of recently published papers that use the WR dataset so as to experimentally verify the applicability and study the performance of the proposed research methods. For example, [7] presents a method whose goal is to robustify tracking-by-detection algorithms without learning specific object models. Regarding behavior recognition, [6] describe activity and workflow recognition methods based on different sets of holistic features and fused Hidden Markov Models and exploiting the redundancies from multiple cameras. When employing the proposed PCH-Zernike features, we attain recognition rates of approximately 92% for dataset-1 and 55% for dataset-2. Finally, in [8] the proposed Evaluative Rectification approach aims at dynamically correcting erroneous task classification results to enhance the behavior modeling and therefore the overall classification rates.

5. CONCLUSION - FUTURE ENHANCEMENTS

We presented the WR dataset as a comprehensive resource for researchers in the areas of motion tracking and actvity recognition. To our knowledge this is a novel dataset in the community in that it involves video sequences from the production line of a complex industrial environment. The heavy occlusions, outliers, human-machinery interaction, and visually complicated environment (similar colors, sparks, vibrations) make this dataset a very challenging testbed for computer vision and image processing related algorithms. Nonetheless, the observed process displayed in the sequences remains structured, in contrast to other surveillance footages (e.g. stemming from airport security cameras), which is a significant quality particularly for machine learning research. In the future, additional annotation related information will be made available, such as bounding boxes for humans and objects for part of the dataset, as well as feature vectors that will be based on the 14-micro-task definition, in order to further augment the usefulness of the WR dataset.

6. REFERENCES

[1] R. Fisher, "The pets04 surveillance ground-truth data sets," in *Proc. IEEE PETS*, 2004, pp. 1–5.

[2] M. Tenorth, J. Bandouch, and M. Beetz, "The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Proc. ICCV Workshops*, 2009, pp. 1089 –1096.

[3] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer, "The human id gait challenge problem: data sets, performance, and analysis," *IEEE TPAMI*, vol. 27, no. 2, pp. 162 –177, 2005.

[4] A. Makris, D. Kosmopoulos, S.J. Perantonis, and S. Theodoridis, "Hierarchical feature fusion for visual tracking," in *Proc. IEEE ICIP*, 2007, pp. VI: 289–292.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, 2005, pp. 886–893.

[6] D. Kosmopoulos, A. Voulodimos, and T. Varvarigou, "Robust human behavior modeling from multiple cameras," in *Proc. IEEE ICPR*, 2010, pp. 3575 – 3578.

[7] S. Stalder, H. Grabner, and L. Van Gool, "Cascaded confidence filtering for improved tracking-by-detection," in *Proc. ECCV*, 2010, pp. 369–382.

[8] N. Doulamis, A. Voulodimos, D. Kosmopoulos, and T. Varvarigou, "Enhanced human behavior recognition using hmm and evaluative rectification," in *ACM ARTEMIS held in conjunction with ACM Multimedia*, 2010, pp. 39–44.