A Dataset of Human Manipulation Actions

Alessandro Pieropan Giampiero Salvi Karl Pauwels Hedvig Kjellström

***Abstract*— We present a data set of human activities that includes both visual data (RGB-D video and six Degrees Of Freedom (DOF) object pose estimation) and acoustic data. Our vision is that robots need to merge information from multiple perceptional modalities to operate robustly and autonomously in an unstructured environment.**

I. INTRODUCTION

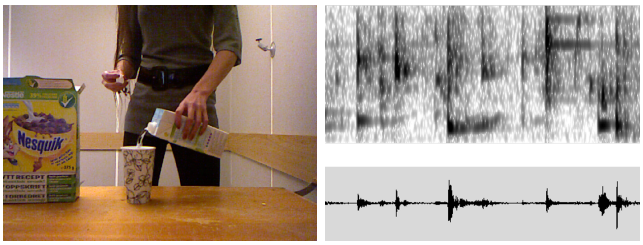
There has been tremendous effort in the robotics community to develop robots able to operate autonomously in unstructured environments. Robots should be able to perceive the world correctly, detect objects, observe and interact with humans, perform activities and understand if the desired outcome has been achieved [1].

In the context of robot learning from demonstration it is essential to understand human activities. Many related data sets have been released in the past decade, however the majority of them focus more on activities that can be best described by looking at the human pose [2]–[4], e.g. running, walking, kicking. Such activities are more suited for video surveillance applications. However in a robotic context it is more interesting to understand task-oriented activities that involve objects (i.e. grasping, pouring, cutting) so that a robot can repeat and fulfill requested tasks autonomously. A few strictly visual data sets have been released in this context [5]–[7]. However, visual cues alone suffer from certain limitations. First, an activity has to be performed within the field of view of the observer. Second, object detection and tracking are sensitive to occlusions while performing activities. Third, some of the objects that are often involved in indoor activities are very hard to detect due to the properties of materials (e.g. shiny or transparent). Last, there are meaningful states induced by an activity that are hard to detect just relying on visual perception. For example, it is very difficult to detect if a person has turned on an oven. We believe that the limitation of visual perception can be compensated by using additional sources of information (Fig. 1). As an example, [5] uses radio-frequency identification (RFID) tags to record the position of objects. However such solution may be intrusive and would not work in a natural environment. We instead propose to rely more on non intrusive sources. Therefore in our data set we provide audio recordings of the sounds produced while the activities are performed. Our dataset

This research has been supported by the EU through TOMSY, IST-FP7-Collaborative Project-270436, Marie Curie FP7-PEOPLE-2011-IEF-301144 and the Swedish Research Council (VR).

The GPU used for this research was donated by the NVIDIA Corporation.

AP and HK are with CVAP/CAS, KTH, Stockholm, Sweden, pieropan,hedvig@kth.se. GS is with TMH, KTH, Stockholm, Sweden, giampi@kth.se. KP is with the University of Granada, Spain, kpauwels@ugr.es.



(a) RGB-D video (b) Audio

Fig. 1. An RGB-D video modality and an audio modality give complementary information about an observed human action. This is a motivation of why to use a recognition method that takes both modalities into account.

enables various experiments. First, it is possible to test object pose detection and tracking algorithms with the baseline provided by our tracker. Second, as the dataset is manually labeled, it is possible to evaluate visual and acoustic cues in the context of activity recognition as we have recently done [8].

II. DATASET

The data set we have collected includes observations of eight subjects fulfilling the task of preparing a milk and cereal breakfast. The actors are not instructed on how to perform the action and, therefore, there is substantial variation in the way they perform it. However, the action can generally be decomposed into 6 different sub-actions: open milk box, pour milk, close milk box, open cereal box, pour cereals and close cereal box. The variability can be observed in the order these sub actions are performed, the distribution of workload between left and right hand, the position of the objects, and so forth. Since the actors perform the action in a natural way the transitions between sub-actions are smooth, some sub actions can be performed in parallel and some may be missing (e.g. sometimes subjects leave the cereal box open at the end of the action.). In details the data set includes:

Video: Calibrated RGB-D video recorded using a Kinect device with 30 Hz framerate and a resolution of 640×480. The time stamp of each frame has been saved so that it is possible to align audio and video correctly. Each frame is saved in a separate matlab file (.mat)

Audio: 4 separate audio tracks using the Kinect microphone array sampled at 16 kHz with 32 bits depth saved as standard waveform audio file.

Object Models: 25 $3D$ object models, built from real images, saved in Wavefront OBJ-files. The models are used to estimate the six DOF pose of the objects detected while

(a) Model shape (b) Model appearance

Fig. 2. Samples of objects used in the dataset. (a) shows the 3D solid models of the objects, (b) shows the models rendered with textures.

performing an activity. The models used in the data set are 4 milk boxes, 2 cereal boxes and 5 cups. We release all objects as they can be used for other purposes such as grasp planning.

Object Pose: To estimate the objects' location and orientation over time we used a real-time method that relies on sparse keypoints for pose detection and dense motion and depth information for pose tracking [9]. This method can simultaneously track the pose of hundreds of arbitrarily-shaped rigid objects at 40 frames per second, with high accuracy and robustness. This is enabled through a tight integration of visual simulation and visual perception that relies heavily on Graphical Processing Units. A detailed 3*D* scene representation, consisting of the textured wireframe models from Fig. 2, is constantly updated on the basis of the observed visual cues. Self-occlusions and occlusions between modeled objects are handled implicitly by rendering the scene through OpenGL. Pose detection runs in parallel with pose tracking, allowing for automatic pose initialization and recovery when tracking is lost. Because the actors were not instructed about the properties of the tracker used, in a few cases some object poses are lost. This happens upon heavy occlusion or fast movements. In spite of these rare cases, the object tracking information included in the database can be used as baseline for future experiments in the context of object detection and tracking.

Manual Labels: manual labels for each sequence including 6 different sub actions: *Open Milk Box*,*Pour Milk*,*Close Milk Box*, *Open Cereal Box*, *Pour Cereals*, *Close Cereal Box*. The labels are provided as Advanced SubStation Alpha (.ass) subtitle files.

Scripts: the information regarding each recorded video frame are stored in a separate matlab file (.mat). A file includes the rgb image, the disparity image, the time stamp and the six DOF pose of each detected object. Python scripts to read the data from the mat file, synchronize the video and audio sources and parse the labels from the subtitle files are provided with the dataset.

III. Conclusions

We present a data set as a resource for studies in the fields of activity recognition, object detection and multi-modal fusion. The data set contains an extended set of examples of *making cereals* action executed by 8 actors in a natural manner (Fig. 3). The data includes 140 RBG-D videos, objects six DOF pose estimates, 3D models of the objects used, acoustic data, manual labels and python scripts to work with it. It will be released publicly and it will be maintained and improved by extracting more features such as hand pose estimation and more activities performed in an indoor kitchen scenario.

References

[1] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Learning object affordances: From sensory motor coordination to imitation. *IEEE TRO*, 24(1):15–26, 2008.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[3] O. Sangmin, A. Hoogs, A. Perera, N. Cuntoor, C. C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011.

[4] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), 2010.

[5] M. Tenorth, J. Bandouch, and M. Beetz. The TUM Kitchen Data Set of everyday manipulation activities for motion tracking and action recognition. In *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences*, 2009.

[6] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[7] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *ECCV 2012*, volume 7572, pages 144–157. 2012.

[8] A. Pieropan, G. Salvi, K. Pauwels, and H. Kjellström. Audio-visual classification and detection of human manipulation actions. In *IROS*, Submitted.

[9] K. Pauwels, L. Rubio, J. Diaz Alonso, and E. Ros. Real-time model-based rigid object pose estimation and tracking combining dense and sparse visual cues. In *CVPR*, pages 2347–2354, 2013.

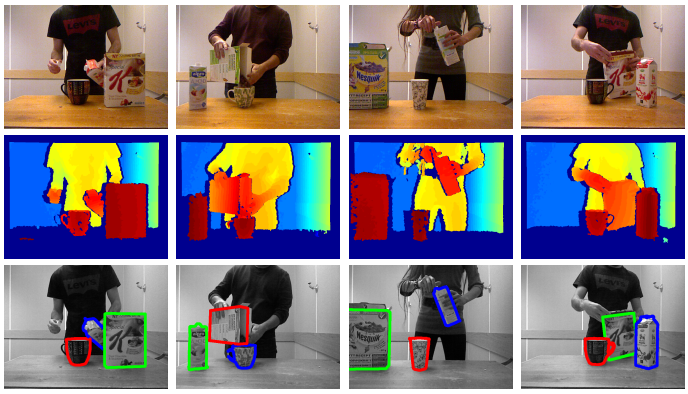


Fig. 3. Examples taken from the dataset. The first two rows show the raw RGB-D data while the last row shows the results of the tracker used. For more examples please refer to the supplemental material.