The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition

Technical Report LIRIS-RR-2012-004
LIRIS Laboratory — UMR CNRS 5205

Christian Wolf, Julien Mille, Eric Lombardi, Oya Celiktutan,
Mingyuan Jiu, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot,
Christophe Garcia, Bülent Sankur

Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR CNRS 5205, F-69621, France

March 28, 2012

Abstract

We describe the LIRIS human activities dataset, the dataset used for the ICPR 2012 human activities recognition and localization competition. In contrast to previous competitions and existing datasets, the tasks focus on complex human behavior involving several people in the video at the same time, on actions involving several interacting people and on human-object interactions. The goal is not only to classify activities, but also to detect and to localize them. The dataset has been shot with two different cameras: a moving camera mounted on a mobile robot delivering grayscale videos in VGA resolution and depth images from a consumer depth camera (Primesense/MS Kinect); and a consumer camcorder delivering color videos in DVD resolution.

1 Introduction

Applications such as video surveillance, robotics, source selection, video indexing and others often require the recognition of actions and activities based on the motion of different actors in a video, for instance, people or vehicles. Certain applications may require assigning activities to several

1	**DI**	Discussion of two or several people	HH
2	**GI**	A person gives an item to a second person	HH, HO
3	**BO**	An item is picked up or put down (into/from a box, drawer, desk etc.)	HO
4	**EN**	A person enters or leaves an room	-
5	**ET**	A person tries to enter a room unsuccessfully	-
6	**LO**	A person unlocks a room and then enters it	-
7	**UB**	A person leaves baggage unattended (drop and leave)	HO
8	**HS**	Handshaking of two people	HH
9	**KB**	A person types on a keyboard	HO
10	**TE**	A person talks on a telephone	HO

Table 1: the behavior classes in the dataset. Some of the actions are human-human interactions (HH) or human-object interactions (HO).

predefined classes, while others may rely on the detection of abnormal or infrequent activities. This task is inherently more difficult than more traditional tasks like object recognition for several reasons. It requires motion information to be extracted from the video and separated from the color and texture information; more importantly, the characteristics of human behavior is less well understood. Previous work and previous datasets focused on very simple actions like running, walking, boxing, hand-clapping etc. in very simple environments. Typically, one video file features a single person performing a single action, e.g. the KTH dataset [5] and the Weizmann dataset [1]. Other datasets focus on more complex activities, but are specialized to broadcast video, for instance the Hollywood 2 dataset [2] or the TRECVid conference series[1]. Others are specialized to other specific situations like the University of Rochester daily living dataset [3] and the different UCF datasets[2]. Datasets oriented on surveillance, including subsets on crowd control, multi-view datasets etc., were introduced as part of the PETS competitions[3]. Other datasets focus on outdoor activities [4]. However, a full and exhaustive review of human action datasets is beyond the scope of this document.

The LIRIS dataset has been designed for the problem of recognizing complex human actions in a realistic surveillance setting and in an office environment. It is used for the ICPR 2012 human activities recognition and localization competition[4] (HARL), a competition organized in conjunction with the *International Conference on Pattern Recognition (ICPR)*. The participants of this competition face the tasks of recognizing actions in a set of videos, where each video may contain one or several actions, eventually at the same time. Table 1 shows the list of actions to be recognized. Note that simple "actions" as walking and running are not part of the events to be detected. The dataset therefore contains motion which is not necessarily relevant for the tasks at hand.

This dataset and the ICPR 2012 HARL competition have been acquired and are organized by the *Imagine* team of the *LIRIS* laboratory[5]. Our research deals with computer vision, in particular recognition: objects activities, faces and emotions.

[1] `http://trecvid.nist.gov`

[2] `http://server.cs.ucf.edu/~vision/data.html`

[3] See, for instance, `http://www.cvg.rdg.ac.uk/PETS2007/data.html` and `http://www.cvg.rdg.ac.uk/PETS2009/index.html`

[4] `http://liris.cnrs.fr/harl2012`

[5] `http://liris.cnrs.fr`

2 The dataset

The dataset is available online on dedicated web site [6]. It is organized into two different and independent sets, shot with two different cameras:

D1/robot-kinect The videos of this set have been shot using our mobile robotics platform VOIR[7], which consists of a mobile robot of model Pekee II manufactured by Wany Robotics[8], see figure 2. During the tests the robot was controlled manually through a joystick. It is equipped with a consumer depth camera of type Primesense/MS Kinect[9], which delivers color images as well as 11bit depth images, both at a spatial resolution of 640×480 pixels, at 25 frames per second (see figures 1a and 1b). In the proposed dataset the RGB information has been converted to grayscale.

The two different sensors of the Kinect module, RGB and depth, produce images whose coordinates will very much differ. We calibrated the Kinect module used during acquisition and we provide information and software allowing to calculate the coordinates in the RGB image (thus, the grayscale image in our dataset) for each pixel of the depth image.

D2/fixed-camcorder The videos of this set have been shot with consumer camcorder (a Sony DCR-HC51) mounted on a tripod. The camera is fixed (zero ego-motion), the videos have been shot in a spatial resolution of 720×576 pixels at 25 frames per second (see figure 1c).

The two sets D1 and D2 are *NOT* completely independent, as most of the D2 videos are shots from the same scenes shot in D1 but taken from a different viewpoint.

Uttermost care has been taken to ensure that the dataset is as realistic as possible:

- As usual, each action has been performed by different people and by different groups of people
- Each action has been shot from different viewpoints and different settings to avoid the possibility of learning actions from background features
- For each video, camera motion tend to be different to avoid the possibility of learning actions from ego motion features

In order to make the dataset more challenging than previous datasets, the actions are less focused on low level characteristics and more defined by semantics and context. The following list gives some examples:

- The discussion action can take place anywhere, either by people standing in some room or in an aisle without any support, or in front of a whiteboard or blackboard, or by people sitting on chairs.
- The action "enter or leave a room" can involve opening a door or passing through an already opened door.
- Three actions involve very similar motion, the difference being the context : "entering a room", "unlocking a door and then entering a room" and "trying to enter a room unsuccessfully".

[6] http://liris.cnrs.fr/voir/activities-dataset
[7] http://liris.cnrs.fr/voir
[8] http://www.wanyrobotics.com
[9] http://www.primesense.com

Figure 1: The dataset has been shot with two different cameras. (a) from the Kinect module we took a grayscale image (left) and a 11bit depth image (middle). Pseudo color videos are provided for better visualization (right); (b) the Sony camcorder delivers RGB color images.

- The action "an item is picked up or put down (into/from a box, drawer, desk etc.)" is very similar to "a person leaves baggage unattended (drop and leave)", as both involve very similar human-object interactions. The difference is mainly defined through the context.
- We took care to use different telephones in the action "telephone conversation": classical office telephones, cell phones, wall mounted phones.
- Actions like "handshaking" and "giving an item" can occur before, after or in the middle of other actions like "discussion", "typing on a keyboard" etc.

The acquisition conditions have *not* been artificially improved, which means that the following additional difficulties are present in the dataset :

- Non-uniform lighting and lighting changes when doors open and close
- The Kinect camera's gain control is rather slow compared to other cameras. This is not the case for the Sony camcorder.
- The depth data delivered by the Kinect camera is disturbed by transparencies like windows etc. This is due to the data acquisition method (shape from structured light).
- The data taken with the mobile robot is subject to vibrations when the robot accelerates or slows down. This reflects realistic conditions in a mobile robotics environment.

The full data set contains 828 actions (subsets D1 and D2) by 21 different people. Each video may contain one or several people performing one or several actions. Some example videos of the dataset can be found on our website[10]. Example images for the different activity classes are shown in figures 3 and 4. Figure 5 shows example frames from longer videos containing several actions, some of which happen in parallel.

3 Groundtruth & evaluation

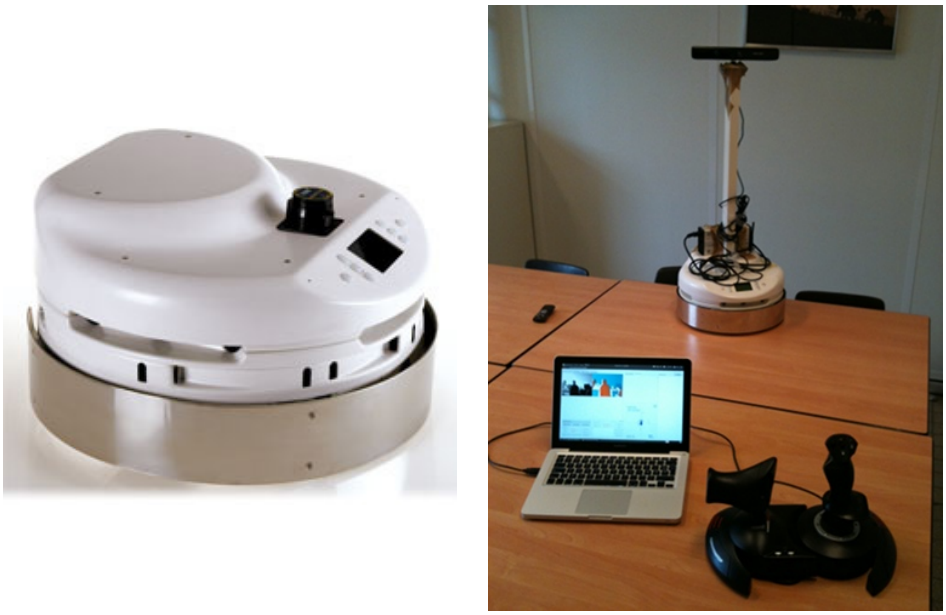
The dataset has been designed for the ICPR HARL competition, whose main objective is

- to detect relevant human behavior in midst of irrelevant additional motion, e.g. other people walking in the environment or performing irrelevant actions;
- to recognize the detected actions among the given action classes
- to localize the actions temporally and spatially

As mentioned previously, different actions may happen in parallel in the same video at the same time. The ground truth data has therefore been annotated by marking labelled bounding boxes for each frame of each action. Figure 6 shows a frame with annotated bounding boxes in a screen shot of the annotation/viewing tool. The viewing tool is available for download on the dataset website. See section 5 for details on the XML format.

The ground truth annotation is segmented into action occurrences regrouping all frames and bounding boxes of the same action. This makes it possible to provide more meaningful recall and precision values — indeed, a recall of 90% is easier to interpret if it tells us that 90% of the actions

[10] `http://liris.cnrs.fr/voir/activities-dataset`



(a) (b)

Figure 2: The Pekee II mobile robot in its standard configuration as delivered by Wany robotics (a) and our setup with the Kinect module during the shooting of the dataset (b).

have been correctly detected, than if its says that, e.g. 90% of the action bounding boxes have been correctly detected on 100% of the activities.

Participants need to report results in the same format — this means that the detection results need to be segmented in the same way: each detected action consists of a list of bounding boxes, where each bounding box corresponds to a frame. Each action must consist of consecutive frames, no holes are allowed in the sequence.

Similar to our object recognition evaluation measure [6], we designed a metric which satisfies the following goals:

1. The metric should provide a quantitative evaluation: the evaluation measure should intuitively tell how many actions have been detected correctly, and how many false alarms have been created.

2. The metric should provide a qualitative evaluation: it should give an easy interpretation of the detection quality.

There is a contradiction between goal (1), to be able to count the number of detected actions, and goal (2), to be able to measure detection quality. Indeed, the two goals are related: the number of actions we consider as detected depends on the quality requirements which we impose for a single action in order to be considered as detected. For this reason we propose a natural way to combine these two goals:

1. We provide traditional precision and recall values measuring detection quantity. An action is considered to be correctly detected or not with two fixed thresholds on the amount of overlap between a ground truth action and a detected action: t_r is a threshold on *area recall*, i.e. it specifies the amount of overlap area which needs to be detected w.r.t. the total area of the ground truth action, whereas t_p is a threshold on *area precision*, i.e. it specifies how much additional detected area is allowed.

Figure 3: Screenshots of the dataset (Kinect grayscale shown only)

Figure 4: Screenshots of the dataset (Kinect grayscale shown only)

Figure 5: Several frames of one of the videos with multiple actions shot from a moving camera. This example contains 3 actions : 2 discussion actions (one on the blackboard, one between two sitting people), and one person typing on a keyboard. Irrelevant motion is produced by other people in the background.

2. We complete the metric with plots which illustrate the dependence of quantity on quality. These performance graphs, similar to the graphs proposed in [6], visually describe the behavior of a detection algorithm.

A detailed and precise description of the metric will be published in a forthcoming publication after the end of the ICPR HARL competition.

4 Dataset set file formats and distribution modes

Care has been taken when the dataset was split into two parts 1) training+validation, and 2) test, such that the same scene is not split over different sets. It is therefore impossible to train on a scene filmed with one camera and to test on the same scene filmed with a different camera.

All videos have been coded as sets of single frames, each video being organized into a different directory. The database is available in two versions :

The standard version, lossy encoded

Color frames of the camcorder are encoded as lossy JPEG images in 75% quality.

Grayscale frames of the Kinect module are encoded as lossy JPEG images in 75% quality.

Depth frames of the Kinect module are encoded in lossy 16bit JPEG2000 images with a compression factor of 20, resulting in 30KB per frame. This image format is supported by Matlab and open-cv, amongst others.

This version of the dataset is of around 20GB size and is available for download.

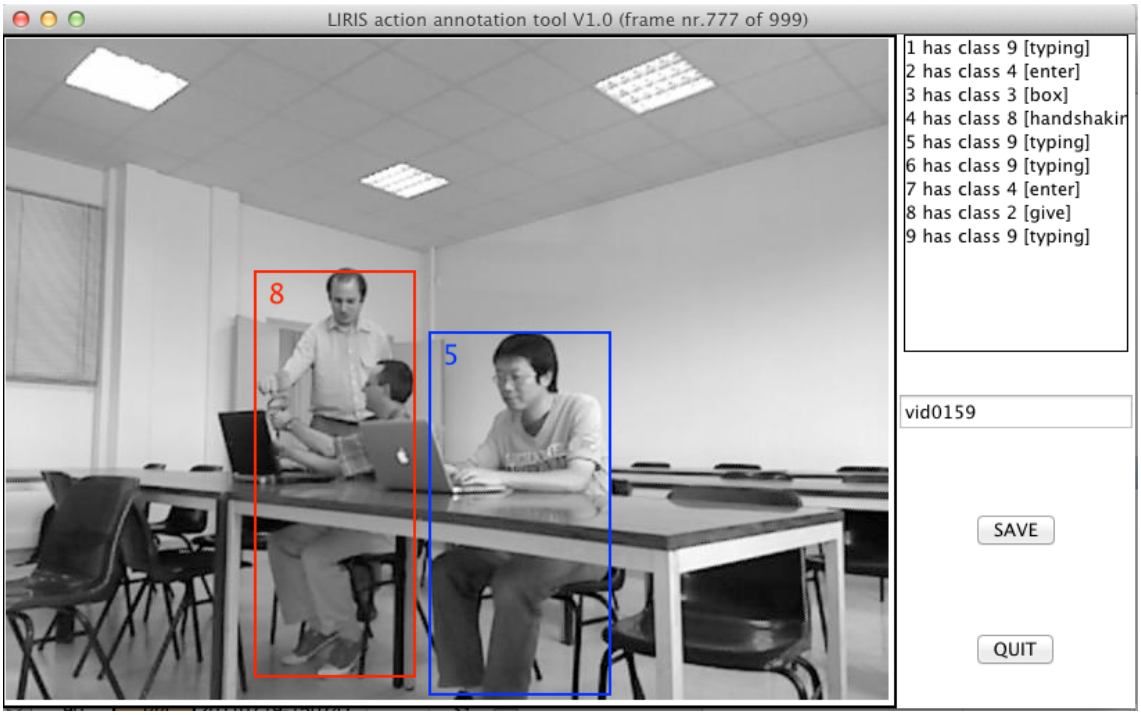


Figure 6: The annotation tool used for the creation of the ground truth XML files.

The (partially) lossless version

Color frames of the camcorder are still encoded as lossy JPEG images in 75% quality, as in the lossy dataset. The camcorder does not provide access to the uncompressed data.

Grayscale frames are encoded in standard PGM format.

Depth frames are encoded in 16bit PGM format. This format is supported by Matlab and open-cv, amongst others.

This version of the dataset is of around 250GB size and will be distributed on request. Participants may send us a portable USB hard drive together with a prepaid envelope and will receive the data per (non electronic) mail.

5 The detection / ground-truth XML file format

The XML file format for the ground truth data and for the detection result is identical (see figure 7). It contains a tag for each video and property video name, which will be used to match groundtruth videos to detection result videos. Participants are free to put the detection results for each video in a separate file or to combine all results in a single file with multiple video tags.

Each video also contains one or several action tags with a running number, the annotated action class, and the list of bounding boxes for the different frames in which the action occurs.

6 Conclusion

Beyond the ICPR 2012 HARL competition, we propose the LIRIS HARL dataset as a new standard dataset, which allows to benchmark activity recognition algorithms based on realistic and difficult

```
<?xml version="1.0" encoding="UTF-8"?>
<tagset>
  <video>
    <videoName>dataset/video1</videoName>
    <action nr="1" class="3">
      <bbox x="40" y="30" width="50" height="101" framenr="34"/>
      <bbox x="41" y="31" width="51" height="105" framenr="35"/>
      <bbox x="41" y="29" width="52" height="101" framenr="36"/>
      <bbox x="41" y="30" width="51" height="104" framenr="37"/>
      <bbox x="42" y="31" width="49" height="102" framenr="38"/>
      <bbox x="42" y="33" width="51" height="103" framenr="39"/>
      <bbox x="42" y="32" width="51" height="100" framenr="40"/>
      <bbox x="42" y="33" width="52" height="100" framenr="41"/>
      <bbox x="41" y="29" width="51" height="101" framenr="42"/>
      <bbox x="41" y="31" width="51" height="101" framenr="43"/>
      <bbox x="41" y="30" width="52" height="100" framenr="44"/>
    </action>
    <action nr="2" class="9">
      <bbox x="212" y="43" width="120" height="87" framenr="342"/>
      <bbox x="211" y="42" width="121" height="86" framenr="343"/>
      <bbox x="209" y="43" width="119" height="86" framenr="344"/>
      <bbox x="208" y="42" width="123" height="86" framenr="345"/>
      <bbox x="209" y="44" width="124" height="87" framenr="346"/>
      <bbox x="207" y="42" width="123" height="87" framenr="347"/>
      <bbox x="210" y="42" width="123" height="87" framenr="348"/>
      <bbox x="212" y="43" width="124" height="86" framenr="349"/>
      <bbox x="211" y="43" width="119" height="88" framenr="350"/>
    </action>
  </video>
</tagset>
```

Figure 7: The XML file format for the ground truth data and the detection results.

data. The target activities are more complex as they need more complex motion modelling and since they contain human-human and human-object interactions. Hopefully this will spur research in the recognition of higher level human behavior.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 1395–1402 Vol. 2, 2005.

[2] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Computer Vision and Pattern Recognition*, pages 2929–2936, 2009.

[3] Ross Messing, Chris Pal, and Henry Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV '09: Proceedings of the Twelfth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2009. IEEE Computer Society.

[4] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, C.-C. Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2011.

[5] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, volume 3, pages 32–36 Vol.3, September 2004.

[6] C. Wolf and J.-M. Jolion. Object count/Area Graphs for the Evaluation of Object Detection and Segmentation Algorithms. *International Journal on Document Analysis and Recognition*, 8(4):280–296, 2006.