JAR-Aibo: A Multi-View Dataset for Evaluation of Model-Free Action Recognition Systems

Marco Körner and Joachim Denzler

Friedrich Schiller University of Jena
Computer Vision Group
Ernst-Abbe-Platz 3, 07743 Jena, Germany
{marco.koerner,joachim.denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. We present a novel multi-view dataset for evaluating model-free action recognition systems. Superior to existing datasets, it covers 56 distinct action classes. Each of them was performed ten times by remotely controlled Sony ERS-7 AIBO robot dogs observed by six distributed and synchronized cameras at 17 fps and VGA resolution. In total, our dataset contains 576 sequences. Baseline results show its applicability for benchmarking model-free action recognition methods.

Keywords: action recognition, behaviour understanding, dataset

1 Introduction and Recent work

The automatic recognition of action and behaviour from video streams gained more and more scientific interest during the last decades, as pointed out by recent reviews[14,1,3]. In order to evaluate and compare algorithms for action recognition or behavior understanding, open-access datasets of high complexity are evidently needed. During the recent years of research on this topic, numerous of those datasets were published and used by the community. The vast majority is designed for single-view approaches, while datasets for multi-view scenarios are rare and only cover a small number of distinct action classes.

We present a multi-view dataset for evaluating model-free action recognition systems. To especially assess the performance of model-free approaches, 56 remotely triggered actions performed by Sony ERS-7 AIBO robot dogs were captured by six synchronized cameras resulting in 576 multi-view sequences.

1.1 Single-View Datasets

As the scientific efforts started to concentrate on recognition of actions and activities captured by single cameras, most of the early datasets show single persons performing basic actions captured from only one view in front of simple and static backgrounds. The most prominent are the `Weizmann`[7] and the `KTH`[16] dataset, where the latter shows varying clothing of the actors.

Table 1: Comparison of recent publicly available datasets for multi-view action recognition. The `JAR-Aibo` dataset mentioned in the last column will be presented in this paper.

	Dataset					
	`IXMAS`	`i3dPost`	`MuHAVi`	`VideoWeb`	`CASIA Action`	`JAR-Aibo`
Year	2006	2009	2010	2010	2007	2013
Application	Human Action Recognition	Human Movement Recognition, 3d Human Action Recognition	Human Action Recognition	Complex Human Activity Recognition	Human Behaviour Analysis	Action and Activity Recognition
Published in	[19]	[6]	[17]	[4]	[18]	—
Number of references[3]	59	10	11	11	18	—
Technicals						
Cameras	5	8	8	4,7,8	3	6
Format	390×291 px, png	1920×1080 px, png	720×576 px, jpg	640×480 px, mpeg1/jpg	320×240 px, avi	640×480 px, png
Frequency	23 fps	25 fps	25 fps	30 fps	25 fps	17 fps
Synchronized	(✓)	✓	✗	✗	(✓)	✓
Content						
Scenery	indoor	indoor	indoor	outdoor	outdoor	indoor
Number of actions	11	8	14	10	8	56
Interactions	none	none, Person-to-Person	none	none, Person-to-Person	none, Person-to-Person, Person-to-Object	none, Actor-to-Actors
Number of actors	13	11	17	2	24	1 (up to 4 in interactions subset)
Repetitions per action and actor	3	1	*several*	*several*	*several*	10
Ground truth data						
Action labels	✓	✓	✓	✓	✓	✓
Calibration	✓	✓	✓	✗	✗	✓
Silhouettes	✓	✗	✓	✗	✗	✗
Bounding boxes	✗	✗	✓	✗	✗	✓
3d models	✓	✓	✗	✗	✗	✗
Background images	✓	✓	✗	✗	✗	✓

As the recognition rates of many approaches obtained for these data got reasonably high, many other datasets were developed over time, *e.g.* `CAVIAR`[5], `Hollywood 1/2`[9,11], `UCF Sports`[15], `UCF Youtube`[10], *etc.* They concentrate on more realistic actions captured in uncontrolled environments showing changing lighting conditions, background, and activities.

Furthermore, datasets like `BEHAVE`[2], `TV Human Interaction`[13], *etc.* were designed to capture person-to-person interactions specifically.

1.2 Multi-View Datasets

After years of research, the interest today moves towards the detection and recognition of actions simultaneously captured by multiple cameras. Nevertheless, only a few datasets with specific limitations exist so far, as summarized and compared in Tab. 1.

The most commonly used is the `IXMAS`[19] dataset, which contains sequences of 11 types of actions performed three times by 13 actors in total. Images were recorded roughly synchronized at a resolution of 390×291 px at 23 fps and saved with lossless png compression. Background images, action labels, as well as body silhouettes and 3d models are delivered with the dataset. The `i3DPost`[6] dataset synchronously captured high-definition videos (1920×1080 px) at 25 fps in lossless png format from 8 points of view. A selection of 8 real-life actions and interactions was performed by 11 actors only once. The distributors provide background images, action labels, and 3d body models. The `MuHAVi`[17] dataset contains 8 views (720×576 px) with 14 actions performed several times by 17 actors. The images were recorded non-synchronously and stored in lossy jpeg format. Only action labels and body silhouettes are available with the data.

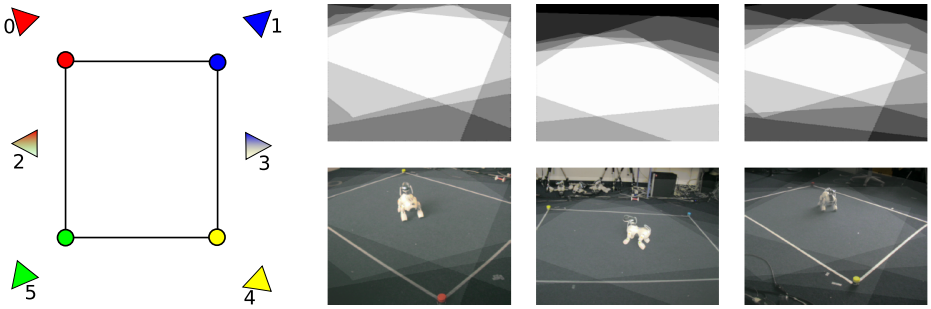
While these datasets mentioned so far were captured under controlled conditions and show a static and simple background, there are also some less commonly used outdoor datasets available, like `VideoWeb`[4] and `CASIA action`[18].

2 Multi-View Action Recognition Dataset

As can be seen, each of the already published datasets shows benefits and drawbacks, which makes them suitable or unsuitable for specific applications and problems. For this reason, we aim to fill a gap by providing a new dataset for evaluation of action recognition systems, especially for the case of appearance-based approaches without any higher-order model knowledge. The selection of actions recorded for our dataset includes well-distinguishable as well as rather similar actions. In this section we will introduce our setup and the provided data.

2.1 Camera Setup and Calibration

We created a setup of six interconnected and calibrated RGB Sony DFW-L500 FireWire cameras distributed around a rectangular region of size $2\,\text{m} \times 3\,\text{m}$ at a height of $90 - 100\,\text{cm}$ as sketched in Fig. 1a. All cameras were oriented to



(a) Six Sony DFW-L500 cameras are distributed around a rectangular scene

(b) All camera views share a common region of the scene (superimposed)

Fig. 1: The setup of the dataset: (a) distribution of cameras, (b) example views with superimposed camera fields of view.

(a) (b) (c) (d)

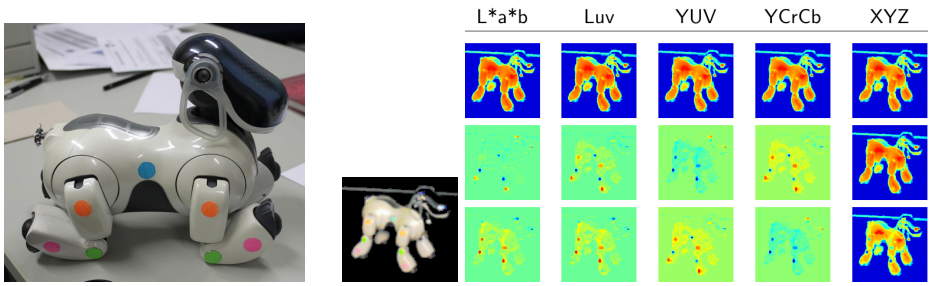
Fig. 2: Images were captured synchronously: (a),(c) and (b),(d) show succeeding frames of views 2 and 4 of the dataset, respectively.

observe a common area of the scene, as displayed in Fig. 1b. Images were captured synchronously (*c.f.* Fig. 2) at VGA resolution (640×480 px) and a frame rate of approximatively 17 fps. We used the png image file format in order to avoid compression artifacts and loss of quality. Further camera parameters, *e.g.* the shutter speed, aperture size, and gain, were adapted once in the beginning of our recordings and kept further untouched. Fig. 4 visualizes the different lightning conditions per camera.

Calibration of the intrinsic and extrinsic parameters of our camera system, was done using the OpenCV library and a "circular grid" calibration pattern and resulted in RMS errors of about 0.2 px (intrinsic) and 0.18 cm (extrinsic).

2.2 Individuals

We decided to use up to four Sony ERS-7 AIBO robot dogs (*c.f.* Fig. 3a) due to their ability to perform a variety of actions in different poses triggered remotely and in order to specifically benchmark model-free approaches. Comparable to



(a) Artificial markers attached to the body surface

(b) Image captured by a Sony DFW-L500 camera and the corresponding channels of different color spaces

Fig. 3: In order to assist model-based approaches, colored markers were attached to the Aibo body. These colors were chosen to be easily distinguishable in appropriate color spaces.

human actors, their anatomy offers many degrees of freedom, which enables them to perform complex actions and to move smoothly. All robot dogs were wireless connected to a central computer, which was used to trigger certain actions. The body surface of the Aibo used in this dataset is bright, glossy and almost untextured. To allow comparisons between model-free and model-based approaches, we applied markers roughly at locations of anatomical joints. Marker colors where chosen to be easily detectable in various color spaces, as illustrated in Fig. 3b. The right half of the body was indicated by an additional blue marker.

2.3 Recorded Actions

While many of the existing datasets only cover a small number of different classes, our dataset was designed to show a high variety of activities. In total, 36 actions performed in up to 3 poses and additionally 6 pose transitions were recorded, which results in a total number of 56 different action classes. Each of them was performed 10 times at different locations and orientations within the scene. Our selection contains rather simple actions (*e.g.* `bow`, `stretch`) as well as complex activities (*e.g.* `dance*`, `lookaround*`). Some actions are easy to distinguish, while others only differ slightly in their type, speed, or order of execution (*e.g.* `hello`, `greeting`). Tab. 2 shows a summary of all recorded pose-action combinations and pose transitions included in our dataset. Additionally, we recorded 16 sequences of interactions between up to 4 dogs, some of them operating in a fully autonomous mode, others acting triggered by the operator.

A selection of actions included in the dataset is shown in Fig. 4.

2.4 Ground Truth Data

All sequences are distributed as frame-wise png images within an unique path such as `$DATAHOME/<pose>_<action>/<sequence>/<camera>_<frame-id>.png`.

Table 2: Overview of all pose-action combinations and pose transitions recorded for the Aibo dataset. Each class was recorded 10 times performed in different positions and orientations. ✓– availlable, ✗– not availlable

Action	Pose			Action	Pose			Action	Pose		
	sit	stand	lie		sit	stand	lie		sit	stand	lie
sit	✗	✓	✓	knockdown	✗	✓	✗	dance2	✗	✓	✗
stand	✓	✗	✓	angry	✓	✓	✗	dance3	✗	✓	✗
lie	✓	✓	✗	disappointed	✓	✓	✗	dance4	✗	✓	✗
				stretch	✓	✗	✗	dance5	✗	✓	✗
greeting	✓	✓	✗	yawn	✓	✓	✗	liftleg	✗	✓	✗
hello	✓	✓	✗	scratch	✓	✓	✗	header	✗	✓	✗
welcomeback	✓	✓	✗	lookaround1	✓	✓	✗	kickright	✗	✓	✗
goodnight	✓	✓	✗	lookaround2	✓	✓	✗	scootleft	✓	✗	✗
bow	✓	✓	✗	snif	✗	✗	✓	scootright	✓	✗	✗
comehere	✓	✓	✗	struggle	✓	✓	✗	pickupbone	✓	✓	✗
yes	✓	✓	✗	bark2	✗	✓	✗	releasebone	✓	✓	✗
no	✓	✓	✗	dance1	✗	✓	✗	touchmyback	✓	✓	✗

We additionally provide background images for each view and bounding boxes of foreground detections. For the interaction subset, a list of action labels sorted by their temporal occurrences is provided for each sequence.

3 Baseline Results

In order to show the applicability of our dataset, we present baseline results for model-free action recognition. For this reason, we used *Temporal Self-Similarity Maps (SSM)* as recently proposed by Körner *et al.*[8], where image sequences are represented by variations of frame-wise extracted low-level features. Within this framework, a SSM is a square-shaped matrix, which entries represent the pairwise similarity (or dissimilarity) of all frames. As can be seen in Fig. 5a, different atomic action primitives induce specific pattern structures in the corresponding SSM. Furthermore, these structures can be assumed to be stable under viewpoint changes. For a more detailed description of this method, we refer to [8].

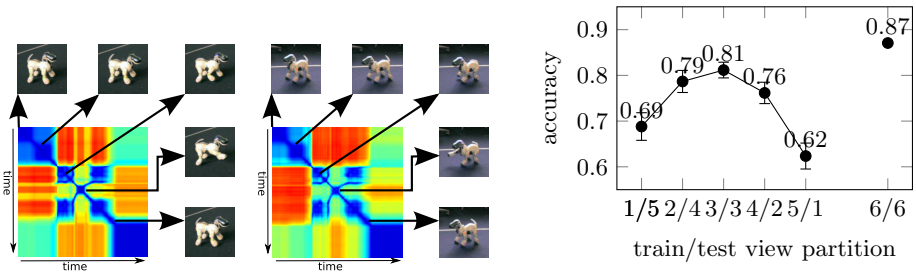
In our experiment, we created SSMs by comparing truncated Fourier descriptors of the single frames. SIFT features were extracted from the diagonal lines of each SSM. After generating a global dictionary of features seen in the testing set, each SSM can be represented by a *Bag of Words* histogram. For training and testing we used disjoint partitions of all camera views. Classification was performed following a 10-fold cross validation scheme by using a *Gaussian Process* classifier and a histogram intersection kernel. Fig. 5b shows the performance of this approach applied to the JAR-Aibo dataset. When applied to the IXMAS[19] dataset, the same method produced recognition rates of about 79%, which is competitive to other model-free methods. This shows that our dataset can be used to benchmark a wide range of appearance-based methods for action recognition.

Fig. 4: Example images from the dataset. Each column represents one camera view, each row show one Aibo action exemplar.

4 Summary

We presented a new extensive dataset for automatic evaluation of appearance-based action recognition approaches. It contains a total number of 576 sequences showing 56 actions performed by remotely triggered Sony ERS-7 AIBO robot dogs observed by 6 synchronized cameras including 16 sequences showing interactions between several Aibos. This dataset shows some challenging properties, which have to be faced:

- Since the dataset was recorded in a windowed lab, the illumination conditions change from view to view as well as from sequence to sequence (*c.f.* Fig. 4).
- While numerous approaches for action recognition operate model-based, there are few standard techniques to extract the body pose of non-human actors [12]. Hence, this dataset is suitable to evaluate model-free approaches.
- Due to the large number of action classes included in the dataset, the chance to confuse semantically related actions is higher compared to other datasets with less, well-distinguishable actions.



(a) Two SSMs obtained for an `stand_kickright` action captured from two viewpoints. Cold and warm colors represent high and low self-similarities, respectively. Action primitives induce similar local patterns in the corresponding SSM even under changes of viewpoint, illumination, or image quality.

(b) Results obtained on JAR-Aibo dataset: averaged recognition rates for different $n_{\text{training}}/n_{\text{testing}}$ view partitions. Errorbars show one standard deviation. Best results were obtained when using the same amount of cameras for training and testing.

Fig. 5: Multi-View Action Recognition by Temporal Self-Similarity Maps: (a) main idea of the SSM approach, (b) performance on JAR-Aibo dataset.

We also gave baseline results to show the applicability of our dataset for benchmarking a wide range of generic model-free action recognition approaches, as they are not limited to the case of recognizing actions performed by human actors.

We hope that this dataset is of use for the research community and can help to further improve the development of this pulsating and important field of research. The complete dataset can be downloaded from `http://www.inf-cv.uni-jena.de/JAR-Aibo`.

Acknowledgements

The authors would like to thank Anna Balbekova for technical assistance during acquisition of this dataset.

References

1. J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16:1–16:43, 2011.
2. S. Blunsden and B. R. Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, (4):1–11, 2010.
3. Jose M. Chaquet, Enrique J. Carmona, and Antonio Fernndez-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.
4. Giovanni Denina, Bir Bhanu, HoangThanh Nguyen, Chong Ding, Ahmed Kamal, Chinya Ravishankar, Amit Roy-Chowdhury, Allen Ivers, and Brenda Varda. Videoweb dataset for multi-camera activities and non-verbal communication. In Bir

Bhanu, Chinya V. Ravishankar, Amit K. Roy-Chowdhury, Hamid Aghajan, and Demetri Terzopoulos, editors, *Distributed Video Sensor Networks*, pages 335–347. 2011.

5. Robert B. Fisher. The pets04 surveillance ground truth data set. In *Proceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS04)*, pages 1–5, 2004.
6. N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas. The i3dpost multi-view and 3d human action/interaction database. In *Proceedings of the 2009 Conference for Visual Media Production*, pages 159–168, 2009.
7. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(12):2247–2253, 2007.
8. M. Körner and J. Denzler. Temporal self-similarity for appearance-based action recognition in multi-view setups. In *Proceedings of the 15th International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2013. (to appear).
9. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the 21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.
10. J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *Proceedings of the 2nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1996–2003, 2009.
11. M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Proceedings of the 22nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2929–2936, 2009.
12. T. Nierobisch and F. Hoffmann. Appearance based pose estimation of aibo's. In *IEEE Conference on Mechatronics and Robotics*, volume 3, pages 942–947, 2004.
13. Alonso Patron, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *Proceedings of the 21st British Machine Vision Conference (BMVA)*, pages 50.1–50.11, 2010.
14. Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.
15. M.D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of the 21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.
16. Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 32–36, 2004.
17. S. Singh, S.A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 48–55, 2010.
18. Y. Wang, K. Huang, and T. Tan. Human activity recognition based on r transform. In *Proceedings of the 20th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.
19. D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV)*, pages 1–7, 2007.