Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection*

Mohammadamin Barekatain[1], Miquel Martí[2,3], Hsueh-Fu Shih[4], Samuel Murray[2], Kotaro Nakayama[5], Yutaka Matsuo[5], Helmut Prendinger[6]

[1]Technical University of Munich, Munich, [2]KTH Royal Institute of Technology, Stockholm, [3]Polytechnic University of Catalonia, Barcelona, [4]National Taiwan University, Taipei, [5]University of Tokyo, Tokyo, [6]National Institute of Informatics, Tokyo

m.barekatain@tum.de, miquelmr@kth.se, r03945026@ntu.edu.tw, samuelmu@kth.se, nakayama@weblab.t.u-tokyo.ac.jp, matsuo@weblab.t.u-tokyo.ac.jp, helmut@nii.ac.jp

Abstract

Despite significant progress in the development of human action detection datasets and algorithms, no current dataset is representative of real-world aerial view scenarios. We present Okutama-Action, a new video dataset for aerial view concurrent human action detection. It consists of 43 minute-long fully-annotated sequences with 12 action classes. Okutama-Action features many challenges missing in current datasets, including dynamic transition of actions, significant changes in scale and aspect ratio, abrupt camera movement, as well as multi-labeled actors. As a result, our dataset is more challenging than existing ones, and will help push the field forward to enable real-world applications.

1. Introduction

With the increased use of unmanned aerial vehicles (UAVs) for tasks such as surveillance, delivery and search and rescue, we believe that a better understanding of human actions from an aerial view is important. For example, in surveillance tasks, it can be essential to recognise actions and track the actors in order to detect anomalies. Likewise, for search and rescue missions, being able to distinguish a person's action could help the system understand if that person is in need of help. Although several new action recognition datasets have been presented over the last years, we argue that none is suitable for being used with UAVs. Not only is the view angle and scale of objects different from UAVs cameras, but available datasets also suffer from not being representative of common outdoor actions.

To address this, we present a new dataset, Okutama-Action, captured from UAVs flying at different altitudes and at different angles, to get a diverse set of sequences. Each sequence is much longer than those in other datasets, which makes them more similar to real world tasks where objects must be tracked over extensive time periods. In total, 12 action classes are used, deemed to be typical outdoor actions. Since basic actions like *sitting* and *walking* are annotated, all humans are labeled in each frame and they may have more than one labeled action. Compared to previous datasets, our action classes are more difficult to tell apart visually, because there are less distinguishing features exterior to the actor, such as change in environment. This is also shown by training and evaluating a state-of-the-art action detection model, which performs worse on Okutama-Action than on other datasets. This indicates that better action detection models have to be developed for use in real-world applications. Figure 1 illustrates some examples of our dataset.

The outline of this paper is as follows: first, a review of currently available datasets for spatio-temporal human action detection is presented. Second, the details and the design choices behind our Okutama-Action dataset are presented, with comparisons to the reviewed datasets. Last, an action detection model is trained and evaluated on our dataset, and its performance is compared to that on other datasets.

*The work was conducted while the first 3 authors were participating in the NII International Internship Program, and the 4th was an Invited Researcher at NII.

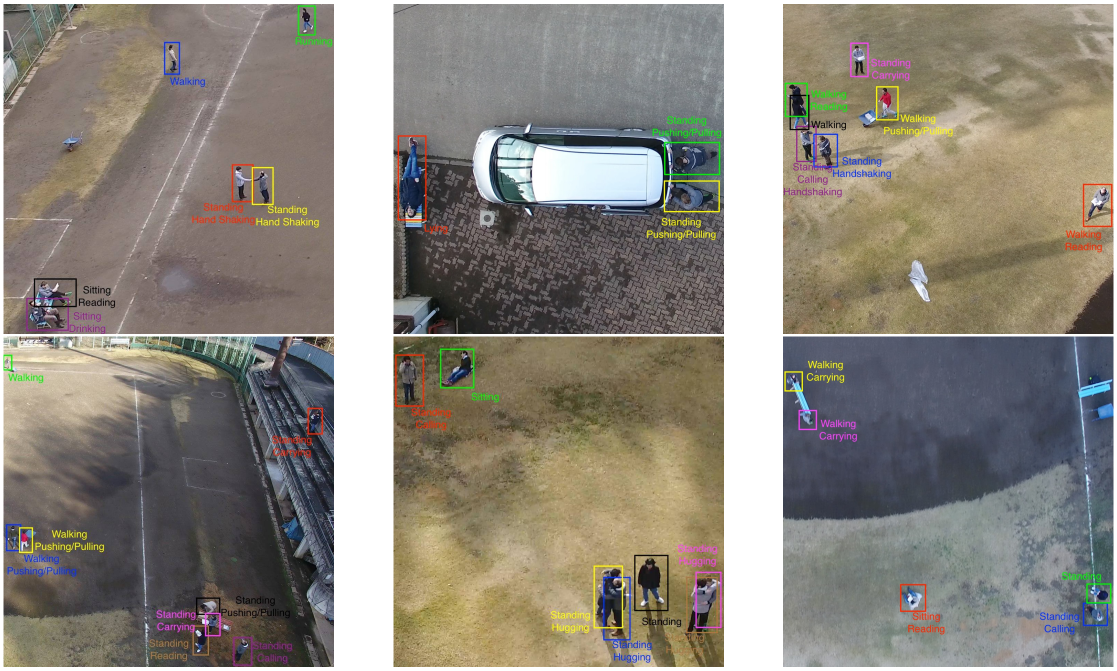

Figure 1: Sample frames of Okutama-Action dataset. These are cropped versions of original frames.

2. Related work

Reviews of current datasets for human action detection [6, 2] indicate a lack of aerial video sequences captured from a mobile platform. The same reviews also ascertain the lack of several challenges existing in real-world airborne scenarios, including dynamic transition of actions, significant changes in scale and aspect ratio and abrupt camera motion. Furthermore, currently available datasets are limited in at least one of these aspects: action types, number of concurrent actors, temporal length of actions, diversity of concurrent actions, video resolution and sequence duration.

We briefly review some of the relevant action detection datasets which include spatial annotation of actions. The UCF Sports dataset [16, 25] contains 150 sports videos for 10 action categories and the J-HMDB dataset [4] consists of 928 videos with 21 actions. Although these datasets include certain challenges such as camera motion, it is still relatively easy to recognize the actions by observing only the scene or the pose of the actor in a single frame [28]. In addition, videos are of low resolution, contain only a single action and are trimmed to the action's duration, which is generally short.

The UCF-101 dataset[1] [26] with 24 action classes and 3207 sequences, is the largest and most diverse dataset to date which includes challenges such as camera motion, and changes in scale and viewpoint. Unfortunately, the actions covered in this dataset are not typical in aerial scenarios as it mainly contains indoor and sports actions. Additionally, although video sequences may contain concurrent actors, they all perform the same action. Furthermore, videos are of low resolution (320x240), have short duration, and in most of them, the actions lasts more than 80% of the video duration [29].

Two of the few datasets which includes video samples of concurrent actions from different categories are LIRIS-HARL [30] and DALY [29] dataset, both with 10 action classes. LIRIS-HARL is captured in an office environment with different cameras, including a moving camera mounted on a mobile robot; videos in DALY are from YouTube. Unfortunately, in both datasets action categories are not common to aerial scenarios and there is no dynamic transition of actions, i.e. each actor performs a single action in each video clip. Moreover, in LIRIS-HARL, videos have low resolution (720x576) and short duration, and DALY is weakly-annotated, meaning that for each temporal action instance, only 5 uniformly sampled frames are annotated.

Another work worth mentioning is UT-Interaction dataset [19], which consists of 20 video sequences of continuous executions of 6 action classes. Actions from different categories may occur concurrently, there is dynamic transition of actions and videos are not trimmed to action duration. Unfortunately, the total number of videos and frames is low, which makes the dataset inappropriate for deep learning models, and various real-world challenges are

[1] The original UCF-101 dataset, that contains over 10K videos for 101 actions does not include spatio-temporal annotations. This annotation was later provided in THUMOS13 challenge [5] for a subset of UCF101.

missing. For example, action types are limited to interaction between two humans, the camera is static, and there is no partial occlusion of actions.

In recent years, various datasets for human action recognition have been constructed, such as ActivityNet [1], Sports-1M [7], HMDB51 [8], MPII Cooking [18, 17], Olympic Sports [12], Hollywood [11, 9] and MERL Shopping [23]. However, the annotation of these datasets does not include spatial location of actions. Moreover, instead of localization of actions for each person, some video benchmarks [13, 21] are intended for assessing the performance of event recognition algorithms, where an event can be composed of multiple complex human activities. Also related to our work is Stanford drone dataset [15] intended for trajectory forecasting and multi-target tracking in aerial view videos.

Okutama-Action provides 43 fully-annotated sequences which are of use to train and evaluate models for spatio-temporal detection of multiple concurrent human actions from different categories in the video footage of a mobile aerial platform. Our dataset is unique in the following aspects: an aerial view dataset that contains representative samples of actions in real-world airborne scenarios; dynamic transition of actions where, in each video, up to 9 actors sequentially perform a diverse set of actions; and a real-world challenge of multi-labeled actors where an actor performs more than one action at the same time. Additionally, our dataset has a significant increase compared to previous datasets, in number of actors and concurrent actions (up to 10 actions/actors), as well as video resolution (3840x2160) and sequence length (one minute on average).

3. Okutama-Action development

All videos of Okutama-Action are captured from UAVs (DJI Phantom 4) at a baseball field in Okutama, Japan. In this section, we describe the selected action categories as well as the data collection settings. We then explain the annotation process and summarize the properties of our dataset. See Figure 1 for example frames of our dataset.

3.1. Dataset design and collection

Action selection In order to collect video samples of human actions that are representative of everyday outdoor actions, we analyzed the video footage of low-altitude UAVs which led to a selection of 12 actions, including *Reading*, *Handshaking*, *Drinking* and *Carrying*. Inspired by [26], we group these actions into 3 types: 1) Human to human interaction; 2) Human to object interaction; 3) None-interaction. Figure 2 displays all action classes and their corresponding groups.

UAVs configuration We experimented with different altitudes and camera angles for capturing videos of these actions, in order to find the proper settings for our UAVs during the data collection, ensuring that the actions are clearly visible and distinguishable. Based on this experiment, we decide the altitude range to be from 10 to 45 meter and camera angle to be either 45 or 90 degree.

Data collection In order to ensure that the sequences of our dataset are representative of real-world aerial applications, scripts for 22 scenarios were written in which up to 9 actors participate. We attempt to include various transitions of actions that can happen in real-world, while also having variability in execution style and in number of actors. For example, actors carry different items and read books of different sizes, and there exists crowded frames with 9 actors as well as deserted frames with no actor. Furthermore, in some scenarios, the actors were asked to perform random actions of their choice in order to increase the diversity.

In addition, a separate set of scenarios were written for our 2 UAV pilots in order to make sure we have variety in viewpoint. For example, in some sequences the UAV is still and only spinning while in others it may be moving with a top-down camera angle. The dataset also includes some metadata for each video sequence, namely camera angle, speed and altitude. Furthermore, it was our goal to include common challenges existing in the video footage of airborne platforms such as partial occlusion of actors as well as changes in scale, aspect ratio and camera speed.

Each scenario, with the exception of one, was captured using 2 UAVs (of different configuration) at the same time, which, together with the metadata, are of use for action detection algorithms comparison. Data collection was done in two different lighting conditions (sunny and cloudy) at a baseball field and the actors were a group of researcher at our lab. Video sequences were recorded at 4K resolution and 30 FPS using a high performance camera mounted on an adjustable, integrated gimbal system on the UAVs.

3.2. Dataset annotations

We use VATIC [27], an open source video annotation tool, integrated with Amazon Mechanical Turk to manually annotate the videos at 10 FPS. The annotations are then linearly interpolated to 30 FPS. The bounding boxes and their corresponding action labels are reviewed and adjusted by

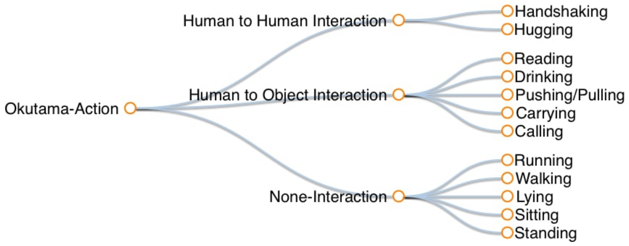

Figure 2: Categorization of action classes in Okutama-Action dataset

members of our group.

In the original annotation, bounding boxes may have more than one label, since an actor may naturally perform multiple actions at the same time. However, current action detection algorithms are limited to detect a single action; hence, we also provide an annotation set in which the bounding boxes have a single label. To do so, we first ranked the action categories based on their group type and number of instances that they have in our dataset (the lower the number of instances is, the higher its rank). The None-interaction actions have the lowest rank since by default an actor is always performing an action of this type (e.g. *Reading* while *Sitting*). Finally, for each bounding box, we keep the action with the highest rank.

3.3. Dataset summary and statistics

Our new Okutama-Action dataset contains a total of 43 video sequences at 30 FPS and 77365 frames in 4K resolution. These sequences were recorded using 2 UAVs flying at altitudes varying between 10-45 meters and with camera angle of 45 or 90 degrees. Figure 3, on the left side, demonstrates the number of samples per action class, and the chart shown on the right side illustrates total duration of actions (blue) and the average action duration (green) for each action category. These statistics are calculated from the single-label annotation of the dataset.

Dataset split for training and testing set For evaluation purposes, we split our dataset into two distinct sets: 1) train-val set, consisting of 33 video sequences; 2) test set, consisting of 10 video sequences. This split is designed in a way to make sure that the two sequences from the same scenario are in the same set. If this were not the case, the test data would not be completely unseen - for example, a model could memorize the relative positions of actors, and the specific action transitions for a given sequence. Moreover, the split ensures that diverse challenges exists in the test set, which is important for evaluating the robustness of action detection models. For example, not all sequences have a change in camera angle or altitude, but the test set should properly assess a model's performance under these circumstances. Lastly, since the configuration of the UAVs (speed, altitude and camera angle) differs between sequences, it is possible to identify the UAV configuration that gives the best performance for a given model, so that during deployment in real-world applications the optimal configuration can be used.

Comparison with other datasets Some features of our dataset is compared to existing datasets in Table 1. Okutama-Action is the second largest fully-annotated dataset for concurrent human action detection task after UCF101. We would like to emphasise that no other action detection dataset includes samples which are representative of real-world airborne scenarios. To the best of our knowledge, Okutama-Action is the first dataset for spatio-temporal action detection that includes multi-labeled actors. Moreover, our dataset has made significant progress in terms of number and diversity of concurrent actions as well as video resolution and sequence duration.

4. Experimental results

In this section, we show how to adapt a simple model designed for object detection to the tasks of action detection, and evaluate it on both tasks with our Okutama-Action dataset. First, we describe the Single Shot MultiBox Detector (SSD) [10], we explain how we use it for detecting pedestrians in our dataset and show the results we get. Second, we explain how we use the same model for action detection and show our results. The results reported in this section were computed using the split detailed in Section 3.3. All experiments were carried out using the SSD Caffe fork with CUDA 7 and two Nvidia K40 GPUs.

4.1. Pedestrian detection

Model description SSD is a unified object detector implemented in a single network. Having inference speed in mind, the space of possible boxes is discretized coarsely into a set of default boxes and then their coordinates are refined to tightly surround the object. This is done with different aspect ratios and at different scales by taking features from different stages of the network. For each (refined) default box, the network predicts the likelihood of having an object of a given class in it. As other single-shot approaches, e.g. YOLO [14], SSD avoids having a separate region proposal generation step and gives a prediction with a single forward pass of the network.

Training strategy We train the SSD model based on VGGnet [22] with an image input size of 512x512 pixels, following the original strategy of 20000 iterations with a learning rate of 10^{-4} and the rest of hyperparameters set to the default values of SSD. In order to use our dataset for the pedestrian detection task, we give all bounding boxes the label *Pedestrian*. This way we validate our choice as we expect this to be an easier task; we believe that if good performance on pedestrian detection is not achieved, the model can not successfully be used for action detection. The resulting detections could be used by a Multiple Object Tracking algorithm directly.

Results We use mean Average Precision at 0.5 IoU threshold (mAP@0.5) as an evaluation metric as is commonplace in object detection tasks [14, 10, 3]. In our case, with only one class, it is simply the Average Precision for the class *Pedestrian* and we get a value of 72.3% on the test set of Okutama-Action dataset. Figure 4 shows two selected frames with the detection results of the best model. We note that as reported in [3], the model performs poorly when pedestrians are too small, which we determine hap-

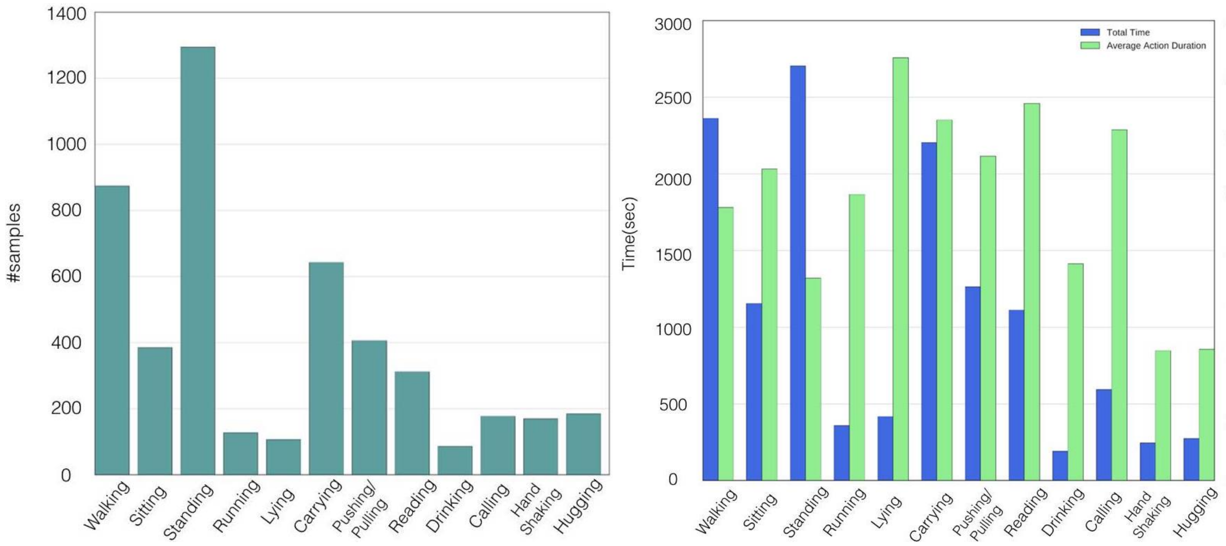

Figure 3: Okutama-Action statistics. Left: Number of samples (instances) per action class. Right: Total duration of actions for each action class is illustrated using the blue bars. The average action duration for each class is depicted in green.

Dataset	Year	Number of Actions	Total Frames	Average Video Dur.	Resolution	Concurrent Actions	Resource
UCF Sports [16, 25]	2008	10	10K	5.8s	690x450	No	TV, Movies
UT-Interaction [19]	2010	6	36K	**60s**	720x480	**Yes**	Actor Staged
UCF-101 [26]	2012	**24**	**558K**	5.8s	320x240	**Yes**	YouTube
J-HMDB [4]	2013	21	32K	1.4s	320x240	No	Movies, YouTube
LIRIS-HARL [30]	2014	10	64K	15.2s	720x576	**Yes**	Actor Staged
Okutama-Action	2017	12	77K	**60s**	**3840x2160**	**Yes**	Actor Staged

Table 1: Comparison of Okutama-Action dataset with current fully-annotated spatio-temporal human action detection datasets.

pens when the altitude of the UAVs is higher than approximately 30 meters.

4.2. Action detection

Model description The action detection model we use follows [24] to get spatio-temporal action localisation and prediction. This model follows a two-stream approach which can be divided in three steps: SSD is the object detector of choice used in the first step of [24] to get the location and class of the actions as detection boxes, in both natural RGB and optical flow streams, for each frame. The second step merges the detections and classification scores of both streams to combine the appearance and motion cues from the natural and optical flow images. In the third step, the sequences of detections are used to incrementally construct action tubes - sequences of detections pertaining to a single action - giving temporal and spatial consistency of the actions across frames and delimiting their duration in an online fashion. Here, for simplicity, we limit our evaluation to the first step. We use the annotation set with only one concurrent action, described in Section 3.2, as SSD cannot handle multiple labels.

Training strategy We train the model with different input image sizes, as we expect a larger size to result in increased accuracy, though at the cost of longer training and inference times. When setting the input image size to 512x512 we train for 34000 iterations with an initial learning rate of 10^{-3}, which we divide by 10 after 19000 and 30000 iterations. For the larger input image size of 960x540, which has the same aspect ratio as the original video, we train the model for 60000 iterations and use an initial learning rate of 10^{-4}, which is divided by 10 after 40000 iterations. We also train a model on the optical flow images of size 512x512, with the same parameters as the RGB 512x512 model.

Results Table 2 shows the mAP of action detection on the test set of Okutama-Action. As in object detection, mAP@0.5 is commonly used as an evaluation metric for action detection [20]. Figure 5 shows the results for each class for the models trained on the natural RGB images, compar-

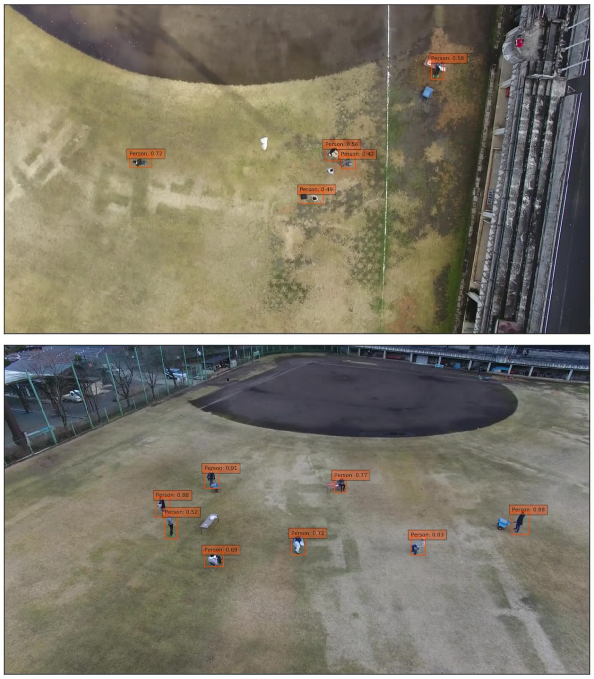

Figure 4: Sample detections of SSD model for pedestrian detection. Detections with confidence score higher than 0.4 are shown. Frames from Okutama-Action test set.

ing both input sizes. An improvement with increased input size can be seen for most of the classes. Figure 6 shows selected detection results from the SSD 960x540 model for the action detection task. By visual inspection we observe that the performance is better when the camera angle is 45 degrees. The reason for this may be that with a lower angle, each actor occupies more pixels of the image. As pointed out in [10], our models are good at localizing objects, but worse at distinguishing classes. The gap between mAP for pedestrian detection and action detection confirms this. We see that the actions strongly related to temporal aspects have a low accuracy, e.g. *Running* often being confused with *Walking*. This is likely because we only distinguish classes at a frame-by-frame level. On the other hand, both *Pushing* and *Carrying* are more easily classified, which we believe is due to the fact that there is a large object next to the actor. Table 3 shows how the results we achieve on Okutama-Action compare to the best reported results on other datasets [24].

5. Conclusions

In this paper, we present Okutama-Action, a new dataset for concurrent human action detection. It is a high resolution aerial view dataset consisting of 43 minute-long sequences, with 12 different action classes. By comparing performance in action detection with existing datasets, we show that Okutama-Action is more challenging, due to having non-static carema with abrupt motion, dynamic transition of actions, multiple concurrent actions and multi-labeled actors. An action detection model is trained and evaluated on our dataset, the performance of which demonstrates the difficulty of the dataset. In order to motivate more research in this area, we plan to make our dataset publicly available at `okutama-action.org`.

As future work, we wish to adopt deep learning models that can handle multi-labeled outputs, to address the multiple-action annotation set provided for Okutama-Action. Another area worth investigating is to evaluate the performance of Multiple Object Tracking algorithms on our dataset, which should not be trivial considering the existing challenges.

Image	Size	mAP (%)
RGB	512x512	15.39
RGB	960x540	**18.80**
Optical Flow	512x512	6.47

Table 2: Results of different SSD models for action detection on Okutama-Action test set. The mAP is computed at 0.5 IoU threshold.

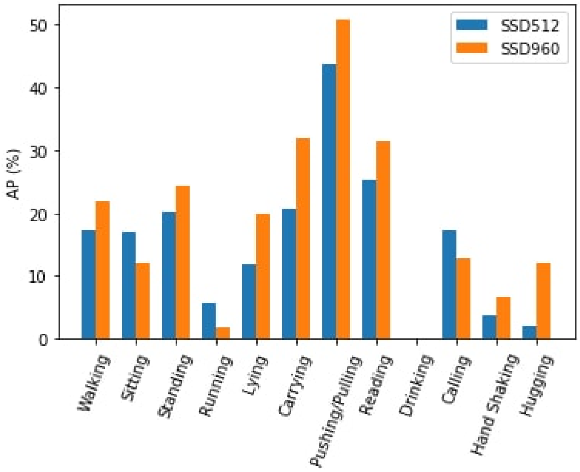

Figure 5: Per-class Average Precision for the models trained for action detection evaluated on Okutama-Action test set. Blue for the model with input size 512x512; Orange for the model with input size 960x540.

Dataset	Model description	mAP (%)
UCF101	SSD, 300x300, RGB [20]	37.93
J-HMDB	SSD, 300x300, RGB [20]	48.39
Okutama-Action	SSD, 960x540, RGB	18.80

Table 3: Best reported results for different action detection datasets. The mAP is computed at 0.5 IoU threshold.

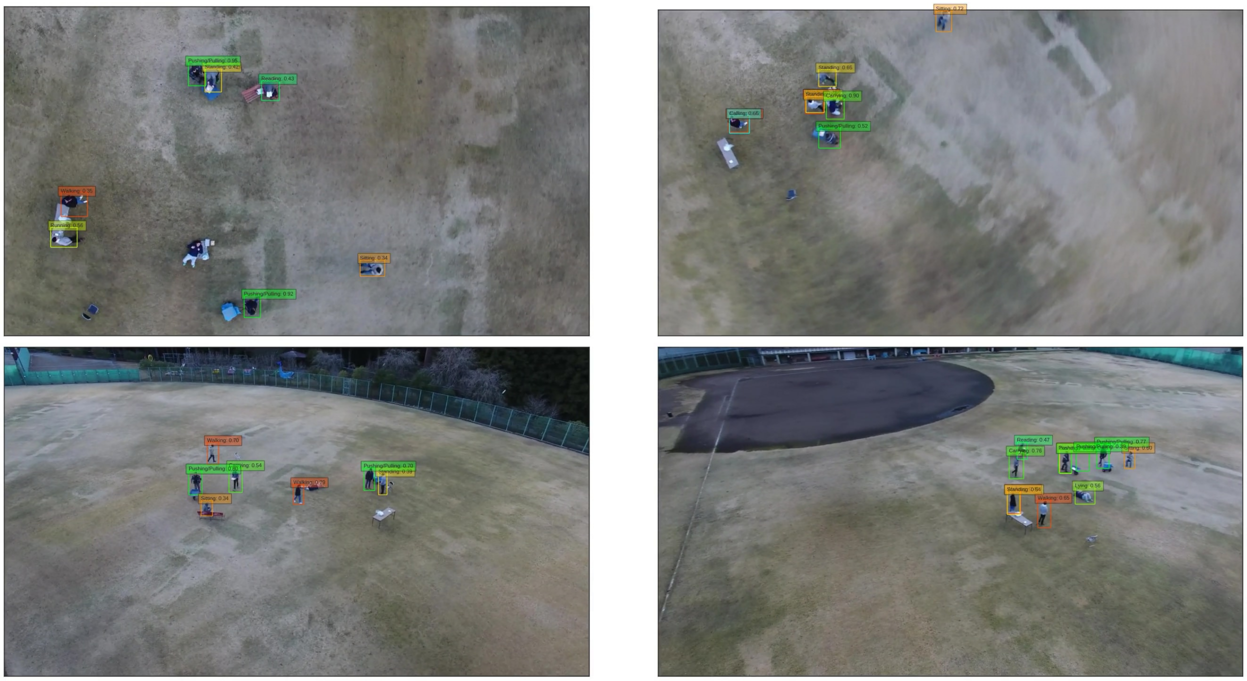

Figure 6: Sample detections of SSD 960x540 model for action detection. Detections with confidence score higher than 0.3 are shown. Each color corresponds to an action category. Frames from Okutama-Action test set.

Acknowledgement

Research in this paper was supported by National Institute of Informatics, Tokyo. We would like to thank researchers at Prendinger and Inamura laboratory for their participation in our data collection. Thanks also to Mr. Eduardo Cordeiro who contributed to the annotation of the dataset and to Ms. Marcia Baptista and Mr. Sergi Caelles for their feedback. The work was partially supported by Prof. Matsuos Grant-in-Aid for Scientific Research on Innovative Areas on Improvement of Predictability by Integrating Deep Learning with Symbol Processing.

References

[1] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[2] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

[3] J. Huang, V. Rathod, C. Sun, M. Zhu, A. K. Balan, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012, 2016.

[4] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3192–3199, 2013.

[5] Y. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014.

[6] S. M. Kang and R. P. Wildes. Review of action recognition and detection methods. *arXiv preprint arXiv:1610.06906*, 2016.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[8] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

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

[12] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity

classification. In *European conference on computer vision*, pages 392–405. Springer, 2010.

[13] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 3153–3160. IEEE, 2011.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[15] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision*, pages 549–565. Springer, 2016.

[16] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[17] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1194–1201. IEEE, 2012.

[18] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373, 2016.

[19] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *Computer vision, 2009 ieee 12th international conference on*, pages 1593–1600. IEEE, 2009.

[20] S. Saha, G. Singh, M. Sapienza, P. H. Torr, and F. Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*, 2016.

[21] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4576–4584, 2015.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2016.

[24] G. Singh, S. Saha, and F. Cuzzolin. Online real time multiple spatiotemporal action localisation and prediction on a single platform. *arXiv preprint arXiv:1611.08563*, 2016.

[25] K. Soomro and A. R. Zamir. Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer, 2014.

[26] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[27] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013.

[28] T.-H. Vu, C. Olsson, I. Laptev, A. Oliva, and J. Sivic. Predicting actions from static scenes. In *European Conference on Computer Vision*, pages 421–436. Springer, 2014.

[29] P. Weinzaepfel, X. Martin, and C. Schmid. Towards weakly-supervised action localization. *arXiv preprint arXiv:1605.05197*, 2016.

[30] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.-E. Bichot, et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014.