UAV-GESTURE: A Dataset for UAV Control and Gesture Recognition

Asanka G Perera, Yee Wei Law, and Javaan Chahl

1 School of Engineering, University of South Australia, Mawson Lakes, SA 5095, Australia
asanka.perera@mymail.unisa.edu.au, {yeewei.law,javaan.chahl}@unisa.edu.au
2 Joint and Operations Analysis Division, Defence Science and Technology Group, Melbourne, Victoria 3207, Australia

Abstract. Current UAV-recorded datasets are mostly limited to action recognition and object tracking, whereas the gesture signals datasets were mostly recorded in indoor spaces. Currently, there is no outdoor recorded public video dataset for UAV commanding signals. Gesture signals can be effectively used with UAVs by leveraging the UAVs visual sensors and operational simplicity. To fill this gap and enable research in wider application areas, we present a UAV gesture signals dataset recorded in an outdoor setting. We selected 13 gestures suitable for basic UAV navigation and command from general aircraft handling and helicopter handling signals. We provide 119 high-definition video clips consisting of 37151 frames. The overall baseline gesture recognition performance computed using Pose-based Convolutional Neural Network (P-CNN) is 91.9 %. All the frames are annotated with body joints and gesture classes in order to extend the dataset’s applicability to a wider research area including gesture recognition, action recognition, human pose recognition and situation awareness.

Keywords: UAV · Gesture dataset · UAV control · Gesture recognition

1 Introduction

Unmanned aerial vehicles (UAVs) can be deployed in a variety of applications such as search and rescue, situational awareness, surveillance and police pursuit by leveraging their mobility and operational simplicity. In some situations, a UAV’s ability to recognize the commanding actions of the human operator and to take responsive actions is desirable. Such scenarios might include a firefighter commanding a drone to scan a particular area, a lifeguard directing a drone to monitor a drifting kayaker, or more user-friendly video and photo shooting capabilities. Whether for offline gesture recognition from aerial videos or for equipping UAVs with gesture recognition capabilities, a substantial amount of training data is necessary. However, the majority of the video action recognition datasets consist of ground videos recorded from stationary or dynamic cameras.
Different video datasets recorded from moving and stationary aerial cameras have been published in recent years [15,6]. They have been recorded under different camera and platform settings and have limitations when used with a wide range of human action recognition behaviors demanded today. However, aerial action recognition is still far from perfect. In general, the existing aerial video action datasets are lacking detailed human body shapes to be used with state-of-the-art action recognition algorithms. Many action recognition techniques depend on accurate analysis of human body joints or body frame. It is difficult to use the existing aerial datasets for aerial action or gesture recognition due to one or more of the following reasons: (i) severe perspective distortion – camera elevation angle closer to 90° results in a severely distorted body shape with large head and shoulder, and most of the other body parts being occluded; (ii) the low resolution makes it difficult to retrieve human body and texture details; (iii) motion blur caused by rapid variations of the elevation and pan angles or the movement of the platform; and (iv) camera vibration caused by the engine or the rotors of the UAV.

We introduce a dataset recorded from a low altitude and slow flying mobile platform for gesture recognition. The dataset was created with the intention of capturing full human body details from a relatively low altitude in a way that preserves the maximum detail of the body position. Our dataset is suitable for research involving search and rescue, situational awareness, surveillance, and general action recognition. We assume that in most practical missions, the UAV operator or an autonomous UAV follows these general rules: (i) it does not fly so low that it poses danger to the civilians, ground-based structures, or itself; (ii) it does not fly so high or so fast that it loses too much detail in the images it captures; (iii) it hovers to capture the details of an interesting scene; and (iv) it records human subjects from a viewpoint that causes minimum perspective distortion and maximum body details. Our dataset was created following these guidelines to represent 13 command gesture classes. The gestures were selected from general aircraft handling and helicopter handling signals [32]. All the videos were recorded at high-definition (HD) resolution, enabling the gesture videos to be used in general gesture recognition and gesture-based autonomous system control research. To our knowledge, this is the first dataset presenting gestures captured from a moving aerial camera in an outdoor setting.

2 Related work

A complete list and description of recently published action recognition datasets is available in [15,6], and gesture recognition datasets can be found in [25,21]. Here, we discuss some selected studies related to our work.

Detecting human action from an aerial view is more challenging than from a fronto-parallel view. Created by Oh et al. [18], the large-scale VIRAT dataset contains about 550 videos, recorded from static and moving cameras covering 23 event types over 29 hours. The VIRAT ground dataset has been recorded from stationary aerial cameras (e.g., overhead mounted surveillance cameras) at
multiple locations with resolutions of 1080×1920 and 720×1280. Both aerial and ground-based datasets have been recorded in uncontrolled and cluttered backgrounds. However, in the VIRAT aerial dataset, the low resolution of 480×720 precludes retrieval of rich activity information from relatively small human subjects.

A 4K-resolution video dataset called Okutama-Action was introduced in [1] for concurrent action detection by multiple subjects. The videos have been recorded in a relatively clutter-free baseball field using 2 UAVs. There are 12 actions under abrupt camera movements, altitudes from 10 to 45 meters and different view angles. The camera elevation angle of 90 degrees causes a severe distortion in perspective and self-occlusions in videos.

Other notable aerial action datasets are UCF aerial action [30], UCF-ARG [31] and Mini-drone [2]. UCF aerial action and UCF ARG have been recorded using an R/C-controlled blimp and a helium balloon respectively. Both datasets contain similar action classes. However, UCF aerial action is a single-view dataset while UCF ARG is a multi-view dataset recorded from aerial, rooftop and ground cameras. The Mini-drone dataset has been developed as a surveillance dataset to evaluate different aspects and definitions of privacy. This dataset was recorded in a car park using a drone flying at a low altitude and the actions are categorized as normal, suspicious and illicit behaviors.

Gesture recognition has been studied extensively in recent years [21,25]. However, the gesture-based UAV control studies available in the literature are mostly limited to indoor environments or static gestures [16,10,19], restricting their applicability to real-world scenarios. The datasets used for these works were mostly recorded indoors using RGB-D images [13,24,27] or RGB images [17,5]. An aircraft handling signal dataset similar to ours in terms of gesture classes is available in [28]. It has been created using VICON cameras and a stereo camera with a static indoor background. However, these gesture datasets cannot be used in aerial gesture studies. We selected some gesture classes from [28] when creating our dataset.

3 Preparing the dataset

This section discusses the collection process of the dataset, the types of gestures recorded in the dataset, and the usefulness of the dataset for vision-related research purposes.

3.1 Data collection

The data was collected on an unsettled road located in the middle of a wheat field from a rotorcraft UAV (3DR Solo) in slow and low-altitude flight. For video recording, we used a GoPro Hero 4 Black camera with an anti-fish eye replacement lens (5.4mm, 10MP, IR CUT) and a 3-axis Solo gimbal. We provide the videos with HD (1920×1080) formats at 25 fps. The gestures were recorded on two separate days. The participants were asked to perform the gestures in a
selected section of the road. A total of 13 gestures have been recorded while the 
UAV was hovering in front of the subject. In these videos, the subject is roughly 
in the middle of the frame and performs each gesture five to ten times.

When recording the gestures, sometimes the UAV drifts from its initial hov-
ering position due to wind gusts. This adds random camera motion to the videos 
making them closer to practical scenarios.

3.2 Gesture selection

The gestures were selected from general aircraft handling signals and helicopter 
handling signals available in the Aircraft Signals NATOPS manual [32, Ch. 2–3]. 
The selected 13 gestures are shown in Fig. 1. When selecting the gestures, we 
avoided aircraft and helicopter specific gestures. The gestures were selected to 
meet the following criteria: (i) they should be easily identifiable from a moving 
platform, (ii) the gestures need to be crisp enough to be differentiated from each 
another, (iii) they need to be simple enough to be repeated by an untrained in-
dividual, (iv) the gestures should be applicable to basic UAV navigation control, 
and (v) the selected gestures should be a mixture of static and dynamic gestures 
to enable other possible applications such as taking “selfies”.

3.3 Variations in data

The actors that participated in this dataset are not professionals in aircraft 
handling signals. They were shown how to do a particular gesture by another 
person who was standing in front of them, and then asked to do the same towards 
the UAV. Therefore, each actor performed the gestures slightly differently. There 
are rich variations in the recorded gestures in terms of the phase, orientation, 
camera movement and the body shape of the actors. In some videos, the skin 
color of the actor is close to the background color. These variations create a 
challenging dataset for gesture recognition, and also makes it more representative 
of real-world situations.

The dataset was recorded on two separate days and involved a total of eight 
participants. Two participants performed the same gestures on both days. For a 
particular gesture performed by a participant in the two settings, the two videos 
have significant differences in the background, clothing, camera to subject dis-
tance and natural variations in hand movements. Due to these visual variations 
in the dataset, we consider the total number of actors to be 10.

3.4 Dataset annotations

We used an extended version of online video annotation tool VATIC [33] to an-
notate the videos. Thirteen body joints are annotated in 37151 frames, namely 
ankles, knees, hip-joint, wrists, elbows, shoulders and head. Two annotated im-
gages are shown in Figure 2. Each annotation also comes with the gesture class, 
subject identity and bounding box. The bounding box is created by adding a 
margin to the minimum and maximum coordinates of joint annotations in both 
$x$ and $y$ directions.
Fig. 1. The selected thirteen gestures are shown with one selected image from each gesture. The arrows indicate the hand movement directions. The amber color markers roughly designate the start and end positions of the palm for one repetition. The Hover and Land gestures are static gestures.
Subject S5, Move to left
Subject S11, Wave off
Subject S5, Move to left
Subject S11, Wave off

Fig. 2. Examples of body joint annotations. Image on the left is from the Move to left class, whereas the image on the right is from the Wave off class.

3.5 Dataset summary

Fig. 3. The total clip length (blue) and the mean clip length (amber) are shown in the same graph in seconds. Note the former is one order of magnitude higher than the latter.

The dataset contains a total of 37151 frames distributed over 119, 25 fps, 1920×1080 video clips. All the frames are annotated with the gesture classes and body joints. There are 10 actors in the dataset, and they perform 5-10 repetitions of each gesture. Each gesture lasts about 12.5 sec on average. A summary of the dataset is given in Table 1. The total clip length (blue bars) and mean clip length (amber bars) for each class are shown in Figure 3.

In Table 2 we compare our dataset with eight recently published video datasets. These datasets have helped to progress research in action recognition, gesture recognition, event recognition and object tracking. The closest dataset in terms of the class types and the purpose is the NATOPS aircraft signals dataset that was created using 24 selected gestures.
Table 1. A summary of the dataset.

| Feature            | Value          |
|--------------------|----------------|
| # Gestures         | 13             |
| # Actors           | 10             |
| # Clips            | 119            |
| # Clips per class  | 7-11           |
| Repetitions per class | 5-10         |
| Mean clip length   | 12.5 sec       |
| Total duration     | 24.76 mins     |
| Min clip length    | 3.6 sec        |
| Max clip length    | 23.44 sec      |
| # Frames           | 37151          |
| Frame rate         | 25 fps         |
| Resolution         | 1920×1080      |
| Camera motion      | Yes, slight    |
| Annotation         | Bounding box, body joints |

Table 2. Comparison with recently published video datasets.

| Dataset            | Scenario          | Purpose              | Environment | Frames | Classes | Resolution         | Year  |
|--------------------|-------------------|----------------------|-------------|--------|---------|---------------------|-------|
| UT Interaction     | Surveillance      | Action recognition  | Outdoor     | 36k    | 6       | 360×240             | 2010  |
| NATOPS             | Aircraft signaling| Gesture recognition | Indoor      | N/A    | 24      | 320×240             | 2011  |
| VIRAT              | Drone, surveillance| Event recognition   | Outdoor     | Many   | 23      | Varying             | 2011  |
| UCF101             | YouTube           | Action recognition  | Varying     | 558k   | 24      | 320×240             | 2012  |
| J-HMDB             | Movies, YouTube   | Action recognition  | Varying     | 32k    | 21      | 320×240             | 2013  |
| Mini-drone         | Drone              | Privacy protection  | Outdoor     | 23.3   | 3       | 1920×1080           | 2015  |
| Campus             | Surveillance      | Object tracking     | Outdoor     | 11.2k  | 1       | 1414×2019           | 2016  |
| Okutama-Action     | Drone              | Action recognition  | Outdoor     | 70k    | 13      | 3840×2160           | 2017  |
| UAV-GESTURE        | Drone              | Gesture recognition | Outdoor     | 37.2k  | 13      | 1920×1080           | 2018  |
4 Experimental Results

![Diagram of P-CNN feature extraction](image)

**Fig. 4.** The P-CNN feature descriptor [9]: The steps shown in the diagram correspond to an example P-CNN computation for body part *left hand*.

We performed an experiment on the dataset using Pose-based Convolutional Neural Network (P-CNN) descriptors [9]. A P-CNN descriptor aggregates motion and appearance information along tracks of human body parts (right hand, left hand, upper body and full body). The P-CNN descriptor was originally introduced for action recognition. Since our dataset contains gestures with full-body poses, P-CNN is also a suitable method for full-body gesture recognition.

In P-CNN, the body-part patches of the input image are extracted using the human pose and corresponding body parts. For body joint estimation, we used the state-of-the-art OpenPose [4] pose estimator which is an extension of Convolutional Pose Machines [34]. Similar to the original P-CNN implementation, the optical flow for each consecutive pair of images was computed using Brox et al.’s method [3].

A diagram showing P-CNN feature extraction is given in Figure 4. For each body part and full image, the appearance (RGB) and optical flow patches are extracted and their CNN features are computed using two pre-trained networks. For appearance patches, the publicly available “VGG-f” network [7] is used, whereas for optical flow patches, the motion network from Gkioxari and Malik’s Action Tube implementation [12] is used. Static and dynamic features are separately aggregated over time to obtain a static video descriptor $v_{stat}$ and a dynamic video descriptor $v_{dyn}$ respectively. The static features are the (i) distances between body joints, (ii) orientations of the vectors connecting pairs of joints, and (iii) inner angles spanned by vectors connecting all triplets of joints. The dynamic features are computed from trajectories of body joints. We select the Min and Max aggregation schemes, because of their high accuracies over other schemes when used with P-CNN [9] on the JHMDB dataset [14] for action recognition. The Min and Max aggregation schemes compute the minimum and maximum values respectively for each descriptor dimension over all video frames.
frames. The static and dynamic video descriptors can be defined as

\[ v_{\text{stat}} = [m_1, \ldots, m_k, M_1, \ldots, M_k]^T, \]  
\[ v_{\text{dyn}} = [\Delta m_1, \ldots, \Delta m_k, \Delta M_1, \ldots, \Delta M_k]^T, \]

where, \( m \) and \( M \) correspond to the minimum and maximum values for each video descriptor dimension \( 1, \ldots, k \). \( \Delta \) represents temporal differences in the video descriptors. The aggregated features \( (v_{\text{stat}} \text{ and } v_{\text{dyn}}) \) are normalized and concatenated over the number of body parts to obtain appearance features \( v_{\text{app}} \) and flow features \( v_{\text{of}} \). The final P-CNN descriptor is obtained by concatenating \( v_{\text{app}} \) and \( v_{\text{of}} \).

The evaluation metric selected for the experiment is accuracy. Accuracy is calculated using the scores returned by the action classifiers. There are three training and testing splits for UAV-GESTURE dataset. In Table 3, the mean accuracy is compared with the evaluation results reported in [9] for the JHMDB [14] and MPII Cooking [23] datasets. For the JHMDB and MPII Cooking datasets, the poses are estimated using the pose estimator described in [8]. However, we use OpenPose [4] for UAV-GESTURE, because OpenPose has been used as the body joint detector in notable pose-based action recognition studies [11,35,20], and has reportedly the best performance [4].

| Dataset         | Remarks                     | Accuracy(%) |
|-----------------|-----------------------------|-------------|
| JHMDB           | Res: 320 × 240, Pose estimation: [8] | 74.2        |
| MPII Cooking    | Res: 1624 × 1224, Pose estimation: [8] | 62.3        |
| UAV-GESTURE     | Res: 1920 × 1080, Pose estimation: OpenPose [4] | 91.9        |

5 Conclusion

We presented a gesture dataset recorded by a hovering UAV. The dataset contains 119 HD videos lasting a total of 24.78 minutes. The dataset was prepared using 13 selected gestures from the set of general aircraft handling and helicopter handling signals. The gestures were recorded from 10 participants in an outdoor setting. The rich variation of body size, camera motion, and phase, makes our dataset challenging for gesture recognition. The dataset is annotated for human body joints and action classes to extend its applicability to a wider research community. We evaluated this new dataset using P-CNN descriptors and reported an overall baseline action recognition accuracy of 91.9 %. This dataset is useful for research involving gesture-based unmanned aerial vehicle or unmanned ground vehicle control, situation awareness, general gesture recognition, and general action recognition. The UAV-GESTURE dataset is available at https://github.com/asankagp/UAV-GESTURE.
Acknowledgement

This project was partly supported by Project Tyche, the Trusted Autonomy Initiative of the Defence Science and Technology Group (grant number myIP6780).

References

1. Barekatain, M., Mart, M., Shih, H.F., Murray, S., Nakayama, K., Matsuo, Y., Prendinger, H.: Okutama-action: An aerial view video dataset for concurrent human action detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2153–2160 (July 2017). https://doi.org/10.1109/CVPRW.2017.267
2. Bonetto, M., Korshunov, P., Ramponi, G., Ebrahimi, T.: Privacy in mini-drone based video surveillance. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). vol. 04, pp. 1–6 (May 2015). https://doi.org/10.1109/FG.2015.7285023
3. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J. (eds.) Computer Vision - ECCV 2004. pp. 25–36. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
5. Carol Neidle, A.T., Sclaroff, S.: 5th workshop on the representation and processing of sign languages: Interactions between corpus and lexicon (May 2012)
6. Chaquet, J.M., Carmona, E.J., Fernandez-Caballero, A.: A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding 117(6), 633 – 659 (2013). https://doi.org/10.1016/j.cviu.2013.01.013, http://www.sciencedirect.com/science/article/pii/S1077314213000295
7. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. CoRR abs/1405.3531 (2014), http://arxiv.org/abs/1405.3531
8. Cherian, A., Mairal, J., Alahari, K., Schmid, C.: Mixing body-part sequences for human pose estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
9. Cheron, G., Laptev, I., Schmid, C.: P-cnn: Pose-based cnn features for action recognition. In: The IEEE International Conference on Computer Vision (ICCV) (December 2015)
10. Costante, G., Bellocchio, E., Valigi, P., Ricci, E.: Personalizing vision-based gestural interfaces for hri with uavs: a transfer learning approach. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3319–3326 (Sept 2014). https://doi.org/10.1109/IROS.2014.6943024
11. Girdhar, R., Ramanan, D.: Attentional pooling for action recognition. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 34–45. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/6609-attentional-pooling-for-action-recognition.pdf
12. Gkioxari, G., Malik, J.: Finding action tubes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
13. Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H.J.: The chalearn gesture dataset (cgd 2011). Machine Vision and Applications 25(8), 1929–1951 (Nov 2014)
14. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: 2013 IEEE International Conference on Computer Vision. pp. 3192–3199 (Dec 2013). https://doi.org/10.1109/ICCV.2013.396
15. Kang, S., Wildes, R.P.: Review of action recognition and detection methods. CoRR abs/1610.06906 (2016). http://arxiv.org/abs/1610.06906
16. Lee, J., Tan, H., Crandall, D., Sabanović, S.: Forecasting hand gestures for human-drone interaction. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. pp. 167–168. HRI ’18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3173386.3176967 http://doi.acm.org/10.1145/3173386.3176967
17. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 444–451 (Sept 2009). https://doi.org/10.1109/ICCV.2009.5459184
18. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J.K., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., Desai, M.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR 2011. pp. 3153–3160 (June 2011). https://doi.org/10.1109/CVPR.2011.5995586
19. Pföhl, K., Koh, S.L., LaViola, J.: Exploring 3d gesture metaphors for interaction with unmanned aerial vehicles. In: Proceedings of the 2013 International Conference on Intelligent User Interfaces. pp. 257–266. IUI ’13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2449396.2449429 http://doi.acm.org/10.1145/2449396.2449429
20. Piergiovanni, A.J., Ryoo, M.S.: Fine-grained activity recognition in baseball videos. CoRR abs/1804.03247 (2018). http://arxiv.org/abs/1804.03247
21. Pisharady, P.K., Saerbeck, M.: Recent methods and databases in vision-based hand gesture recognition: A review. Computer Vision and Image Understanding 141, 152 – 165 (2015). https://doi.org/10.1016/j.cviu.2015.08.004 http://www.sciencedirect.com/science/article/pii/S1077314215001794
22. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 549–565. Springer International Publishing, Cham (2016)
23. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1194–1201 (June 2012). https://doi.org/10.1109/CVPR.2012.6247801
24. Ruffieux, S., Lalanne, D., Mugellini, E.: Chairgest: A challenge for multimodal mid-air gesture recognition for close hci. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. pp. 483–488. ICMI ’13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2522848.2532590 http://doi.acm.org/10.1145/2522848.2532590
25. Ruffieux, S., Lalanne, D., Mugellini, E., Abou Khaled, O.: A survey of datasets for human gesture recognition. In: Kurosu, M. (ed.) Human-Computer Interaction. Advanced Interaction Modalities and Techniques. pp. 337–348. Springer International Publishing, Cham (2014)
26. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1593–1600 (Sept 2009). https://doi.org/10.1109/ICCV.2009.5459361

27. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

28. Song, Y., Demirdjian, D., Davis, R.: Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In: Face and Gesture 2011. pp. 500–506 (March 2011). https://doi.org/10.1109/FG.2011.5771448

29. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. Tech. rep., UCF Center for Research in Computer Vision (2012)

30. University of Central Florida: UCF aerial action dataset. [http://crcv.ucf.edu/data/UCF_Aerial_Action.php](http://crcv.ucf.edu/data/UCF_Aerial_Action.php) (November 2011)

31. University of Central Florida: UCF-ARG Data Set. [http://crcv.ucf.edu/data/UCF-ARG.php](http://crcv.ucf.edu/data/UCF-ARG.php) (November 2011)

32. U.S. Navy: Aircraft signals natops manual, navair 00-80t-113 (1997), [http://www.navybmr.com/study%20material/NAVAIR_113.pdf](http://www.navybmr.com/study%20material/NAVAIR_113.pdf)

33. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision 101(1), 184–204 (Jan 2013). [https://doi.org/10.1007/s11263-012-0564-1](https://doi.org/10.1007/s11263-012-0564-1)

34. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

35. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. CoRR abs/1801.07455 (2018), [http://arxiv.org/abs/1801.07455](http://arxiv.org/abs/1801.07455)