An Online Approach for Gesture Recognition Toward Real-World Applications

Zhaoxuan Fan[1], Tianwei Lin[1], Xu Zhao[1](✉), Wanli Jiang[2], Tao Xu[2], and Ming Yang[1]

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai, China
zhaoxu@sjtu.edu.cn

[2] Connected and Automated Driving Lab, BMW China Services Ltd., Shanghai, China

Abstract. Action recognition is an important research area in computer vision. Recently, the application of deep learning greatly promotes the development of action recognition. Many networks have achieved excellent performances on popular datasets. But there is still a gap between researches and real-world applications. In this paper, we propose an integrated approach for real-time online gesture recognition, trying to bring deep learning based action recognition methods into real-world applications. Our integrated approach mainly consists of three parts. (1) A gesture recognition network simplified from two-stream CNNs is trained on optical flow images to recognize gestures. (2) To adapt to complicated and changeable real-world environments, target detection and tracking are applied to get a stable target bounding box to eliminate environment disturbances. (3) Improved optical flow is introduced to remove global camera motion and get a better description of human motions, which improves gesture recognition performance significantly. The integrated approach is tested on real-world datasets and achieves satisfying recognition performance, while guaranteeing a real-time processing speed.

Keywords: Gesture recognition · Action recognition · Online Real-time

1 Introduction

Human action recognition is a very energetic area in computer vision field, with a great potential in practical applications. In recent years, deep learning has shown its superior capability over conventional methods in many image based tasks, in terms of both performance and speed. Same trend also can be seen in video based problems [1–4]. Specifically, recently some deep learning based methods [5–9] are proposed to solve action recognition problem.

Among these works, two-stream CNNs [10] is a typical and successful method, which achieves excellent performance on mainstream datasets. Stacked optical

© Springer International Publishing AG 2017
Y. Zhao et al. (Eds.): ICIG 2017, Part I, LNCS 10666, pp. 262–272, 2017.
https://doi.org/10.1007/978-3-319-71607-7_23

flow images are used to represent temporal information and the still frame images are used to represent spatial information of action sequences. Based on this idea, two CNNs are trained separately on temporal and spatial data. The scores of these two CNN streams are combined by late fusion. Because of the salient performance of two-stream CNNs, some related improved method are proposed in different ways. Feichtenhofer *et al.* [11] propose to use CNNs to fuse outputs of two CNN streams. Zhang *et al.* [12] significantly improve the recognition speed by replacing optical flow with enhanced motion vector. Singh *et al.* [13] add two additional CNN streams in target bounding box and combine the multi-stream CNNs with LSTM network. These state-of-the-art deep learning methods have achieved excellent performances on public datasets like UCF101 and HMDB51. However, most of the current methods are offline and concentrate on limited public datasets, so they are not completely applicable for real-world applications.

In this paper, we propose an integrated real-time online gesture recognition approach toward real-world applications. Our initial motivation is to recognize traffic police gestures from a moving vehicle. Gesture can be considered as a subset of action. For gesture recognition method, we adopt the concepts of temporal and spatial information of actions from two-stream CNNs [10], and further simplify the network to fit our requirements. In real applications, major difficulty comes from the uncertain environments, which bring about many disturbances from the background and other objects in scene. So target localization is a natural choice to reduce these disturbances. In our approach, target detection and tracking are utilized to allow us focusing on the bounding area of the target. In this way, the disturbances from external background are maximally eliminated. Another problem raised in many scenarios is the global change in optical flow caused by camera motion, which is disadvantageous for gesture recognition. We introduce the improved optical flow method from iDT [14] as assistance to fix this problem.

In sum, there are three main contributions in this paper. Firstly, we propose a framework for real-world applications of gesture recognition. Target localization and optical flow improvement are introduced to solve real-world difficulties efficiently. Secondly, the integrated approach is designed and streamlined to perform real-time online recognition. In our test, it costs about 55 ms–85 ms per recognition for videos of different resolutions, which is qualified for real-time recognition with commonly used frame sampling method. Thirdly, we collect a new action dataset for traffic police gestures, which will be released with this paper later.

2 Approach

Our integrated approach mainly consists of three parts, namely, gesture recognition network, target localization and optical flow improvement, as shown in Fig. 1. The following sections will discuss about these three parts respectively.

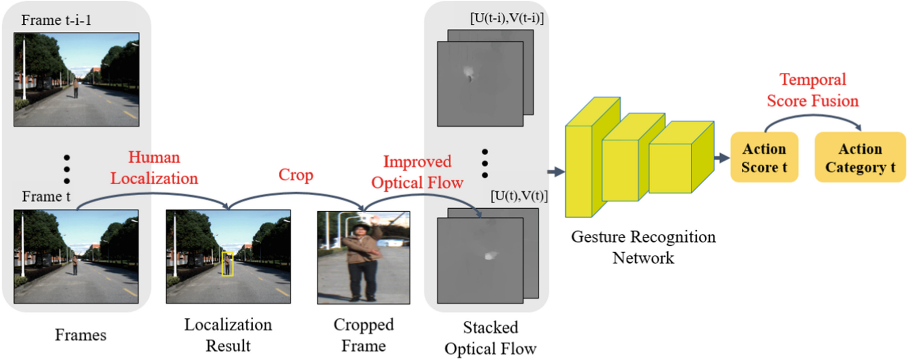


Fig. 1. Overall framework of the integrated approach. For a raw frame, target localization is performed to get target area. In target area, improved optical flows are computed as the features for recognition. Optical flow images are then stacked to feed into gesture recognition network. The network's output is post-processed by temporal fusion to give a final gesture label.

2.1 Gesture Recognition Network

Two-stream CNNs method fuses spatial and temporal information simultaneously to recognize actions and achieves state-of-the-art performance on popular public datasets. Its architecture contains two separate recognition networks, which perform spatial and temporal recognition separately and are then combined by certain fusion method. The spatial stream uses still frames, while the temporal stream uses optical flow image sequence to recognize actions [10].

Although the combination of spatial and temporal information achieves good performance, in our work, we decide to only use the temporal stream to recognize gestures in our approach, motivated by the following two reasons. **(1)** The temporal network alone is able to give a relatively satisfying recognition performance. The reduction of spatial network and late fusion network causes a small decrease on accuracy but saves a lot of time. **(2)** In popular datasets like UCF101, the videos are collected from all kinds of sources, and the difference of appearance features between different videos are significant, so spatial stream works well in this situation. But in real-world environments and especially in our target scenarios, the difference between still frames of different gestures are not significant and the spatial stream usually performs not as good as expected.

So the final recognition network only takes optical flow stream for gesture recognition. The deep CNN network in our approach is implemented with a VGG-16 net. It takes stacked optical flow images of $L = 10$ frames as a $2L$ dimension input (x, y directions). For one frame F_t in a video, the optical flow images $[U_i, V_i]$ of its previous L frames, $[F_{t-L+1}, F_{t-L+2}, ..., F_t]$, are stacked to represent the temporal feature of this frame, as $[U_{t-L+1}, V_{t-L+1}, U_{t-L+2}, V_{t-L+2}, ..., U_t, V_t]$. The stacked optical flow images are then fed into the network for recognition. To further improve recognition performance, temporal pooling is used to post-process the output scores. The gesture label of current frame is determined by the average score of last K frames.

2.2 Target Localization

In real-world applications, the environment is usually very complicated and noisy, making it difficult to recognize gestures from the whole image. Target localization is a good solution to this problem. In our approach, target detection and tracking methods are used to make sure only focusing on the target. In this way, recognition is performed only in the bounding box of the target. Target location is beneficial to gesture recognition in two ways. **(1)** Target location makes it possible to focus on the target, therefore maximally eliminates the impacts from environment. This guarantees the applicability of recognition algorithm in various environments. **(2)** Focusing on target area makes the motions of body parts more significant, therefore is helpful to improve recognition performance.

For now, our methods are based on single target scenarios. A Single Shot MultiBox Detector (SSD) [15] network is applied to detect target in scene. SSD is a superior state-of-the-art detection network and is very efficient in terms of speed and accuracy, therefore is very suitable for our task. Then KCF [16] algorithm is used to track the detected target. KCF is a tracking algorithm with extremely high speed and satisfying performance. In application, when tracking moving targets for a long time, micro errors may gradually accumulate, causing inaccurate tracking results. To solve this, SSD detector is set to re-detect target and re-initialize KCF tracker with refreshed target location at a certain time interval. The choice of SSD and KCF reduces time cost in target detection and tracking procedure at a large extent, while giving a satisfactory tracking performance.

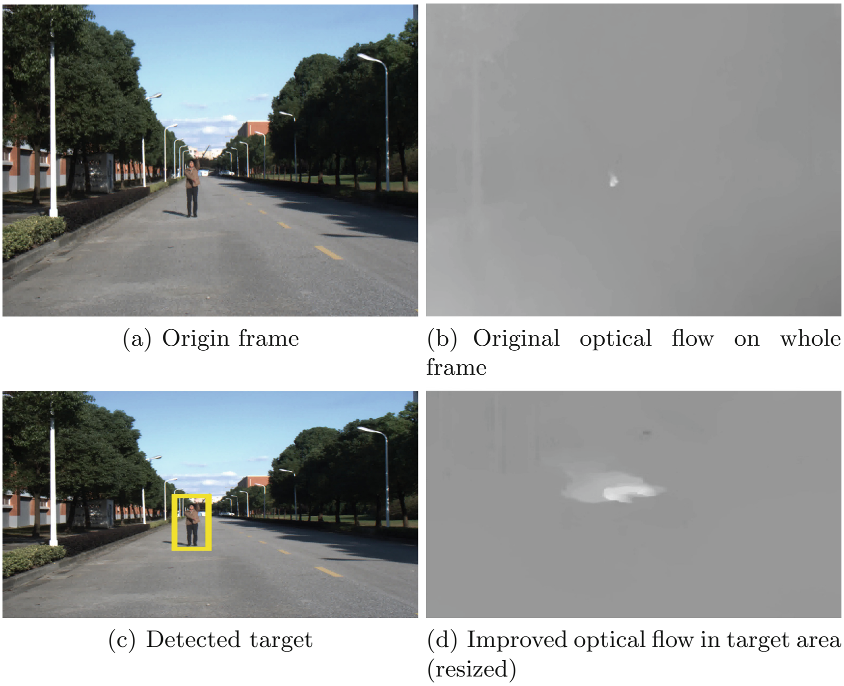
2.3 Optical Flow Improvement

In many real scenarios, the camera is not fixed, which causes global changes in optical flow inevitably. We introduce the optical flow improvement method from iDT [14] to eliminate the influences from global camera motion in order to get a better description of actions from optical flow images. By doing so, the algorithm adapts well to moving camera scenarios.

Usually, the global motion between two adjacent frames are small, and the relation between two frames can be assumed as a homography. To estimate the homography, SURF features are extracted to find the correspondence between these two frames, after which the homography is estimated robustly using RANSAC algorithm [14].

The SURF matches obtained on the whole image include matches from human motions, which actually are disturbances for estimating global camera motion. So we use the bounding box obtained from target localization to remove the matches from human region. This reduces disturbances from moving targets and guarantees the correctiveness of homography computation, which gives an accurate description about global motion maximally.

With the homography matrix, the two frames are rectified to eliminate camera motion before calculating optical flow. Figure 2 shows the comparison of optical flow images. We can see that the application of target localization and optical flow improvement significantly enhance the final optical flow. The optical flow of target motion is subtle in the original whole image, and interfered by optical flows caused by global motion, e.g. the optical flows of the lamp and trees. After target localization and optical flow improvement, the optical flow images distinctly describes the target motion in target area, while other irrelevant optical flows are eliminated.


(a) Origin frame

(b) Original optical flow on whole frame

(c) Detected target

(d) Improved optical flow in target area (resized)

Fig. 2. Comparison of optical flow images. In the original optical flow image, the optical flow of target is subtle, and there are false optical flows caused by camera motion, such as the optical flow of the lamp and trees on the left. After target localization and optical flow improvement, the optical flow of target is obviously enhanced.

2.4 Integration

The methods presented above are integrated to make an applicable real-time online gesture recognition algorithm. The overall procedure for online gesture recognition is described in Fig. 1. For a new frame in a video to be recognized, the following procedure is applied: (1) Localize target in the frame. (2) Compute improved optical flow in target bounding box. (3) Feed stacked optical flow images into gesture recognition network. (4) Temporal fuse the outputs over previous K results and to get the final recognition result.

3 Experiments

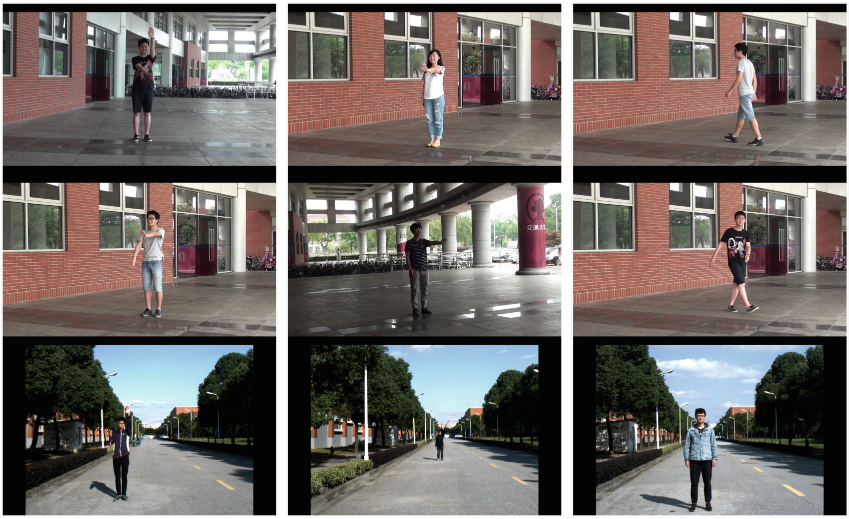
In this section, the components involved in our approach are evaluated separately in several aspects. The experiments are implemented based on Caffe [17] and MatConvNet [18].

3.1 Datasets

The experiments are conducted on 3 datasets, TPG (Traffic Police Gesture), UCF101 [19] and NATOPS (Naval Air Training and Operating Procedures Standardization) [20].

TPG. TPG is a dataset collected by us for our task of recognizing Chinese traffic police gestures. For now, it contains two traffic police gestures, *stop* and *pullover*. Other random gestures are also collected as an *other* gesture class, which is mainly used as negative samples. The dataset consists of videos collected from still cameras and moving cameras. For convenience, in our experiments the gestures are further divided into 3 subclasses, HandUp, HandDown and WaveHorizon. Including "Other" class of random gestures, the dataset consists of over 2 h videos of 4 classes. These videos are split into more than 2000 clips, with approximately 500 clips per class. Videos in TPG have a high resolution with 1440 × 1080. Figure 3 shows some example frames in TPG dataset.

UCF101. UCF101 is a popular dataset for action recognition research. It contains 13320 realistic videos of 101 classes, which are mostly of 341 × 256



(a) Pullover (b) Stop (c) Other

Fig. 3. Example frames from TPG dataset

resolution. UCF101 is a universal action dataset and not specifically collected for gesture recognition. So we select 24 classes from its original 101 classes, making it more appropriate for gesture recognition. The selected UCF24 mainly contains subjects with a clear body or upper body in video.

NATOPS. NATOPS is a dataset of US naval aircraft handling signals. It contains 24 gestures, with 400 samples for each gesture. The videos are recorded with a stereo camera, which produces 320 × 240 resolution images. The dataset contains RGB images, depth maps and mask images.

3.2 Evaluation on Online Method

Our online method is evaluated on TPG and UCF101 dataset. The experiment is conducted without target localization and optical flow improvement. It simply evaluates online method's performance. For TPG, training and test sets are split as 5:1, in terms of subject. The recognition network is trained on extracted optical flow data. For UCF101, training and test sets are split according to its official split 1 and the model provided by [11] is used in this experiment.

For online test, the recognition is performed at every incoming frame in a test video, and recognition result of every frame is counted for final accuracy calculation. Different averaging schemes are tested. For offline test, as the baseline, each test video gets a single recognition result and this result is counted for accuracy calculation. In offline test, 25 random frames are sampled uniformly in a test video and recognized, and the scores of these 25 frames are averaged to give a single label for the whole test video.

The experiment results are shown in Table 1. As can be seen, the online method causes a decrease in accuracy, but is still acceptable. The reason for the decrease is that, offline test samples 25 frames evenly in the video and gets a comprehensive description about the whole action. While for the online method, only a limited clip of action information is available at every time instant for recognition. Another results is that, for online test, the average schemes could improve accuracy but the improvement is not salient. So, to speed up response time for online recognition, $K = 5$ score averaging scheme is adopted in our approach.

Table 1. Accuracy results of online method experiment

Method	TPG	UCF101
Offline method	0.9238	0.8625 [11]
Online method (K = 1)	0.8124	0.7000
Online method (K = 5)	0.8137	0.7197
Online method (K = 10)	0.8151	0.7386

3.3 Evaluation on Target Localization and Optical Flow Improvement

In this part, the effectiveness of target localization (TL) and optical flow improvement (OFI) in gesture recognition is evaluated. The experiments are performed on TPG and UCF24. Training and test sets of UCF24 are still split according to its official split 1.

For each dataset, two CNNs are trained and compared. One network, as the baseline, is trained and tested on the original optical flow images of the raw videos. The other network is trained on optical flow images obtained with target localization and optical flow improvement. The test is performed offline, same as the offline test method in Sect. 3.2.

The experiment results are shown in Table 2. From the comparison, we can see that target localization and optical flow improvement effectively increases accuracy on TPG dataset, but causes a small decrease on UCF24. The main reason for the decrease on UCF24 is that most of the actions in UCF24 are not exactly gestures but interactions with objects, such as balls and instruments. In this case, the application of target localization only concentrates on humans but leaves out the objects in the interactions, which causes the information deficiency. But for TPG dataset, the target localization brings significant performance improvement, because target localization effectively extracts target from environment and enhances optical flow from target motion. The result indicates that target localization is an effective method to enhance recognition performance in gesture area, especially in real-world scenarios.

Table 2. Accuracies of target localization (TL) and optical flow improvement (OFI) experiment

Method	TPG	UCF24
Offline test on original video	0.9238	0.9824
Offline test with TL and OFI	0.9887	0.9573

3.4 Evaluation on Overall Integrated Approach

The integrated online gesture recognition approach, which combines online recognition pipeline, target localization and optical flow improvement, is evaluated on three datasets, TPG, UCF24 and NATOPS, respectively. The online and offline tests are executed in the same way as introduced in Sect. 3.2.

For TPG and UCF24, our approach is compared with the temporal stream (TS) of two-stream CNNs [10], which is an offline method. For NATOPS dataset, our approach is compared with the extended online LDCRF method proposed in [21]. The LDCRF method utilizes all three kinds of data in NATOPS (RGB image, depth image and mask image) to recognize gestures, while only the RGB images of NATOPS are used in our approach.

Table 3 shows the comparison results, we can see that our integrated approach achieves satisfactory performance compared with baselines. The performance decrease in UCF24 is reasonable as we have analyzed in the previous sections. The reason for the relatively poor performance on NATOPS is that the gestures in this dataset contains many subtle hand movements and many pairs of gestures only differ in minor ways like thumb up or down. In this case, the optical flow feature used in our approach is not micromesh enough to capture these indistinctive differences. So our approach is better at gestures with big body motions, like traffic police gestures.

Table 4 shows our speed performance on three datasets. What is important is that our approach guarantees a real-time implementation of the algorithm, which is critical for online recognition in real applications. The average time cost per frame of our approach are about 55 ms–85 ms for low resolution videos (UCF24 and NATOPS) and high resolution videos (TPG) respectively. The processing speed is qualified for real-time recognition with frame sampling, which is a commonly used method to eliminate information redundancy in gesture sequences. The result shows that our integrated approach works well for real-time online gesture recognition task.

Table 3. Accuracy comparison on three datasets

Method	TPG	UCF24	NATOPS
TS of two-stream CNNs [10]	0.9238	0.9824	-
Extended online LDCRF [21]	-	-	0.7537
Our integrated approach	0.9498	0.8927	0.7277

Table 4. Speed performance of our approach on different datasets

Dataset	TPG	UCF24	NATOPS
Video resolution	$1440\times,1080$	$341,\times,256\,(\mathit{most})$	$320,\times,240$
Time cost (ms/frame)	85	67	55

4 Conclusion

In this paper, we propose an integrated online gesture recognition approach. The approach assembles online gesture recognition network, target localization and optical flow improvement. It is applicable in real-world environment and moving camera situation. Evaluations on several datasets show that the integrated approach gives a satisfying recognition accuracy while guaranteeing real-time processing speed.

Acknowledgement. This research has been supported by the funding from NSFC (61673269, 61375019, 61273285).

References

1. Karpathy, A., Toderici, G., Shetty, S., et al.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
2. Xu, Z., Yang, Y., Hauptmann, A.G.: A discriminative CNN video representation for event detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1798–1807 (2015)
3. Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Learning to track for spatio-temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3164–3172 (2015)
4. Park, E., Han, X., Berg, T.L., et al.: Combining multiple sources of knowledge in deep cnns for action recognition. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8. IEEE (2016)
5. Ji, S., Xu, W., Yang, M., et al.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
6. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with R*CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1080–1088 (2015)
7. Ijjina, E.P., Mohan, C.K.: Human action recognition based on motion capture information using fuzzy convolution neural networks. In: 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), pp. 1–6. IEEE (2015)
8. Sun, L., Jia, K., Yeung, D.Y., et al.: Human action recognition using factorized spatio-temporal convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4597–4605 (2015)
9. Song, J., Shen, H.: Beyond frame-level CNN: saliency-aware 3D CNN with LSTM for video action recognition. IEEE Signal Process. Lett. **24**(4), 510–514 (2016)
10. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp. 568–576 (2014)
11. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016)
12. Zhang, B., Wang, L., Wang, Z., et al.: Real-time action recognition with enhanced motion vector CNNs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2718–2726 (2016)
13. Singh, B., Marks, T.K., Jones, M., et al.: 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, pp. 1961–1970 (2016)
14. Wang, H., Schmid, C.: Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
16. Henriques, J.F., Caseiro, R., Martins, P., et al.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2015)
17. Jia, Y., Shelhamer, E., Donahue, J., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)

18. Vedaldi, A., Lenc, K.: Matconvnet: convolutional neural networks for MATLAB. In: Proceedings of the 23rd ACM international conference on Multimedia, pp. 689–692. ACM (2015)
19. Soomro, K., Zamir, A.R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
20. Song, Y., Demirdjian, D., Davis, R.: Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database. In: 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011), pp. 500–506. IEEE (2011)
21. Song, Y., Demirdjian, D., Davis, R.: Continuous body and hand gesture recognition for natural human-computer interaction. ACM Trans. Interact. Intell. Syst. (TiiS) **2**(1), 5 (2012)