Hand-hygiene activity recognition in egocentric video

Chengzhang Zhong, Amy R. Reibman, Hansel Mina Cordoba and Amanda J. Deering
School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA
Department of Food Science, Purdue University, West Lafayette, Indiana, USA

***Abstract*—Food safety is affected by the conditions and practices during different manufacturing steps to prevent contamination and food-borne illnesses. In this paper, we focus on detecting hand-hygiene actions in Egocentric videos. We create a two-stage system to localize and recognize all the hand-hygiene actions in each untrimmed video. In the first stage, we apply a low-cost hand mask and motion histogram features to localize the temporal regions of hand-hygiene actions. In the second stage, we use the two-stream network model combined with a search algorithm to recognize all types of hand-hygiene actions that happen in the untrimmed video. The system achieves a detection accuracy close to 80% on our dataset with 100 participants.**

***Index Terms*—egocentric video, activity recognition, deep learning, temporal segmentation**

I. Introduction

Food safety focuses on the implementation of conditions and practices to prevent contamination and food-borne illnesses at different stages of food production. In recent years the burden of outbreaks of food-borne illnesses has increased. Evidence suggest that one of the main causes of contamination on the food production chain is inappropriate food handling practices by workers and consumers [1]. For food handlers employed in this industry, the application of good manufacturing practices (GMPs) is critical. For instance, hand washing is one of the most efficient steps to prevent contaminating food with human pathogens. According to World Health Organization, there are 12 steps for effective hand-hygiene with soap and water [2]. These techniques include but are not limited to: rinse hands, apply soap, rub hands with different poses, and dry hands.

Egocentric video is recorded by mounting wearable cameras on human body. This video type contains rich body and camera motion. In recent years, the study of egocentric video on personal living activities has become popular. Pirsiavash *et al*. [3] investigated egocentric videos of daily living. These videos are recorded with wearable cameras and represent a person's daily activities, such as eating and working inside home. The video content usually contains hand/object interactions, and the video is usually varied in length and includes multiple objects.

In this paper, we focus on classification of hand-hygiene actions from egocentric videos. Compared to the existing egocentric videos in the daily living categories [3], egocentric recording of hand-hygiene actions provides richer details of the subtle motions of hands [2] and lack hand/object interactions. This is because the standard procedure of hand washing requires people not to touch any objects other than soap or towel.

Due to the absence of publicly available data for hand-hygiene, we introduce our new hand hygiene-egocentric dataset, which consists of 100 polled participants. We design a two-stage system to localize the temporal regions of hand-hygiene actions and recognize them in untrimmed hand-hygiene videos. In the first stage, we extract a low-cost hand mask and motion histogram feature, and process the entire video to localize temporal regions which contain potential hand-hygiene actions. In the second stage, we use the temporal regions detected from the first stage as input. In this stage, we apply a two-stream network model combined with our searching algorithm to recognize all hand-hygiene actions that happen in the input untrimmed video.

In Section 2, we introduce previous work on activity recognition, deep learning, and egocentric videos. In Section 3, we introduce our new hand-hygiene egocentric dataset. In Section 4, we describe our rationale for the two-stage system design. In Section 5, we apply two-stream network model for trimmed hand-hygiene video classification, and in Section 6, we explain the implementation details of the two-stage system.

II. Related Work

Activity recognition is an important research area in computer vision. Traditionally, researchers search through the spatial and temporal dimensions to build representative features of the video and then apply machine learning algorithms for recognition. Features such as STIP [4], HOG [5], HOF and MBH [6] have proved their effectiveness in recognizing human activities viewed from the third-person. Wang *et al*. [7] propose the Improved Dense Trajectory (IDT) algorithm that combines HOG, HOF, MBH features and achieves the state-of-art in many third-person datasets.

In recent years, deep learning methods are widely used to solve activity recognition problems. Deep learning structures like AlexNet [8], VGG Net [9] and ResNet [10] take single images as inputs for image classification. Other structures such as LSTM [11], two-stream network [12] and C3D [13] consider both spatial and temporal information from video frames for activity recognition.

In addition, many researchers focus on recognizing activities in egocentric videos. Singh *et al*. [14] consider hand poses and

This work was supported by Purdue University's Colleges of Agriculture and Engineering Collaborative Projects Program 2018

978-1-7281-1817-8/19/$31.00

optical flow information as important features. By constructing a EgoConvnet with a few layers which takes a stack of hand masks, head motion and saliency maps as inputs and further fuses with two stream network, they achieve promising detection accuracy. Ma *et al*. [15] propose to use appearance and motion information for egocentric activity recognition. They believe that the object and hand interaction area provide crucial appearance information, and they construct networks to locate this area.

However, all of the egocentric video research focuses on testing on the same category of daily living videos [3] [16] [17]. The majority of the content in these videos involves participant's hands interacting with various objects. We believe none of these researches discuss egocentric video that focuses on hand motion only, as we do.

III. HAND-HYGIENE EGOCENTRIC DATASET

A. Data recording procedures

Many publicly available egocentric datasets involve only a few participants and the recording environments are usually inside each participant's home apartment, especially for daily activity videos. For our hand-hygiene actions, the cooking tools or food in a home kitchen environment depend on the specific kitchen and most of them should not appear in a standard industrial food handling factory. Moreover, every participant has his/her own style of hand-hygiene actions. Involving only few participants recording hand-hygiene actions multiple times could easily result in too much the similarity in video content.

To ensure our dataset includes enough variations of hand-hygiene action samples, we invited 100 participants of various ages and races to record their hand-hygiene actions. All participants are allowed to wear watches and jewelries on their wrists. Each person is recorded twice, once in each of two adjacent public restrooms. Both of these rooms have similar environments.

Each participant is asked to wear a GoPro camera with a harness on his/her chest for recording. We record two videos from each participant. In the beginning, a participant performs a naive hand washing, following his/her typical hand washing habit in the first room. After finishing the first hand washing, we ask the participant to read instructions for hand-hygiene [2]. When he/she finishes reading, the participant records another hand washing in the second room. All videos are recorded under 1080p resolution with 30 FPS and wide viewing angle. To increase processing speed, we further down-sampled these videos to 480 270 resolution.

B. Hand-hygiene action definition

We define salient hand-hygiene action classes which can reflect hand washing quality. A subject should not touch the faucet with their hands, to avoid re-contamination [2]. Therefore, we need to distinguish whether the subject touched the faucet with hand or with an elbow. Moreover, it is also important to detect the strength used to rub hands. We enable this by labelling an action of rinse hand, where the subject rubs hands with little strength. Furthermore, the subject should apply soap before hand washing and dry their hands after hand washing. When soap is applied, the subject needs to rub hands without keeping their hands in water. Based on these principles, we define and label 8 actions as indicated in Figure 1, which are: touch faucet with elbow, touch faucet with hand, rinse hands, rub hands without water, rub hands with water, apply soap, dry hands, and a background non-hygiene action. All 8 actions are manually labelled at the frame-level.

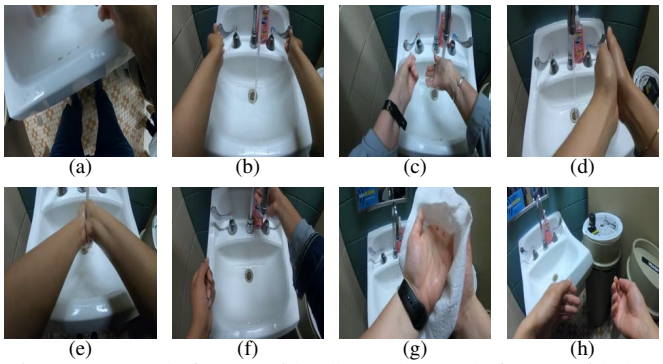

Fig. 1: (a) touch faucet with elbow (b) touch faucet with hand (c) rinse hands (d) rub hands without water (e) rub hands with water (f) apply soap (g) dry hands with paper towel (h) non-hand hygiene action

IV. TWO STAGE HAND-HYGIENE SYSTEM

A. System design background

Activity recognition for untrimmed video clips is often termed temporal action proposals or temporal action localization [18]. For hand-hygiene videos, our goal is to localize temporal regions where hand-hygiene actions happen in untrimmed videos. Then, by applying an action classifier on these targeted short segments, we will be able to identify what hand-hygiene actions have been performed by a participant.

Our hand-hygiene videos contain densely-distributed hand actions with an average of 5 different types of actions per video. Non-hygiene actions such as standing or walking around can happen anytime during the video. Thus, it is difficult for coarse-level temporal proposal methods [19] [20] to localize hand-hygiene actions in our videos. Moreover, the average duration of an untrimmed hand-hygiene video is around 1 minute. Therefore, the temporal segmentation method [21] designed for long duration egocentric videos is also not applicable here.

B. System basic description

We propose a two-stage system to localize and detect hand-hygiene actions from untrimmed videos as shown in Figure 2. In the first stage of our system, we localize the temporal interval where hand-hygiene actions happen inside the untrimmed video. Hand-hygiene actions are dominated by hand and arm motion, which can be interpreted as the appearance of hands, arms and their related motion patterns. We divide our 8 types of actions into two categories. First, actions containing strong motions, including rinse hands, rub hands without water, rub hands with water and wipe hands, are considered as action class "1". The other four types of actions,

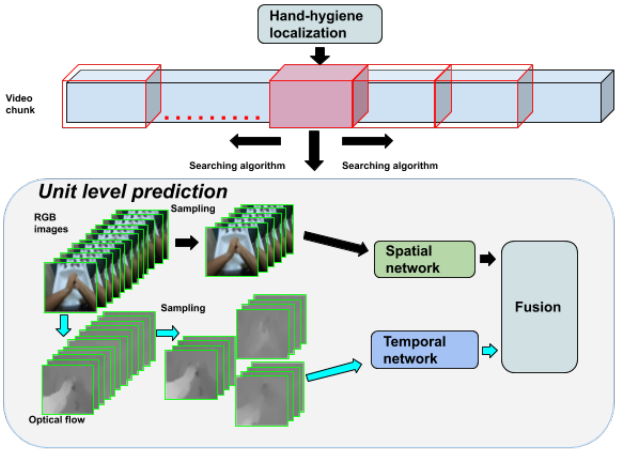

Fig. 2: Two-stage prediction system pipeline

including non-hygiene actions, are labelled as action class "0". We apply a low cost hand mask and motion histogram features to process the input video. And the goal of this first stage is to correctly predict these labels at a frame-level inside the whole untrimmed video. The implementation details are explained in Section VI-A and VI-B.

In the second stage, we use a two-stream network to make a fine-level prediction on all 8 action classes in our data. Using the location information from the first stage, we apply a deep learning model to predict on unit of 30 frames. By combing this model with a certain searching algorithm, we are able to localize and identify all the hand-hygiene actions that happen inside the input video. The construction of deep learning model is introduced in Section V. The implementation detail of the system's second stage and overall performance are explained in Sections VI-C and VI-D.

V. Hand-hygiene action classification

Based on our system design, we need to construct a robust model which is capable to recognize all hand-hygiene actions. In this section, we explore the performance of the two-stream network on recognizing actions in trimmed hand-hygiene video clips.

A. Two-stream convolutional neural network

Hand-hygiene actions are composed of hand and arm motions, which lack meaningful objects that might reveal clues about action itself [15]. In this Section, we would like to apply a deep learning based model to learn deep feature representations to distinguish all 8 types of actions.

The two-stream network has demonstrated its effectiveness in activity recognition in third-person videos [12]. The two-stream network considers both appearance and motion information by separately constructing a spatial-stream ConvNet and a temporal-stream ConvNet. The spatial-stream ConvNet takes RGB images as inputs, which provides appearance information in the scene. On the other hand, the temporal-stream ConvNet takes chunks of optical flow images between consecutive frames as inputs. These optical flow images provide strong clues to the motion information that exists in the video. The final prediction result is generated from a score fusion of these two individual networks.

For our experiment, we use the method of Wang *et al.* [22] with implementation [23], which applies deeper network structures and takes advantage of a small learning rate and more data augmentation techniques.

B. Experiments on two-stream network

We split the 200 videos in our dataset into training and testing sets with 135 videos and 65 videos respectively. All videos are trimmed into clips where each clip includes only one action from beginning to end, which result in 1380 training video clips and 675 testing video clips.

For training, we use the pre-trained ResNet 152 [10] from ImageNet [24] for both the spatial and temporal networks with fine-tuning on the 8 action classes. Input video with 480 270 are down-sampled to resolution 224 224 to fit the ResNet.

For testing, we apply both the sparse [22] and dense sampling strategies. For the sparse sampling, only 25 frames with equal distance step are selected from each input video clip. For dense sampling, all frames are selected. The two-stream network model predicts each selected frame individually and uses the average prediction score from these frames as the prediction for the input video.

Model	Accuracy
Spatial Network sparse	85.3%
Spatial Network dense	86.4%
Temporal Network sparse	84.4%
Temporal Network dense	86.8%
Fusion sparse	87.3%
Fusion dense	87.7%

TABLE I: Two-stream network performance

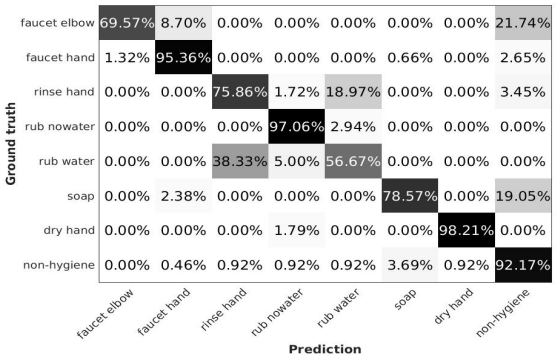

Fig. 3: Confusion matrix for two-stream network fusion, dense sampling.

The results in Table I show the average detection accuracy among all 675 video clips. The dense sampling only outperforms the sparse sampling by 0.4 % after score fusion. Therefore, sparse sampling is a better strategy for its faster processing speed and minor sacrifice on detection accuracy.

A prediction confusion matrix for dense sampling after score fusion is shown in Figure 3. We observe that the trained deep model performs well on several of the actions with over 90% accuracy. However, for the action pair of rinse hands and rub hands with water, many participants switch between these

two actions in a short period of time, which caused difficulty in creating ground truth labels. Therefore, the trained model makes mistakes on recognizing these two actions.

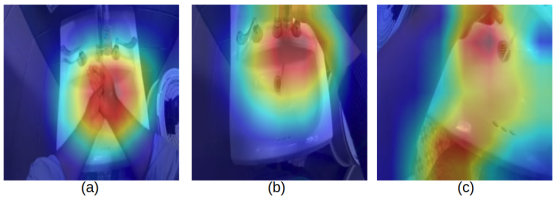

Fig. 4: Grad cam [25] results of (a) rub hands with water (b) apply soap (c) touch faucet with elbow

To understand what the two-stream model has learned, we use Grad-cam [25]. Figure 4 shows these heat maps, where the highlighted region indicates saliency for a target class. In Figure 4 (a), the trained model successfully captures hand related regions to recognize rub hand with water. In Figure 4 (b)(c), however, the chest camera angle hasn't completely captured the entire action of applying soap or touching faucet with elbow. As a result, the trained model makes mistakes by recognizing these two actions as non-hygiene actions.

VI. System implementation details

A. Hand-hygiene localization

Hand mask Hand poses are good indicators of hand-hygiene actions. Especially when rinsing or rubbing hands, the two hands overlapping each other create distinguishable patterns. In our work, we applied the pixel-level hand detection method [26] to generate hand masks. We train the model with a set of 134 images of manually labelled segmented hand regions under different illumination conditions. The resulting hand masks are gray-scale images with size S_m S_n.

Inspired by the work of Singh *et al.* [14], we create a network structure to predict frame-level "0" or "1" action using hand mask as features. The network is composed of 2 Conv layers followed by RELU, max pooling and LRN (local response normalization) and 2 fully connected (fc) layers. The network takes L hand masks as input. In our training stage, a cross entropy loss is applied as well as a dropout 0.5 to avoid over-fitting. In testing, the softmax score from the last fc layer indicates the prediction result.

Motion histogram Motion is also a good indicator of hand-hygiene actions. We create a optical flow histogram feature within the hand mask region to represent motion patterns. Applying the hand masks generated on dense optical flow images, we create two optical flow histograms with bin size B for both region inside and outside hand mask. Within each region, we count the magnitude and angle of optical flow for each pixel i.

$$M_i = \sqrt{gx_i^2 + gy_i^2}, \theta_i = tan^{-1}(\frac{gy_i}{gx_i}) \quad (1)$$

The pixel with θ_i angle that falls into the range of $[\frac{b-1}{B}\pi, \frac{b}{B}\pi)$ contributes to the bin b with magnitude M_i, where 1 b B. To overcome the problem of hand mask size variation, the final sum value for each bin b is normalized by dividing the total number of pixels in its corresponding region. The result histograms for hand masked region and outside hand mask region at frame t are represented as $H_{ht} = [h_{ht,1}, h_{ht,2}, ..., h_{ht,B}]$ and $H_{bt} = [h_{bt,1}, h_{bt,2}, ..., h_{bt,B}]$. The concatenation of these two histograms creates a motion representation at frame t. We also compute the ratio $R_t = \frac{\sum_{i=1}^{B} h_{ht,i}}{\sum_{j=1}^{B} h_{bt,i}}$ and hand motion sum $S_t = \sum_{i=1}^{B} h_{ht,i}$ as two additional features. The final representation of motion histogram at frame t is $H_t = [H_{ht}, H_{bt}, R_t, S_t]$ with size 1 $2B + 2$.

For classification, we apply a Random Forest classifier with 30 estimators and max depth 40 to learn the motion histogram patterns.

B. Hand-hygiene localization testing

In this section, we test the performance of the hand mask and motion histogram feature on localizing hand-hygiene actions from untrimmed videos.

Training For the efficiency of system design, we split the untrimmed 65 videos, with resolution 480 270, from 100 people's testing dataset into 43 and 22 videos for training and testing the hand localization system. To increase the processing speed of the hand-hygiene localization, the hand masks are generated with size of 32 18 and 64 36 in this experiment. Motion histogram features are generated on 480 270 dense optical flow images and applied previous generated hand masks,which resized to 480 270, on it. The hand mask network is trained under batch size 128 and learning rate $1e^{-5}$ with a stack of $L = 5$ hand masks. The Random Forest classifier is trained with three bin size options: 9,12 and 16.

Testing The testing experiment is done on 22 untrimmed videos with the label "0","1" as positive and negative labels on every frame. The hand mask network slides through the whole video and predicts using an overlapped stack of hand masks. The Random Forest classifier predicts on every frame of each video. For each testing video, we count the TP (true positive),TN (true negative),FP (false positive) and FN (false negative) at the frame-level. The performance of each classifier is measured by the accuracy = $\frac{tp+tn}{tp+tn+fp+fn}$ and true negative ratio = $\frac{tn}{tn+fp}$.

Model	Accuracy	True negative ratio
9 bins motion hist	73.7%	74.0%
12 bins motion hist	74.7%	71.6%
16 bins motion hist	75.0%	70.4%
32x18x5 hand mask network	78.9%	74.1%
64x36x5 hand mask network	80.7%	76.8%

TABLE II: Classifier comparison

Table II indicates the average accuracy and true negative ratio among 22 testing videos. We notice that hand mask network outperforms the combination of motion histogram with Random Forest, and the input stack with hand-mask size 64 36 5 is the best option. However, since the hand mask feature only reveals appearance information, mistakes can be made when a participant holds his/her hands in a overlapped manner without motion.

In the final design of hand-hygiene localization, we first apply the hand mask network to predict frame-level negative

label "0" and positive label "1". Then we re-check the positive predicted frames with motion histogram and Random Forest classifier. A frame is predicted as positive label "1" only when it is confirmed by both classifiers. Otherwise, a frame is marked as negative label "0". The detailed performance of this structure will be explained in the next section.

C. Hand-hygiene search and detection

In this Section, we describe the second stage of our two stage hand-hygiene system. As it has been shown that two-stream network has a reasonable performance on recognizing trimmed hygiene videos, we would like to use this model as a unit level detector to further process untrimmed hand-hygiene videos.

Location unitization We consider an untrimmed video composed by non-overlapped units. Each unit has 30 frames, which is 1 second in under 30 FPS. We start by assigning each unit with a unified label of "0" or "1". Based the frame-level prediction from the first system stage, if a unit contains more than 15 frames of positive label "1", the unit is marked with "1". Otherwise, it will be marked as "0".

Unit level prediction The unit with positive label "1" indicates those actions with strong hand motion. We start to check these locations first. To recognize all 8 action classes, we employ the pre-trained two stream network in Section V-A with a sparse sampling strategy. We sample 10 RGB images and 3 non-overlapped 10 pairs of optical flow images for each frame unit. The spatial network and temporal network individually predict using their sampled inputs and fuse the results with equal weights for the final prediction.

Searching algorithm There exist 7 types of hand-hygiene actions to recognize. However, due to short duration and indistinctive motion patterns, actions of applying soap, touching faucet with hands and touching faucet with elbow are categorized into class "0" in the localization step. These actions normally happen before or after the actions labeled in "1". Therefore, we designed a searching algorithm to find all 7 types of hand-hygiene actions. The algorithm iteratively searches the surrounding unit of each label "1" unit and makes predictions using the two-stream network model. The algorithm stops when it reaches non-hygiene actions on both left and right-side unit. After finishing the searching algorithm, each unit visited has been predicted with a result label and the unvisited units are automatically considered as non-hygiene actions.

D. System testing

Testing of the overall two-stage system is applied on the same 22 untrimmed videos in Section VI-B. To evaluate a video's prediction accuracy, we compare the prediction result with our frame-level ground truth labels. We map unit-level prediction result into a frame-level result by replicating each unit's result by 30 times.

The system performance is evaluated by frame-level accuracy = $\frac{tp+tn}{tp+tn+fp+fn}$ and the percentage of units visited (PV). We introduce the PV to measure the system efficiency. A high frame-level accuracy with a low PV value indicates the system was effective at localizing hand-hygiene actions and avoiding non-hygiene regions. For comparison, we create a baseline by applying two-stream network model to densely predict all non-overlapping units in each untrimmed video.

Methods	Accuracy	PV
Baseline	79.3%	100.0%
H+S	79.3%	81.5%
H+M+S	78.6%	76.4%

TABLE III: Average performance on 22 untrimmed videos. H: **H**and mask network localization, M: **M**otion histogram localization, S: **S**earching algorithm with two-stream network recognition

As indicated in Table III, the baseline system that checked every unit in each video obtains an average accuracy of 79.3%, which is lower than the performance on Section V-B due to the strict frame-level comparison. When applying the hand mask network only on the first system stage of localization, the PV drops from 100% to 81.5% while maintaining the same accuracy as the baseline system. This proves that the hand-hygiene localization stage helps to avoid processing the non-hygiene action. By applying the hand mask network with motion histogram, the PV further drops to 76.4% while sacrificing 0.7% detection accuracy. It is worth to note that the average percentage hand-hygiene actions occupied in the 22 untrimmed videos is 71.3%, which is the upper bound for the PV value.

VII. CONCLUSION

In this paper, we introduced our new hand hygiene-egocentric dataset. The dataset contains video samples from 100 participants, which were recorded in two public restrooms. We manually labeled all videos with 8 action categories. Moreover, we designed a two stage system to localize and recognize hand-hygiene actions in untrimmed hand-hygiene video. The system consists of two stages. In the first stage, our system takes of the hand mask and motion histogram feature to localize hand-hygiene actions temporally. In the second stage, we expanded the two-stream network model to combine with a searching algorithm to recognize all the hand-hygiene actions in the video. The system has achieved an acceptable performance. In the future, we plan to explore the effect of using multiple camera views to recognize hand-hygiene actions.

REFERENCES

[1] L. McIntyre, L. Vallaster, L. Wilcott, S. B. Henderson, and T. Kosatsky, "Evaluation of food safety knowledge, attitudes and self-reported hand washing practices in foodsafe trained and untrained food handlers in British Columbia, Canada," *Food Control*, vol. 30, no. 1, pp. 150–156, 2013.

[2] F. G. P. S. Challenge, "WHO guidelines on hand hygiene in health care: a summary," *World Health Organization, Geneva, Switzerland*, 2009.

[3] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2847–2854.

[4] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 1, 2005, pp. 886–893.

[6] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*. Springer, 2006, pp. 428–441.

[7] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *In ICLR*, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[14] S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2620–2628.

[15] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1894–1903.

[16] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3281–3288.

[17] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *IEEE Computer Society Conference On Computer Vision and Pattern Recognition Workshops, CVPR*, 2009, pp. 17–24.

[18] B. Ghanem, J. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, V. Escorcia, R. Krishna, S. B, and C. D. Dao, "The activitynet large-scale activity recognition challenge 2018 summary," *CoRR*, vol. abs/1808.03766, 2018. [Online]. Available: http://arxiv.org/abs/1808.03766

[19] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Q. Chen, "Temporal context network for activity localization in videos," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5727–5736.

[20] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," *IEEE International Conference on Computer Vision (ICCV), Oct*, vol. 2, 2017.

[21] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2537–2544.

[22] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *Computing Research Repository*, vol. abs/1507.02159, 2015. [Online]. Available: http://arxiv.org/abs/1507.02159

[23] Y. Zhu, "Pytorch implementation of popular two-stream frameworks for video action recognition," https://github.com/bryanyzhu/two-stream-pytorch, 2017.

[24] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[26] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3570–3577.