



Convolutional Long Short-Term Memory Networks for Recognizing First Person Interactions

Swathikiran Sudhakaran^{1,2} and Oswald Lanz²

¹University of Trento, Trento, Italy

²Fondazione Bruno Kessler, Trento, Italy

{sudhakaran, lanz}@fbk.eu

Abstract

In this paper, we present a novel deep learning based approach for addressing the problem of interaction recognition from a first person perspective. The proposed approach uses a pair of convolutional neural networks, whose parameters are shared, for extracting frame level features from successive frames of the video. The frame level features are then aggregated using a convolutional long short-term memory. The hidden state of the convolutional long short-term memory, after all the input video frames are processed, is used for classification in to the respective categories. The two branches of the convolutional neural network perform feature encoding on a short time interval whereas the convolutional long short term memory encodes the changes on a longer temporal duration. In our network the spatio-temporal structure of the input is preserved till the very final processing stage. Experimental results show that our method outperforms the state of the art on most recent first person interactions datasets that involve complex ego-motion. In particular, on UTKinect-FirstPerson it competes with methods that use depth image and skeletal joints information along with RGB images, while it surpasses all previous methods that use only RGB images by more than 20% in recognition accuracy.

1. Introduction

A new and vast array of research in the field of human activity recognition has been brought about with the recent technological advancements in wearable cameras. Majority of the human activity recognition techniques till now were concentrated on videos captured from a third person view. Developing techniques for the automatic analysis of egocentric videos has immense application potential ranging from assistance during surgical procedure, law and order,

assisting elderly citizens, video summarization, etc. But most of the existing research deal with activities carried out by the camera wearer, not interactions and reactions of others to the observer (person wearing the camera). The interaction recognition problem is different and difficult compared to the action recognition problem. This is because egocentric video captures a huge variety of objects, activities, and situations, and sometimes the person interacting with the observer can move out of the field of view as the person approaches the observer. Presence of ego-motion adds more complexity to the problem as it can interfere with the analysis of motion taking place in the scene. The applications of a method capable of automatically understanding interactions from egocentric view include surveillance, social interaction for robots, human machine interaction, etc.

Regardless of the recent advancements in deep learning techniques for addressing problems such as object recognition, caption generation, action recognition, etc., it has never been applied in the problem of first person interaction recognition. All the existing methods use hand-crafted features or processing stages for recognizing interactions in first person videos. End-to-end deep learning techniques require no prior information regarding the data and it is capable of achieving high degree of generalization compared to hand-crafted features based approaches. As a result, we propose to follow a deep learning approach for addressing the problem under study. To the best of our knowledge, this is the first time a deep learning based method is proposed for solving first person interaction recognition problem.

Our contributions can be summarized as follows:

- We develop an end-to-end trainable deep neural network model for recognizing interactions from a first person perspective
- We propose a novel architecture that can utilize convolutional neural networks pre-trained on image datasets, for processing video data

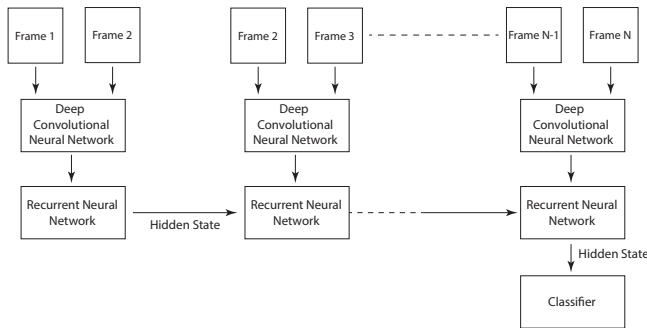


Figure 1. Basic block diagram showing the proposed method

- We experimentally evaluate the effectiveness of the proposed method on four publicly available first-person interaction recognition datasets

The remaining of the paper is organized as follows. A brief idea about the relevant state of the art techniques is provided in section 2. The proposed deep neural network architecture for addressing the concerned problem is discussed in section 3. Section 4 discusses the experimental results obtained as part of the performance evaluation and the document is concluded in section 5.

2. Related Works

An extensive amount of exploratory studies has been carried out in the area of first person activity or action recognition. They concentrate on the activity that the camera wearer is carrying out. These methods can be divided in to two major classes: activities involving object manipulation such as meal preparation [7, 20, 8, 19, 16, 3] and activities such as running, walking, etc. [30, 32, 18, 33, 27]. The former relies on information about objects present in the scene for classifying the activity while the latter concentrates on the ego-motion and the salient motion in the scene.

More recent studies have focussed on the analysis of interaction recognition in the egocentric vision context. One of the pioneering work in this area is carried out by Ryoo & Matthies in [29]. They extracted optical flow based features and sparse spatio-temporal features from the videos and a bag of words model as the video feature representation. Narayan *et al.* [22] use the TrajShape [37], MBH [4] and HOF [37] as the motion features followed by bag of words approach or Fisher vector encoding [26] for generating the feature descriptor. Wray *et al.* [40] propose graph-based semantic embedding for recognizing egocentric object interactions. Instead of focussing on the objects [5], Bambach *et al.* [2] investigate the use of strong region proposals and CNN classifier to locate and distinguish hands involved in an interaction.

Another line of research in this area was influenced by the development of Kinect device which is capable of capturing depth information from the scene together with the

skeletal joints of the humans present in the scene. Methods exploiting this additional modality of data have been proposed by Gori *et al.* [11], Xia *et al.* [43] and Gori *et al.* [10]. Gori *et al.* [11] use the histogram of 3D joints, histogram of direction vectors and depth images along with the visual features. Xia *et al.* propose to use the spatio-temporal features computed from the RGB and depth images [41] together with the skeletal joints information as the feature descriptor. Gori *et al.* [10] propose a feature descriptor called relation history image which extracts information from skeletal joints and depth images.

Several deep learning based approaches have been developed by researchers for action recognition from third person view. Simonyan & Zisserman [31] use raw frames and optical flow images as input to two CNNs for extracting the feature descriptor. Donahue *et al.* [6] and Srivastava *et al.* [34] use an architecture consisting of CNN followed by long short-term memory (LSTM) RNN for action recognition. A variant of the LSTM architecture, in which the fully-connected gates are replaced with convolutional gates (convLSTM), have been proposed by Shi *et al.* [44] for addressing the problem of precipitation nowcasting prediction from radar images. The convLSTM is found to be functioning with improved performance compared to the fully-connected LSTM. The convLSTM model has been later used for predicting optical flow images [25] and anomaly detection [21]. The results show that the convLSTM model suits applications involving spatio-temporal data such as videos. For this reason, we propose to use convLSTM as one of the important blocks in the proposed model for first person interaction recognition.

3. Proposed Method

Figure 1 illustrates the basic idea behind the proposed approach. During each iteration two successive frames from the video will be applied to the model, i.e., frames 1 and 2 in the first time step, frames 2 and 3 in the second time step, etc. until all the video frames are inputted in to the network. The hidden state of the RNN in the final time step is then used as the feature representation of the video and is fed to the classifier.

Simonyan & Zisserman [31] have previously proposed to use optical flow images along with RGB images, as the inputs to CNN, for performing action recognition which has resulted in improved performance. The optical flow images will provide information regarding the motion changes in the frames which will in turn help the network in learning discriminating features that can distinguish one action from another. Inspired from this, we also evaluate the performance of the network when the difference between adjacent frames are applied as input. The difference of frames can be considered as an approximate version of optical flow images. Sun *et al.* [35] and Wang *et al.* [39] have also ex-

explored the possibility of using difference images as inputs to a deep neural network for action recognition. By applying frame difference as the input, the network is forced to learn the changes taking place in the video, i.e., the motion changes which define the actions and interactions occurring inside the video.

3.1. Convolutional Long Short-Term Memory

Convolutional long short-term memory (convLSTM) is an extension of the popular long short-term memory (LSTM) RNN [12]. In this the fully-connected gates of the LSTM module are replaced by convolutional gates thereby making it capable of encoding spatio-temporal features. The equations governing the functioning of convLSTM are:

$$\begin{aligned}
 i_t &= \sigma(w_x^i * I_t + w_h^i * h_{t-1} + b^i) \\
 f_t &= \sigma(w_x^f * I_t + w_h^f * h_{t-1} + b^f) \\
 \tilde{c}_t &= \tanh(w_x^{\tilde{c}} * I_t + w_h^{\tilde{c}} * h_{t-1} + b^{\tilde{c}}) \\
 c_t &= \tilde{c}_t \odot i_t + c_{t-1} \odot f_t \\
 o_t &= \sigma(w_x^o * I_t + w_h^o * h_{t-1} + b^o) \\
 h_t &= o_t \odot \tanh(c_t)
 \end{aligned}$$

where ‘*’ and ‘ \odot ’ represent the convolution operation and the Hadamard product. In this all the gate activations and the hidden states are 3D tensors as opposed to the case with standard fully-connected LSTM which are vectors.

3.2. Network Architecture

The architecture of the entire model used in the proposed approach is given in figure 2. The model can be divided in to the following parts:

1. **Deep feature extractor:** A Siamese like architecture constituting of the AlexNet model [14] is used for extracting the frame level features. The fully-connected layers of the AlexNet model are removed. The two branches of the AlexNet model share their parameters and are pre-trained using the ImageNet dataset.
2. **Spatio-temporal feature aggregator:** This consists of a 3D convolution layer followed by the convLSTM explained in section 3.1. The output of the two branches of the previous module are concatenated together and is applied to the 3D convolution layer. The 3D convolutional layer uses a kernel of size $2 \times 3 \times 3$ and a stride of 1 in both the temporal and spatial dimensions. The output of the 3D convolution layer (256 feature planes) is then applied to the convLSTM layer for feature aggregation. The convLSTM layer uses 256 filters in all the gates with a filter size of 3×3 and stride 1. The hidden state, which contains the representation of the input video, consists of 256 feature planes.

3. **Classifier:** Once all the input frames are applied to the first two stages of the network, the video representation, which is the hidden state of the convLSTM, is applied to the classifier. The classifier consists of a spatial convolution layer followed by global average pooling. The convolution layer generates an output with feature planes corresponding to the number of classes. It uses 3×3 kernels with stride 1. Global average pooling is performed on each of the feature map generated by the convolutional layer and is applied to the softmax layer. This is inspired from the network in network (NIN) model [17] and the SqueezeNet model [13] proposed for image classification task.

The two branches of the AlexNet extract relevant frame level information consisting of the objects present in the frame. This will help the network in understanding a frame level context of the video. By using two branches of the AlexNet together with the 3D convolution layer, information in a short temporal duration is encoded. An advantage of following this approach evolves from the fact that in this way existing powerful CNN architectures that are pre-trained on large scale image datasets can be used, along with the capability to take in to consideration the temporal context. The convLSTM layer maintains a memory of all the previous information it has seen which will result in the encoding of information in a longer extend of time. This is very relevant in the case of ego-centric videos. This is because, unlike in the case of third person views observing the scene from a distance, the objects or humans present in the video can move out of the field of view once they approach closer to the observer for physical interaction, which necessitates the requirement for a memory. Also, the convolutional gates present in the convLSTM helps the model to analyze the local spatial and temporal information together as compared to the fully-connected LSTM. This has a significant importance in the analysis of videos because the network should be capable of understanding the spatial and temporal changes occurring in the video. By removing the fully-connected layers of the AlexNet and using the convLSTM, the network achieves this capability. More specifically, we maintain spatial structure in the propagated memory tensor and generated video descriptor. This is in contrast e.g. to LRCN [6] and the many CNN-LSTM architectures that fully convolve into a vector before temporal aggregation with standard LSTM, propagating a memory vector. This late spatio-temporal aggregation is distinguishing and more effective in representing spatial structure in videos. Furthermore, prior to classification we maintain the spatial structure till the very final layer of our architecture by applying spatial convolution to obtain class-specific activations that are directly associated to local receptive fields in the input. By following the classifier used in the proposed approach, a significant reduction in the number of param-

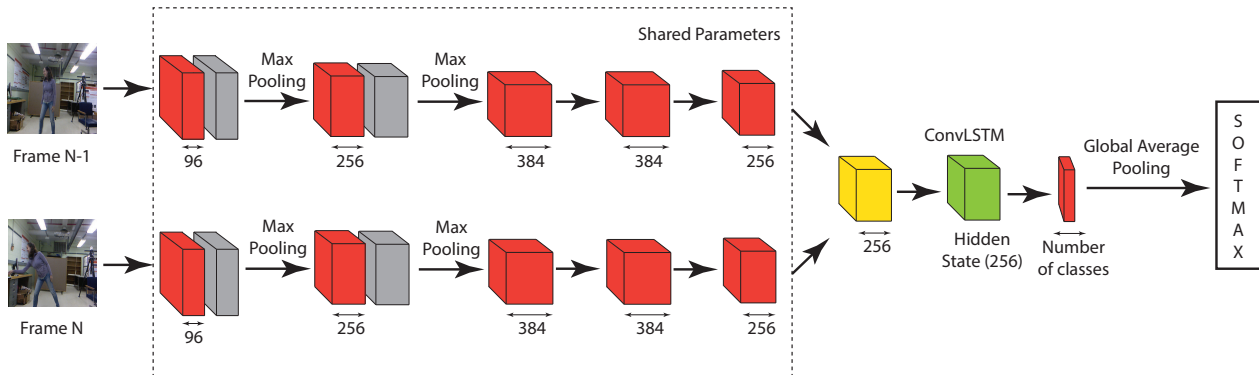


Figure 2. The architecture of the network. The convolutional layers are shown in red followed by normalization layer in gray. The 3D convolutional layer is shown in yellow and the convLSTM layer in green. We also experiment with a variant where, instead of raw frames we input difference-images obtained from pairs of successive frames.

ters is achieved thereby minimizing the propensity to overfit. On top of that, the global average pooling by itself can be considered as a regularization [17], thereby avoiding the usage of other regularization techniques.

3.3. Training

The implementation of the proposed method is done using the Torch library. We used RMSProp [36] optimization algorithm for training the network and a fixed learning rate of 10^{-5} with a batch size of 12. The network is run for 10k iterations. The weights of the 3D convolution, convLSTM and the final spatial convolution layer are initialized using Xavier method [9]. The whole network is trained together to minimize the cross-entropy loss.

Since it is a generally accepted consensus that deep neural networks tend to overfit in the absence of large amount of training data, we used several data augmentation techniques for minimizing the possibility of overfitting. In the proposed approach, all the video frames are first rescaled to a dimension of 256×340 . We follow the corner cropping and scale jittering techniques proposed by Wang *et al.* [39]. In this, during each training iteration, a portion of the frame (four corners or the centre) is cropped and is applied as input to the network. The dimension of the cropped region is selected randomly from the following set $\{256, 224, 192, 168\}$. The cropped region is then rescaled to 224×224 . The selected regions are then horizontally flipped randomly during each training iteration. It is to be noted that the same augmentation techniques are applied to all the frames coming from a video during a particular training iteration. All the video frames in the training set are normalized to make their mean zero and variance unity.

4. Experiments and Results

The performance of the proposed approach is evaluated by testing it on four publicly available standard datasets.

From all the videos, 20 frames equidistant in time are selected as the input to the network. The performance is evaluated in terms of recognition accuracy. During evaluation, the mean of the training frames are subtracted from the testing frames and is divided by the variance of the former. Image crops of dimension 224×224 are extracted from all four corners and the centre of each frame and are applied to the network. The horizontally flipped versions of the frames are also used during evaluation. Thus, for each video in the test set, 10 different samples (5 crops and their horizontal flips) are generated during inference. The outputs of all the crops and their flipped versions are then averaged together to find the class of the input video.

4.1. Datasets

The following four datasets are used for validating the proposed first person interaction recognition approach.

JPL First-Person Interaction Dataset (JPLFPID) [29] : The dataset consists of videos of humans interacting with a humanoid model with a camera attached on top of its head. The dataset consists of 7 different interactions which include friendly (hugging, shaking hands, etc.), hostile (punching and throwing) and neutral (pointing fingers) behaviors. There are 84 videos in the dataset and the evaluation method is done following existing approaches. During each evaluation instance, half of the videos are chosen as training and the remaining half as testing. Mean of the recognition accuracy obtained after each evaluation is reported as final accuracy.

NUS First Person Interaction Dataset (NUSFPID) [22] : This dataset, captured by placing the camera on the head of a human, consists of both actions (opening door, operating cell phone, etc.) and interactions (shaking hands, passing objects, etc.) with other humans from a first person view point. The dataset contains 152 videos. Random train-test split method is used for evaluation and the mean of the accuracies obtained after all the runs is used for reporting.

	JPLFPID	NUSFPID	UTKinect First Person Dataset	
			Humanoid	Non-humanoid
Ryoo & Matthies [29]	89.6	-	57.1	58.48
Ryoo [28]	87.1	-	-	-
Abebe <i>et al.</i> [1]	86	-	-	-
Ozkan <i>et al.</i> [24]	87.4	-	-	-
Narayan <i>et al.</i> [22]	96.7	61.8	61.9	57.6
Wang and Schmid [38]	-	58.9	-	-
HOF [29]	-	-	45.92	-
Laptev <i>et al.</i> [15]	-	-	48.46	50.83
LRCN [6] (raw frames)	59.5	68.9	72.6	71.4
LRCN [6] (difference of frames)	89.0	69.1	63.1	67.8
Proposed Method (raw frames)	70.6	69.4	79.6	78.4
Proposed Method (difference of frames)	91.0	70.0	66.7	69.1

Table 1. Comparison of the proposed method with existing techniques, on various datasets. Results are reported in terms of recognition accuracy in %. [22] on UTKinect are our reproduced results following the paper description. LRCN [6] results are produced from the authors’ code following same training/testing protocol used with our method. All other results are those available from the authors’ papers.

UTKinect First Person Dataset-humanoid (UTKFPD-H) [43] : In this dataset, a kinect sensor is mounted on a humanoid robot and the video of various humans interacting with it is captured. This dataset also contains friendly (hugging, shaking hands, etc.), hostile (punching, throwing objects, etc.) and neutral behavior (standing up, running, etc.) behaviors. There are 177 video samples from 9 different classes present in the dataset. The results reported are using the provided train-test split. For the experiments related to the proposed method, only the RGB videos are used and the depth information is not used.

UTKinect First Person Dataset-non-humanoid (UTKFPD-NH) [43] : This is similar to the above dataset with exception of the type of the robot used for capturing the video. A non-humanoid robot is used for capturing the videos in this dataset. The types of the interactions are also different from the previous dataset. There are 189 videos in the dataset and 9 different classes. Here also, the depth information available with the dataset is not used in the experiments.

All four datasets considered in this work is of significant difference from each other. For instance, the observer in each case is very much different so that the ego-motion that can occur to the video is different from one another. In the first dataset, the observer is stationary so that ego-motion can take place only if someone interacts with it physically. The NUSFPID and the UTKFPDH are more or less similar to each other but still the extend to which the interactions can affect their respective ego-motion is different, since the robot being more stable compared to a human. The type of interactions that are carried out in the datasets are also different. In this way, we did the experiments to evaluate the performance of the proposed method in identifying interactions with different levels of ego-motion and different variety of interactions.

4.2. Results and Discussions

The results obtained for each of the dataset is reported and compared with the state of the art techniques in table 1. As already mentioned, this is the first time a deep learning technique is proposed for recognizing first person interactions. Most of the existing deep learning based methods use gaze information or hand segmentations for identifying the objects being handled by the user which in turn is used for recognizing the action [16, 32]. Since first person interaction videos differ from such action recognition videos drastically, we are not considering these methods for comparing the performances. It is evident from the table that the proposed method is suitable for recognizing interactions from first person videos. Our method outperforms the state of the art methods except in one dataset (JPL first-person interaction dataset), where it came as the second best method. All the state of the art methods listed in table 1, except LRCN, use hand-crafted features as opposed to the proposed approach which is based on deep learning. As mentioned in the previous section, the videos in the JPL first-person interaction dataset contain limited ego-motion as the camera is placed on a static object. Therefore, the ego-motion present in the videos is strongly correlated with the action that is occurring (for example, vertical motion when shaking hands, a sudden horizontal jerking motion in the punching action, etc.). Thus the ego-motion present in videos in the JPL dataset is part of the motion pattern that defines the action. In contrast to this, the rest of the datasets contain videos with strong ego-motion that may also arise occasionally and independently from the performed action. This may explain our findings that on the JPL dataset, we obtain improved performance when the difference of frames are applied as the input to the network since the difference operation will force the network to model the motion changes between frames (including the ego-motion that depends on

the action class). On the other hand, the performance of the network on the UTKinect dataset is the highest when the raw frames are applied as input. The videos in the UTKinect dataset contains strong ego-motion. The presence of ego-motion that is uncorrelated with the action being performed will act as a sort of noise when the frame difference is taken as input: the network should be able to distinguish between the random ego-motion and the motion caused by the action, thereby making it more difficult for the network to isolate the salient motion patterns occurring in the video. The network performs in a comparable way on the NUSFPID dataset when either types of inputs are applied. Even though the NUSFPID dataset contains ego-motion, some of the classes include actions performed by the observer such as using cell phone, typing on the keyboard, writing, etc. For discriminating these types of actions that contain limited amount of motion, the network needs to learn the objects handled by the observer and hence the raw frames can be deemed as the most suited input modality. As opposed to this, interactions such as shaking hand, passing object, etc. contains motion and in this case, difference of frames performs the best.

We have also compared the proposed method with another deep learning approach proposed for action recognition, LRCN [6], in Table 1. As mentioned previously, LRCN uses an AlexNet model for extracting frame level features followed by an LSTM for temporal aggregation. The results show that the proposed method is more suitable for first-person interaction recognition compared to the LRCN model. The primary reason is the usage of convLSTM in the proposed model as opposed to the fully-connected LSTM used in the LRCN. In LRCN, the input of the LSTM is the output of the penultimate fully-connected layer (fc6). The local spatial information in the video frame is lost by using the fully-connected layer, where as the proposed model preserves the spatial information by forwarding the convolutional features of the input. The convolutional gates of the convLSTM layer then performs information aggregation across both spatial and temporal domain. This is very important in the case of first-person interaction recognition since it is required to memorize the changes occurring in both the spatial and temporal domains. Another feature of the proposed model that results in its improved performance is the utilization of the second AlexNet branch so that the model can process information from a broader temporal context. Since the parameters of both the branches are shared, there is no additional complexity added to the network and in addition to this, the same configuration can be used with any pre-trained CNN models. It should also be noted that, the proposed model achieves this improved performance with fewer number of parameters as compared to the LRCN model (21.8M ours vs 61.3M LRCN). This is achieved by eschewing from using fully-connected layers

	UTKFPD-H	UTKFPD-NH
Xia <i>et al.</i> [41]	72.83	53.25
Oreifej & Liu [23]	52.54	45.55
Xia <i>et al.</i> [43]	85.6	83.7
Xia <i>et al.</i> [42]	-	70
Gori <i>et al.</i> [10]	-	85.94
Proposed Method	79.6	78.4

Table 2. Recognition accuracy obtained for the UTKinect first person dataset when the raw frames are applied as input to the proposed method. All other methods listed in the table uses either the depth image or skeletal joints information or both along with the RGB images for feature descriptor generation

(both in the convLSTM and the classification layers). This will also enable the proposed model to be trained on smaller datasets without getting overfit.

Table 2 compares the performance of the proposed method on the UTKinect dataset with state of the art methods that use other modalities of information such as depth image and skeletal joints. It is clearly evident from the table that the proposed method performs comparably to these methods without using any other additional data. The proposed method is able to perform competitively to these methods by using only the RGB video as input.

5. Conclusions

The paper proposes a novel deep neural network architecture for recognizing interactions from first person videos. The proposed model extracts frame level features using a convolutional neural network followed by feature aggregation using a convolutional long short-term memory. Two different types of input modalities have been tested on the network. The first one uses adjacent frames as input and the second uses the difference between consecutive frames as input to the network. The proposed model is evaluated using four publicly available datasets. The network was able to outperform existing state of the art techniques. Raw frames is found as the suitable input to the network for videos with strong ego-motion, where as the difference of frames resulted in better performance for videos with less but action-correlated ego-motion. We also show that our proposed method performs competitively to the techniques that use depth image and skeletal joints information along with RGB images. As a future work, we will explore the possibility of training the network jointly on both raw frames and difference of frames, thereby enabling the network to adapt by itself depending on the input videos, regardless of the presence or absence of strong ego-motion.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPU used for this research.

References

- [1] G. Abebe, A. Cavallaro, and X. Parra. Robust multi-dimensional motion features for first-person vision activity recognition. *Computer Vision and Image Understanding*, 149:229–248, 2016. 5
- [2] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015. 2
- [3] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *CVPR Workshops*, 2014. 2
- [4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006. 2
- [5] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014. 2
- [6] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2, 3, 5, 6
- [7] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding ego-centric activities. In *ICCV*, 2011. 2
- [8] A. Fathi and J. M. Rehg. Modeling actions through state changes. In *CVPR*, 2013. 2
- [9] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, 2010. 4
- [10] I. Gori, J. Aggarwal, L. Matthies, and M. S. Ryoo. Multitype activity recognition in robot-centric scenarios. *IEEE Robotics and Automation Letters*, 1(1):593–600, 2016. 2, 6
- [11] I. Gori, J. Sinapov, P. Khante, P. Stone, and J. Aggarwal. Robot-centric activity recognition in the wild. In *International Conference on Social Robotics*, 2015. 2
- [12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3
- [13] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 3
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3
- [15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 5
- [16] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *CVPR*, 2015. 2, 5
- [17] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 3, 4
- [18] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016. 2
- [19] K. Matsuo, K. Yamada, S. Ueno, and S. Naito. An attention-based activity recognition for egocentric video. In *CVPR Workshops*, 2014. 2
- [20] T. McCandless and K. Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In *BMVC*, 2013. 2
- [21] J. R. Medel and A. Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016. 2
- [22] S. Narayan, M. S. Kankanhalli, and K. R. Ramakrishnan. Action and interaction recognition in first-person videos. In *CVPR Workshops*, 2014. 2, 4, 5
- [23] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 2013. 6
- [24] F. Ozkan, M. A. Arabaci, E. Surer, and A. Temizel. Boosted multiple kernel learning for first-person activity recognition. *arXiv preprint arXiv:1702.06799*, 2017. 5
- [25] V. Pătrăucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. In *ICLR Workshops*, 2016. 2
- [26] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 2
- [27] Y. Poley, A. Ephrat, S. Peleg, and C. Arora. Compact cnn for indexing egocentric videos. In *WACV*, 2016. 2
- [28] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011. 5
- [29] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 2, 4, 5
- [30] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *CVPR*, 2015. 2
- [31] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2
- [32] S. Singh, C. Arora, and C. Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016. 2, 5
- [33] S. Song, N.-M. Cheung, V. Chandrasekhar, B. Mandal, and J. Liri. Egocentric activity recognition with multimodal fisher vector. In *ICASSP*, 2016. 2
- [34] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 2
- [35] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *ICCV*, 2015. 2
- [36] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012. 4
- [37] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013. 2
- [38] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 5
- [39] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In *ECCV*. Springer, 2016. 2, 4

- [40] M. Wray, D. Moltisanti, W. Mayol-Cuevas, and D. Damen. Sembed: Semantic embedding of egocentric action videos. In *ECCV Workshops*, 2016. 2
- [41] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, 2013. 2, 6
- [42] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPR Workshops*, 2012. 6
- [43] L. Xia, I. Gori, J. K. Aggarwal, and M. S. Ryoo. Robot-centric activity recognition from first-person rgb-d videos. In *WACV*, 2015. 2, 5, 6
- [44] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 2