Multimodal Multi-stream Deep Learning for Egocentric Activity Recognition

Sibo Song[1], Vijay Chandrasekhar[2], Bappaditya Mandal[2], Liyuan Li[2], Joo-Hwee Lim[2], Giduthuri Sateesh Babu[2], Phyo Phyo San[2], and Ngai-Man Cheung[1]

[1]Singapore University of Technology and Design
[2]Institute for Infocomm Research

Abstract

In this paper, we propose a multimodal multi-stream deep learning framework to tackle the egocentric activity recognition problem, using both the video and sensor data. First, we experiment and extend a multi-stream Convolutional Neural Network to learn the spatial and temporal features from egocentric videos. Second, we propose a multi-stream Long Short-Term Memory architecture to learn the features from multiple sensor streams (accelerometer, gyroscope, etc.). Third, we propose to use a two-level fusion technique and experiment different pooling techniques to compute the prediction results. Experimental results using a multimodal egocentric dataset show that our proposed method can achieve very encouraging performance, despite the constraint that the scale of the existing egocentric datasets is still quite limited.

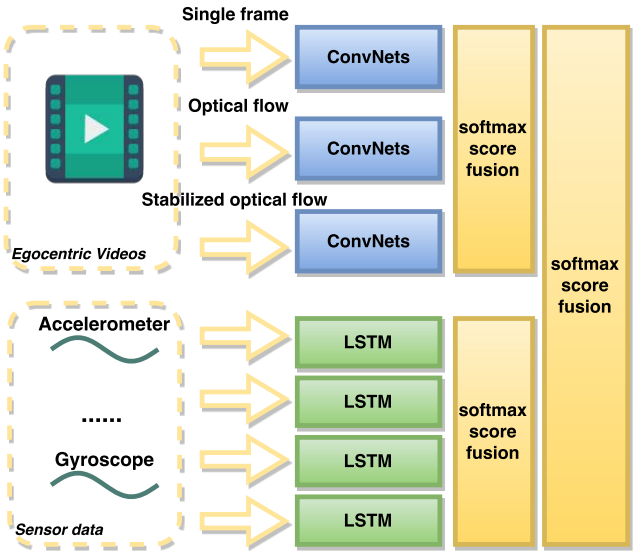

Figure 1: Architecture of our proposed Multimodal Multi-stream Deep Learning for Egocentric Activity Recognition.

1. Introduction

The last several years has witnessed a fast-growing market of wearable devices and there is an increasing interest in understanding egocentric actions. The ever-increasing adoption of those devices such as Google Glass, Microsoft SenseCam, Apple Watch and Mi band enables low-cost, unobtrusiveness collection of rich egocentric or first-person view activity data. This makes possible the monitoring of all-day and any-place activities. Automated analyzing and understanding of egocentric multimodal data (i.e., first person videos, wristband sensors, etc.) is very important for many applications ranging from military, security applications, health monitoring, lifestyle analysis, to stimulation for memory rehabilitation for dementia patients.

Currently, research on automatic egocentric activity recognition is mainly based on two broad categories of data: low-dimensional sensor data and high-dimensional visual data. Low-dimensional sensor data such as GPS, light, temperature, direction or accelerometer data has been found to be useful for activity recognition [17, 12, 7, 15, 11]. Low-dimensional sensor data can be collected and stored easily, and the computational complexity of the recognition is usually low. On the other hand, high-dimensional visual data is information-rich, and egocentric activity recognition based on visual data has achieved encouraging results using hand-crafted video features that encode both appearance [16, 6] and motion information [19].

During the last several years, deep learning has been successfully applied to many problems from various areas, like image or video classification, speech recognition. With deep learning approaches, multiple layers of feature hierarchies can be learned and also high-level representations of raw inputs can be automatically built. Some successful methods on video analysis include 3D ConvNets [9], Deep ConvNets [10], Two-Stream ConvNets [18], C3D

[22]. In order to take advantage of temporal information and improve the accuracy, some existing works combine Recurrent Neural Network with ConvNets [1, 4] to tackle the activity recognition problem. These approaches aim to automatically learn high-level semantic representation for raw videos by utilizing deep architectures discriminatively trained end-to-end in a supervised way.

To the best of our knowledge, there is only limited and preliminary effort to study multimodal egocentric activity recognition using the sensor and visual data simultaneously. In [20], the authors proposed a novel approach for generating sensor feature representation using Fisher vector and sliding window technique which also incorporates temporal information by introducing temporal order into trajectory-like sensor data. Furthermore, they also propose to apply Fisher Kernel framework in order to fuse sensor and video features for multimodal egocentric activity recognition.

In this paper, we propose a multi-stream deep architecture (see Figure 1) to recognize activities from multimodal egocentric data: video data and sensor data. In particular, we make the following contributions:

- For video data, we extend the two-stream ConvNets [18] to a three-stream ConvNets for spatial, optical flow and stabilized optical flow data. Egocentric videos captured from wearable devices usually contain a significant amount of camera motion. We propose to compute and analyze the stabilized optical flow extracted from the videos.
- For sensor data, we propose a new multi-stream Long Short-Term Memory (LSTM) framework to analyze multiple-axis sensor measurements: accelerometer, gyroscope, magnetic field and rotation. We leverage LSTM [1, 4] to capture long-term temporal information in the sensor streams.
- To fuse the results of multiple streams (spatial, optical flow and stabilized optical flow for video data, various sensor measurements), we examine average pooling and maximum pooling and a two-level fusion approach.
- We evaluate our proposed framework in detail on a Multimodal Egocentric Activity dataset. We show that our multimodal deep learning has comparable performance with state-of-the-art hand-crafted feature approach, despite the constraint that the size of the egocentric dataset is quite limited.

The rest of this paper is organized as follows. In Section 2 we present related works on egocentric activity recognition and also some state-of-the-art deep architectures. In Section 3 we discuss a Multimodal Egocentric Activity dataset for experiment. The proposed multimodal multi-stream deep learning framework is discussed in Section 4.

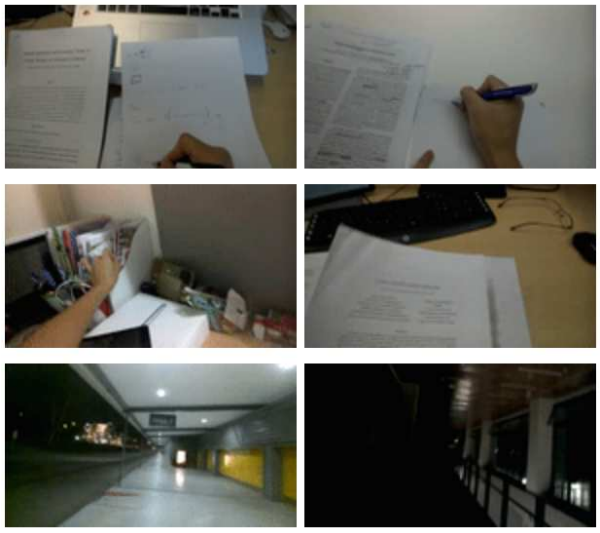

Figure 2: Sample frames of video data from Multimodal Egocentric Activity Dataset

Experimental evaluation is presented in Section 5 and we conclude the work in Section 6.

2. Related Work

For low-dimensional sensor data classification, [12] proposes features for egocentric activity recognition computed from cell-phone accelerometer data. They reported over 90% accuracy for 6 simple activities. [7] reported more than 80% accuracy for 9 activities with sensors located at two legs. And more recently, there is a lot of interests to perform egocentric activity recognition using high-dimensional visual streams recorded from individuals' wearable cameras. Compared to low-dimensional sensor data, visual data captures much richer information: scene details, people or objects the individual interacts, for example. Therefore, several egocentric video datasets and approaches have been proposed to recognize complex activities. Among them, some previous works focus on extraction of egocentric semantic features like object [16, 6], gestures [13] and object-hand interactions [5] or discriminative features[14]. Recently, trajectory-based approach [23] has been applied to characterize ego-motion in egocentric videos, and encouraging results have been obtained for activity classification [19].

Inspired by the breakthroughs from image domain, many deep architectures for automatically learning video features are proposed. Two-stream ConvNets [18] is one of the most successful architecture. It matches the state-of-the-art performance of trajectory based algorithm on large scale video datasets like UCF101 and HMDB51. The two streams consist of spatial and temporal networks. Spatial net is able

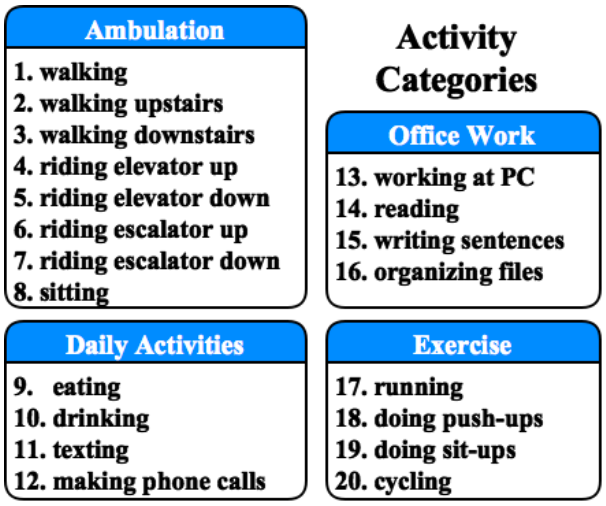

Figure 3: Activity categories of Multimodal Egocentric Activity dataset

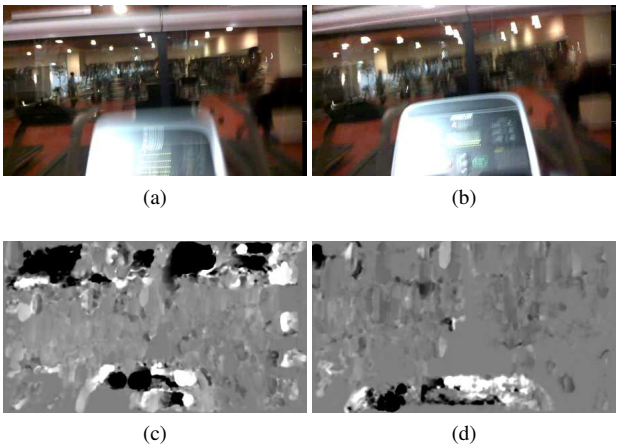

Figure 4: (a) (b) Two successive frames of *Running* activity (c) Optical flow (horizontal) (d) Stabilized optical flow (horizontal)

to extract semantic features from static information. While temporal net aims to learn motion features from stacking optical flows. Our work in this paper is also inspired by this two-stream architecture. C3D [22] propose a 3D convolution operation to learn spatial-temporal features which utilizes 3D volume filters instead 2D filters. [1, 4] introduce Recurrent Neural Network or Long-Short Term Memory to take advantage of long-term temporal information since most of the existing ConvNets are incapable of capturing long-term sequential information.

3. Dataset

The following types of sensor data are included: accelerometer, gravity, gyroscope, linear acceleration, magnetic field and rotation vector. Each sensor signal has a duration of 15 seconds and sampling rate of 10.

To evaluate the effectiveness of multimodal deep learning on egocentric activity recognition, we conduct experiments on a Multimodal Egocentric Activity dataset [20][1]. The dataset contains 20 distinct life-logging activities performed by different human subjects. The data is captured using a Google Glass that records high-quality synchronized video and sensor streams. The dataset has 200 sequences in total and each activity category has 10 sequences of 15 seconds each. The categories of egocentric activity are presented in Figure 3. Furthermore, the categories can also be grouped into 4 top-level types: *Ambulation*, *Daily Activities*, *Office Work*, *Exercise*.

The dataset has the following characteristics. Firstly, it is the first life-logging activity dataset that contains both egocentric video and sensor data which are recorded simultaneously. The dataset enables evaluation of multimodal approach. Secondly, the dataset is collected in various environment which contains large variability in background and illumination. Egocentric videos are recorded both indoor and outdoor with significant changes in the illumination conditions and from different human subjects. Thirdly, the dataset has a taxonomy based on the categories as shown in Figure 3. All 20 fine-grained activities can be grouped into 4 top-level categories to allow evaluation of new visual analysis approaches against different levels of life-logging activity granularity.

[1]The dataset is publicly available at `http://people.sutd.edu.sg/~1000892/dataset`

4. Approach

In this section, we describe the deep learning framework for multimodal egocentric activity recognition. Firstly, we extend the multi-stream ConvNets for recognizing activities in egocentric videos. Secondly, we propose a new multi-stream Long Short-Term Memory (LSTM) for classifying wearable sensor data. Finally, we experiment average pooling and maximum pooling techniques to fuse the video and sensor softmax scores to obtain the final predictions.

4.1. Multi-stream ConvNets on Egocentric Videos

Firstly, we review the original architecture of the two-stream ConvNets that was proposed for third-person videos [18]. Then we discuss how to use a small-scale egocentric video dataset to fine-tune the two-stream ConvNets that was pre-trained on a large third-person video dataset. We also discuss how to extend the ConvNets model with stabilized optical flows suitable for egocentric video recognition.

The original two-stream ConvNets consist of spatial

Layer	conv1	pool1	conv2	pool2	conv3	conv4	conv5	pool5	full6	full7	full8
size	7×7	3×3	5×5	3×3	3×3	3×3	3×3	3×3	-	-	-
stride	2	2	2	2	1	1	1	2	-	-	-
channel	96	96	256	256	512	512	512	512	4096	2048	101
receptive field	7×7	11×11	27×27	43×43	75×75	107×107	139×139	171×171	-	-	-

Table 1: Details about the pre-trained ConvNet model.

nets and temporal nets which are two separate ConvNets streams. The spatial streams are similar with deep architectures that used for image classification task. It is designed for capturing scenes or objects information. The input of spatial nets is single frame images of $(224\times224\times3)$. While temporal nets are built to extract dynamic motion information with the input of stacking optical flow fields volumes ($224 \times 224 \times 2F$, F is the number of stacking flows) with horizontal and vertical components of the vector field.

Egocentric videos are usually captured by wearable devices. Camera motions are usually prominent and pronounced due to the movement of the subjects. These varying and unintended camera motions may not be the representative of the motion present in an action. To capture the foreground motion information, we extend the two-stream ConvNets to three-stream ConvNets by introducing another stream. This additional stream consists of stacking optical flows extracted from stabilized egocentric videos. From Figure 4, we can see that the stabilized optical flow contains less shaking compared to the original optical flow. Then it can provide cleaner foreground information. In our experiment, we use the same temporal pre-trained model for stabilized optical flows.

The ConvNet architecture used is original from the Clarifai networks. The detailed architecture of ConvNets is shown in Table 1. A small modification is made to adapt to action recognition with fewer filters in conv4 layer and lower-dimensional full7 layer. We use the pre-trained model for single frame images, optical flows and stabilized optical flows. We fine-tune the last fully-connected layer for three streams.

After that, we average the softmax class scores for frames or optical flows extracted from the same video to be the final score of prediction. Furthermore, to fuse the three streams (spatial, optical flow and stabilized optical flow), we apply average pooling and maximum pooling to predict the labels of activities.

4.2. Multi-stream LSTM on Sensor data

In this section, we first describe an existing approach based on Convolutional Neural Network as a benchmark. Then we propose a new Multi-stream Long Short-Term Memory (LSTM) framework for classifying multi-channel sensor data which tackles the issue of missing temporal information and adopts the late fusion of prediction scores.

In [25], an approach inspired by CNN is used for classifying multi-channel time-series. Firstly, they use sliding window strategy to segment the time-series into short pieces of the signal. To be specific, each piece used by CNN is a two-dimensional matrix containing r raw samples with D channels. After that, 3 temporal convolution layers and 2 pooling layers are applied to learn the temporal features and reduce the temporal dimensionality. In this way, the number of channels D keep unchanged until the last unification layer which is a fully connected layer to fuse all channels data. Finally, a pooling technique is utilized to determine the label of the whole time-series by examining labels of all segments.

One limitation of this CNN-based method is that it is incapable to capture the temporal relationship for a long time-series data. Although CNN-based approach can learn some short patterns through the learning process, it still lacks long-term temporal information. Therefore, we propose our own multi-stream LSTM for egocentric activity recognition using different sensors (e.g., accelerometer, gyroscope, etc.)

We choose LSTM unit instead of traditional RNN in this work. The basic RNN takes in sequential input and for each data in the sequence, it calculates hidden states which take part in predicting the next data in the sequence. In this way, the RNN performs prediction or classification for a certain data point by finding the temporal relationship from the previous data point in the sequence. The LSTM unit is a popular variant of the basic RNN unit. Recently, for most of the problem, the LSTM is preferred because the gating mechanism of LSTM allows it to explicitly model long-term dependencies. Also, the calculation of hidden states involves adding previous hidden state information, rather than multiplying it which causes the values stored to blow up and the gradients to vanish.

To exploit information from all sensors, we propose a multi-stream LSTM which is inspired by the multi-stream ConvNets work. A maximum pooling is adopted to choose the maximum score of different sensor data as the prediction of multiple sensors for egocentric activity recognition.

5. Experiments

In this section, we first describe the experiment setup and implementation details for multi-stream ConvNets and

LSTM for classifying egocentric videos and sensor data. Then, we discuss the results of individual data modalities and also the fusion of video and sensor data.

5.1. Experiment setup

In our experiment, we downsampled the egocentric videos to the size of 454×256. We applied random cropping to the video when fine-tuning multi-stream ConvNets. The frame rate is 10 fps. And we selected 4 types of sensor data from the dataset: accelerometer, gyroscope, magnetic field and rotation vector. The sampling rate is 10 Hz. There are 10 splits in the dataset. The accuracy is obtained by averaging precisions over 10 splits with Leave-One-Out cross-validation (recall that there are 10 sequences in each activity class).

5.2. Implementation details

Fine-tuning multi-stream ConvNets. Training deep ConvNets is more challenging for egocentric activity recognition as activity is more complex (considerable camera motion in addition to the object motion in the scene). Furthermore, the size of egocentric video dataset is extremely small compared to the ImageNet dataset [3]. In addition, due to the privacy concern, egocentric video samples are difficult to collect and is scarce compared to traditional third-person view videos. In this work, we choose to fine-tune the ConvNet model which is trained on UCF101 dataset, a large-scale action recognition dataset. The pre-trained model is provided in [24]. We use Torch library [2] to fine-tune the last layer of the network with egocentric video data. The batch size is set to 128, the number of units of last fully-connected layer is set to 2048 and dropout is set to 0.8. Momentum is set to 0.9 and decay is set to 10^{-6} for stochastic gradient descent (SGD). For single frame stream, we first resize the frame by making smaller side to 256, and then randomly crop a region of 224×224. The learning rate is set as 10^{-2} initially and fine-tuning is stopped after 40 epochs. For both optical flow stream and stabilized optical flow stream, the input is 3D volumes of stacking optical flows fields. TVL1 optical flow algorithm [26] is chosen with OpenCV implementation, since it makes a good balance between efficiency and accuracy. We discretize the floating number of optical flow fields into integers in $0-255$ like color channels of images for fast computation. Specifically, we stack 10 frames of optical flow fields of horizontal and vertical direction. Then we also fine-tune the last layer of the pre-trained temporal net which is provided in [24]. The fine-tuning process of the temporal net is similar with spatial net and a $224 \times 224 \times 20$ volume is randomly selected from training video and also flipped for augmentation purpose. Video stabilization is done by using *ffmpeg vidstabdetect* and *vidstabtransform* filters.

Training multi-stream LSTM. For training multi-stream LSTM using multi-channel sensor data, we train from scratch using Torch library. A single-layer LSTM with 128 hidden units is used. Following the LSTM layers, a softmax classifier makes a prediction. The dimensionality of the sensor data is 3 or 4 which depends on the type of sensor used. Four types of sensor data are used in our experiment: accelerometer (3-axis), gyroscope (3-axis), magnetic field (3-axis) and rotation vectors (3-axis and magnitude). The batch size is set to 16 and RMSProp [21] is chosen as the optimizer. Training is stopped after 100 epochs.

Based on the scores of different sensor data, a maximum pooling is used to select the maximum score from all sensors to make a more accurate prediction.

Algorithm	Accuracy
Single frame	72.4%
Optical flow	48.9%
Stabilized optical flow	45.2%
average pooling	68.5%
maximum pooling	75.0%
Trajectory + Fisher Vector [20]	78.4%

Table 2: Result on egocentric videos using multi-stream ConvNets.

5.3. Results

In this section, we evaluate the performance of improved multi-stream ConvNets on video data and proposed multi-stream LSTM on sensor data. Then we apply late fusion for our method and compare it to state-of-the-art result.

Result on egocentric videos Table 2 provides the results of egocentric videos using multi-stream ConvNets. The single frame, optical flow and stabilized optical flow are the three streams that we use to classify egocentric videos. From this table, we can observe the performance of individual streams and also fusion results using average and maximum pooling at the video level. Surprisingly, single frame stream has the highest accuracy which is much higher than the two optical flow streams. This is not expected since the optical flow is considered to be important in recognizing activities from first-person view videos. Also, from the previous work on third-person view videos, optical flow usually has better performance than single frame stream [20, 18]. We believe the reason is that in our case, the pre-trained model for learning optical flows feature is trained on third-person view videos. Therefore, the filters are less meaningful and useful for distinguishing egocentric videos, which usually contain a large amount of shaking and noise.

Then, we evaluate two fusion techniques: average pooling and maximum pooling. For average pooling, we sum up all scores from three streams and the choose the maximum score as the final prediction. For maximum pooling, we

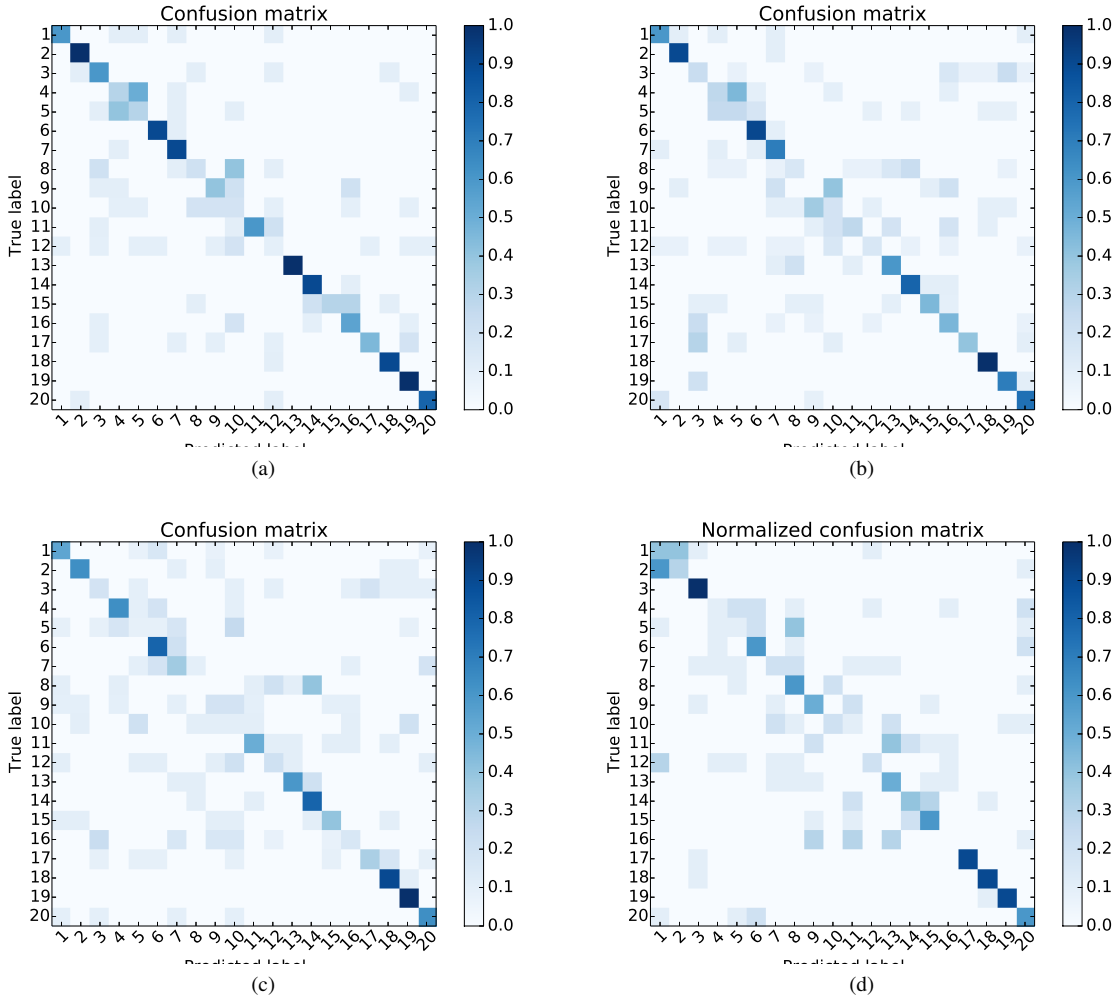

Figure 5: Confusion Matrices of (a) single frame stream (b) optical flow stream (c) stabilized optical flow stream (d) multi-stream LSTM

directly choose the maximum score from all sensor scores and take that score as the prediction. It turns out that maximum pooling has better performance. And we also compare our result with methods using hand-craft features. The trajectory-based approach outperforms our approach which suggests that hand-craft features would still be useful when the datasets are small.

Result on sensor data Result of the proposed multi-stream LSTM is reported in Table 3. We have evaluated the CNN-based approach from [25]. Our proposed approach outperforms the CNN-based approach which proves that it is beneficial to utilize temporal information with LSTM network. By examining the confusion matrix of multi-stream LSTM and video-based results in Figure 5, we can find that they both perform well for *Exercise* activities which makes sense since they are easily distinguished no matter from video or sensor data. While, for activity *sitting*, *eating* and *drinking*, multi-stream LSTM outperforms the video-based results. And multi-stream ConvNet on egocentric video has better performance for *Office work* activity since high-level information is beneficial and also it is extremely hard to recognize activity like *organizing files* with sensor data only.

Algorithm	Accuracy
CNN-based [25]	47.8%
Proposed multi-stream LSTM	49.5%
Temporal enhanced FV [20]	69.0%

Table 3: Comparison of different approaches on sensor data.

Fusion results After generating scores from multi-stream ConvNets and LSTM model, we also need to fuse

Algorithm	Accuracy
Proposed method with average pooling	76.5%
Proposed method with maximum pooling	80.5%
Multimodal Fisher Vector [20]	83.7%

Table 4: Comparison of different approaches on all data.

them together. Similarly, we choose average pooling and maximum pooling to evaluate. From Table 4, we can find that maximum pooling achieves an accuracy of 80.5% and outperforms average pooling by 4%.

As we have shown, a two-level fusion strategy is applied in our experiment. We first fuse video and sensor result respectively and then fuse their scores together. The reason is that the three streams in video-based ConvNets provide similar information, and the information is extremely different from the one learned from sensor data because of the different modalities.

The performance of our multimodal deep learning approach is slightly worse compared to the Multimodal Fisher Vector approach proposed in [20]. It is not unusual for hand-crafted features to work better than learned representations when the amount of training data is small. Such observations have been reported in other domains like speech recognition [8]. Tagging millions of egocentric videos is challenging and might not be feasible. Future work will focus on how the performance can be further improved by combining the merits of different approaches, given limitations in the size of the training data.

6. Conclusion

This paper proposed a new multimodal multi-stream deep learning framework to recognize egocentric activities. In our proposed framework, multi-stream ConvNets and multi-stream LSTM architectures are utilized to learn discriminative spatial and temporal features from the egocentric video and sensor data respectively. Two different fusion techniques are evaluated on two different levels. Our comparison with state-of-the-art results shows that our proposed method achieves very encouraging performance, despite the fact that we have only very limited egocentric activity samples for training. Future work investigates different data argumentation approaches to improve the networks.

References

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *Human Behavior Understanding*, pages 29–39. Springer, 2011.

[2] R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. Technical report, IDIAP, 2002.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[5] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 407–414. IEEE, 2011.

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

[7] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, and S. Lee. Activity recognition based on semi-supervised learning. In *Embedded and Real-Time Computing Systems and Applications, 2007. RTCSA 2007. 13th IEEE International Conference on*, pages 469–475. IEEE, 2007.

[8] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[9] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, 2013.

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

[11] A. M. Khan, A. Tufail, A. M. Khattak, and T. H. Laine. Activity recognition on smartphones via sensor-fusion and kda-based svms. *International Journal of Distributed Sensor Networks*, 2014, 2014.

[12] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[13] S.-R. Lee, S. Bambach, D. J. Crandall, J. M. Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 557–564. IEEE, 2014.

[14] B. Mandal and H.-L. Eng. 3-parameter based eigenfeature regularization for human activity recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 954–957. IEEE, 2010.

[15] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. Activity recognition and monitoring using multiple sensors on different body positions. In *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*, pages 4–pp. IEEE, 2006.

[16] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision*

and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2847–2854. IEEE, 2012.

[17] H. Rezaie and M. Ghassemian. Implementation study of wearable sensors for activity recognition systems. *Healthcare Technology Letters*, 2(4):95–100, 2015.

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

[19] S. Song, V. Chandrasekhar, N.-M. Cheung, S. Narayan, L. Li, and J.-H. Lim. Activity recognition in egocentric lifelogging videos. In *Computer Vision-ACCV 2014 Workshops*, pages 445–458. Springer, 2014.

[20] S. Song, N.-M. Cheung, V. Chandrasekhar, B. Mandal, and J. Lin. Egocentric activity recognition with multimodal fisher vector. *arXiv preprint arXiv:1601.06603*, 2016.

[21] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop. *COURSERA: Neural networks for machine learning*, 2012.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *arXiv preprint arXiv:1412.0767*, 2014.

[23] 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.

[24] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2015.

[25] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), Buenos Aires, Argentina*, pages 25–31, 2015.

[26] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition*, pages 214–223. Springer, 2007.