Multimodal Egocentric Action Recognition

Giorgio Mongardi
Politecnico di Torino
s292490@studenti.polito.it

Luca Calo'
Politecnico di Torino
s294828@studenti.polito.it

Abstract

The following project, carried out for the course of Advanced Machine Learning of Politecnico di Torinio, deals with Multimodal Action Recognition applied to the field of Egocentric Vision. We first interacted with the Epic Kitchens dataset, using a pretrained model to extract features from portions of videos taken from it, and exploiting these features to train a classifier for action recognition task. After that we worked on ActionNet dataset, analyzing the EMG modality inside it (given by a body-trackment system using muscle activity sensors on the forearms of subjects) and then working on a variation of ActionNet, in order to build a new model to exploit CNN over EMG features. Project Source Code: https://github.com/BlankTo/aml23-ego-master.git

1. Introduction

Egocentric vision is an emerging branch of computer vision referring to the use of visual data captured from a first-person perspective. This offers a unique perspective including hand movements, object interactions and a direct view of the environment. Given that, egocentric vision is particularly useful for tasks such as action recognition [1, 2, 5–9], human-object interaction, robot manipulation [9] and others. One of the main challenges research community noted in this field is to reach the capability to understand the dynamic nature of the scene: using only RGB video data can result in poor performances because models may be led to try to detect more the objects in the scenes than the performed actions themselves [6]. With the emergence of multi-sensor wearable devices, depth cameras and microphones [7], new modalities have been explored alongside RGB to achieve better results. As a result, there has been a renewed interest from the computer vision com- munity on collecting large-scale datasets as well as developing new or adapting existing methods to the first- person point-of-view scenario [8]. In this paper we first explore EPIC-Kitchens, extracting the RGB features using an I3D pretrained backbone, analyzing them through clustering and using them to train a classifier. Secondly we retrieve the EMG features from ActionNet, resample them and train a LSTM based on ActionNet model. Finally from the EMG features we compute the spectrograms in order to train a simple CNN.

2. Related Work

Datasets: In literature there are several datasets useful for the egocentric vision action recognition. One of the most famous is EPIC-Kitchens [3] : it captures everyday activities from a first-person perspective in a video format only, specifically focusing on kitchen environments where users perform various tasks such as cooking, cleaning, and food preparation. The dataset contains more than 55 hours of video for a total of 39.6k action segments. In the context of using other modalities than RGB, Del Preto *et al.* introduced ActionNet [7], a dataset that includes 12 hours of data from cameras and microphones recordings, IMU data, finger-tracking data, tactile sensors data, EMG armband recordings and eye-tarcking data.

Multimodal egocentric action recognition: Various approaches can be employed when attempting to integrate different data sources associated with the same action. Kazakos *et al.* addressed the challenge by introducing EPIC-Fusion [8]. This architecture integrates RGB, audio, and optical flow modalities with varying temporal offsets, leading to state-of-the-art results in classifying egocentric tasks on the EPIC-Kitchens [3] dataset. DelPreto *et al.*, instead, focused on a multi stream network composed of four parallel LSTMs, each responsible for one modality, that are concatenated and then passed through another LSTM, dropout layer and finally a dense layer [4].

Spectrogram as Modality: Working on EPIC-Fusion [8], Kazakos *et al.* used the audio data to generate spectrograms as 2D images to be used to try and improve the results obtained with audio modality.

3. Method

In this section we describe the methodologies we used from dataset extraction to classification.

3.1. RGB Features from EPIC-Kitchens

We start from data relating to the subject P08 of the reduced version of EPIC-Kitchens dataset, which includes 28 videos, each divided in frames of shape 456x256. Although the EPIC-Kitchens annotations already categorized the segments of videos by action, we further divide these actions into clips of varying lengths using uniform and dense sampling to create a larger and more manageable dataset. Uniform sampling refers to the selection of evenly spaced frames in the clips, focusing on images that may be further apart, better highlighting the temporal dynamics. On the other hand dense sampling focuses on the appearance of the video as the images are close in time by taking frames spaced by a small stride. For each video segment we extract 5 clips using these two samplings and a varying number of frames per clip (5-10-25). For each clip we then employ an I3D backbone pretrained on ImageNet to extract the features before the fully connected layer of the model. To make some evaluations on sampling methods and clip lengths, we proceed to plot the PCA-reduced extracted features and perform clustering. Having 5 clips per sample, we try both models that process one clip at a time, models that perform temporal aggregation and models that make use of the temporal dimension.

3.2. EMG Features from ActionNet

The ActionNet dataset is composed of 4 modalities, but we use only the Electro-Myography (EMG) one. The EMG were recorded by DelPreto *et al.* using two Myo Gesture Control armbands, one per arm. The recordings consist of 8 signal channels per device sampled at 160 Hz and downsampled to 10 Hz. We preprocess the EMG features following [4] by rectifyng each channel with its absolute value, applying an absolute filter to exclude frequencies under 5 Hz, joint-normalizing the channels and rescaling them to [-1, 1] using min and max across all channels. Since the number of samples is very limited, after the preprocessing, we proceed to divide the samples in 5s/10s/25s segments and then resample them in order to have the same number of readings for each segment. Having obtained fixed-size matrices for EMG data, we use the ActionNet model, but removing the other modalities' branches. Every matrix is passed through two LSTMs followed by one dropout layer and finally a dense layer. For the labeling, we remove some error obtaining 20 narration classes and then we categorize the narrations into 8 verb classes. We than tried some ablations for the classifier.

3.3. Spectrograms and CNN

In this part, we start from the EMG data obtained in the previous step and use Short-Time Fourier Transform (STFT) to compute the spectrograms of each channel (Fig. 1) using torchaudio.transform.Spectrogram. STFT is applied to data from body-tracking systems (or other signal modalities) to analyze how the frequency components of body movements change over time. This analysis reveals how the dynamics of body movements vary temporally, allowing for a better understanding of patterns and periodicities in actions. Then, we employ a simple CNN model to classify the 16 channel images representing the EMG of a clip.

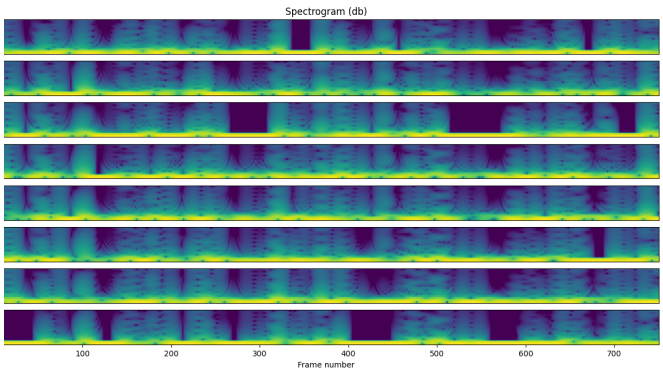


Figure 1. Spectrograms obtained from EMG data

4. Experiments

In this section we describe the experiments made, from sampling to classification. All trials were carried without using the GPU on a local machine and the program flow can be found in the project.ipynb file. Some ablations were tested and reported below.

4.1. Clustering EPIC-Kitchens Extracted Features

After extracting the features from EPIC-Kitchens using different clip lengths and either dense or uniform sampling, we reduce the dimensionality to 2 using PCA and plot the central frame of each clip so that we could visualize the distribution of actions and objects. K-means is also applied to the features and the clusters are shown on figures as colored polygons (Fig. 4, 5). For what we can see, the frames are distributed more according to the object depicted in the image than the action performed in that clip. This behaviour was expected and shows the limitations of using only RGB features for action recognition.

4.2. RGB Action Recognition and Temporal Aggregation

The temporal dimension represented by the clips in a sample allow us to explore various methods that make use of different temporal aggregation strategies or that instead accept the temporal dimension as it is. We tested all models

for 1000 epochs with a learning rate of 0.001 (reduced after 750 iterations), weight decay of 1e-7, momentum of 0.9 and Cross Entropy Loss:

- a simple MLP feeding one clip at the time as a baseline
- a simple MLP that before feeding the clip to the net performs average pooling to condense the temporal dimension
- a simple MLP that before feeding the clip to the net performs max pooling to condense the temporal dimension
- a simple MLP that accepts the concatenation of all clips by flattening the temporal dimension
- an LSTM that accepts the temporal dimension
- an TRN that accepts the temporal dimension

Overall, the dense sampling gave better results than uniform sampling and leveraging the temporal dimension proved to to enhance the performances with respect to the case where only one clip at the time was passed (Tab. 1).

	5		10		25	
	Dense	Uniform	Dense	Uniform	Dense	Uniform
MLP single clip	51.86 54.25	38.76 47.59	52.63 55.17	44.99 52.18	59.38 58.62	51.01 53.56
MLP flatten	57.56 59.08	51.17 54.02	55.97 57.24	55.22 57.70	57.99 59.54	53.62 58.16
MLP avg pooling	58.49 58.16	48.09 50.57	57.63 57.01	54.87 54.02	56.86 59.54	53.77 56.09
MLP max pooling	57.33 57.70	45.76 49.66	57.00 57.24	48.02 53.33	58.06 59.77	51.69 54.48
Temp Conv	57.29 57.01	50.42 50.57	54.90 55.86	55.39 54.94	58.10 59.31	56.55 56.78
LSTM	53.82 55.17	45.88 49.20	57.31 55.63	52.11 54.71	56.51 58.85	55.61 56.32
TRN	57.26 57.01	45.29 51.49	59.76 59.08	56.16 56.78	56.71 58.62	53.63 57.47

Table 1. Action Recognition on EPIC-Kitchens extracted features

Figure 2. base LSTM model

4.3. LSTM on EMG data

We started from the suggested LSTM based model composed of one LSTM that reduces the 16 input dimensions to 5 and then a second LSTM that expand the dimensionality to 50. We also tried to remove the first LSTM to check if it was losing information and it seems to be the case. Finally we also tried to increment the number of neurons in LSTM; the results improved, but the model became much more computationally expensive. The results improved considerably with respect to RGB models, showing that the use of other modalities can be beneficial. It seems that the 10 seconds sampling doesn't have enough samples to train viable model, in particular for the classification over narrations (Tab. 2).

	5		10	
	Verb	Narration	Verb	Narration
LSTM (16⇒ 5 ⇒ 50)	51.48 83.20	81.27 91.93	53.64 83.23	9.68 56.71
LSTM (16⇒ 50)	65.90 91.27	40.19 73.97	63.84 89.02	33.62 57.62
LSTM (16⇒ 512)	85.21 96.05	12.39 32.62	31.66 57.32	12.94 29.88

Table 2. Action Recognition on ActionNet EMG features

Figure 3. CNN model

4.4. CNN on Spectrograms

We tested our CNN over the spectograms obtaining the following results:

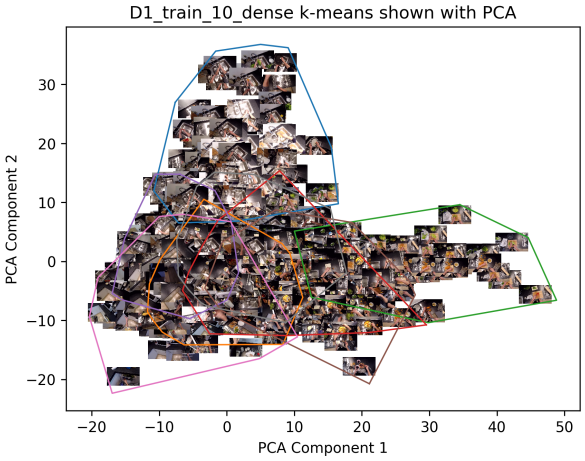
	5		10	
	Verb	Narration	Verb	Narration
CNN	81.60 95.55	81.35 92.42	65.90 90.24	65.66 85.37

Table 3. Action Recognition on Spectrograms from EMG data

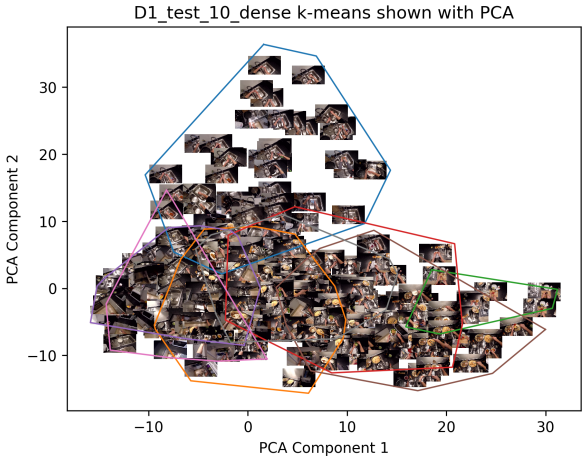
The CNN seems to show some grade of improvement with respect to LSTM and also seems to be less effected by the scarcity of samples (Tab. 3).

5. Conclusions

In this paper, we explored the task of egocentric action recognition for actions carried out in the context of a kitchen environment. We started by comparing dense and uniform sampling on EPIC-Kitchens. The results show that dense sampling seems to perform better on all models, probably meaning that spatial information is more meaningful in action recognition than temporal information, at least for RGB data. We did not notice much differences between various clip lengths, but 25 frames per clip highlights slightly better results in most cases. The temporal dimension seems to make a difference for shorter clips.

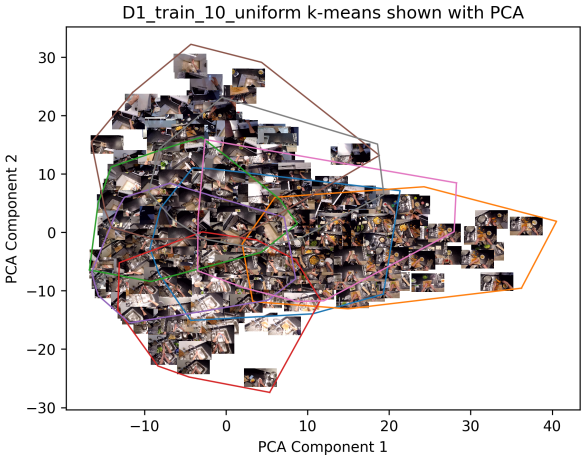


(a) Training Set

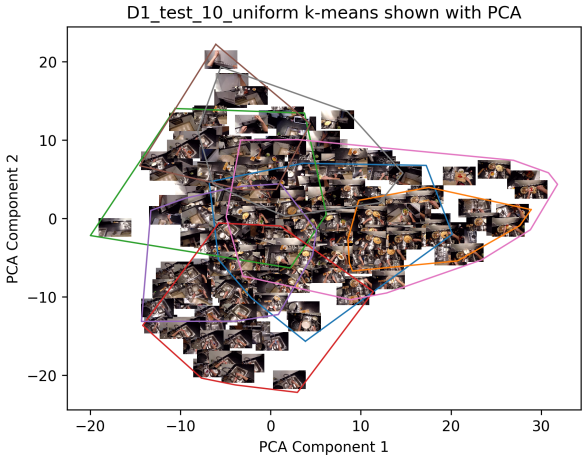


(b) Test Set

Figure 4. K-Means Clustering for Dense Sampling (10 Frames per clip)



(a) Training Set



(b) Test Set

Figure 5. K-Means Clustering for Uniform Sampling (10 Frames per clip)

Even if the results are not directly comparable cause they come from different datasets, the use of EMG modality as described in ActionNet gives much better results, further improved by removing the first layer LSTM.
Finally, the results obtained with the CNN over the spectrograms are similar to ones obtained with the EMG data. As a future work, we plan to test the mid-fusion and late-fusion of modalities to try to improve our results.

References

[1] Gabriele Goletto Marco Cannici Emanuele Gusso Matteo Matteucci Barbara Caputo Chiara Plizzari, Mirco Planamente. E2(go)motion: Motion augmented event stream for egocentric action recognition. *CINI*, 2021. 1

[2] Randan Ramakrishnan Rogerio Feris John Conh Aude Oliva Quanfu Fan Chun-Fu Chan, Rameswar Panda. Deep analysis of cnn-based spatio-temporal representations for action recognition. *CVPR*, 2021. 1

[3] Giovanni Maria Farinella Sanja Fidler Antonino Furnari Evangelos Kazakos Davide Moltisanti Jonathan Munro Toby Perrett Will Price Michael Wray Dima Damen, Hazel Doughty. Scaling egocentric vision: The epic-kitchens dataset. *ECCV*, 2018. 1

[4] Simone Alberto Peirone Gabriele Goletto. Multimodal egocentric action recognition. 2023. 1, 2

[5] Andrew Zisserman Joao Carreira. Quo vadis, action recognition? a new model and the kinetics dataset. *CVPR*, 2017. 1

[6] Dima Damen Jonathan Monroe. Multimodal domain adaptation for fine-grained action recognition. *CVPR*, 2020. 1

[7] J. (n.d.) Joseph DelPreto. Actionnet: A multimodal dataset for human activities using wearable sensors in a kitchen environement. *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1

[8] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. *ICCV*, 2019. 1

[9] Vikash Kumar Chelsea Finn Abhinav Gupta Suraj Nair, Aravind Rajeswaran. R3m: A universal visual rapresentation for robot manipulation. *CoRL*, 2022. 1