Multimodal Egocentric Action Recognition

Filippo Greco
Politecnico di Torino
s309529@studenti.polito.it

Stiven Hidri
Politecnico di Torino
s315147@studenti.polito.it

Gabriele Ferro
Politecnico di Torino
s308552@studenti.polito.it

Abstract

Egocentric vision is a field of computer vision that focuses on action recognition from a first person point of view. Many researches tried to exploit different data sources other than video acquisition, such as audio information and signals recorded by various sensors placed on the actor's body. In this work, we investigate the differences between frames sampling techniques such as dense and uniform sampling. We extract, through the employment of a pre-trained I3D network, the features from different video recordings of the Epic-Kitchens dataset, in order to address which sampling works better in the egocentric vision environment. We then move to the ActionSense dataset and propose the utilization of the ElectroMyoGraphy reading as an additional data source. Finally we explore the application of this new information alongside the video recording data through the use of late and mid-level fusion merging the best performing classifiers, highlighting the potential of multimodal approaches that enhances egocentric action recognition systems. Source code can be found at: https://github.com/figimodi/AML-egovision

1. Introduction

Egocentric vision poses the challenge of recognizing an action using its video recording from the point of view of the actor. Given the fact that the principal source of information is constituted by a video, it is of relevant importance the consideration of both the spatial information present in each frame and the temporal information available between them.

While the simple RGB information available via video seems enough, other data modalities could contribute to the action recognition task including audio sources, ElectroMyoGraphy (EMG) sensors and event cameras.
When using all these kind of data together it is important to consider the synchronisation among them. We chose to sample the sources from the same time instant. Others adopted a much more articulated extraction strategy like Kazakos et al. [7], who consider different instants for the different sources, to emphasize for each modality the most relevant moment.

In this project we start considering the RGB video source alone, which allows a simple study of the trade-off between spatial and temporal information. In this phase we analyze the results obtained using the frame sampling techniques known as **dense sampling**, which captures frames close to the central one, favoring spatial information, and **uniform sampling**, which takes evenly spaced frames from the start to the end of the video, favoring instead temporal information.

Once the frames are sampled into clips, they are then passed through a CNN known as I3D [1] that extracts their features. The obtained features are then visually analyzed with the help of techniques such as K-Means clustering and dimensionality reduction algorithms including t-SNE and PCA, subsequently plotted to study the presence of patterns that can guide the feature extraction. In the last step of this phase, the features coming from the two different sampling techniques are mutually used to train a classifier in order to check which sampling method gives the best performances.

Additional information streams are taken into consideration, where the principal supplementary source of information is obtained by the employment of EMG sensors, able to capture the electrical signals from the arms of the actor that performs a variety of actions.
Starting from the raw data a pre-processing procedure is applied and the product is later utilized for the EMG and RGB classifiers. Moreover, we also managed to use the spectrogram representation of the EMG signals as an additional data modality. Finally, the best performing models for each modality are chosen and then used together with late and mid-level fusion approaches to take advantage of all of the information available.

2. Related works

Many studies have been made in the action recognition domain which enabled to transfer the obtained knowledge to multiple challenges such as augmented reality, robotics [11] and surveillance.

Alternative data modalities might help to improve the action recognition task. **Plizzari et al.** [12] explore the potential of event cameras for egocentric action recognition, highlighting the advantages with respect to traditional RGB cameras. It is important to note that these cameras are very well suited for egocentric vision applications since they perform well when fast camera motions are present and when there is a limited power consumption restriction, often present when wearable devices are employed.

Researchers have developed sophisticated methods for interpreting human actions combining different data sources. **Kazakos et al.** [7] propose a new method that combines visual (RGB, optical flow) and audio information with mid-level fusion alongside sparse temporal sampling of fused , demonstrating the complementarity of audio and appearance information.

Techniques that are able to adapt models to new environments without requiring new data have also been explored since the variability of the surroundings is a significant obstacle in action recognition. **Munro et al.** [10] worked on a solution that combines images and motion information to improve model adaptability. It has been proved that aligning these data sources across different environments promotes generalization in activity recognition throughout different environments.

Action recognition applications are computationally demanding and trying to maintain high accuracies while limiting the computational cost is surely one of the top necessities in the field. **Lin et al.** [9] address this challenge by proposing the Temporal Shift Module (TSM) which is able to achieve the performance of 3D-CNNs while maintaining the lower complexity of 2D-CNNs.

Moreover a comprehensive analysis done by **Chen et al.** [2], which consisted in building an unified framework for both 2D-CNN and 3D-CNN action models, reveals that advances in action recognition have improved efficiency rather than accuracy, and that both kind of models behave similarly in terms of spatio-temporal representation abilities and transferability.

3. Working on Epic-Kitchens

3.1. Dataset

Epic-Kitchens [3] is a large dataset that includes the first person video together with the audio and the annotation for each action performed by the actor. The actions are non-scripted and performed in different environments to reduce the bias that the kitchen's surroundings can inject in the data. Figure 1 shows some examples of frames extracted from the Epic-Kitchens dataset.

Figure 1. Frame samples of recordings from the Epic-Kitchens dataset

3.2. Sampling Techniques

As previously said, the first part of the project concerns the decision of which sampling modality to adopt for the RGB video frames.
We starts from a bunch of different videos, of a subset of actors, included in the Epic-Kitchens dataset [3] of which we already have the frames available. The annotations for these videos are included in a different file that provides the label, together with the start and end timestamp of the clip that corresponds to the action.

In videos, the relevant information is distributed among different frames, so we need to consider not only a specific image but a different set of them, sampled from the original clip.
The choice of the sampling technique concerns the trade-off between the focus on spatial and temporal information.
The spatial information is the one that captures the appearance of the frames themselves, giving more attention to the presence and position of an object and its surrounding.
The temporal information, on the other hand, instead of focusing on the position of the items in the environment, captures their movement, highlighting the evolution of the position over time.
To understand better the importance of these two types of information, and which one gives the best grasp of the action, we concentrated on two main strategies: the dense sampling and the uniform sampling.
The first step and the final product of both techniques are the same.
The first step is taking a randomly sampled frame that will be the center of a clip. A clip is then composed by a subset

of frames that lies within an action and those frames are chosen by the sampling technique. Meanwhile, the final product is a set of clips, composed as previously explained, that will represent a subset of the frames of the entire action. Different clips could even be overlapped.

With **dense** sampling we refer to the technique that selects adjacent frames spaced with a stride of 1 or 2 frames at most. This technique focuses more on the appearance of the video, given that it considers frames close to each other in the temporal sequence, thus it favors capturing the spatial information of the video.

The **uniform** sampling, on the other hand, selects evenly spaced frames within the action's video segment, pointing its attention more towards the temporal information rather than the spatial ones.
The differences between the two adopted techniques can be perceived more on longer action's sequences, given the fact that for shorter samples the stride adopted in dense sampling can be equal to the spacing adopted in the uniform sampling. Even when the number of frames is not enough to allow the presence of spacing, the two techniques will output the same clips. When the number of frames of the action are lower than the number of frames per clip we want to extract, some copies of the already selected frames are randomly sampled, in order to match the required clip length.

Figure 2 visually show the main differences among the two techniques.

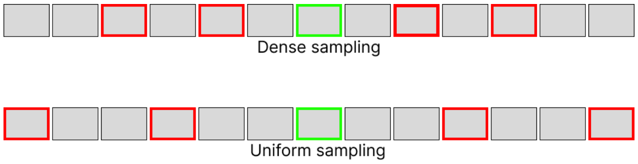

Figure 2. Example for dense and uniform sampling techniques

3.3. Features Extraction

For the feature extraction it is possible to rely on a pre-trained network that has already demonstrated its potential in video classifications tasks, the **I3D** (Inflated 3D Convolutional Networks) [1].
Its main idea consists in leveraging the power of 3D convolutions to capture both spatial and temporal information in videos.

The clips, previously obtained by dense or uniform sampling, are fed to the I3D network, which will produce the corresponding extracted features. After the extraction, the features are used to train and test a MLP, RNN and a LSTM model for classification, in this way is possible to test the impact of the sampling techniques employed and consequentially chose which model will perform the best for action recognition on Epic-Kitchens.

Morover a visual analyisis is performed using clustering techniques, where the details are explained in detail in Section 5.

4. Working on ActionSense

4.1. Dataset

ActionSense [4] is a less diversified dataset with respect to the environment of the actions compared to Epic-Kitchens, but this is justified by the amount and diversity of data modalities that are collected for each single registration.
In fact, in this dataset, there are different kind of data sources such as the first person video, the eye tracking of the actor's eyes movements and different sensor along the actor's body. Among the body sensors there are finger-tracking gloves with tactile sensors on both hands, inertial sensors for the body movements and Myo Armband sensors that collects the muscles' activation signals. Figure 3 shows some examples of frames extracted from the ActionSense dataset along with the corresponding labels.

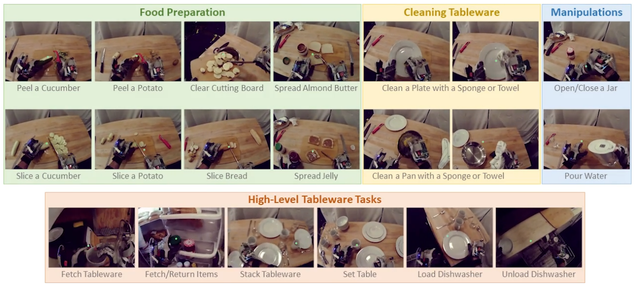

Figure 3. Frame samples of recordings alongside corresponding labels from the ActionSense dataset

4.2. Multimodality

From the variety of data that the Action Sense dataset offers we chose to utilize three different modalitie: a)EMG, representing the ElectroMyoGraphy of the arms' sensors; b)RGB, that are the frames of the egocentric video of subject 4; c)Spectrograms, which are the product of the spectrogram representation of the EMG data. Multimodality is applied by the means of fusion models, which put together information coming from different data sources. For this purpose mid-level fusion and late fusion, with all the possible combinations of the aforementioned sources, are tested to analyse if and how much the adoption of fusion models can help the classification.

4.2.1 EMG

The EMG data, present in the dataset, are obtained by two Myo Armband sensors, one for each forearm. Each sensor registers 8 values for each acquisition timestamp, which yield in total 16 values at each time step.

In order to correctly pre-process such values we consulted the ActionNet documentation [4] and tested many combinations of different pre-processing techniques that led us to choose the following steps for each action: a low-pass filter with cut frequency of 5Hz, followed by the application of the absolute value and a scaling between -1 and 1 and, finally, the readings are sampled at a lower frequency of 10Hz. Due to the lack of many samples, data augmentation has been adopted by subdividing each sample in 20 sub-samples of 5 and 10 seconds (which seemed enough time to represent the action). Each sample can eventually overlap with the others if the original action's sample is not long enough. Zero padding have been applied on both sides of the readings to obtain actions with the same dimensionality. Other padding techniques have been tried, including mean and Gaussian noise, but they gave worse results.

We noticed that in the pre-defined splits some classes didn't have any samples. In order to obtain a fairer analysis we moved a fraction of the train samples to the test split to test all classes. Finally, the pre-processed information are fed into a LSTM network. We tested both a 2-layer and a single layer LSTM networks, composed as it is shown in Figure 4.

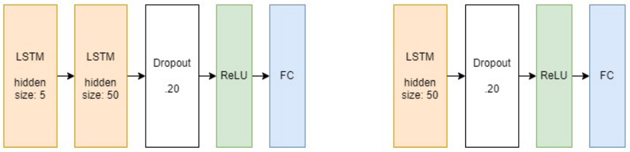

Figure 4. 2-layer LSTM (left) and single layer LSTM (right) used for EMG modality.

4.2.2 RGB

First of all we downloaded the video recording regarding Subject 4 from the ActionSense dataset and extracted all the frames. We then adopted the dense sampling strategy with 16 frames per clip (that gave the best result for the Epic-Kitchens RGB classification), which will be used for the feature extraction task.
The chosen model for the classification of RGB extracted features is a network composed with an LSTM block, followed by a dropout layer and a ReLU as activation function.
The input of the classifier is composed by the extracted feature via the I3D network, as done for Epic-Kitchens dataset.

For this project, RGB video frames are available only for subject 4, which causes a reduced number of samples for this modality classification, as well as for the other combinations of RGB data in fusion models, i.e. the accuracies when RGB data is an input are to be considered the product of a different dataset than EMG and spectrograms alone.

4.2.3 Spectrograms

Additionally we have extracted the spectrograms, represented in Figure 5, from the pre-processed EMG signals using `torchaudio.transforms.Spectrogram` and fed them into a CNN. Initially, we tried with ResNet18 [5], but we ended up opting for a simpler and faster network: LeNet5 [8] shown in Figure 6.
This modality is worth considering, given that a CNN might be able to find correlations between the signals frequencies, plotted from lighter to darker colors, and their respective class labels.

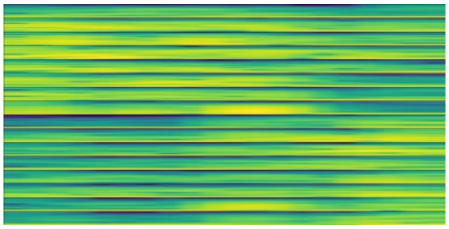

Figure 5. Example of extracted spectrogram. Each row represents a different channel

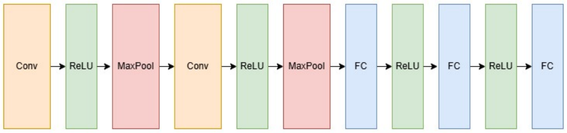

Figure 6. LeNet5 architecture, employed for spectrogram modality on ActionSense dataset.

4.2.4 Fusion

To leverage the aforementioned modalities (RGB, EMG and spectrograms), we tried to put together these different sources through the use of model fusion. We experimented two main fusion techniques: late fusion and mid-level fusion.

The **late fusion** is the simplest and fastest approach because it does not require any specific training for each combination of the modalities. The models (one for each modality) are pre-trained and their individual outputs are combined directly at inference time. The combination happen through a weighted mean or a sum. We investigated, for each combination, the best weights for the different modalities.

For what regards the **Mid-Level Fusion** strategy, we took inspiration from [6]. Different combinations of EMG, RGB and spectrograms data are fed to a network, composed as shown in the left side of Figure 7. The network pass the raw data of each modality to the relative model and then it combines the different mid-level features altogether. In this way the different modalities are jointly trained using the same loss, so the models learn how to cooperate with each other, hoping to improve the overall action recognition performance.

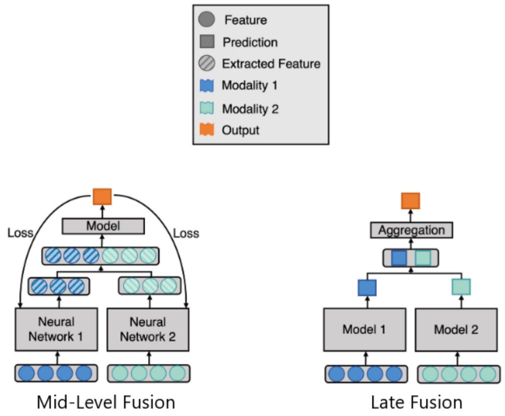

Figure 7. Fusion modalities architectures. Mid-level fusion on left side, late fusion on right side [6]

5. Experiments

This section is meant to explain in detail all the experiments performed, along with the respective results. The first metric of our interest is the top-1 accuracy, followed by the top-5 accuracy in case of deeper analysis, which are respectively the accuracy when guessing the correct label with one single guess and within the first five most likely predictions.

5.1. Sampling techniques

We tested both dense and uniform sampling, with different number of frames per clip (5, 10, 16, 25 frames). In both cases the number of clips extracted remained 5 for each action sample. Dense sampling adopts a stride of 2.

5.1.1 Feature visualization

After the extraction, we can use a visual analysis representing the extracted feature, to eventually spot the presence of a visual pattern that drives the extraction.
We first tried to apply K-Means to the extracted features (dense sampling, 16 frames per clip) and plotted the samples with the help of PCA as it is shown in Figure 8. We also added a symbol for each true centroid of each class. It can be seen that, for some few cases, the centroids almost represent the center of a cluster generated by K-Means. For example the upside down triangle on the right (which represent the class that comprehend the actions of whisk, mix-around, stir and mix) well defines the center of the green points. Also the standard triangle on the top (rinse, wipe-of, wash, clean, wipe) is leaning towards the blue points. This could mean that these classes are well separated from the others and could be potentially easily recognized by the classifier. On the other hand some other classes, such as the rhombus, the hexagon and the circle are very close to each other and don't fully match with any cluster or a combination of them, leaving the possibility to be an hard job for the classifier.

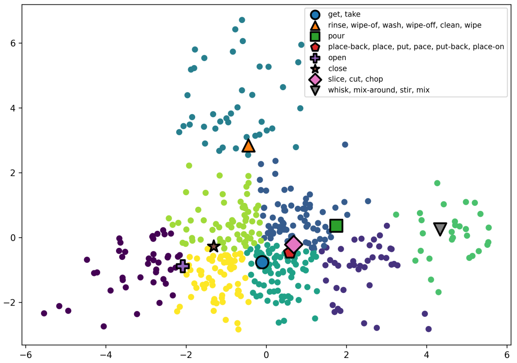

Figure 8. Extracted features using dense sampling, with 16 frames per clip, clustered with K-Means, after applying PCA with 2 dimensions. The 2d-shapes represents instead the real centroids of each class.

In Figure 9 instead, we show the result of t-SNE, which is another algorithm for dimensionality reduction. The test samples, are extracted with dense sampling and 16 frames per clip. For each sample we show the central frame of the action. It is possible to spot that the shape of the scene influences the position of the sample. In fact, on the right side is easy to identify the presence of an area where the pan is the central element, meanwhile on the bottom left

Sampling technique	Frames per clip	MLP		RNN		LSTM	
		Dropout	Top-1 Accuracy	Dropout	Top-1 Accuracy	Dropout	Top-1 Accuracy
Dense	5	0.7	55.63	0.7	54.48	0.7	54.48
	10	0.5	59.54	0.5	59.08	0.7	59.54
	16	0.7	**60.00**	0.7	58.62	0.5	**60.23**
	25	0.7	57.01	0.7	55.40	0.7	56.55
Uniform	5	0.7	54.48	0.7	50.80	0	51.95
	10	0	54.94	0	54.71	0.5	54.25
	16	0.5	**60.00**	0.5	**59.31**	0.5	58.39
	25	0.5	57.70	0.7	57.47	0.7	57.70

Table 1. Sampling techniques comparison for MLP, RNN and LSTM models, using Top-1 Accuracy metric

corner, the sink is the key subject. In this particular case, the different scenes represented in the frames can indicate different kind of actions - classes. The clear separation can mean that scenes with pans or sinks as the central element can be easily recognized by the classifier.

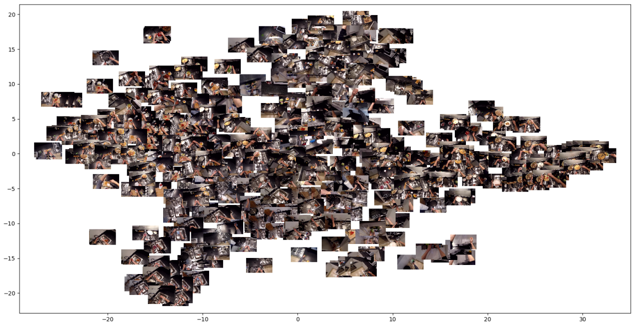

Figure 9. Extracted features using dense sampling, with 16 frames per clip, clustered with t-SNE. Each sample is represented by the central frame of the action

5.2. Classification on Epic-Kitchens

The extracted features coming from the previously mentioned sampling techniques are input of an MLP, RNN and an LSTM model. Each model is tested with different dropout rates (0, 0.5 and 0.7) to eventually spot some differences. In all the models the learning rate is fixed at 0.01 and is updated every 3000 epochs.
The training is composed of 5000 epochs, with an evaluation frequency of 50 epochs, the stochastic gradient descent momentum set at 0.9 and a weight decay of 10^{-7}.

In Table 1 we report the results obtained from the training of the MLP, the RNN with 1 layer and the LSTM with 1 layer for the different sampling techniques. For each row we only show the dropout that concerns the best result obtained among the different tested values. The best results are highlighted for each model.
It can be spotted that the best result is achieved by the LSTM model with dense sampling and 16 frames per clip, reaching a top-1 accuracy of 60.23%, followed by the MLP, with the same sampling technique, that stops at 60%. It is to mention that also uniform sampling with 16 frames per clip and MLP as a classifier gives 60% of accuracy, however from now on we will consider the dense sampling with 16 frames per clip the best technique together with the LSTM model, which will be also used for the RGB video of the ActionSense dataset.

	Sub-sampling	
	Top-1 Accuracy	
LSTM	5s	10s
1 Layer	**56.97**	48.75
2 Layer	35.45	54.46

Table 2. Top-1 accuracy for different sub-sampling for EMG modality

5.3. Multimodality

For what concerns the multimodal part, we firstly trained each model independently for EMG, RGB and spectrograms data and then use them jointly to analyze multimodal tenchniques. We expect in this section to show that fusion models can help the classification.

5.3.1 RGB

For the **RGB** modality, the LSTM model has been trained for 5000 epochs, with a batch size of 32 and a learning rate of 0.1. As previously said, the chosen sampling strategy is dense sampling with 16 frames per clip. With this configuration, we managed to reach a top-1 accuracy of 78.57%.

5.3.2 EMG

For what concerns the **EMG** modality, the LSTM is trained for 5000 epochs, again with a batch size of 32 but a learning rate of 0.2. Considering 5s long actions when augmenting the dataset and the 2-layer LSTM we obtained an accuracy of just 35.45%, so we decided to test also a single layer architecture. Passing from the 2-layer LSTM to the single layer one, we managed to increase the top-1 accuracy of around 20%, reaching an accuracy of 56.97%. As shown in Table 2, we also tried to augment the dataset dividing each sample in sub-samples of 10s each, which gave a slightly different behavior. Using a single layer LSTM we achieve an accuracy of 48.75% that, this time, is worse with respect to the 2-layer architecture, that on the other hand reaches 54.46%. Given those results we chose to proceed the analysis with the augmentation of sub-samples long 5s each and the employment of a single-layer LSTM.

5.3.3 Spectograms

Using **spectrograms**, obtained from the EMG signals, we managed to reach a 56.97% top-1 accuracy. This time the training epochs where again 5000, the batch size 32, but the learning rate was set to 0.005.

5.3.4 Fusion

Finally we chose the best performing models and used them for the fusion models. We start analyzing the late fusion, where we combined all the modalities in different combinations and with different weights. In table 3 we report the weights that lead to the best results. We noticed that the best top-1 accuracy never goes beyond the one achieved by RGB modality alone. It can also be seen that, as far as the combination of RGB with EMG, the best weights do not take in consideration EMG at all. Meanwhile the combination of EMG and spectrograms produces an improvement with respect to the two single modalities taken alone, and we even reached a top-5 accuracy of 90%. We also tested the mid-level fusion, again with all the combination between the 3 considered modalities. In this case we obtained a peak accuracy of 80.92% with the combination of RGB and EMG. The results are in line with our expectation, with the mid-level fusion increasing the performances when using two or more modalities combined. We noticed that during the training, the considered models tend to overfit, leaving us with potentially reduced performances. We also tried different learning rates to try to escape the local optimality, but with no improvements at all.

In Table 3 we report just the best result achieved, obtained with the models with the parameters as they where in the first place.

Modality	Model	Accuracy %	
		Top-1	Top-5
RGB	**LSTM**	**78.57**	**89.29**
EMG	LSTM-1L	56.97	87.81
Spec.	LeNet5	54.50	88.89
EMG + RGB	Mid-Level Fusion	**80.92**	**86.18**
EMG + Spec.		57.24	66.45
RGB + Spec.		79.61	86.84
All		80.26	86.18
EMG + RGB (0.0, 1.0)	Late Fusion	78.57	86.18
EMG + Spec. (0.58, 0.42)		63.82	90.79
RGB + Spec. (0.32, 0.68)		**78.57**	**88.82**
All (0.0, 0.32, 0.68)		78.57	88.82

Table 3. Performances recorded in different modalities using the ActionSense database.

6. Conclusions

In this work we investigated the egocentric action recognition task for activities performed in a kitchen environment.
In the first part we explored the trade-offs between capturing spatial and temporal information by comparing dense and uniform sampling techniques for RGB video frames in the Epic-Kitchens dataset. We have shown that dense sampling is the most promising technique, with 16 frames per clip to be the best option. This demonstrates that spatial information is mostly preferred by a classifier when recognizing kitchen actions. As far as the classifier the study shows that LSTM and MLP are slightly better than RNN and gives top-1 accuracies respectively of 60.23% and 60.00% when using the mentioned sampling method.

Whereas in the second part we focused on the ActionSense dataset, which incorporates additional data modalities like EMG sensors. We developed separate models for RGB video frames using an single layer LSTM, preprocessed EMG signals fed into an single and double layer LSTM network, and spectrograms of EMG signals classified by a LeNet5 CNN. The individual models achieved top-1 accuracies of 82.24% (RGB), 60.72% (1-layer LSTM for EMG), and 56.97% (spectrograms).

Finally, we explored mid-level and late fusion approaches that combined the best performing models of each modality alone. Late fusion did not increase the performances overall, but kept the best scores already reached without fusion, probably due to the oversimplified combination of logits. Mid-Level Fusion instead, demonstrates the benefit of incorporating complementary information sources for egocentric action recognition, with EMG+RGB and RGB+EMG+spectrograms increasing the top-1 accuracies by 2% with respect to RGB modality alone, reaching 80.92% and 80.26% respectively. For future studies, we suggest to incorporate multiple subjects for the RGB videos, as well as additional sources of data such as audio and eye-tracking cameras, that might reduce overfitting and lead to an increase of performances with respect to the ones described in this paper.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 1, 3

[2] Chun-Fu Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition, 2021. 2

[3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[4] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. ActionSense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 3, 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[6] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136, 2020. 5

[7] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition, 2019. 1, 2

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998. 4

[9] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[10] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[11] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation, 2022. 2

[12] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E^2(go)motion: Motion augmented event stream for egocentric action recognition, 2022. 2