Multimodal Action Recognition

Federico Buccellato
Politecnico di Torino
s309075@studenti.polito.it

Raffaele Viola
Politecnico di Torino
s309063@studenti.polito.it

Nunzio Messineo
Politecnico di Torino
s315067@studenti.polito.it

***Abstract*— The rapid advancement of wearable technology and egocentric vision systems have opened new opportunities for understanding human behavior through automatic action recognition. This study focuses on the analysis of RGB video data from the EPIC-Kitchens dataset [1] and the integration of RGB video data and electromyography (EMG) signals from the ActionNet dataset [2], with the goal of developing robust models for action recognition in everyday environments.**

In our approach, the RGB data from the EPIC-Kitchens and ActionNet dataset are preprocessed by segmenting the videos into clips of fixed length and applying dense and uniform sampling strategies to select a K-number of frames [5-10-25]. For the ActionNet dataset, we utilize both RGB data and EMG signals. The EMG signals are filtered, normalized, resampled, and combined from both forearms. Additionally, we generate spectrograms from the EMG signals to enhance feature extraction and analysis. We trained and evaluated several models, including the MultiScaleTRN, LSTM, and MLP [3][4][5] classifiers, using different configurations of the preprocessed data. For the integration of multimodal data in the context of ActionNet, we implemented both late fusion and middle fusion techniques, analyzing their impact on model performance.

Github code: github.com/RaffaeleViola/aml23-ego.git

Keywords—Action Recognition, Egocentric Vision, Multimodal Data, Electromyography (EMG), RGB Video Analysis, Preprocessing, Feature Extraction, Late Fusion, Middle Fusion, Spectrograms, EPIC-Kitchens, ActionNet, Human Activity Recognition.

I. Introduction

The growing prevalence of wearable technology and egocentric vision systems has created new possibilities for automated human behavior analysis, particularly through action recognition in everyday environments. These systems provide a first-person perspective and collect sensory data that prove particularly valuable in applications such as activity monitoring and assistive technologies. However, recognizing actions from multimodal data of this nature presents complex challenges due to the dynamic and intricate nature of human activities.

Egocentric vision, which captures the scene from the first-person viewpoint, offers a unique perspective closely aligned with the user's attention and actions. This approach is particularly effective for analyzing daily activities, characterized by frequent and varied interactions with objects and environments. However, thanks to the temporal dynamics present in video data and the multimodal nature of wearable sensors, action recognition is expanding new horizons, allowing for the exploration of innovative solutions to the problem. Both spatial and temporal information are essential for accurately interpreting video content, requiring advanced techniques capable of effectively integrating these dimensions. In this study, we utilize existing models and datasets to tackle the challenge of action recognition in egocentric vision. We focus on two benchmark datasets: EPIC-Kitchens and ActionNet. EPIC-Kitchens provides a vast collection of egocentric video data in kitchen environments, while ActionNet extends this perspective by incorporating a multimodal methodology that includes data from wearable sensors such as EMG and first-person cameras. Our approach is divided into two main parts. In the first part, we focus on extracting features from the datasets to isolate the most useful information for action recognition. Once we have obtained these features, we move on to the second part, where we experiment with using networks such as LSTM, TRN, or MLP, applying them to both single modalities and combinations of multiple modalities. This allows us to determine which approach is most effective for action recognition, exploring various strategies to optimize the integration of multimodal information.

II. Related works

Action recognition in videos has made significant progress thanks to the introduction of large-scale video datasets such as Kinetics [6], Sports1M [7], and Moments-In-Time [8]. These datasets have provided a solid foundation for the development of advanced models capable of efficiently capturing the complex spatiotemporal information present in videos, leading to a series of innovations in the field of deep learning applied to computer vision.

A. Approaches for Egocentric Action Recognition

One the most important work for action recognition was realized by Kazakos [9]. This work introduces a novel multimodal architecture for egocentric action recognition, fusing RGB, Flow, and Audio modalities through mid-level fusion before temporal aggregation. Trained end-to-end, the approach outperforms individual modalities and late fusion methods, demonstrating the importance of audio for action and object recognition. The method achieves state-of-the-art results on the EPIC-Kitchens dataset across seen and unseen test sets.

Another important work in the field of action recognition is made by Plizzari [10]. It explores the potential of event cameras, bio-inspired sensors that capture pixel-level intensity changes with high temporal resolution and minimal motion blur, for egocentric action recognition. Despite their advantages—such as low power consumption and resistance to motion blur—event cameras have been underutilized in this

field. The authors introduce N-EPIC-Kitchens, the first event-camera extension of the EPIC-Kitchens dataset and propose two strategies: processing event data with traditional video architectures and using it to distill optical flow information. The study shows that event data performs comparably to RGB and optical flow, improving performance by up to 4% over RGB alone.

B. 3D CNN Approaches

In parallel the use of 3D CNNs has gained popularity as a powerful alternative approach for action recognition in videos. Chen work [11] presents an in-depth comparative analysis of 2D and 3D convolutional neural networks (CNNs) for video action recognition, using a unified framework to ensure fair comparison across models. The study analyzes over 300 action recognition models, revealing that while significant progress has been made in efficiency, accuracy improvements have lagged. Additionally, the analysis shows that both 2D and 3D CNN models exhibit similar spatio-temporal representation capabilities and transferability.

C. Domain Adaptation Problem

Munro and Damen [12] addresses environmental bias in fine-grained action recognition datasets, where domain shifts between training and deployment environments lead to performance drops. Traditional Unsupervised Domain Adaptation (UDA) approaches use adversarial training but overlook the multi-modal nature of video data. This study introduces a self-supervised alignment method that leverages the correspondence between RGB and Optical Flow modalities, alongside adversarial alignment, for UDA. Tested on the EPIC-Kitchens dataset, the proposed approach improves performance by 2.4% over source-only training and outperforms other UDA methods by 3%.

D. Leveraging Human Videos for Robotic Learning

Finally, the use of human videos for robotic learning has gained increasing interest. Nair [13] explores how visual representations pre-trained on diverse human video data can enhance data-efficient learning for robotic manipulation tasks. The authors pre-train a visual representation, R3M, using the Ego4D dataset, combining time-contrastive learning, video-language alignment, and an L1 penalty for sparse representations. R3M serves as a frozen perception module for downstream policy learning. Across 12 simulated robot manipulation tasks, R3M improves task success by over 20% compared to training from scratch and outperforms visual models like CLIP and MoCo by over 10%. Additionally, R3M enables a robotic arm to learn manipulation tasks with just 20 demonstrations.

III. DATA PREPROCESSING

In any machine learning project, especially in the context of action recognition, data preprocessing is a crucial step that significantly impacts the model's performance. This chapter details the preprocessing and analysis methods applied to the different types of data used in this study: RGB video data, EMG signals, and their transformation into spectrograms. Each type of data requires specific handling to ensure that the extracted features are both meaningful and conducive to effective learning, while also reducing the overall data size to improve the model's efficiency.

A. RGB Data Preprocessing

RGB data preprocessing is a fundamental step to ensure that the action recognition model can effectively learn from the spatial and temporal information present in the videos. This process has been applied to both the EPIC-Kitchens (EK) dataset and the ActionNet dataset, adopting a systematic approach that includes segmenting the videos into clips and applying different sampling strategies.
We explore three different sampling sizes: 5, 10, and 25 frames. This variation in frames number allows us to better understand how different amounts of temporal information affect the model's performance. Shorter sizes, composed of 5 frames, are useful for analyzing rapid actions or specific movement segments, while longer clips, up to 25 frames, provide a broader view of the actions, capturing more temporal information. For each clip, we employ two different sampling strategies: uniform sampling and dense sampling.

- Uniform Sampling: this strategy involves selecting frames that are evenly distributed throughout the entire clip. Uniform sampling is designed to capture a complete and consistent representation of the clip, providing an overview of the temporal dynamics of the action. For example, in a clip composed of 25 frames, uniform sampling selects evenly spaced frames, allowing the model to capture the progressive changes in the action as it unfolds over time.

- Dense Sampling: this strategy focuses on selecting consecutive or nearly consecutive frames with a small stride. Dense sampling is particularly useful for capturing spatial details and rapid variations in the visual appearance, as the selected frames are temporally close to each other. This strategy is ideal for analyzing detailed movements or fast actions that require special attention to spatial continuity.

Once the videos are segmented in clips and the frames are sampled, we use a pretrained model to effectively extract features from the selected frames.
The information, extracted from a pre-trained I3D network model with an Inception-V1 backbone, represents compressed representations of the video data.
These representations contain the essential information needed for action recognition and are stored in a reduced-dimensional space, making the subsequent classification phase faster and more efficient. This process ensures that the model can accurately identify and distinguish actions while optimizing the use of computational resources.
Given C as the number of clips, the output produced by the backbone network is C×1024. This output size corresponds to the dimensionality of the feature representation from the last layer of the backbone, just prior to the layers that are dedicated to the classifier.

By combining the segmentation of videos into clips with both dense and uniform sampling strategies and using pre-trained model to extract and compress key features from the video data, the RGB data preprocessing in the EPIC-Kitchens and ActionNet datasets allows us to assess how each technique impacts the model's ability to learn from spatial and temporal variations. This approach is crucial for identifying which method provides the most effective representation of actions, thereby improving the model's accuracy in recognizing and distinguishing various activities in the videos.

B. EMG Data Preprocessing

Among the various modalities used for action classification, EMG data from the ActionNet dataset are also employed. The process begins with loading the raw EMG data from the corresponding files. To remove high-frequency noise and preserve the useful components of the signal, a low-pass filter with a cutoff frequency of 5 Hz is applied. This filtering is essential to reduce interference that could distort the muscle signals, thereby improving the overall quality of the data.

After filtering, the EMG signals are normalized to ensure they are scaled within a common range (between -1 and 1). Normalization is a crucial step to prevent signal intensity variations from disproportionately affecting the model's performance. This process makes the signals more homogeneous and comparable, both across different recording sessions and between different subjects.

Subsequently, the EMG data are resampled to standardize the sampling frequency to a target value of 10 Hz. Resampling is necessary to ensure temporal consistency in the data, regardless of any variations that may have occurred during recording. This temporal consistency facilitates learning, as the signals are uniformly sampled.

Finally, to increase the amount of data in the dataset, the EMG signals are segmented into 10-second clips. During this segmentation process, the signals from both the right and left forearms are combined to create a unified representation of the muscle activity that provides a more comprehensive view of the actions being performed.

C. Spectrogram Data Preprocessing

To gain a deeper understanding of the EMG signals, we go beyond basic processing by converting these signals into spectrograms. Spectrograms provide a detailed visual representation of how the frequency content of the signal evolves over time, allowing us to capture both temporal and frequency-based information. This advanced perspective could prove useful in identifying interesting patterns in muscle activity and enhancing action recognition. The preprocessing begins by dividing the filtered, normalized, and resampled EMG signals into small time segments. For each of these segments, we apply a method that reveals how the different frequencies in the signal change over time, thus generating a spectrogram.

After creating the spectrograms, we ensure they are consistent across all samples by normalizing them.

These spectrograms, along with their corresponding action labels, are then ready for analysis, providing a richer dataset that captures the complexity of muscle activity over time and that could facilitate the recognition and classification of actions.

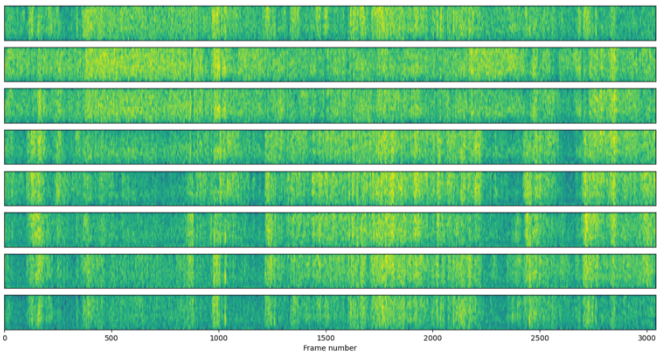

Figure 1 – Representation of a spectrogram

IV. Data analysis

The data analysis phase is focused on discovering consistent patterns within the features that are extracted during the preprocessing stage. This involves examining the relationships and structures present in the data to identify any recurring trends or similarities. The goal is to understand how these features are related to each other, how they might group together, and whether certain characteristics are consistently linked to specific actions or labels. By identifying these patterns, the analysis aims to gain deeper insights into the underlying structure of data and improve the effectiveness of subsequent classification tasks. To achieve a clearer representation of the data, the features from each video clip are condensed into a single one-dimensional vector by averaging the values across the feature set. This reduces the complexity of the data while preserving the essential information. The resulting 1D vectors are then visualized in a lower-dimensional space using UMAP (Uniform Manifold Approximation and Projection) [14]. UMAP is chosen because it excels at maintaining both the local and global structures of the dataset, meaning it captures the relationships between points more effectively than other methods like t-SNE [15].

In the analysis, UMAP is used to project the high-dimensional feature vectors into a two-dimensional space. Each point in this UMAP visualization represents a video clip, and points are colored based on their corresponding verb class, allowing for a visual comparison of how different actions are distributed in the feature space. Next, K-Means clustering [16] is applied to these reduced-dimension features to identify groups or clusters of similar actions. The resulting clusters are plotted alongside the original verb class labels, allowing for a comparison between the true labels and the clusters identified by the algorithm. This process helps to assess how well the clustering aligned with the actual verb classes, revealing whether similar actions are grouped together or if there is significant overlap among different actions.

However, the results are not ideal. Features from clips with similar actions tends to overlap, making it difficult to achieve clear separation of points. In the case of features derived from RGB, particularly for EK, similarities arise

from actions like "take" and "put-down" or "open" and "close", as illustrated in Figure 2. Additionally, different actions performed in the same location within the kitchen contribute to these overlaps. This is evident in the plots where the central frames of the clips are displayed instead of colored points, as shown in Figure 3. In the AN dataset with RGB features, where labels also specify objects, overlaps are noticed in different actions involving the same objects. In contrast, EMG features generally show better distinction between points, though issues with overlapping labels persist. The analysis concluded that clustering algorithms alone are insufficient for effectively and automatically classifying the data from the extracted features. This is primarily due to the presence of multiple actions with overlapping characteristics. As a result, additional steps and techniques were necessary to enhance the classification process and improve accuracy.

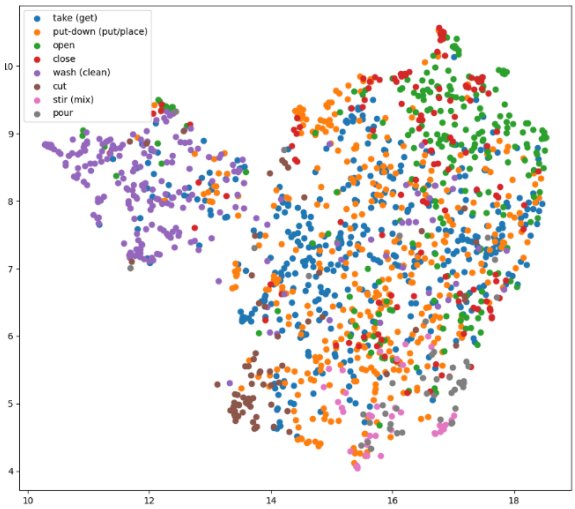

Figure 2 - UMAP Visualization with labels for EK 5 frames – Uniform sampling

V. Models Description

We begin by exploring different models and techniques to find an effective approach for egocentric action recognition using RGB data. Afterward, we investigate the ElectroMyoGraphy (EMG) modality, introducing also a multimodal approach trying to improve performance.

A. RGB Modality

Using the RGB features extracted from the EpicKitchen dataset, as described in Section III.A, we train and evaluate three classifiers: MLP, LSTM, and TRN [17].
We begin with a simple approach, starting with the MLP due to its faster training and ease of implementation compared to the other networks. Before passing the data into the classifier, we aggregate them along the temporal axis using a pooling layer.
Next, we move to the LSTM network, leveraging its ability to process and learn from sequences of data.
Finally, we analyze the Temporal Relation Network (TRN), aiming to utilize its strength in reasoning about temporal dependencies between video frames at multiple time scales.
Given TRN better performances compared to the previous models, we train only this model on the ActionNet dataset for the second part of our analysis.

B. EMG Modality

For the EMG modality, we employ a model based on an LSTM recurrent neural network, as shown in Figure 4. The network takes EMG-extracted data with a shape of 100 x16 as input and applies two LSTM layers with hidden sizes of 5 and 50, respectively. The final output is passed through a dropout layer with a probability of 20% before reaching the final fully connected (FC) layer. Similar to the approach used for the audio modality in [18], we also train a 2D CNN on the spectrogram data obtained as described in Section III.C. We select SqueezeNet as the 2D CNN due to its low computational cost, allowing us to quickly validate our approach.

C. Multimodal Model

Integrating data from various modalities can greatly enhance model performance. As the final step of our work, we train a multimodal classifier that leverages the previously described modalities. For the RGB modality, we use TRN, the best-performing model. Initially, we apply late fusion as the fusion strategy (Figure 5). Next, we explore the impact of mid-level fusion (Figure 6), which is performed by taking the features from both networks before their final FC layers, concatenating them, and passing them to a fully connected (FC) layer for the actual fusion, followed by the final FC layer.

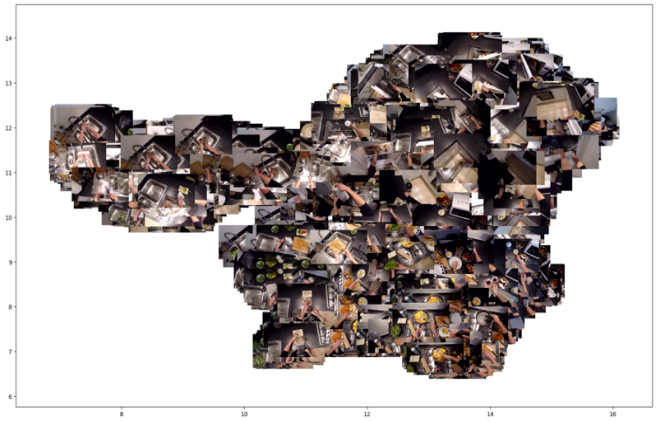

Figure 3 - UMAP visualization with frames for EK 5 frames – Dense sampling

Figure 4 – LSTM based network for EMG data

VI. Experiments

In this section, we first present the results obtained from training the models on the RGB modality using the EPIC-Kitchens dataset. Following this, we showcase the outcomes from the ActionNet dataset, starting with the RGB modality. We then present the results from the EMG modality, including the EMG spectrograms and, finally, the performance of the multimodal fusion. Specifically, we analyze the outcomes of both middle fusion and late fusion strategies.

A. Parameters

We train both the RGB and EMG models using the same set of parameters. The training is conducted for up to 3000 epochs with the Stochastic Gradient Descent (SGD) optimizer, using a momentum of 0.9, a learning rate of 0.01, and a weight decay of 10^{-7}. The batch size is set to 32 for both modalities, ensuring consistent training conditions.

B. Epic Kitchens RGB

We trained each model described in Section V.A using both dense and uniform sampling strategies, evaluating them across samplings of 5, 10, and 25 frames. The MLP and TRN models are assessed with dropout rates of 0.2 and 0.5, while the LSTM models are tested with both 1 and 2 layers. The results for each model are presented in Tables 1, 2, and 3. For the LSTM models, dense sampling consistently outperforms uniform sampling. The 1-layer LSTM configuration yields better performances across all frame lengths compared to the 2-layer variant, with the highest accuracy achieved using 10 frames and dense sampling, as shown in Table 1. In the MLP models, a dropout rate of 0.2 consistently resulted in higher accuracy compared to 0.5. Dense sampling also provided better results than uniform sampling, particularly with fewer frames, as reported in Table 3. Similarly, in TRN models, a dropout rate of 0.2 offers superior performance, and dense sampling again outperforms uniform sampling across the board (Table 2).

SAMPLING	FRAMES	NUM LAYERS	BEST TOP1 (%)
DENSE	5	1	56.55
		2	53.23
	10	1	57.47
		2	55.16
	25	1	56.55
		2	54.03
UNIFORM	5	1	50.80
		2	49.74
	10	1	54.94
		2	54.48
	25	1	54.94
		2	55.40

Table 1 - Results EpicKitchens dataset with LSTM

The best performance for TRN is observed with 10 frames and dense sampling, achieving the highest accuracy overall. In general, dense sampling consistently provides more useful temporal information, leading to better model performance compared to uniform sampling across all tested architectures (LSTM, MLP, TRN). Increasing the number of frames from 5 to 10 generally improves accuracy, particularly in the TRN model. However, increasing the frame count further to 25 results in a slight drop in performance, indicating diminishing returns beyond 10 frames.

SAMPLING	FRAMES	DROPOUT	ACCURACY (%)
DENSE	5	0.2	60.09
		0.5	59.31
	10	0.2	64.62
		0.5	63.45
	25	0.2	59.85
		0.5	60.02
UNIFORM	5	0.2	59.31
		0.5	50.80
	10	0.2	62.76
		0.5	54.94
	25	0.2	60.00
		0.5	57.93

Table 2 - Results EpicKitchens dataset with TRN

SAMPLING	FRAMES	DROPOUT	ACCURACY (%)
DENSE	5	0.2	57,91
		0.5	56.78
	10	0.2	57.31
		0.5	57.24
	25	0.2	56.40
		0.5	54.71
UNIFORM	5	0.2	51.72
		0.5	50.57
	10	0.2	55.17
		0.5	52.45
	25	0.2	56.09
		0.5	54.65

Table 3 - Results EpicKitchens dataset with MLP

C. ActionNet RGB

Based on the results from Section VI.B, we selected the top-performing model, the TRN, for further evaluation on the ActionNet dataset. The model is trained using dense sampling with K set to 5, 10, and 25 frames to assess how temporal variation impacts performance across different frame counts.

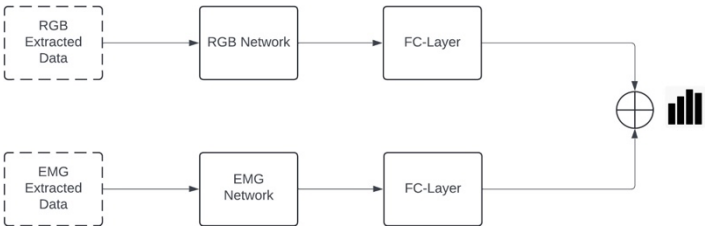

Figure 5 – Late fusion strategy

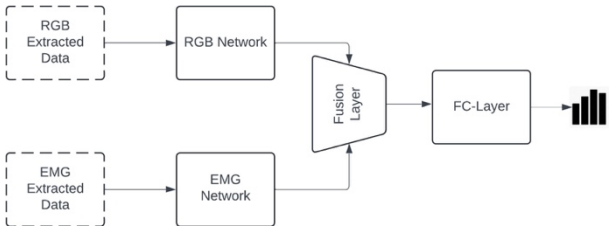

Figure 6 – Middle fusion strategy

This decision is made to leverage the strengths of the TRN in capturing spatiotemporal patterns effectively, as evidenced by its superior performance on the EPIC-Kitchens dataset. As shown in Table 4, the TRN's performance on the ActionNet dataset, while respectable, did not reach the accuracy levels achieved on the EPIC-Kitchens dataset. This difference could be attributed to the smaller portion of the RGB data used in ActionNet. However, similar to our findings in Section VI.B, using 10 frames once again emerged as the optimal configuration for this task, outperforming both the 5 and 25 frame setups. The 10-frame configuration likely provides a balance between capturing sufficient temporal information and avoiding overfitting or redundancy, which can occur with higher frame counts.

Modality	Frames	Accuracy (%)
RGB	5	37.83
	10	41.56
	25	38.04

Table 4 - Results ActionNet dataset RGB Modality TRN network

D. EMG and Multimodal analysis

In this section, we present the performance results of models trained using EMG data, spectrograms, and multimodal fusion approaches, as summarized in Table 5. We first evaluate the LSTM model using EMG data from S04 and the broader S00-S09 dataset. The LSTM trained on the S04 subset achieves an accuracy of 47.67%, while training on the larger S00-S09 dataset results in a higher accuracy of 52.77%. This indicates that expanding the training data to include multiple subjects (S00-S09) helps the model capture more generalizable features, leading to improved performance. To explore the potential of 2D convolutional neural networks for EMG data, we converted the EMG signals into spectrograms and trained a SqueezeNet model. However, the accuracy achieved using the EMG spectrograms was 36.26%, which was lower than the LSTM model. This suggests that LSTM, which is designed to capture sequential data, is better suited for handling the temporal dynamics present in the EMG signals compared to a 2D CNN. We then investigate the benefits of combining RGB and EMG data through multimodal fusion techniques. Two fusion strategies are employed: late fusion and mid-fusion. In the late fusion approach, we combine the predictions of the TRN model, trained on RGB data, with those of the LSTM model, trained on EMG data. This results in an accuracy of 56.05%, significantly improving upon the individual EMG model's performance. In mid-fusion, features from both the TRN and LSTM models are combined earlier in the network, before classification, yielding an accuracy of 55.67%. Although slightly lower than the late fusion method, mid-fusion still demonstrates the value of integrating RGB and EMG features. Overall, these results highlight the advantage of multimodal approaches, where combining RGB and EMG data provides a notable performance boost compared to using either modality alone. The late fusion method produces the highest accuracy, indicating that integrating modalities at the decision level allows for more effective utilization of complementary features from both data types.

Modality	Model	Accuracy (%)
EMG(S04)	LSTM	47.67
EMG(S00-S09)	LSTM	52.77
EMG 2D	SqueezeNet	36.26
RGB + EMG(late fusion) †	TRN + LSTM	56.05
RGB + EMG(mid-fusion) †	TRN + LSTM	55.67

Table 5 – Results ActionNet dataset EMG, Spectrogram and multimodalities (†=train only on S04)

VII. Conclusion

In this work, we analyze the impact of different sampling strategies and classifiers on the Egocentric Action Recognition task. One of the key findings is that using a pretrained network to extract representative RGB features is crucial for reducing the dimensionality of the input data, thus facilitating the training process. When combined with an appropriate sampling strategy and classifiers like TRN, this approach leads to promising results on the EPIC-Kitchens dataset. Moreover, our analysis of the results on the ActionNet dataset reveals that EMG data, when combined with RGB modality, can significantly enhance model performance, resulting in improved accuracy for multimodal

models. Although we use only a small portion of the RGB dataset, the results are still promising. As a potential future direction, a full evaluation of the ActionNet dataset could provide further insights into the model's capabilities when using the entire dataset. Future work could also explore the integration of additional modalities beyond EMG to assess their impact on the overall model performance. This could further enhance the robustness and accuracy of action recognition systems in real-world applications.

REFERENCES

[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, and Will Price. Scaling egocentric vision: The EPIC-Kitchens dataset. In European Conference on Computer Vision (ECCV), 2018. 1, 2, 3, 5

[2] DelPreto, J., Liu, C., Luo, Y., Foshey, M., Li, Y., Torralba, A., Matusik, W., and Rus, D. "ActionNet: A multimodal dataset for human activities using wearable sensors in a kitchen environment." MIT CSAIL, Cambridge, MA, 2019.

[3] Dominey, P. F., and Ramus, F. "Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant." LANGUAGE AND COGNITIVE PROCESSES", 2000.

[4] Sepp Hochreiter and Ju ̈rgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. 4

[5] G. Singh and M. Sachan, "Multi-layer perceptron (MLP) neural network technique for offline handwritten Gurmukhi character recognition," *2014 IEEE International Conference on Computational Intelligence and Computing Research*, Coimbatore, India, 2014, pp. 1-5, doi: 10.1109/ICCIC.2014.7238334.

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv:1705.06950, 2017. 1, 2, 13

[7] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Computer Vision and Pattern Recognition (CVPR), 2014.

[8] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. IEEE TPAMI, 2019. 1, 2, 7, 13

[9] EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition - Kazakos, Evangelos and Nagrani, Arsha and Zisserman, Andrew and Damen, Dima, ICCV 2019.

[10] E(GO)^2MOTION: Motion Augmented Event Stream for Egocentric Action Recognition - Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, Barbara Caputo, 2021

[11] Deep Analysis of CNN-based Spatio-temporal Representations for Action Recognition - Chun-Fu Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, Quanfu Fan CVPR2021

[12] Multi-Modal Domain Adaptation for Fine-Grained Action Recognition - Jonathan Munro, Dima Damen, CVPR 2020

[13] R3M: A Universal Visual Representation for Robot Manipulation - Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, Abhinav Gupta 2022

[14] McInnes, L., & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

[15] T. Tony Cai, Rong Ma (2021). Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data.

[16] Wu, J. (2012). Advances in K-means clustering: a data mining thinking. Springer Science & Business Media.

[17] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European conference on computer vision (ECCV), pages 305–321, 2018. 2

[18] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European conference on computer vision (ECCV), pages 305–321, 2018. 2