Integrating RGB and sEMG Signals for Egocentric Action Recognition: A Multimodal Approach

Jacopo Bracci
Polytechnic of Turin
Turin, Italy
s314674@studenti.polito.it

Alessandro Caruso
Polytechnic of Turin
Turin, Italy
s317028@studenti.polito.it

Simone Clemente
Polytechnic of Turin
Turin, Italy
s309291@studenti.polito.it

Abstract

The Inflated 3D ConvNet (I3D) has shown effectiveness in egocentric action recognition due to its capability of capturing spatio-temporal features. However, the computational demand of this model poses challenges to its deployment. Therefore, it is crucial to explore alternative approaches that can reduce computational cost while maintaining satisfactory performance. The proposed approach involves utilizing a pre-trained I3D model as a backbone to extract RGB features, eliminating the need to train the network, and enhancing the results using Surface Electromyography (sEMG) signals in a multi-modal classification framework. This multi-modal classification is carried out by exploring both a late-fusion and mid-fusion approach, and finally by implementing an attention module. Our approach achieves a low computation cost while yielding promising results.

1. Introduction

In the modern era of technology, the growing popularity of affordable and lightweight wearable cameras has led to a remarkable increase in the creation and sharing of videos capturing first-person (egocentric) perspectives. Additionally, the progress in technology led to the development of various smart accessories, including body-tracking sensors, which generate valuable and informative data. Hence, it is important to devise systems capable of both accurately and efficiently recognizing actions performed by users and captured by these devices. 3

In this work, our initial approach focuses on using visual RGB data alone for action classification. To achieve this, we use pre-trained I3D checkpoints provided to extract intermediate features for the RGB stream of two datasets (EPIC-Kitchens and ActionSense). By analyzing these features, some interesting patterns emerge. We reduce the dimensionality of RGB features by projecting them onto a

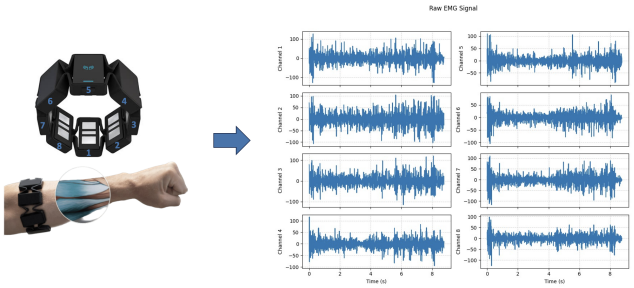

Figure 1. Visualization of 8-channel sEMG features extracted from a Myo armband on an arm.

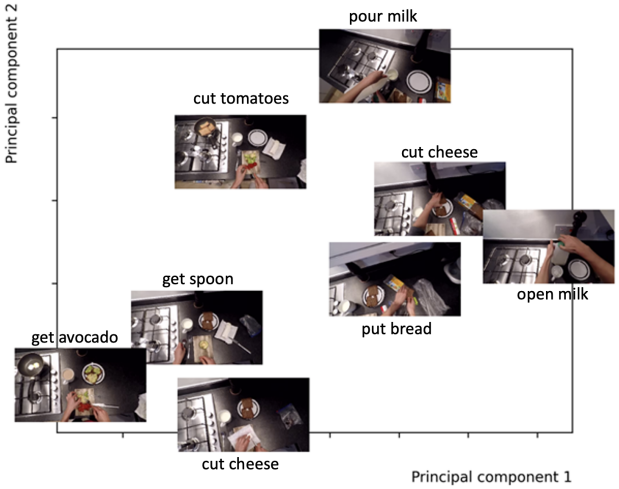

Figure 2. Visualization of the extracted features for the EPIC-Kitchens dataset using k-nearest neighbor and PCA. The central frame of each sample is used to represent the points.

2-dimensional space using PCA. Most neighborhoods of images appear to rely predominantly on visual similarities rather than other factors. For instance, as shown in Figure 2, samples containing the stove in the same position exhibit greater similarity, despite having potential variations in the

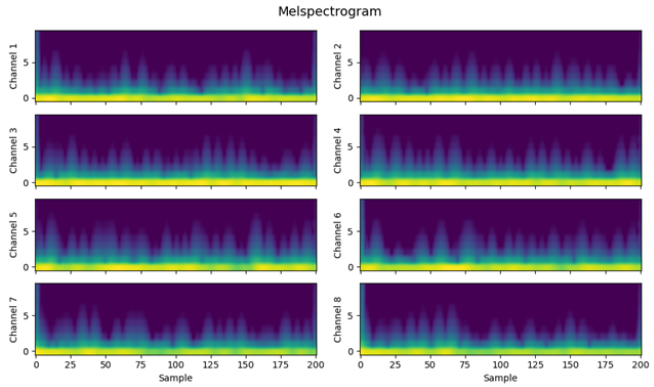

Figure 3. Visualization of the 8 Mel-spectrograms, representing each channel of a single arm, depicting a specific action.

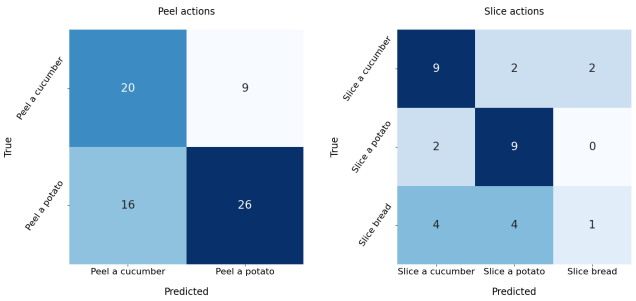

Figure 4. Confusion matrix illustrating classification performance and misclassifications of the model for "peel" and "slice" actions using a 2D CNN.

associated actions, due to the stove's prominent presence in the background. This finding highlights the limitation of relying solely on RGB features for action classification. To evaluate the performance of our RGB-only approach, we employ a linear classifier and a Long Short-Term Memory (LSTM) network. However, in order to overcome the limitations associated with using only RGB frames, we explore the incorporation of sEMG signals.

Surface Electromyography (sEMG) refers to the recording and analysis of the electrical signals generated by muscles using surface electrodes placed on the skin. It is used to gather information during contractions when flexing or extending an articulation. We first utilize the raw sEMG signal provided by ActionSense dataset. These signals were captured for both arms using a Myo Gesture Control Armband, which has 8 channels for recording sEMG data. Therefore, we have a total of 16 channels for both arms. Figure 1 illustrates the visual representation of the sEMG recording setup. We train a LSTM classifier using this sEMG data. Subsequently, we pursue a more advanced approach by utilizing spectrograms for improved representation, as depicted in Figure 3. These spectrograms are then input of a 2D CNN, which performs classification tasks and generates embeddings. As shown in Figure 4, we observe that certain actions, such as "peel a potato" or "peel a cucumber" as well as "slice a cucumber", "slice a potato" or "slice bread", which exhibit a similar pattern of muscular activity between each other, are misclassified due to the absence of visual data. The sEMG signal alone lacks the necessary information to make precise differentiations between these actions. Consequently, relying solely on the sEMG signal results in misclassifications for this type of actions.

In our work, we propose the inclusion of RGB frames alongside the sEMG signal. By incorporating visual information alongside the sEMG signals, we expect to enhance the accuracy of the classification process. We proceed with our research focusing on various approaches. Initially, we run a single modality test in order to establish a baseline. Subsequently, two separate networks are trained together, and their outputs are fused together in a late-fusion manner. Following that, we adopt a mid-fusion approach, where a single network is trained to provide unified predictions using mid-fusion features extracted by a Fusion Layer. Finally, we adopt an attention based classifier. This attention mechanism allows the model to focus in a different manner on each clip, providing a weighted average for each clip.

2. Related Works

2D CNNs. In the field of action recognition, 2D Convolutional Neural Networks (CNNs) approaches have gained significant popularity and have been widely studied [11]. These methods have notable advantages in terms of efficiency compared to their 3D counterparts. A limitation of these approaches is their inability to capture temporal order or temporal connections [7] [16].

sEMG Signal Processing with 2D CNNs. The effectiveness of 2D CNNs for sEMG signal classification has been investigated in recent literature [4]. These studies have consistently demonstrated the valuable role of pre-processing techniques in the analysis of sEMG data, emphasizing the importance of employing suitable pre-processing methods. Furthermore, the application of spectrograms in audio classification has yielded promising outcomes [3].

3D CNNs and Transfer Learning Approaches. 3D Convolutional Neural Networks (CNNs) have been acknowledged for their effectiveness in learning spatio-temporal features [12] [8]. However, the high computational complexity associated with 3D CNNs poses deployment challenges. To address these concerns, transfer learning has emerged as a promising solution [7]. Transfer learning leverages the knowledge acquired from a pretraining phase on a large dataset to improve the performance on a target dataset for a similar task. In our work, we utilize a pretrained 3D CNN model trained on the Kinetics dataset to extract RGB features for the two datasets we are using:

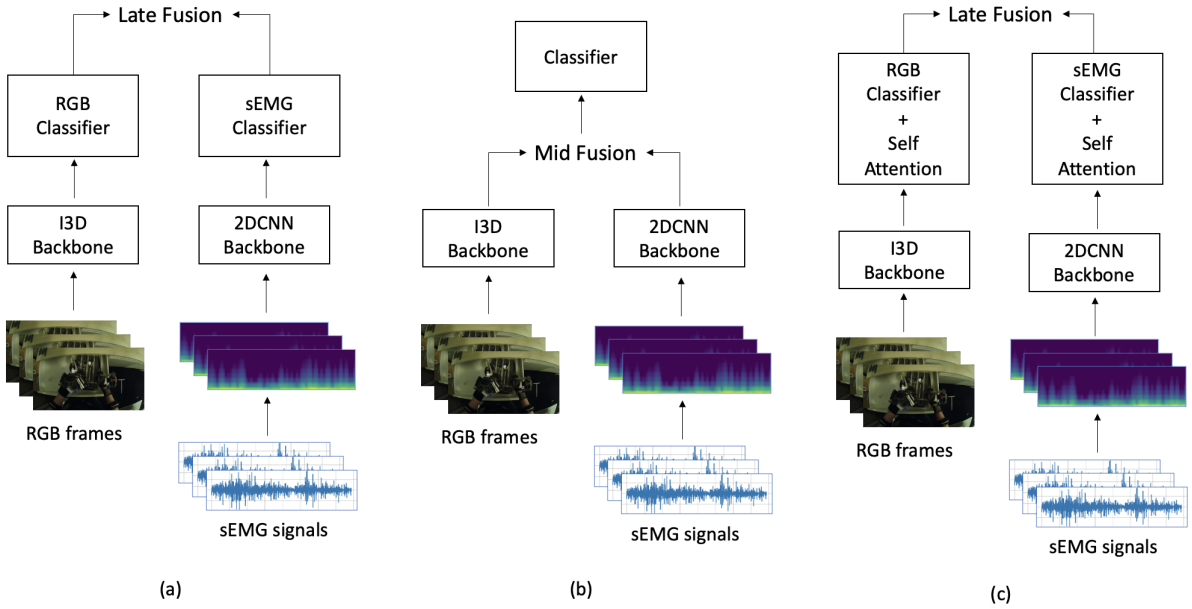

Figure 5. Visualization of Proposed Fusion Architectures: (a) Late-Fusion, (b) Mid-Fusion, and (c) Late-Fusion with Self-Attention

EPIC-Kitchens and ActionSense. Our goal is to take the most of the advantages presented by 3D CNNs by implementing transfer learning.

Multimodal action recognition. Multimodal action recognition is a field of research focused on the recognition and comprehension of human actions by utilizing data from multiple sources. Traditional approaches primarily rely on visual information alone, but its limitations have prompted the development of a new approach that incorporates complementary data modalities [14] [15] [17]. This integration of multiple modalities aims to enhance the performance of action recognition systems. By incorporating additional data modalities, multimodal models address the shortcomings of relying solely on visual information. They can improve recognition performance in scenarios where the dynamics of an action are not adequately captured by the visual signal alone. This approach reduces the computational cost by minimizing reliance on expensive visual data, making it more efficient and practical for real-world applications.

3. Proposed method

Our goal is to determine the most effective approach for combining multiple modality inputs for egocentric action recognition, specifically RGB and sEMG. The extracted RGB and sEMG features, respectively from the I3D backbone and 2D CNN backbone, serve as inputs for our models. We explore various methods to identify the optimal model in terms of performance, aiming to leverage the valuable insights provided by both RGB and sEMG signals, as they offer complementary information. All models use a clip-based approach which analyzes the single clips averaging their prediction in order to obtain the final output. A high-level representation of the proposed models is shown in Figure 5.

3.1. Late-fusion

As first implementation, we propose an approach based on the integration of late-fusion technique. This model combines the power of two different networks by integrating their outputs in a late-fusion manner. Features are fed to the two classifiers and, with the obtained outputs from each network, we compute the mean of the predictions $\hat{Y}_{RGB}$, $\hat{Y}_{EMG}$ and the shared loss.

$$\hat{Y} = mean(\hat{Y}_{RGB}, \hat{Y}_{sEMG})$$

$$\mathcal{L}(\hat{Y}, Y) = -\sum_{i=1}^{N} P(i) \cdot log\hat{P}(i)$$

In this way, the architecture allows the two networks to learn complementary representations from different sources of information leveraging the unique strengths of each network. Furthermore, another version of late-fusion is employed, in which the final prediction is obtained by taking a weighted average of the outputs from the two networks.

$$\hat{Y} = p \cdot \hat{Y}_{RGB} + (1 - p) \cdot \hat{Y}_{sEMG}$$

The parameter p can be either a constant or learnable, allowing the model to determine the optimal balance.

3.2. Mid-fusion

We then present a classifier with mid-fusion, utilizing a fusion layer to combine sEMG and RGB signals. Our proposed model aims to exploit the synergistic benefits of integrating sEMG and RGB modalities in order to generate new mid-level features. The fusion layer has the role of generating new features Z_{fusion} starting from the original ones: it can be implemented concatenating the features Z_{RGB}, Z_{sEMG} and applying a series of linear layers [10].

$$Z_{RGB}, Z_{sEMG} \in R^{n,d}$$

$$Z_{fusion} = fusion(Z_{RGB}, Z_{sEMG}) \in R^{n,k}$$

By leveraging the fusion layer, the model effectively merges the distinctive features extracted from both sEMG and RGB data sources at a mid-level representation, enabling the classifier to capture a complete view of the context of each clip.

3.3. Attention based classifier

As last model we present an attention-based fusion classifier, which exploits a self-attention module in order to weight the prediction for each clip. In the previous configuration, we assign equal importance to each clip, obtaining the final prediction through a simple average. This approach involves assigning attention scores to individual clips to determine their importance. These scores are used as weights in the final prediction to prioritize certain clips over others as shown in Figure 6. The employed self-attention module [10] takes a series of features as input, where each feature vector is divided into n clips, each with dimension d.

$$Z_{RGB}, Z_{sEMG} \in R^{n,d}$$

The attention module outputs a square matrix $S \in R^{n,n}$ which contains the attention scores: each coefficient represents the attention given to the clip i from the clip j. The attention scores are obtained through a dot attention mechanism [19]:

$$S = softmax(\alpha \cdot \frac{QK^T}{\sqrt{d_k}})$$

The α parameter is called temperature and it serves the purpose of augmenting the distinctions among various scores within the softmax function: by adjusting the temperature, the relative differences between attention scores can be magnified. Specifically, as the temperature increases, the disparities between the scores in the softmax function become more pronounced. To enhance the desired impact, we apply a masking technique to restrict the attention allocated to an individual clip by itself. This approach ensures that when calculating the average of the attention scores, only the scores provided by other clips are taken into consideration. Once we obtained the scores, we output the class probabilities computing the weighted average of the single clips predictions:

$$C = \sum_{i=1}^{n} avg(S_i) \cdot P_i$$

The aforementioned module is integrated into the same late-fusion architecture described previously: every network within the system outputs its own attention-weighted prediction which can be averaged with the others to obtain the final prediction. This configuration offers the flexibility to compute attention scores for each source's clips individually, or alternatively, to leverage the clips from the complementary signal to enable each source to determine the weights assigned to the other sources.

4. Implementation Details

4.1. Datasets

We experiment on two large-scale egocentric action recognition datasets: EPIC-Kitchens-55 [9] and ActionSense [13] datasets. Additionally, we employed a pre-trained I3D model that had been trained on the Kinetics [6] dataset. EPIC-Kitchens is widely recognized as a comprehensive and extensively annotated dataset that captures a diverse range of daily activities set within a kitchen environment. The EPIC-Kitchens dataset was specifically utilized for RGB classification tasks. On the other hand, the ActionSense dataset is a multimodal dataset that focuses on wearable sensing in a kitchen setting. We utilized both sEMG signals and RGB streams from the ActionSense dataset.

4.2. Clip-based approach

In our work, we employed a clip-based approach to conduct our experiments. Clip-based action recognition allows for the efficient processing of video data. We have implemented two different sampling methods: uniform sampling and dense sampling. In uniform sampling, we evenly select clips from the video, ensuring an equal time interval between each clip. By distributing the clips evenly throughout the video, we can reduce redundancy and obtain a diverse set of frames. This approach is useful for obtaining a general idea of the video's content. Instead, dense sampling involves selecting data points with higher density in specific regions. This method focuses on areas of the video where more information is desired or where key action moments occur. By densely sampling these regions, we aim

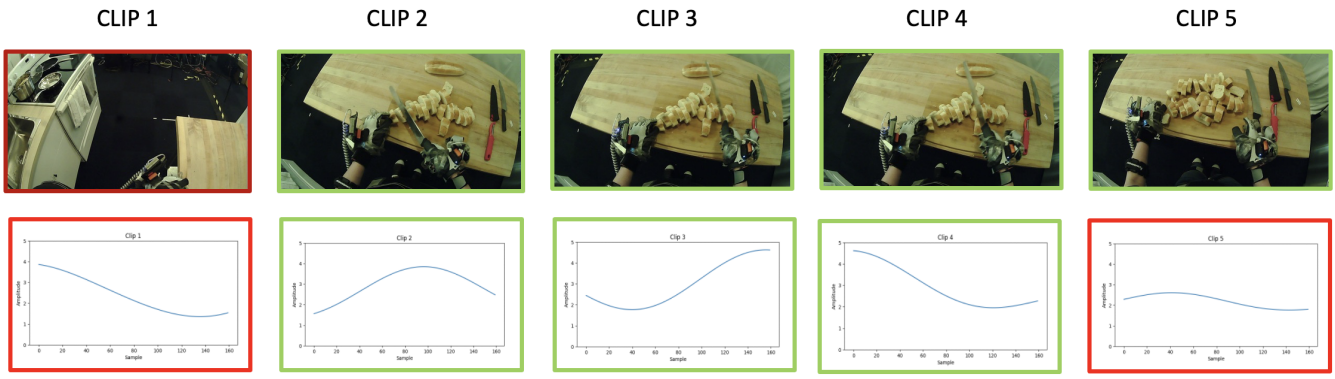

Figure 6. Visualization of the attention scores related to the RGB and sEMG clips in a given action, with green indicating important clips and red denoting less interesting ones. Analysis of these visualizations reveals that the initial clip lacks significance in terms of both RGB and sEMG perspectives. On the contrary, while the last clip appears visually interesting, there is an absence of motion, resulting in a lower attention score for the sEMG.

to capture more fine-grained details. In light of some comparison tests, we decide to use 5 clips for each action, each composed of 16 frames sampled in a dense manner.

4.3. sEMG signal processing

The processing of the sEMG signals is conducted by implementing the recommended methodology proposed by [18] and [5].

RMS. The first step of the pipeline is the smoothing of the signal through the computation of the Root Mean Square (RMS) value over a sliding window, whose size is set to 32 samples. The RMS value it's calculated by taking the square root of the average of the squared values of the signal in the specified window and it provides an indication of the signal's overall strength or energy. This process allows to calculate a smoothed version of the signal that is simpler but still meaningful as we are keeping important local variations thanks to the moving window.

Normalization. Since sEMG signals can vary significantly in amplitude and duration between individuals or different recording sessions, each single channel is normalized with a min-max normalization, shifting the amplitude of the signal to the range [0, 1].

LPF. Lastly, the signal is passed through a low-pass filter (LPF) allowing low-frequency signals to pass through while attenuating higher-frequency components. This is done with the purpose of removing high-frequency noise or unwanted high-frequency components. Specifically, we used a Butterworth filter of 5^{th} order with a cut-off frequency of 5Hz as this kind of filter provides a maximal flat frequency response in the passband.

Frequency-domain. As time-domain data doesn't carry as much information as frequency-domain data we computed the Mel-spectrogram of the sEMG preprocessed signals. We followed and adapted the proposed method for audio data given by [3]. The Mel-spectrograms are computed with the number of Mel frequency bands, `n_mels`, set to 10, and the number of samples of the STFT, `n_fft`, set to 32. This will lead to the Mel-spectrograms having all fixed size. This step is crucial when we will feed them to the 2D-CNN model.

The obtained Mel-spectrograms for one action and one arm are shown in Figure 3.

4.4. Features extraction

In this section, we discuss the feature extraction process for two modalities: RGB and sEMG.

RGB For the RGB modality, we employ transfer learning using the I3D architecture as the backbone. The I3D model has been pre-trained on the Kinetics dataset, a large collection of videos, enabling it to learn general visual representations. This model takes the frames associated with a video clip as its inputs and produces as outputs embeddings of dimensionality 5x1024.

sEMG In the case of the sEMG modality, the backbone utilized for this modality is based on a simple 2D-CNN architecture. The architecture takes as inputs the spectrograms of the 16 channels for every action. The convolutional layers of the model consist of two convolutional filters with kernel size of 3x3 that are followed by ReLU activation functions. After the activation function, a max pooling layer with a kernel size of 2x2 and a stride of 2 is applied. Successively a flattening layer the tensor is then passed through two fully connected layers with the last one having 20 outputs, representing the predicted class probabilities. After training the architecture on the spectrograms derived from the preprocessed sEMG signals, the configuration associated with the highest accuracy is saved and the model is switched to eval-

Table 1. Results Table EPIC-Kitchens-55

Model	Features	Accuracy
Linear	RGB	55.6 %
LSTM	RGB	54.5 %

Table 2. Results Table ActionSense

Model	Features	Accuracy
Linear	RGB	80.2 %
LSTM	RGB	81.6 %
LSTM	sEMG (RAW)	28.3 %
Linear	sEMG	67.9 %
LSTM	sEMG	69.4 %
Late-Fusion	RGB + sEMG	**87.2 %**
Late-Fusion (PAR)	RGB + sEMG	86.3 %
Late-Fusion LSTM	RGB + sEMG	84.3 %
Mid-Fusion	RGB + sEMG	86.6 %
Self-Attention Late-Fusion	RGB + sEMG	82.5 %
Self-Attention LF LSTM	RGB + sEMG	82.8 %

uation mode. During this phase, the deep convolutional part of the network is preserved, while all other layers, except for a single fully connected layer, are discarded. By keeping this layer we are able to extract the embeddings from the original input and reduce the dimensionality of the resulting feature vector to 5x1024, matching the RGB feature dimension.

4.5. Configuration

For the experiments, we use a 70-30 splitting for ActionSense on the splits S04, S07 and S08. We divide each action into sub-actions of 5 seconds obtaining 799 training samples and 343 items for the test. In EPIC-Kitchens instead, we work on the D1 split with 1543 actions used for the training and 435 elements in the test set. We run the networks with similar parameters and resources in order to obtain comparable results. Here the list of parameters: $\texttt{learning\_rate} = 0.01$, $\texttt{momentum} = 0.9$, $\texttt{weight\_decay} = 10^{-7}$.

5. Results

In this section, we compare the results obtained from training several models. The results related to EPIC-Kitchens are represented in Table 1 while the ones from ActionSense are listed in Table 2.

Single modality Even though RGB features extracted from I3D provide good results on ActionSense, reaching 80.2% accuracy, this kind of approach focuses only on the visual aspects. Surface electromyography (sEMG), on the other hand, offers insights into movement classification. However, it lacks visual perception, rendering it "blind" in terms of visual observations. This leads to classes with similar movement being confused with each other: "Slice a potato" is classified correctly only 46% of the time, which is worse than the general accuracy which is 67.9%. Using only RGB in EPIC-Kitchens, which is a bigger dataset compared to ActionSense, we reach lower values around 55.6% as peak accuracy. However, this result is important since it proves once again the transfer learning capabilities of I3D: as a matter of fact Carreira *et al* showed that an I3D trained on Kinetics reaches an accuracy of around 60% [6]; we obtain comparable results even though we use another dataset different from the one used for training.

Classic multimodal approaches Late-fusion approach is overall the best performer with an 87.2% accuracy. This approach outperforms all other single and multi-modal approaches, despite being a quite simple solution. The insertion of a trainable parameter to weigh the sum of the two predictions doesn't seem to provide any advantage. It's interesting to highlight the fact that at convergence the model tends to give more weight to RGB predictions (around 80% of the final result). The mid-fusion approach reaches comparable results to late-fusion: the fusion layer is able to generate new features which still embed valuable information and it is able to provide an 86.6% accuracy through a linear layer.

Self-attention mechanism Despite improving over the single modality approach, the self-attention model is not able to reach the other multimodal approaches, with an accuracy slightly below 83%. Through the implementation of the self-attention module, we are able to give a score to each clip. Despite adding flexibility in terms of classification, we don't experience a big leap in performance. This observation could be partially attributed to the fact that the attention scores of each clip are linked with their similarity to other clips of the same action. This characteristic renders this approach appropriate for actions predominantly comprised of "interesting" clips, as self-attention aids in reducing the significance of certain portions of the action that lack direct relevance to the specific label. However, our scenario involves sub-actions that consist mostly of noisy clips: in such conditions, the model tends to penalize the segments that, in reality, hold the utmost importance.

sEMG spectrogram-based 2D CNN The outcomes of this study highlight the significance of our sEMG feature extraction pipeline. The traditional signal-based approach, which directly utilizes raw signals, is shown to be inadequate in effectively capturing the underlying patterns, resulting in an accuracy of around 28%. In contrast, the spectrogram-based approach, coupled with a 2D CNN, demonstrates superior performance by leveraging the spectral and temporal information and it provides a 40% leap in accuracy.

Temporal context In order to evaluate the importance of the temporal order of the action we trained some alternative

models using LSTM: in this way, each clip can embed the temporal context of the action. The performance improvement achieved through this approach is found to be minimal, and in certain cases, even yields inferior results: as a matter of fact, we work with clips that inherently contain a temporal description. On the contrary, other approaches which exploit frames directly can experience higher benefits [7].

Computational cost Our approach adopts a modular framework, wherein we leverage features extracted by the backbone networks as outlined in Section 4. By adopting this strategy, we circumvent the need for re-training the computationally expensive I3D network. Instead, we only require training a straightforward 2D CNN for extracting sEMG features. Notably, all the classifiers employed in our study demonstrate efficient resource utilization, with each testing phase completing within a time frame of less than 10 minutes on an Nvidia GTX 1050M. This performance sets the stage for further exploration of our approach on devices with average computational resources.

6. Conclusion

Our experiments prove that sEMG signals embed useful information and they can be used in order to improve egocentric video recognition task: all the proposed multimodal models outperform the single modality alternatives. The simplest fusion models are the ones that reach the highest accuracy. However, self-attention has potential and further research could lead to better results. Its effect could be more visible in particular with datasets composed of untrimmed videos or, as mentioned above, with longer actions that can overcome the limits of the used data.

References

[1] Repository for classification models. https://github.com/simoclemens/MLDL23_classifiers.git, 2023. 7

[2] Repository for preprocessing and feature extraction. https://github.com/al3ssandrocaruso/MLDL2023.git, 2023. 7

[3] Tomás Arias-Vergara, Patrick Klumpp, Juan Carlos Vasquez-Correa, et al. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Analysis and Applications*, 24:423–431, 2021. 2, 5

[4] Manfredo Atzori, Matteo Cognolato, and Henning Müller. Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands. *Frontiers in Neurorobotics*, 10:9, Sep 2016. 2

[5] Müller H. Atzori M, Cognolato M. Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands. *Front Neurorobot*, 2016. 5

[6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset, 2022. 4, 6

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 7

[8] Chun-Fu Richard 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021. 2

[9] 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, 2018. 4

[10] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. 2020. 4

[11] Marjan Gholamrezaii and Seyed Mohammad Taghi Almodarresi. Human activity recognition using 2d convolutional neural networks. In *2019 27th Iranian Conference on Electrical Engineering (ICEE)*, pages 1682–1686, 2019. 2

[12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 2

[13] Yiyue Luo Michael Foshey Yunzhu Li Antonio Torralba Wojciech Matusik Joseph DelPreto, Chao Liu and Daniela Ru. 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. 4

[14] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition, 2021. 3

[15] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition, 2019. 3

[16] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding, 2019. 2

[17] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text, 2022. 3

[18] Bruno Cornelis Panagiotis Tsinganos, Athanassios Skodras and Bart Jansen. *Learning Approaches in Signal Processing*, volume 2 of *Pan Stanford Series on Digital Signal Processing*. 5

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 4