EPIC-Muscle: cross-modality EMG generation for EPIC-Kitchen dataset

Salvatore Adalberto Esposito
Politecnico di Torino
s304800@studenti.polito.it

Gabriele Iurlaro
Politecnico di Torino
s294917@studenti.polito.it

Abstract

Understanding video representations and exploit them in the First Person Action Recognition task can be hard. For this reason recently, several papers propose to combine the RGB recordings with other modalities like audio, optical flow, and other non-visual one(e.g. EMG, tactile) This is done in order to help the model being more independent from the environment, allowing them to impose new state of the art results. In this paper we propose and analyze a cross-modal framework able to successfully generate electromyographic (EMG) features for the well-known RGB video EPIC-Kitchen dataset. This is done training Variational Autoencoders on features extracted from RGB and EMG samples. We demonstrate that our method can generate good representation of EMG signal that actually support the classification task. This eventually opens to the possibility of exploiting the same framework to fill the missing modalities of other datasets.

Github pyTorch Implementation

1. Introduction

Egocentric vision refers to a visual perception approach based on a first-person perspective, typically captured using wearable cameras like GoPro. This unique type of data holds significant value for various tasks, particularly in the domain of action recognition. In recent years the research in this domain followed different paths, one of the most promising is multimodal combination of different data models as in [9], [11] and [12]. In our study two datasets have been taken in consideration: the Epic-Kitchen dataset [4] which is a collection of high-scale videos, specifically focused on human cooking activities, recorded by participants within their kitchen environments; the ActionSense dataset [5] which has been recorded in 9 different modalitaties: 19 IMUs, 3D Body Joints, Hand Pose, Gaze, EMG, Tactile, RGB, Depth, Audio. In our context only RGB and EMG modalities are taken into account. Our job aims to leverage deep learning techniques to analyze videos, considering both spatial and temporal information, and develop a robust action recognition system based on compressed features. To do so we extract intermediate features from pre-trained deep convolutional models for this two specific datasets. These features capture the most important information about the appearance and dynamics of the videos. To efficiently process long video sequences, a two-phase strategy is employed. In the first phase, we extract compressed representations from the videos using a pre-trained model [1] for computational reasons. Both dense and uniform sampling strategies are used to select a certain number of frames per video clip. Then features are extracted from each selected frame and saved. In the second phase these features are aggregated on the temporal dimension and fed in a classifier for the action recognition task. Exploiting classification results and plotting features we were able to select the best features-set for each modality and dataset (RGB for Epic Kitchen and RGB, EMG for ActionNet). Since EMG modality seems to provide features considering different aspects of the action, we use those features to train a variational autoencoder, able to create a multimodal latent space, enabling cross modality translation from RGB features to EMG one. This is done, in order to generate the missing modality for the EPIC-Kitchen dataset.

2. Related works

Temporal modeling and video understanding The action recognition task and in general video understanding have become popular over these years because of the revolution introduced by deep models and convolutional neural networks. Several works have presented models that incorporate temporal information in a robust and effective manner, making them well-suited for analyzing the dynamic and complex actions captured in egocentric vision [1]. In [2] authors benchmark the main used techniques and models in the field. They claim both 2D-CNN and 3D-CNN models are able to provide state-of-the-art results in terms of accuracy, but different behaviors in terms of efficiency and temporal modeling. The latter mainly refers to the way temporal information is aggregated, which is a key point in video understanding. Generally, a 2D architecture offers more flexibility in temporal modeling than 3D ones, although an

effective one has to be used in order to guarantee efficiency and good performance. Moreover in [2] a comprehensive analysis of video frames sampling strategies is conducted, comparing uniform and dense sampling. Through their experiments, they showed most recent and robust architectures (i.e. I3D [1], TAM [6] on top of ResNet, TRN [17]) are invariant with respect to the model input, identifying in the uniform sampling the best trade-off between efficiency and performances for 2D architectures.

Multimodal action recognition Egocentric videos are typically captured from a first-person perspective and are subject to motion blur, camera motion, and variations in lighting conditions. Moreover, this modality is highly affected and dependent on the environment in which it is recorded. These factors make it hard to recognize actions in egocentric videos using exclusively traditional methods based on image frames. E^2(GO)MOTION [12] addresses a key challenge in this context by using event cameras which provides a more robust representation of motion information, outperforming RGB models. EPIC-Fusion: Audio-Visual Temporal Binding [9] tries to solve this task by combining audio and visual information in order to better distinguish between different actions and handle challenging cases such as occlusions, background clutter, and viewpoint variations. Multi-Modal Domain Adaptation [11] tackles FPAR task by leveraging on multi-modal self-supervision with RGB and Optical Flow modality, it succeeds in increasing robustness to changes in appearance and context. In recent years, a non-visual based dataset has been relased [5], moving the attention of people to new modalities to make inferences (*e.g.* EMG, tactile sensors, etc.)

Variational Autoencoder In recent years, many works have been published on generative networks, employing Generative Adversarial Networks(GAN) and Variational Autoencoders [10] [8]. In some papers, authors have exploited GANs for generating biomedical signals (in particular EMG signal) [7] [3], in order to enhance the training of different tasks. However, reconstructing a new modality starting from another is a demanding task, since it requires a good amount of paired data and regularization of the latent space. Variational Autoencoders have these characteristics, and in recent years have been used for cross-modal translation alone [13] and in combination with GANs [15] in the context of hand-pose estimation. In particular, [13] proposed a mathematical formulation that demonstrates the possibility of jointly training Variational Autoencoders that share a common latent space among different modalities.

3. Methodology

In the action recognition scenario, and especially in the egocentric one, videos play a significant role in understanding the action behavior. Videos are an ordered collection of frames. One single frame cannot be enough to predict the action, since videos incorporate temporal dynamics that cannot be ignored. In order to teach the network temporal information, different strategies can be used.

3.1. Sampling strategies and clip level samples

We now introduce the notation that we will use in the whole paper. A video v is divided into n clips(randomly selected segments of equal length), and each clip is composed of m frames. Our datasets are collections of pair video, label(v, y).

$$(v, y) \in V$$

Given a video:

$$v = \{c_1, \cdots c_n\}$$

Given a clip, the frames are indicated in this way

$$c_i = f_1, \cdots f_m$$

The label y is associated with the video as well as with its clips. Fig. 1 shows the relation between video, clips and frames. This allows us to make predictions at video level and clip level. There are two ways to extract frames from a clip:

- **Dense sampling**: Given a clip and its center, we select m frames separated by a small stride. It takes relatively near time frames, favoring the appearance over the temporal dynamics.
- **Uniform sampling**: Given a clip, we select m equally spaced frames in the whole clip. It selects more far frames, that can be very different, allowing to exploit better temporal dynamics

It's difficult to establish which is better a priori. Looking at their definitions, one could say that uniform sampling, thanks to the wider window of frames, can achieve better results in action classification, but experimental results depict a more complex behavior. We have also considered different numbers of clips and different numbers of frames within the clips, to choose the best setup in terms of qualitative and quantitative results.

3.2. RGB modality

Since training a deep convolutional video model can be very heavy, we worked on a compressed representation, the so-called *features*. Features are the output of the convolutional backbone of the model(`I3D-Inception-v1` [1]), a very popular backbone for action recognition. `I3D` was

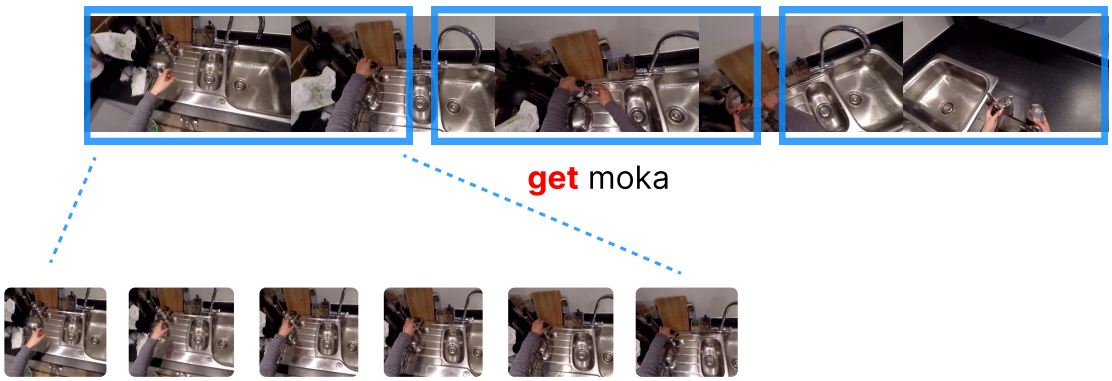


Figure 1. **Difference between video, clips and frames**. A video is a sequence of frames. We select clips by randomly selecting central frames and by sampling frames within each segment. The clips belonging to the same video share the same label.

one of the first deep network that use 3D convolution to exploit temporal dynamics, and is based on `Inception-v1` [14]:

Later, we used the extracted features to train some classifiers, showing that finetuning only a simple classifier can achieve good results. We proposed various architectures:

- `MLP Late Fusion`: it takes as input multiple clips and it uses a fully connected network to classify each clip separately. The predicted logits of each clip are stacked together and the average is taken to produce the final video prediction. This is a late fusion method since we combine the outputs in a "late" stage of processing before making the final prediction.

- `MLP Early Fusion`: We stack the features of the different clips to a unique feature vector and pass it through a fully connected network that outputs the final prediction. In this way, the network uses the whole video information at once to make the prediction.

- `Long Short-Term Memory` (LSTM): It is a type of recurrent neural network(RNN) that can model long-term dependencies between the clips exploiting them to make predictions about the action being performed. In our case, the network is fed with one clip at a time, and the temporal aggregation is done using the internal state.

- `Temporal Relational Network` (TRN) [17]: this network module allows multiple level of relationship beetween clips (from a 2-clip relation to n-clip relation), allowing temporal relation reasoning. Starting from the 2-clip relation

$$T_2(v) = h_\phi \left(\sum_{i<j} g_\theta(c_i, c_j) \right)$$

that express relation between clips by fusing them through a simple multilayer perceptron, it's possible to extend the 2-clip relation to n-clip relation and by combining them we obtain the multi-scale temporal relation as:

$$MT_n(v) = T_2(v) + T_3(v) + \cdots + T_n(v)$$

The classification task has been useful to understand which combination of sampling strategies where more suited for our dataset. Features extracted in this phase (each sample is a vector of 1024 features) will then be used for generating synthetic EMG signals from video RGB features.

3.3. EMG modality

ActionSense [5] is one of the first datasets to give specific attention to other non-visual modalities, thanks to the use of several types of sensors. This allows to analyze different aspects of an activity, recurring to different modalities:

- Motion
- Tactile sensing
- Muscle activity
- Body and Finger tracking
- Eye-tracking
- RGB + D video
- Audio

In our analysis, we will use only **muscle activity**, in the form of EMG samples. These samples are recorded as 8 channels of muscular activity for each forearm. The proposed processing of these recordings in [5] consisted of

rectifying and filtering them, in order to produce a normalization effect on the signal. We choose instead to operate directly on the spectrogram of each channel of the EMG signal in order to exploit both its frequency and temporal components as [9] has done with audio.

Starting from the signal, we compute the spectrogram for each channel, using `n_fft` in order to have a frequency dimension that matches with the chosen temporal granularity(number of frames for each clip). The raw spectrogram is later sampled in order to obtain a tensor of dimension 16×# frames×# frames. After this preprocessing step, samples are employed to perform classification on ActionSense dataset, allowing us to encode the samples through a CNN-based feature extractor.

Feature extraction Encoding EMG (spectrogram) samples in a feature space to reduce their dimensionality proved to be effective for our next steps, allowing for efficient training and simplifying the reconstruction objective. The selected feature extractor is the backbone of our EMG classifier, realized as a regular parametric CNN with limited complexity and depths and small kernel sizes.

Again the classification task helped in suggesting which architectures and sampling strategies are the most suited to extract meaningful features. We also considered the class separability of the extracted features in a 2-D dimensional space through dimensionality reduction approches like PCA and TSNE.

3.4. Visual2Signals

3.4.1 Variational autoencoder

Variational autoencoder(VAE) [10] [8] is a neural network working in an unsupervised way. VAE framework allows us to achieve better latent space properties with a simple idea: mapping the input sample to a distribution instead of a single latent representation. In the generation process, we sample from the obtained distribution in order to reconstruct the original sample (or in our case, a new modality of that sample). This results in the following loss function(ELBO or *Evidence lower bound*):

$$-\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] + \beta\, D_{KL}(q_\phi(z|x)||p_\theta(z)) \quad (1)$$

Where $D_{KL}(\cdot)$ is the Kullback-Leibler divergence(KL divergence), $q_\phi(z|x)$ and $p_\theta(x|z)$ are respectively the conditional distribution represented by the encoder and the decoder, $p_\theta(z)$ is the prior on the latent space, parametrized as $\mathcal{N}(z|0, I)$. The encoder gives out the mean μ and the variance σ^2 of a normal distribution such that $z \sim \mathcal{N}(\mu, \sigma^2)$. The first term represent the likelihood of generating real data. We have introduced in our final loss function a hyperparameter β that acts as a regularizer of the KL divergence [8].

In order to make the VAE framework learn a latent representation, able to make cross-modal translation, we need a latent space capable of capturing action information and of generating a new modality starting from another. Following [13], we re-derive a loss function that leverages multiple modalities:

$$-\mathbb{E}_{z\sim q_\phi(z|x_t)}[\log p_\theta(x_t|z)] + \beta\, D_{KL}(q_\phi(z|x_s)||p_\theta(z)) \quad (2)$$

Where x_t represents the target modality (*i.e.* EMG), and x_s represents the starting modality (*i.e.* RGB). The training procedure is composed of different stages:

1. Extract EMG and RGB features (respectively from ActionSense and Epic-Kitchen).
2. Train a pair of VAE to reconstruct features for each modality.
3. Build a third VAE starting from the RGB encoder and the EMG decoder.
4. Reconstruct the missing EMG modality on Epic-Kitchen dataset.

The complete algorithm for the cross-modal translation can be found in Algorithm 1.

Algorithm 1 Cross-modality translation

Require: $models = (q_{RGB}, p_{RGB}), (q_{EMG}, p_{EMG})$
  for $i < epochs$ **do**
    for data ∈ `S04` **do**
      $x_{RGB} \leftarrow \texttt{data}_{RGB}$
      $x_{EMG} \leftarrow \texttt{data}_{EMG}$
      $(\mu, \sigma) \leftarrow q_{RGB}(x_{RGB})$
      $z \sim \mathcal{N}(\mu, \sigma^2)$
      $\hat{x}_{EMG} \leftarrow p_{EMG}(z)$
      $\mathcal{L}_{MSE} \leftarrow ||\hat{x}_{EMG} - x_{EMG}||^2$
      $\mathcal{L}_{KL} \leftarrow -0.5 \cdot (1 + \log(\sigma^2) - \mu^2 - \sigma^2)$
      $\theta_{p_{RGB}} \leftarrow O(\mathcal{L}_{MSE} + \beta\, \mathcal{L}_{KL})$
      $\theta_{q_{EMG}} \leftarrow O(\mathcal{L}_{MSE} + \beta\, \mathcal{L}_{KL})$
    end for
  end for

Fig. 2 represents our full architecture.

3.4.2 Network details

The 3 VAEs share the same architecture. They are composed of 3 fully connected layers, with Batch normalization and reLU activation function. The encoder has 2 more layers that output μ and σ, while the decoder is specular to the encoder and uses the sampled z to reconstruct the input sample.

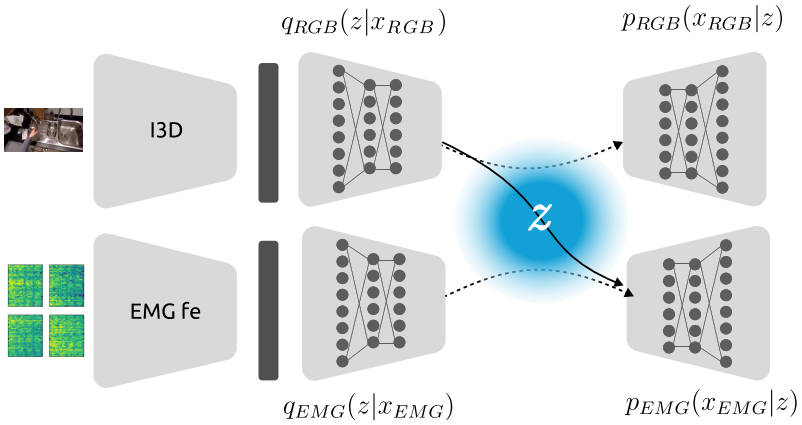


Figure 2. **Schematic overview of our architecture**. The feature extractors (I3D and EMG fe) compute a compressed representation from RGB and EMG modalities in order to train the respective autoencoders. Finally, the RGB encoder and EMG decoder are combined to translate from one modality to another.

4. Experimental results

Our experiment phase is divided into two parts: in the first one, we generate features with different hyperparameter sets, and then we evaluate them in a qualitatively and quantitatively. In the second part of the experiments, we use the best set of extracted features to train and evaluate our `RGB` $\rightarrow$ `EMG` variational autoencoder.

4.1. Datasets

We conduct our experiments on the reduced versions of 2 datasets:

- Epic-Kitchen(EK) [4]. One of the most popular datasets of action recognition based on an egocentric camera, provides different scenes recorded by different characters in their usual environment. We use data from one kitchen only(`P08`, also called `D1`), in particular we focus on the RGB modality. With respect to the original Epic-Kitchen AR task, ours has been simplified, predicting only verb (and not verb-nouns) classes and considering a partition of the available classes(8 verb classes in total). It consists of 1543 train video samples and 435 test samples. The verb classes are highly imbalanced.

- ActionSense(AS) [5], a novel egocentric and multimodal dataset. Though it provides data from several sensors, we used only EMG data from the whole dataset and RGB from `S04` subject. The dataset is quite small compared to EK, nonetheless, video samples are quite long. For the considered subject, we only have 51 train and 8 test videos. While for the EMG data, we can count on 526 train and 59 test samples. As done for EK, we considered a subset of the labels, using only verbs (12 classes in this case).

4.2. Feature analysis and classification

4.2.1 RGB modality

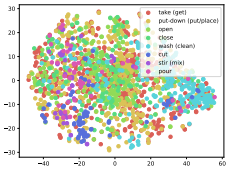
In the first part of our work, we analyze the extracted features from our backbone, pre-trained on the Kinetics [1] dataset. We report how the choice of hyperparameter affects the quality of these middle representations. In particular, we extracted features:

- From the backbone only initialized on Imagenet and pre-trained on kinetics dataset *or* finetuned on Epic Kitchen
- Using dense *or* uniform sampling
- Using 8 *or* 16 frames for each clip

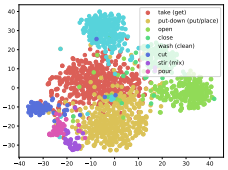
During the train, we choose 10 clips per video. Qualitative results are reported in Fig. 3 using t-SNE to reduce dimensionality for visualization purposes.

Features in Fig. 3(b) are the best in terms of class separability, due to the finetuning on EPIC-Kitchen dataset. We notice that dense sampling appears to be more effective, with slightly better performances when 16 frames are chosen instead of 8. This behavior, in contrast with other papers(in [2] uniform sampling outperforms dense) can be explained since the I3D backbone is pretrained using dense strategy. A confirm can be found in Figures 3(b) and 3(d), which employ a different sampling strategy and a different number of frames 3(c).

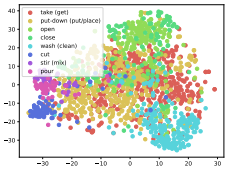
The poor result of the kinetics pre-trained model can be explained by the different temporal dynamics of the two datasets. In the Kinetics dataset, temporal modeling is not so crucial in identifying the action, leading the model to focus on more appearance-based representations, in which context and environment are dominant. While, in EK the appearance is not enough to classify actions. Proof of this



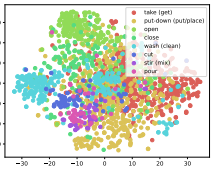
(a) Imagenet + kinetics, 16 frames, dense sampling



(b) Epic Kitchen, 16 frames, dense sampling



(c) Epic Kitchen, 8 frames, dense sampling



(d) Epic kitchen, 16 frames, uniform sampling

Figure 3. Extracted RGB features from Epic kitchen dataset

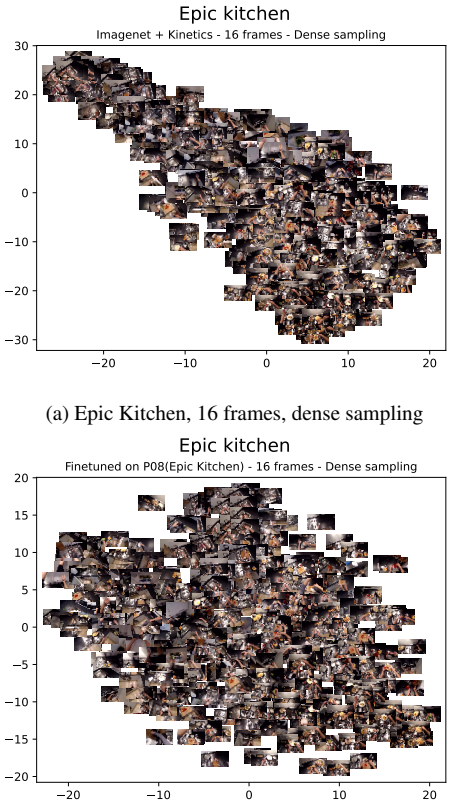
can be found in Fig. 3(a), and in Fig. 4. The latter shows video distribution in the latent space, using the central frame to represent the sample. We can notice, that similar in appearance videos tend to be near in the latent space of Fig. 4(a), while in Fig. 4(b) they are more heterogeneous, aggregated more on the activity itself rather than on the environment.

This analysis is reflected in the quantitive and complete results presented in Tab. 1. The classification task has been performed by the built-in top classifier of the I3D model.

Finetuning	# of frames	Sampling strategy	train accuracy(%)	validation accuracy(%)
✗	16	Dense	12.38	10.34
✗	16	Uniform	11.56	8.74
✗	8	Dense	14	12.18
✗	8	Uniform	11.21	9.89
✓	16	Dense	93.13	57.93
✓	16	Uniform	71.74	48.97
✓	8	Dense	82.31	54.25
✓	8	Uniform	59.95	43.68

Table 1. RGB Video level accuracy on EPIC-Kitchen dataset using different settings.

In order to further model the temporal information, we adopted several temporal aggregation strategies, showing how these affect performances:



(a) Epic Kitchen, 16 frames, dense sampling

(b) Epic Kitchen, 16 frames, dense sampling

Figure 4. Epic Kitchen extracted features, with frame

Model	train accuracy(%)	validation accuracy(%)
LSTM	100	55
TRN	88	59.54
MLP(ef)	100	58.6
MLP(lf)	100	59.1

Table 2. Video level accuracy on RGB data

The TRN module appears to be the most effective due to its temporal relation reasoning capabilities.

4.2.2 EMG modality

The original EMG signals provided in ActionSense are recorded through 8 electrodes on each forearm, for a total of 16 channels. For each channel, we computed the spectrogram of the recording, sampling with the previously reported techniques. So when the dataset is loaded, the shape of a clip-level sample is 16x32x32 (in the case of 32 'frames'). From each clip, a feature vector is then extracted sharing the same shape as the RGB ones, 1x1024. Several values have been tested as *features vector size*, as it moder-

ately affects classification performances. It is inversely proportional to the depth (complexity, number of layers) of the extractor, which implies that learning a (good) smaller encoding of the original data is a more difficult task and, since our EMG dataset is small, larger values should be preferred.

Features Vector Size	# of clips	Sampling strategy	train accuracy(%)	validation accuracy(%)
1024	5	Dense	72.66	56.3
1024	10	Dense	92.2	68.75
1024	10	Uniform	94.5	75
1024	15	Uniform	84.4	71.9
512	20	Uniform	99.2	58.4

Table 3. EMG Video level accuracy on ActionSense

We tried to overcome the initial difficulties related to the limited number of samples, first by exploiting known transformations on the samples in order to augment the dataset. This path, though consistent and promising, had to be abandoned due to the poor improvements provided with respect to the complexity required. Nonetheless, simply using a higher number of clips per video resulted in a more robust model, less prone to overfitting.

Feature extraction is performed by a 2D-CNN in which, at each layer, the number of channels is doubled and batch normalization is applied. Clip results are then aggregated after a 2-layer linear classifier in order to predict class labels, video level accuracies are reported in table 3. As reported, the most relevant parameters proved to be the ones related to temporal modeling. Uniform sampling outperforms dense strategy, confirming what is claimed in [2] for RGB modality: dense sampling models early stop benefiting from increasing clips, while uniform strategy incurs in overfitting, worsening its performances. The extracted features appear clearly separably class, as shown in Fig. 5, this suggests that our model found an effective encoding of these signals.

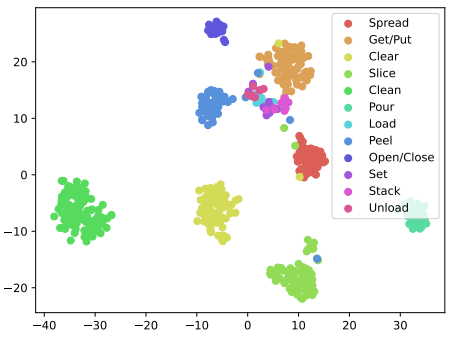


Figure 5. Extracted EMG features from ActionSense

4.3. Visual2Signals

As described in Section 3, the training phase of the VAE for cross-modality translation is composed of 3 different stages:

- Train a VAE for RGB → RGB reconstruction
- Train a VAE for EMG → EMG reconstruction
- Combine the RGB→RGB encoder and the EMG → EMG decoder to train a cross-modality translation VAE

It's important to notice that this approach has as drawback of training more models. Also, since we don't have enough paired modality samples, we decided to train in this way:

- The RGB→RGB VAE is trained on the previously described reduced EPIC Kitchen dataset extracted features.
- The EMG→EMG VAE is trained on the full ActionSense dataset extracted features.
- The RGB→EMG VAE is finetuned on the `S04` subject from ActionSense extracted features.

We trained each VAE for 100 epochs, with Adam optimizer and learning rate of 10^{-3}. We have found that the features were qualitatively better when the hyperparameter β was equal to 10^{-5}. We analyzed the obtained result by inspecting the visualization of reconstructed features(Fig. 6 and 7) and employing them from classification, obtaining similar results in terms of accuracy.

Since our VAE works on features vector of size 1024, we use a simple fully connected structure both for encoder and decoder, with batch Normalization and reLU activation functions. We have used different bottleneck sizes and found that 256 has performance comparable to 512.

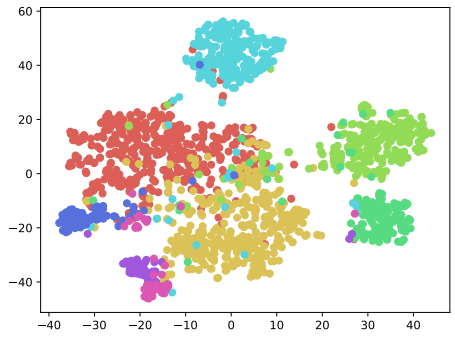


Figure 6. RGB reconstructed features from EPIC-kitchen dataset D1. Different colors represent different labels

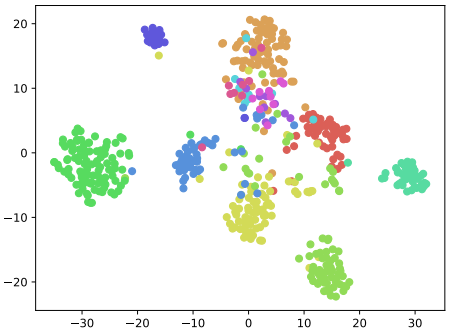
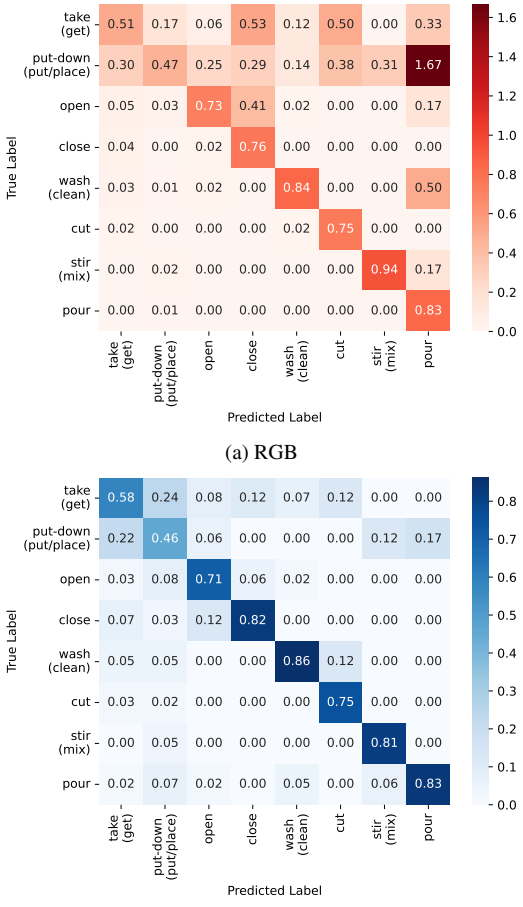


Figure 7. EMG reconstructed features from the ActionSense dataset. Different colors represent different labels



(a) RGB

(b) RGB + EMG

Figure 8. RGB only confusion matrix and the respective multimodal one. Some classes actually benefit from the EMG modality.

Finally, we have trained for a smaller number of epochs and a smaller learning rate the full VAE for cross-modal translation. The latter is used to generate the missing EMG modality for the EPIC-Kitchen dataset.

The newly generated modality is then used to perform multimodal classification. For the RGB modality, we used the previously introduced TRN model, while for the EMG data, we adopt a single 2-layer MLP. The logits from each modality are then fused by taking the mean and used to make predictions. The model benefits form the different type of data, improving generalization and overall validation accuracy(from 59.77% to **60.61 %**).

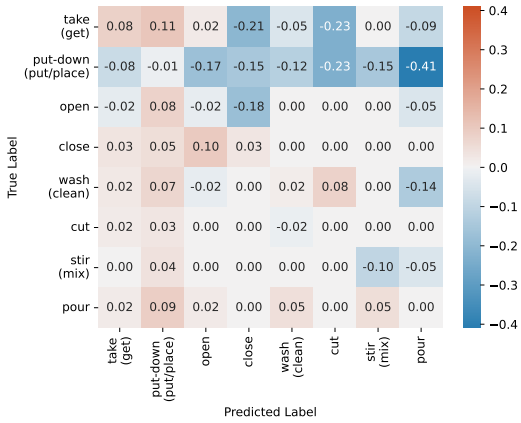


Figure 9. Accuracy gain in the multimodal scenario. The values outside the main diagonal(referring to wrongly classified samples) decrease, showing a general correcting behavior

5. Conclusions

During our analysis, we explored several ways to model temporal dynamics, confirming its importance in study cases where appearence and context are not enough for activity recognition. We were also able to find confirmation on what stated about sampling strategies in [2] and in [17] about temporal relations even also the EMG modality.

Once again, it appeared clear the role of transfer learning in effectively reaching high performances, especially in the context of RGB videoclips due to the high dimensionality and complexity of the data itself. Variational autoencoders, as well, keep demonstrating their usability in a cross-modality setting [13] [16] and their regular latent space makes them suited for our translation task.

Overall, our results cannot be numerically compared with those of other researches employing full datasets [12] [9], nonetheless our framework shows a promising possible path to further improve our knowledge in FPAR field.

In fact generated modalities can actually help in increasing accuracies of the classification task. Further analysis would be required to correctly quantify the concrete contribution of our framework, in particular maybe exploring:

- more complex feature extractor architectures;

- larger datasets;
- more and diverse modalities;

it could be used to reach new state of the art results in the field.

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. 1, 2, 5

[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, 2020. 1, 2, 5, 7, 8

[3] Zihan Chen, Yaojia Qian, Yuxi Wang, and Yinfeng Fang. Deep convolutional generative adversarial network-based emg data enhancement for hand motion classification. *Frontiers in Bioengineering and Biotechnology*, 10:909653, 2022. 2

[4] 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. 1, 5

[5] 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. 1, 2, 3, 5

[6] Quanfu Fan, Chun-Fu (Richard) Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. *Neurips*, 32, 2019. 2

[7] Debapriya Hazra and Yung-Cheol Byun. Synsiggan: Generative adversarial networks for synthetic biomedical signal generation. *Biology*, 9(12), 2020. 2

[8] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 2, 4

[9] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4, 8

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 2, 4

[11] 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. 1, 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, 2021. 1, 2, 8

[13] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation, 2018. 2, 4, 8

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014. 3

[15] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation, 2017. 2

[16] Yixin Wang, Shuang Qiu, Dan Li, Changde Du, Bao-Liang Lu, and Huiguang He. Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition. *IEEE/CAA Journal of Automatica Sinica*, 9(9):1612–1626, 2022. 8

[17] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. *CoRR*, abs/1711.08496, 2017. 2, 3, 8