Anticipating Activity from Multimodal Signals

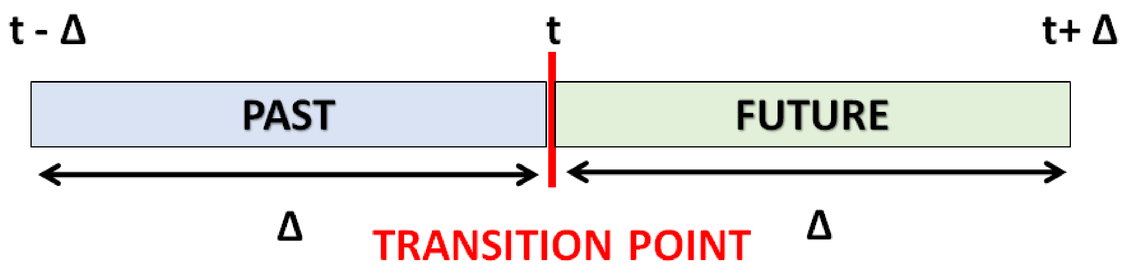
Tiziana Rotondo[1], Giovanni Maria Farinella[1], Davide Giacalone[2],
Sebastiano Mauro Strano[2], Valeria Tomaselli[2], Sebastiano Battiato[1]
[1]Department of Mathematics and Computer Science, University of Catania, Italy
[2] STMicroelectronics, Catania
tiziana.rotondo@unict.it, {davide.giacalone, mauro.strano, valeria.tomaselli} @st.com,
{gfarinella, battiato}@dmi.unict.it

Abstract

Images, videos, audio signals, sensor data, can be easily collected in huge quantity by different devices and processed in order to emulate the human capability of elaborating a variety of different stimuli. Are multimodal signals useful to understand and anticipate human actions if acquired from the user viewpoint? This paper proposes to build an embedding space where inputs of different nature, but semantically correlated, are projected in a new representation space and properly exploited to anticipate the future user activity. To this purpose, we built a new multimodal dataset comprising video, audio, tri-axial acceleration, angular velocity, tri-axial magnetic field, pressure and temperature. To benchmark the proposed multimodal anticipation challenge, we consider classic classifiers on top of deep learning methods used to build the embedding space representing multimodal signals. The achieved results show that the exploitation of different modalities is useful to improve the anticipation of the future activity.

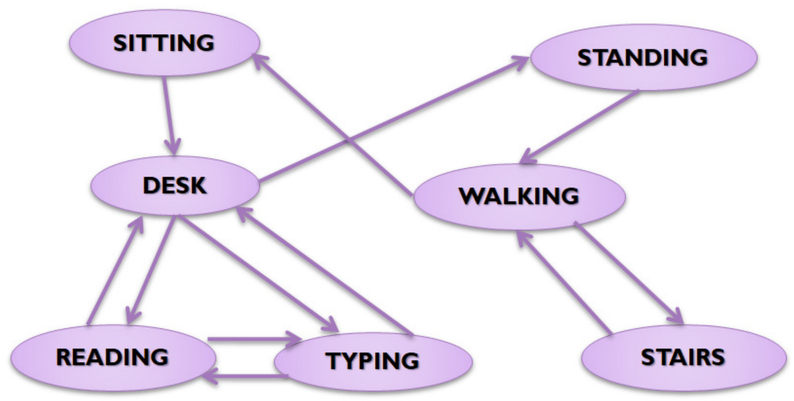
Problem Definition

Problem: Let $\mathbf{y}_t = (\mathbf{v}_t, a_t, \mathbf{s}_t)^T$ be the input vector at time t where $\mathbf{v}_t$ is the video signal, $\mathbf{a}_t$ is the audio signal and $\mathbf{s}_t$ is related to other sensor data, we define the feature representation of video, audio and sensor as $\mathbf{x}_t^v$, $\mathbf{x}_t^a$ and $\mathbf{x}_t^s$. Hence, $\mathbf{x}_t = (\mathbf{x}_t^v, \mathbf{x}_t^a, \mathbf{x}_t^s)^T$ is the features vector at time t. Given the features of the past $(\mathbf{x}_{t-\Delta}, \ldots, \mathbf{x}_t)$ as input, we want to predict the label of the next activity in $]t, t+\Delta[$.



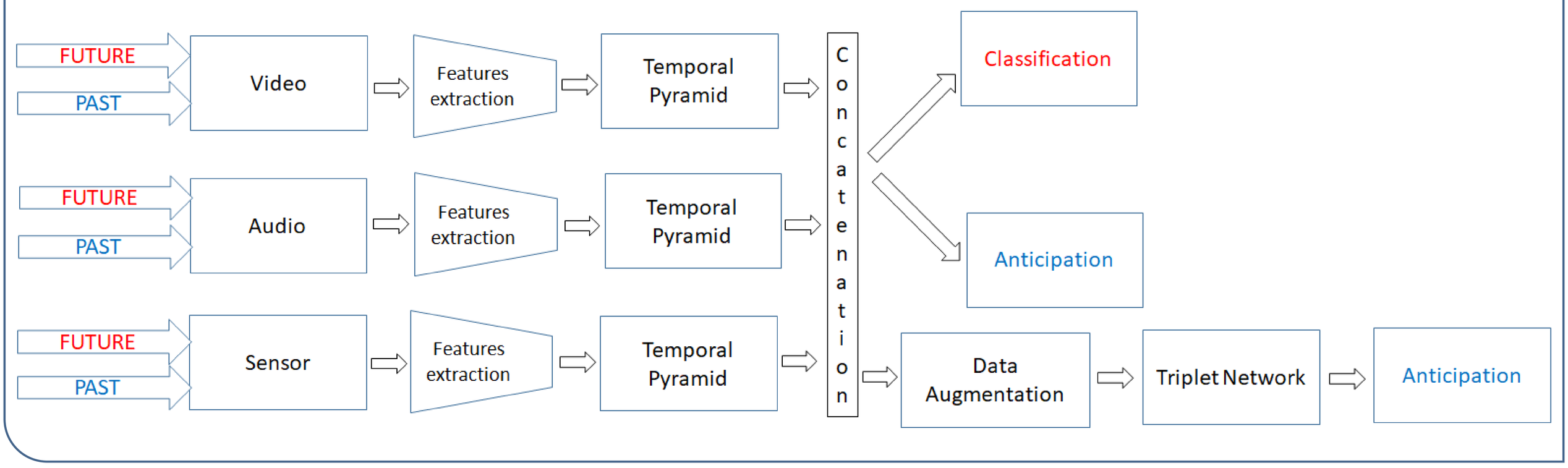
ST Multimodal Dataset

The dataset comprises the following modalities: video (collected with a smartphone camera), audio, tri-axial acceleration, angular velocity, tri-axial magnetic field, pressure and temperature (acquired with a board (called BlueCoin) built by STMicroelectronics.)



Proposed Method

To predict the future activity, a Triplet Network is proposed trained with sequences that represent the past observations. To build an embedding space where past and future sequences semantically correlated are close to each other, a branch of the Triplet network processes the past features (x), the second one processes the features related to the possible (true) future activities (x^+) and the last one considers the negative sample related to the false future activities (x^-). The pipeline is sketched in the following Figure.



Experimental Results: Baseline vs Siamese Network

These tables summarize results obtained with K-NN and SVM classifier for different values of K and different kernels, respectively. As baseline, the combination of all sensors rate results are reported.

	Classification	Anticipation	
K	Baseline	Baseline	Triplet
1	63.24%	64.12%	64.65%
3	62.79%	63.61%	64.73%
5	62.79%	63.14%	64.14%
7	61.97%	62.27%	64.55%
9	61.76%	62.15%	64.18%

Classification		Anticipation			
Baseline		Baseline		Triplet	
Linear Kernel	RBF	Linear Kernel	RBF	Linear Kernel	RBF
69.55%	73.75%	64.06%	70.29%	57.52%	63.32%

Conclusion

- Our results suggest that multi-modality improves both classification and prediction.
- Considered activities can be anticipated with an accuracy close to the one obtained when the signals are fully observed (i.e., classification task).
- Future works could be devoted to collect bigger labelled multimodal datasets considering different environments and activities, as well as to model attention mechanisms among the different modalities.