Action Anticipation from Multimodal Data

Tiziana Rotondo[1], Giovanni Maria Farinella[1,2], Valeria Tomaselli[3] and Sebastiano Battiato[1,2]
[1]*Department of Mathematics and Computer Science, University of Catania, Italy*
[2]*ICAR-CNR, Palermo, Italy*
[3]*STMicroelectronics, Catania, Italy*

Keywords: Action Anticipation, Multimodal Learning, Siamese Network.

Abstract: The idea of multi-sensor data fusion is to combine the data coming from different sensors to provide more accurate and complementary information to solve a specific task. Our goal is to build a shared representation related to data coming from different domains, such as images, audio signal, heart rate, acceleration, etc., in order to anticipate daily activities of a user wearing multimodal sensors. To this aim, we consider the Stanford-ECM Dataset which contains syncronized data acquired with different sensors: video, acceleration and heart rate signals. The dataset is adapted to our action prediction task by identifying the transitions from the generic "Unknown" class to a specific "Activity". We discuss and compare a Siamese Network with the Multi Layer Perceptron and the 1D CNN where the input is an unknown observation and the output is the next activity to be observed. The feature representations obtained with the considered deep architecture are classified with SVM or KNN classifiers. Experimental results pointed out that prediction from multimodal data seems a feasible task, suggesting that multimodality improves both classification and prediction. Nevertheless, the task of reliably predicting next actions is still open and requires more investigations as well as the availability of multimodal dataset, specifically built for prediction purposes.

1 INTRODUCTION

The prediction of the future is a challenge that has always fascinated humans. As reported in (Lan et al., 2014), given a short video or an image, humans can predict what is going to happen in the near future. The overall design of machines that anticipate future actions is still an open issue in Computer Vision. In the state of the art, there are many applications in robotics and health care that use this predictive characteristic. For example, (Chan et al., 2017) proposed a RNN model for anticipating accidents in dashcam videos. (Koppula and Saxena, 2016; Furnari et al., 2017) studied how to enable robots to anticipate human-object interactions from visual input, in order to provide adequate assistance to the user. (Koppula et al., 2016; Mainprice and Berenson, 2013; Duarte et al., 2018) studied how to anticipate human activities for improving the collaboration between human and robot. In (Damen et al., 2018), the authors propose a new dataset, called Epic-Kitchen Dataset, and action and anticipation challenges have been investigated.

In this paper we consider the problem of predicting user actions. Since the information in the real world comes from different sources and can be captured by different sensors, our goal is to predict an action before it happens from multimodal observed data.

Multimodal learning aims to build models that are able to process information from different modalities, semantically related, to create a shared representation of them. For example, given an image of a dog and the word "dog ", we want to project these data in a representation space that takes account of both source domains.

As reported in (Srivastava and Salakhutdinov, 2014), each modality is characterized by different statistical properties, and hence each one of it can add valuable and complementary information to the shared representation. A good model for multimodal learning must satisfy certain properties. In fact the shared representation must be such that resemblance in the shared space of representation implies that the similarity of the inputs can be easily obtained even in the absence of some modalities.

Another aspect, not less important than the previous one, is represented by the data; in particular, they are collected at different sampling frequencies, there-

Rotondo, T., Farinella, G., Tomaselli, V. and Battiato, S.
Action Anticipation from Multimodal Data.
DOI: 10.5220/0007379001540161
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 154-161
ISBN: 978-989-758-354-4

fore, before to features captioning, it is necessary to synchronize the various inputs in order to have all the related modalities properly aligned.

This paper presents a study of predicting a future action from currently observed multimodal data. To this aim, the Stanford-ECM Dataset (Nakamura et al., 2017) has been considered. It comprises video, acceleration and heart rate data. We adapted this dataset to extract transitions from unknown to specific activities. Siamese network with Multi Layer Perceptron and 1D CNN are used for predicting next activity just from features extracted from the previous temporal sequence, labeled as "Unknown". The feature representations obtained with the considered deep architecture are classified with SVM or KNN classifier.

The prediction accuracy of the tested models is compared with respect to the classic action classification which is considered as a baseline. Results demonstrate that the presented system is effective in predic-ting activity from an unknown observation and suggest that multimodality improves both classification and prediction in some cases. This confirms that data from different sensors can be exploited to enhance the representation of the surrounding context, similarly to what happens for human beings, that elaborate information coming from their eyes, ears, skin, etc. to have a global and more reliable view of the surrounding world.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 describes the dataset used in this paper. Section 4 details the building blocks of our system while section 5 presents the experimental settings and discusses the results. Finally, conclusions are given in Section 6.

2 RELATED WORKS

We focus our review to related work which focus on action anticipation and multimodal learning.

2.1 Action Anticipation

The goal of action anticipation is to detect and recognize a human action before it happens. The work of (Gao et al., 2017) proposes a Reinforced Encoder-Decoder (RED) network for action anticipation that takes multiple representations as input and learns to anticipate a sequence of future representations. These anticipated representations are processed by a classification network for action classification. In (Lan et al., 2014), it is presented a hierarchical model that represents the human movements to infer future actions from a static image or a short video clip. In (Ma et al., 2016), the authors proposed a method to improve training of temporal deep models to learn activity progression for activity detection and early recognition tasks. Since in the state of the art there are not sufficiently large datasets for action anticipation task, the work of (Damen et al., 2018) proposes a new dataset, called Epic-Kitchen Dataset. The authors show the great potential of the dataset for pushing approaches that target fine-grained video understanding to new frontiers.

2.2 Multimodal Learning

In this work, we are interested in considering multimodal inputs to address action anticipation. One of the first paper on Multimodal Learning is (Ngiam et al., 2011) where video and audio signals are used as input. The aim of the work is to embed the inputs into a shared representation in order to be able to use only a single modality at test time. The creation of a shared representation has also been treated in other works. In particular (Srivastava and Salakhutdinov, 2014; Aytar et al., 2017) build representations that are useful for several tasks, such as cross-modal retrieval or transferring classifiers between modalities.

In (Nakamura et al., 2017), a model for reasoning on multimodal data to jointly predict activities and energy expenditures is proposed. In particular, they consider Egocentric videos augmented with heart rate and acceleration signals. In (Wu et al., 2017), an on-wrist motion triggered sensing system for anticipating daily intention is proposed. The authors introduce a RNN method to anticipate intention and a policy network to reduce computation requirement.

3 DATASET

There are few publicly available multimodal datasets in literature. Table 1 shows the relevant multimodal datasets together with main characteristics and the presence of transitions between actions. Since sequences with transitions among activities are needed, in our experiments, we considered the egocentric multimodal dataset, called Stanford-ECM Dataset (Nakamura et al., 2017). This dataset comprises 31 hours of egocentric video (113 videos) synchronized with acceleration and heart rate data. The video and triaxial accelerations were captured with a mobile phone equipped with a 720×1280 resolution camera at 30fps and 30Hz, respectively. The lengths of the videos range from 3 minutes to about 51 minutes. The heart rate was collected with a wrist sensor every 5 seconds (0.2 Hz). These multimodal data were time-

Table 1: Relevant multimodal datasets together with main characteristics. The second and third columns indicate the acquisition modality. Fourth column indicates the number of action class, whereas column five is related to the number of subjects involved into the acquisition. The last two columns are related to the resolution of frames and the presence of transitions between actions.

Dataset	First Person	Third Person	♯ Class	♯ Subjects	Resolution	Transition
Multimodal Egocentric Activity Dataset (Song et al., 2016)	✓	✗	20	-	1280x720	✗
Daily Intention Dataset (Wu et al., 2017)	✓	✗	34	3	640x480	✓
Epic-Kitchen Dataset (Damen et al., 2018)	✓	✗	149	32	1920x1080	✗
CMU-MMAC Dataset (Torre et al., 2009)	✓	✓	31	39	800x600	✓
Stanford-ECM Dataset (Nakamura et al., 2017)	✓	✗	24	10	720x1280	✓

Table 2: Activity classes of Stanford-ECM Dataset.

Activity	*Activity*
1.BicyclingUphill	13.Shopping
2.Running	14.Strolling
3.Bicycling	15.FoodPreparation
4.PlayingWithChildren	16.TalkingStanding
5.ResistanceTraning	17.TalkingSitting
6.AscendingStairs	18.SittingTasks
7.Calisthenics	19.Meeting
8.Walking	20.Eating
9.DescendingStairs	21.StandingInLine
10.Cooking	22.Riding
11.Presenting	23.Reading
12.Driving	24.Background

synchronized through Bluetooth. Cubic polynomial interpolation was used to fill any gap in heart rate data. Finally, data have been aligned considering millisecond level at 30 Hz.

The activity classes present in the Stanford ECM-Dataset are listed in Table 2. There are 24 classes in total. "Background" is a miscellaneous activity class which includes activities such as taking pictures or parking a bicycle. The dataset has also an additional class, *unknown*, that is related to part of the data before or after an action occurs.

Since this dataset was created for classification task, we have reviewed it to be compliant to our action prediction task.

We considered a transition, suitable to build training and test sets: Unknown/Activity, where "Activity" means a generic activity different from "background" and "unknown". We cut each video around the Unknown/Activity transitions including 64 frames before and 64 after the transitions point. Since some transitions were represented with few samples, we have concentrated the analysis to the following 9 activities: Bicycling, Playing With Children, Walking, Strolling, Food Preparation, Talking Standing, Talking Sitting, Sitting Tasks and Shopping. Hence, the final dataset contains 309 transitions Unknown/Activity.

4 PROPOSED APPROACH

The proposed approach is synthetically sketched in Figure 1. The model considers the three modalities video, acceleration and heart rate as input after a feature extraction process. Moreover details of the different component of our approach will be given.

4.1 Problem

Let be $\mathbf{y}_t = (\mathbf{v}_t, \mathbf{a}_t, hr_t)^T$ the input vector at time t where $\mathbf{v}_t \in \mathbb{R}^2$ is a video, $\mathbf{a}_t \in \mathbb{R}^3$ is an acceleration signal and $hr_t \in \mathbb{R}$ is a heart rate data, we define the feature representation of video, acceleration and heart rate signal as $\mathbf{x}_t^v$, $\mathbf{x}_t^a$ and $\mathbf{x}_t^{hr}$ and $\mathbf{x}_t = (\mathbf{x}_t^v, \mathbf{x}_t^a, \mathbf{x}_t^{hr})^T$ the features vector at time t. Given $\mathbf{x}_t$ as input, we want to predict the label $label_{t+1}$ of the next action by observing only data before the activity starts.

4.2 Features Extraction

In this section we describe the feature representation $\mathbf{x}_t^v$, $\mathbf{x}_t^a$ and $\mathbf{x}_t^{hr}$ for each signal. The extraction of video and acceleration features is similar to (Nakamura et al., 2017).

For visual data, features are extracted from the pooling layer five of the Inception CNN architecture (Szegedy et al., 2015) pretrained on ImageNet (Deng et al., 2009). Each video frame has been transformed into a $\mathbf{x}_t^v$ feature vector of 1024 dimension. For acceleration data, we extracted features from raw signals through a temporal sliding window process considering a window size of 32fps. Time-domain features and frequency-domain features are extracted from raw signals. For time-domain features, mean, standard deviation, skewness, kurtosis, percentiles (10th, 25th, 50th, 75th, 90th), acceleration count for each axis and correlation coefficients between each axis are computed. For frequency-domain features, we consider the spectral entropy $J = -\sum_{i=0}^{N/2} \bar{P}_i \cdot \log_2 \bar{P}_i$ where $\bar{P}_i$ is the normalized power spectral density computed from Short Time Fourier Transform (STFT). Then,

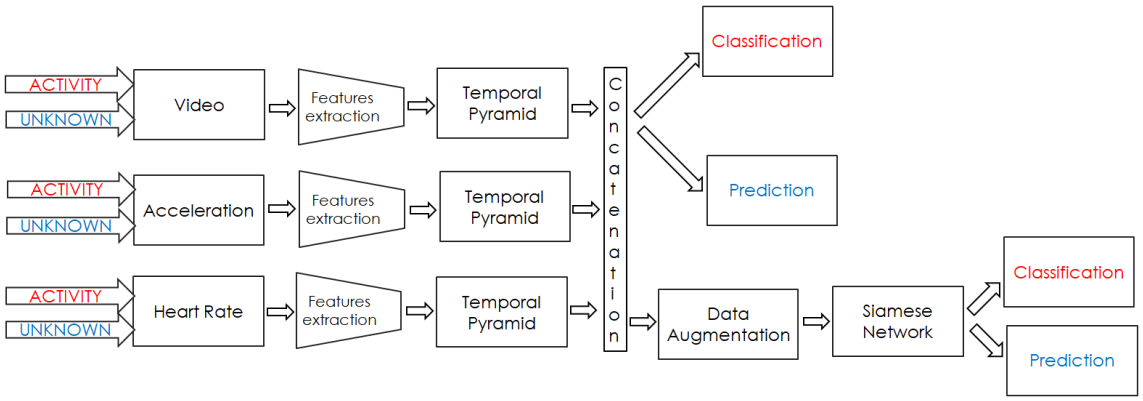

Figure 1: Pipeline of our anticipation approach.

the obtained features from these domains are concatenated and $\mathbf{x}_t^a$ is a 36-dimensional vector.

For heart rate data, the features are extracted from the time-series of the raw signals. Mean and standard deviation are calculated to compute a $\mathbf{x}_t^{hr} \in \mathbb{R}^2$ vector.

4.3 Temporal Pyramid

We represent features in a temporal pyramid fashion (Pirsiavash and Ramanan, 2012) composed by three level. The top level (j=0) is an histogram over the full temporal extent of a data, the next level (j=1) is the concatenation of two histograms obtained by temporally segmenting each modality into two halfs, the last level (j=2) is the concatenation of four histograms obtained by temporally segmenting each previous histogram into two halves. In this way, 7 histograms are obtained corresponding to a 1024×7 visual features, 36×7 acceleration features, and 2×7 heart rate features. All features are concatenated into a single vector $\mathbf{x}_t = (\mathbf{x}_t^v, \mathbf{x}_t^a, \mathbf{x}_t^{hr})^T$ of 7434 components.

4.4 Data Augmentation

Since we have few transition samples, data augmentation technique is used to expand the training set to prevent over-fitting. In this paper, the permutation Unknown/Activity is considered. Each unknown sequence is paired with all the possible sequences of activity. For example, we combined the unknown clip related to “walking” activity with every other activity. The label of each augmented transition is changed from 0-8 to 0-1, as follows: if unknown and the activity belong to the same class (e.g. unknown related to walking and the following activity is walking), we assign a label 1, otherwise a label 0 is assigned if unknown and the activity are different (e.g. unknown related to walking and activity is related to food preparation).

The obtained dataset is strongly unbalanced. Table 3 compares the number of sequences before and after augmentation. Some classes, such as Shopping or Food Preparation, are poorly represented therefore it is necessary to down-sample the dataset. We consider the square of minimum value of the number of original activity transitions ($11^2 = 121$) from sequences with label 1 and 154 sequences from sequences with label 0 for each class, in order to balance activities classes and unknown class. The final dataset has 12177 sequences.

4.5 Learning Approach

Our goal is to build an embedding space where the unknown sequences, which are related to the past, are close to those of future activities. In this regard, we use Siamese networks (Bromley et al., 1993; Koch et al., 2015) which consist of twin networks that share weights and accept two different inputs. After learning process, two similar images should be mapped by the network to close points in the feature space because each network computes the same function. During training the two networks extract features from two inputs, while the final shared neuron measures the distance between the two feature vectors.

In our experiment, since the Siamese network will be trained to make representations of features of ”Unknown” sequences and next ”Activity” very close in the embedding space, one stream of the Siamese network processes the unknown features whereas the other stream processes those related to the activity. Euclidean metric is used as distance between inputs. The contrastive loss function (Hadsell et al., 2006) is used for training purposes:

$$Y\sqrt{D} + (1-Y)\sqrt{\max(1-D,0)} \qquad (1)$$

where Y is the ground truth activity label and D is the euclidean distance between two feature points.

Table 3: Number of sequences for each activity before and after augmentation.

Activity	♯ of original activity transitions	♯ of augmented activity transitions	♯ of final transitions
Bicycling	18	4482	1353
Walking	79	19671	1353
Shopping	11	2739	1353
Talking Standing	26	6474	1353
Sitting Tasks	17	4233	1353
Playing With Children	32	7968	1353
Strolling	32	7968	1353
Food Preparation	14	3486	1353
Talking Sitting	20	4980	1353
TOT	249	62001	12177

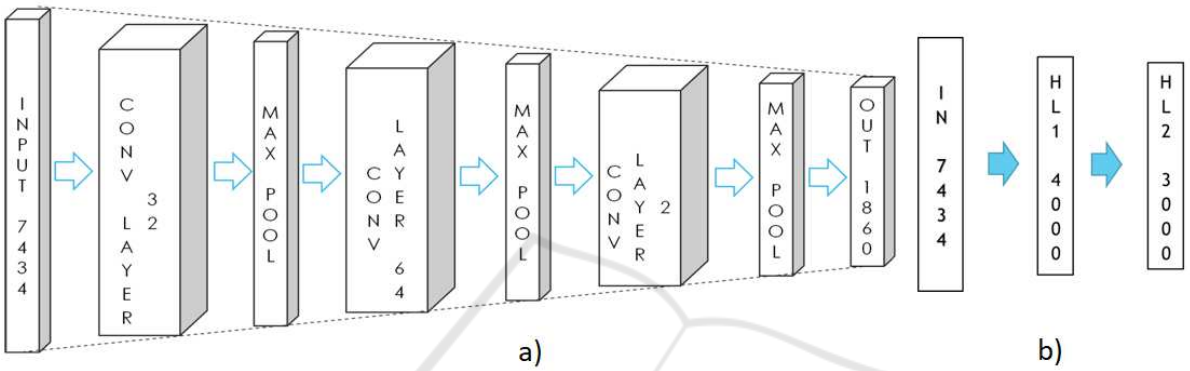

Figure 2: a) 1D CNN Architectures. b) MultiLayer Perceptron architecture.

We consider two different architectures for Siamese network: Multilayer Perceptron (Bishop, 2006) and 1D Convolutional Neural Network (CNN) (Kiranyaz et al., 2016; Lee et al., 2017). Figure 2 shows the architecture of the used networks. For Multilayer Perceptron, two hidden layers are considered with a number of neurons of 4000 and 3000 respectively. For 1D CNN, three convolutional layers are used with a number of filters of 32, 64 and 2, respectively, (all of size 3×1) and a relu activation function. The output of each convolutional layer is reduced in size using a max-pooling layer that halves the number of features.

4.6 Classification and Prediction

Our aim is to predict next activity from an unknown clip. To our knowledge, in the state of the art, there are not results on action anticipation from multimodal data, therefore we consider as baseline the classification of activity sequences and the classification of unknown sequences. A k-nearest-neighbor classification algorithm (K-NN) and a support vector machine (SVM) are used for classification purposes.

5 EXPERIMENTS

In this section, the results of the proposed approach are shown and discussed. Our model is evaluated on Stanford-ECM Dataset. The feature representations obtained with the considered deep architectures are classified with SVM or KNN classifier.

5.1 Setup

We randomly split our data into disjoint training (249 sequences) and testing sets (60 sequences) for training and testing purposes. For Siamese Network, the Adam optimizer is considered with batch size of 249 samples. Variable learning rate is used starting from 0.001. In the Multilayer Perceptron, in order to prevent overfitting, we apply a dropout procedure during training. We evaluate K-NN for different values of k and SVM for different kernels. In K-NN classifier, we consider two different weights: uniform and distance. The first assigns equal weights to all points, while distance weight assigns weights proportional to the inverse of the distance from the query point.

Table 4: SVM Results.

	Classification		Prediction	
Modality (♯ Features)	**Linear Kernel**	**RBF**	**Linear Kernel**	**RBF**
Acceleration (252)	31.67%	46.67%	31.67%	46.67%
Heart rate (14)	33.33%	28.33%	33.33%	35%
Video(7168)	66.67%	68.33%	60%	56.67%
Acceleration+Heart rate (266)	36.67%	50%	38.33%	48.33%
Video+Acceleration (7420)	70%	71.67%	68.33%	68.33%
Video+Heart rate(7182)	66.67%	66.67%	60%	63.33%
Video+Acceleration+Heart rate (7434)	70%	68.33%	68.33%	68.33%

Table 5: K-NN Results.

	Classification									
Modality (♯ Features)	**weights=uniform**					**weights= distance**				
	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Acceleration (252)	41.67%	53.33%	48.33%	45%	45%	41.67%	51.67%	46.67%	50%	50%
Heart rate (14)	23.33%	21.67%	18.33%	23.33%	31.67%	23.33%	20%	15%	15%	15%
Video(7168)	63.33%	61.67%	61.67%	61.67%	60%	63.33%	61.67%	63.33%	66.67%	63.33%
Acceleration+Heart rate (266)	38.33%	46.67%	48.33%	45%	43.33%	38.33%	43.33%	45%	48.33%	46.67%
Video+Acceleration (7420)	61.67%	61.67%	63.33%	63.33%	61.67%	61.67%	58.33%	61.67%	65%	65%
Video+Heart rate(7182)	63.33%	61.67%	61.67%	61.67%	60%	63.33%	61.67%	65%	65%	63.33%
Video+Acceleration+Heart rate (7434)	60%	65%	63.33%	63.33%	58.33%	60%	61.67%	63.33%	65%	63.33%
	Prediction									
Modality (♯ Features)	**weights=uniform**					**weights= distance**				
	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Acceleration (252)	41.67%	53.33%	48.33%	45%	45%	41.67%	51.67%	46.67%	50%	50%
Heart rate (14)	20%	26.67%	26.67%	25%	35%	20%	20%	20%	23.33%	26.67%
Video(7168)	55%	56.67%	60%	56.67%	60%	55%	58.33%	58.33%	56.67%	63.33%
Acceleration+Heart rate (266)	33.33%	46.67%	48.33%	45%	48.33%	33.33%	41.67%	45%	45%	48.33%
Video+Acceleration (7420)	53.33%	58.33%	60%	60%	56.67%	53.33%	61.67%	58.33%	60%	60%
Video+Heart rate(7182)	55%	56.67%	60%	56.67%	60%	55%	58.33%	58.33%	56.67%	63.33%
Video+Acceleration+Heart rate (7434)	53.33%	58.33%	61.67%	60%	56.67%	53.33%	61.67%	60%	60%	58.33%

5.2 Baseline

In order to better evaluate our approach, we define a baseline where the values of accuracy in classification and in prediction are compared. In classification, the features related to activity sequence, extracted as described in Session 4.2, are classified, while in prediction we consider the classification of features related to unknown clips.

The Tables 4 and 5 show the values of accuracy for each signals and combinations of all of them. For example, if we consider the accuracy values of video features, in Table 4, we can see that, with a linear kernel, we obtain an accuracy value of 66.67% in classification and a value of 60% in prediction; if we combine video features with acceleration data, for instance, the values are 70% in classification and 68.33% in prediction. These results suggest two conclusions. The first is that, as it is easily understandable, the values of accuracy in classification are higher than those in prediction, but not so much higher, therefore it is possible to anticipate the future action. The second is that most of the information comes from the video, but if we combine video with another signal, such as acceleration, the value of accuracy increases. The same conclusions are obtained with K-NN classifier.

5.3 Siamese Network

Our goal is to predict the label of the next action by observing only data before the activity starts. Our baseline suggests that it is necessary to fill the gap between the accuracy of classification and that of the prediction. As discussed in previous section 4, we consider a Siamese network for our purpose. Two different architectures are used: Multilayer Perceptron and a 1D CNN. The interesting point is that with a 1D CNN we can consider three convolutional layers therefore our output has dimension of 1860 while with a MLP we have only two layers and the output size is 3000. Table 6 and Table 7 show the results of the Siamese network. The tables list the obtained accuracy with K-NN and SVM classifier both for classification and anticipation. With a Siamese Network composed by a Multilayer Perceptron, results on anticipation are not so good and are even worse, in most cases, than those obtained by the baseline. This could be due to the difficulty of the MLP to learn from a very tiny dataset. More in details, the number of parameters of the network (7434x4000x3000) is too big with respect to the dataset size. Table 7 shows results obtained by training the considered classifiers on the representation learned through a Siamese Network, by exploiting a 1D convolutional layer architecture.

Table 6: Siamese Network Results considering a MultiLayer Perceptron architecture.

	KNN							
	Classification				Prediction			
k	Baseline		Siamese		Baseline		Siamese	
	weights=uniform	weights=distance	weights=uniform	weights= distance	weights=uniform	weights= distance	weights=uniform	weights= distance
1	60%	60%	58.33%	58.33%	53.33%	53.33%	55%	55%
3	65%	61.67%	60%	60%	58.33%	61.67%	55%	55%
5	63.33%	63.33%	58.33%	58.33%	61.67%	60%	55%	55%
7	63.33%	65%	56.67%	56.67%	60%	60%	53.33%	53.33%
9	58.33%	63.33%	56.67%	56.67%	56.67%	58.33 %	53.33%	53.33%
	SVM							
	Linear Kernel	RBF	Linear Kernel	RBF	Linear Kernel	RBF	Linear Kernel	RBF
	70%	68.33%	58.33%	46.67%	68.33%	68.33%	55%	56.67%

Table 7: Siamese Network Results considering a 1D CNN architecture.

	KNN							
	Classification				Prediction			
k	Baseline		Siamese		Baseline		Siamese	
	weights=uniform	weights=distance	weights=uniform	weights= distance	weights=uniform	weights= distance	weights=uniform	weights= distance
1	60%	60%	50%	50%	53.33%	53.33%	55%	55%
3	65%	61.67%	50%	53.33%	58.33%	61.67%	51.67%	58.33%
5	63.33%	63.33%	55%	55%	61.67%	60%	63.33%	66.67%
7	63.33%	65%	55%	58.33%	60%	60%	63.33%	65%
9	58.33%	63.33%	55%	60%	56.67%	58.33 %	58.33%	65%
	SVM							
	Linear Kernel	RBF	Linear Kernel	RBF	Linear Kernel	RBF	Linear Kernel	RBF
	70%	68.33%	71.67%	65%	68.33%	68.33%	60%	60%

The best values of accuracy are obtained with K-NN for k equals 5 and k equals 7. Indeed, if we compare the accuracy values of our baseline in the Table 7 for $k = 5$ and weights=distance, we have 63.33% for classification, 60% for prediction whereas the Siamese network overcomes these values obtaining a 66.67% of accuracy. For k=7, results show that the accuracy value with a Siamese network is equal to 65%, in other words, the same value of accuracy obtained for classification baseline. It is also interesting to note that the representation generated by the Siamese Network is not suitable in this case for classification task; in fact, accuracy achieved in classification is quite lower than that of the simple baseline. This could be due to the fact that the Siamese network has been trained to solve the challenge of making representations of features of "Unknown" sequence and next "Activity" very close in the embedding space with few samples. The results achieved with the SVM classifier do not reach the accuracy of the baseline.

6 CONCLUSION

This work presents preliminary results on action anticipation from multimodal data. In particular, the Stanford-ECM Dataset has been considered to address the problem. We compared the performances of different architecture and classifiers. Our preliminary results suggest that multi-modality improves both classification and prediction, but we couldn't deeply take advantage of deep learning approaches on multi-modal data due to a very limited dataset for training the methods. Future works could be aimed to improve the overall pipeline in order to fill the gap between classification and prediction performances and to test algorithm on bigger multimodal datasets, specifically built with the aim of addressing prediction and anticipation.

ACKNOWLEDGEMENTS

This research is supported by STMicroelectronics and Piano della Ricerca 2016-2018 Linea di Intervento 2 of DMI, University of Catania. We also thank the authors of (Nakamura et al., 2017) for providing the original Stanford-ECM dataset.

REFERENCES

Aytar, Y., Vondrick, C., and Torralba, A. (2017). See, hear, and read: Deep aligned representations. abs/1706.00932.

Bishop, C. M. (2006). *Patter Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1993). Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, pages 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chan, F.-H., Chen, Y.-T., Xiang, Y., and Sun, M. (2017). Anticipating accidents in dashcam videos. In Lai, S.-H., Lepetit, V., Nishino, K., and Sato, Y., editors, *Asian Conference on Computer Vision*, pages 136–153, Cham. Springer International Publishing.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., and Wray, M. (2018). Scaling egocentric vision: The epic-kitchens dataset. *European Conference on Computer Vision*.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Duarte, N., Tasevski, J., Coco, M. I., Rakovic, M., and Santos-Victor, J. (2018). Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters*, abs/1802.02788.

Furnari, A., Battiato, S., Grauman, K., and Farinella, G. M. (2017). Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401 – 411.

Gao, J., Yang, Z., and Nevatia, R. (2017). RED: reinforced encoder-decoder networks for action anticipation. *British Machine Vision Conference*, abs/1707.04818.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Kiranyaz, S., Ince, T., and Gabbouj, M. (2016). Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 63(3):664–675.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*.

Koppula, H. S., Jain, A., and Saxena, A. (2016). cipatory planning for human-robot teams. In *Experimental Robotics: The 14th International Symposium on Experimental Robotics*.

Koppula, H. S. and Saxena, A. (2016). Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):14–29.

Lan, T., Chen, T.-C., and Savarese, S. (2014). A hierarchical representation for future action prediction. In *European Conference on Computer Vision – ECCV*, pages 689–704, Cham. Springer International Publishing.

Lee, S.-M., Yoon, S. M., and Cho, H. (2017). Human activity recognition from accelerometer data using convolutional neural network. In *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 131–134.

Ma, S., Sigal, L., and Sclaroff, S. (2016). Learning activity progression in lstms for activity detection and early detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1942–1950.

Mainprice, J. and Berenson, D. (2013). Human-robot collaborative manipulation planning using early prediction of human motion. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 299–306.

Nakamura, K., Yeung, S., Alahi, A., and Fei-Fei, L. (2017). Jointly learning energy expenditures and activities using egocentric multimodal signals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6817–6826.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 689–696, USA. Omnipress.

Pirsiavash, H. and Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854.

Song, S., Cheung, N., Chandrasekhar, V., Mandal, B., and Lin, J. (2016). Egocentric activity recognition with multimodal fisher vector. abs/1601.06603.

Srivastava, N. and Salakhutdinov, R. (2014). Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Torre, F. D., Hodgins, J. K., Montano, J., and Valcarcel, S. (2009). Detailed human data acquisition of kitchen activities: the cmu-multimodal activity database (cmu-mmac). In *CHI 2009 Workshop. Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research.*

Wu, T., Chien, T., Chan, C., Hu, C., and Sun, M. (2017). Anticipating daily intention using on-wrist motion triggered sensing. *Intenational Conference on Conmputer Vision*, abs/1710.07477.