ANTICIPATING ACTIVITY FROM MULTIMODAL SIGNALS

T. Rotondo[1], G. M. Farinella[1], D. Giacalone[2] , S. M. Strano[2] , V. Tomaselli[2] , S. Battiato[1]

[1]Department of Mathematics and Computer Science, University of Catania, Italy

[2] STMicroelectronics, Catania, Italy

tiziana.rotondo@unict.it, {gfarinella, battiato}@dmi.unict.it, {davide.giacalone, mauro.strano, valeria.tomaselli} @st.com

UNIVERSITÀ degli STUDI di CATANIA

IMAGE PROCESSING LABORATORY

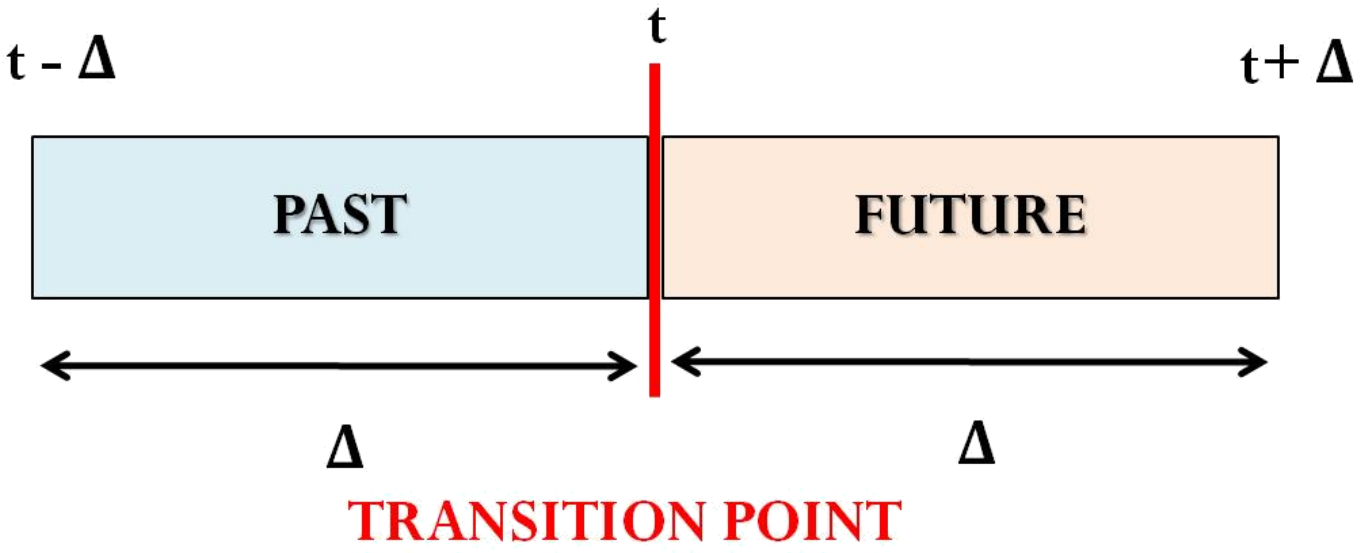
life.augmented

Outline

- Problem definition
- ST Multimodal Dataset
- Proposed Pipeline and Results
- Conclusions and Future work

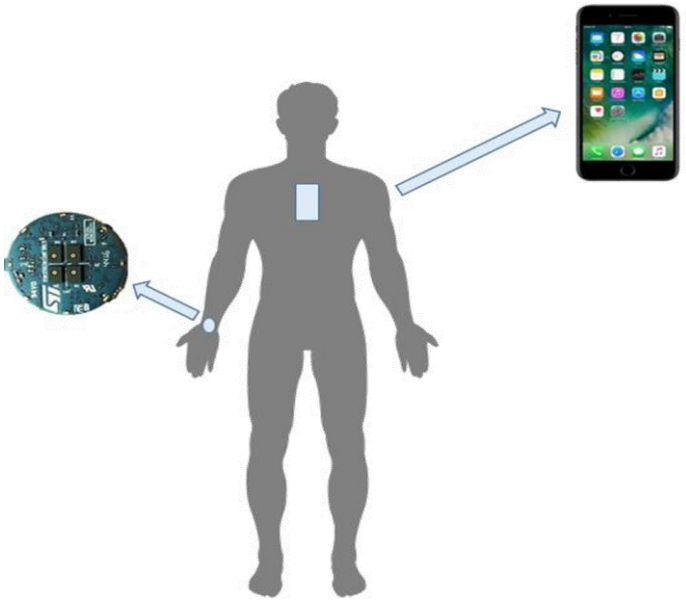
Problem Definition

- Given the features vector x_t at time t as input, the goal is to predict the label of the next action, observing only data before the activity starts.



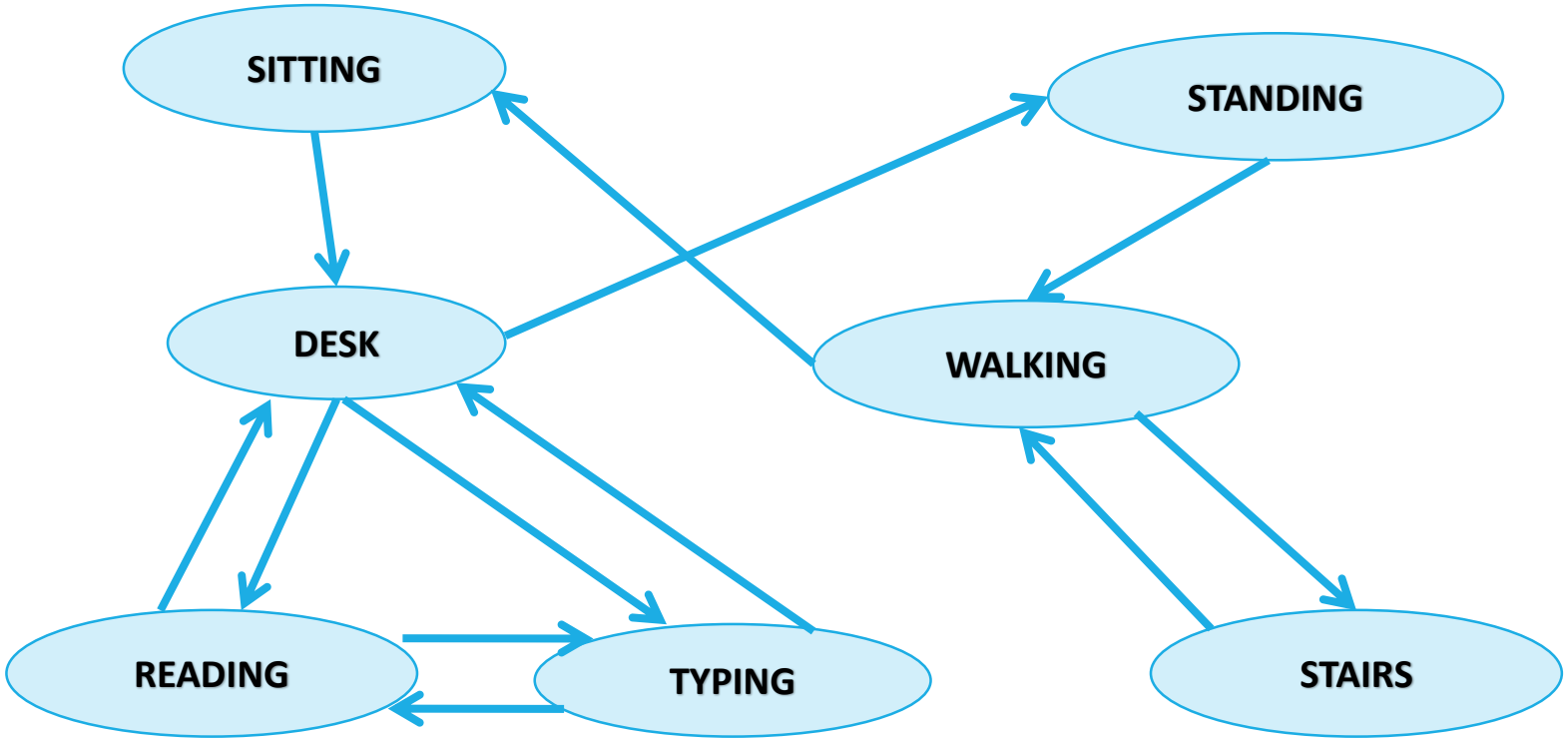
ST Multimodal Dataset - Modalities

- Mobile phone
  - Video
    - Camera Resolution: 720x1280
    - Frame rate: 29.95 fps
- Bluecoin:
  - Audio
    - Sampling rate: 32KHz
  - Acceleration, Gyroscope, Pressure, Temperature, Magnetic Field
    - Sampling rate: 52.63 Hz



ST Multimodal Dataset - Activities

- Activities: Desk, Reading, Sitting, Stairs, Standing, Typing, Walking.



ST Multimodal Dataset - # of sequences

Activity	# of sequences
Desk	2026
Reading	824
Typing	824
Walking	1353
Sitting	405
Stairs	420
Standing	405

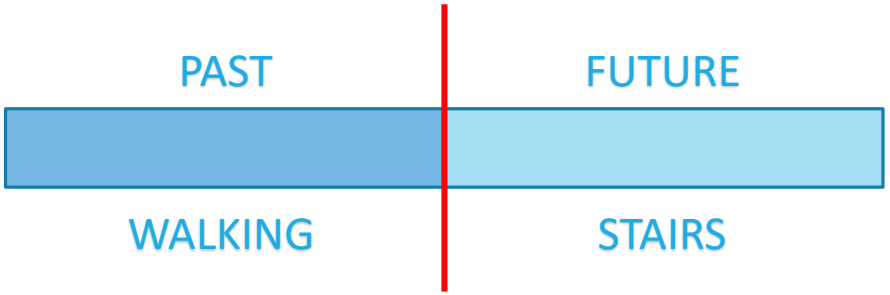
- "Desk" activity is more represented than "Stairs", "Sitting" or "Standing" activities because "Desk" is the past of "Typing", "Reading" and "Sitting" and the future of "Typing", "Reading" and "Standing".

- The collected dataset is balanced, i.e. the same number of sequences for each transition was acquired.

	Desk	Reading	Typing	Walking	Stairs	Sitting	Standing
Desk		418	411				419
Reading	411		418				
Typing	418	411					
Walking					428	419	
Stairs				428			
Sitting	419						
Standing				419			

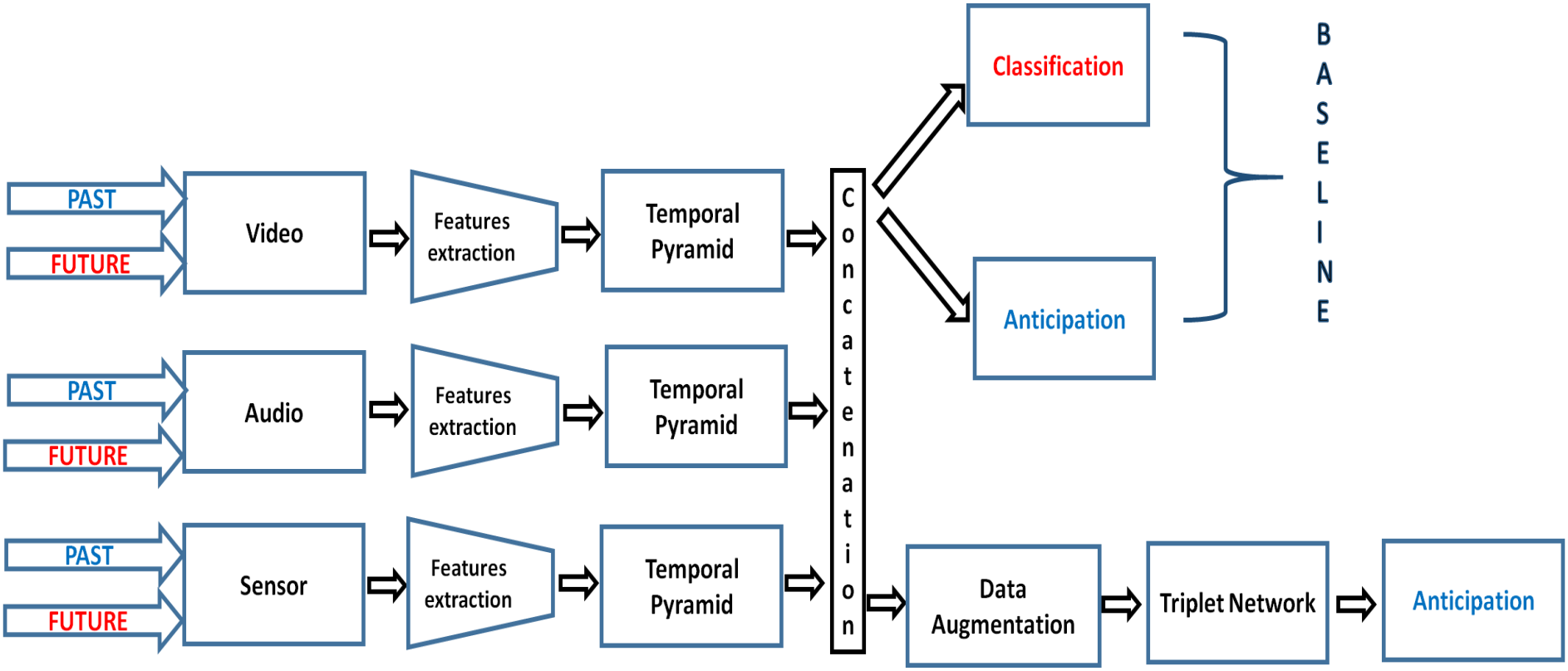
Cut- # of samples

- Time: 1,2160 sec
- Sensor: 64 samples
- Video: 36 frames
- Audio: 32768 samples



The dataset contains 4874 transitions.

Pipeline


PAST
FUTURE
Video
Features extraction
Temporal Pyramid
PAST
FUTURE
Audio
Features extraction
Temporal Pyramid
PAST
FUTURE
Sensor
Features extraction
Temporal Pyramid
Concatenation
Classification
Anticipation
BASELINE
Data Augmentation
Triplet Network
Anticipation

Baseline-SVM

	Classification		Anticipation	
	Linear SVM	Rbf SVM	Linear SVM	Rbf SVM
Video	66.70%	70.47%	61.35%	66.49%
Audio	31.62%	35.49%	32.52%	34.26%
Sensors	46.27%	58.44%	43.89%	53.18%
Video and Sensors	69.37%	72.89%	63.89%	70.57%
Video and Audio	67.27%	70.72%	61.52%	67.58%
Audio and Sensors	46.29%	58.83%	43.03%	54.04%
Video Audio and Sensors	69.55%	73.75%	64.06%	70.29%

Baseline-K-NN

	Classification					Anticipation				
	K=1	K=3	K=5	K=7	K=9	K=1	K=3	K=5	K=7	K=9
Video	60.02%	59.81%	59.51%	59.53%	59.61%	61.78%	61.31%	60.57%	60.45%	59.73%
Audio	29.75%	30.92%	31.54%	32.30%	32.15%	30.94%	31.56%	32.44%	32.42%	32.79%
Sensors	52.44%	52.58%	54.16%	53.42%	53.81%	47.82%	48.28%	48.71%	49.30%	49.82%
Video and Sensors	62.95%	62.66%	62.40%	61.72%	61.58%	64.18%	63.15%	63.09%	62.32%	61.87%
Video and Audio	60.33%	60.00%	59.96%	59.76%	59.61%	62.05%	61.72%	60.72%	60.08%	60.29%
Audio and Sensors	54.49%	55.12%	55.57%	55.27%	54.63%	51.29%	52.11%	52.29%	52.19%	51.02%
Video Audio and Sensors	63.24%	62.79%	62.42%	61.97%	61.76%	64.12%	63.61%	63.14%	62.27%	62.15%

Triplet Network-SVM

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%

Triplet Network-K-NN

	Classification	Prediction	
K	Baseline	Baseline	Triplet
K=1	63.24%	64.12%	64.65%
K=3	62.79%	63.61%	64.73%
K=5	62.42%	63.14%	64.14%
K=7	61.97%	62.27%	64.55%
K=9	61.76%	62.15%	64.18%

Conclusions

- 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

Thank you