Object Recognition and Tracking for Smart Audio Guides

Lorenzo Seidenari, Claudio Baecchi, Tiberio Uricchio, Andrea Ferracani, Marco Bertini, Alberto Del Bimbo

University of Florence

UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

Smart Audio Guides

- Audio Guides are the To-Go for delivering complex information in cultural heritage sites

- They are ofter cumbersome to use and not context sensitive

- Ideally an intelligent agent should:
  - Understand the user interest
  - Provide information at the right time
  - Avoid intrusiveness and be aware of context and distractions

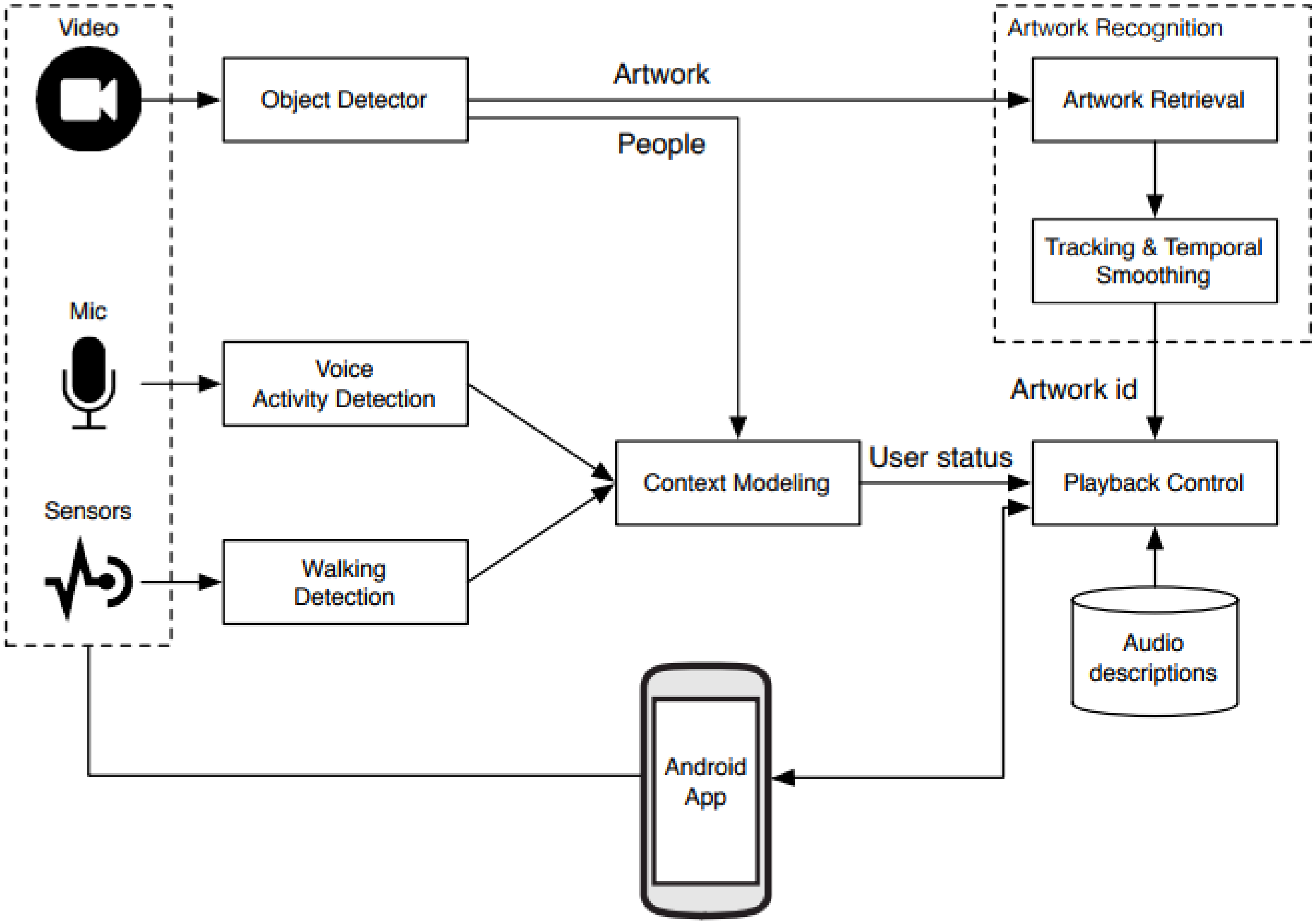
Wearable Computing

Project Goals

- Smart device understanding the environment
- Provide hands-free non-intrusive experience
- Augment reality via audio descriptions

Her, Spike Jonze 2013

System Architecture



SeeForMe Development Timeline

Full System

VAD & Sensors

AudioGuide App

Detection and Recognition

CUDA/GPU
YOLO C/C++
Linux

NVIDIA TK1

NVIDIA SHIELD TK1

CUDA/GPU
YOLO C/C++, JNI
Android

Android

~~CUDA/GPU~~
CPU
TensorFlow
Android

2015

2016

ACM MM16

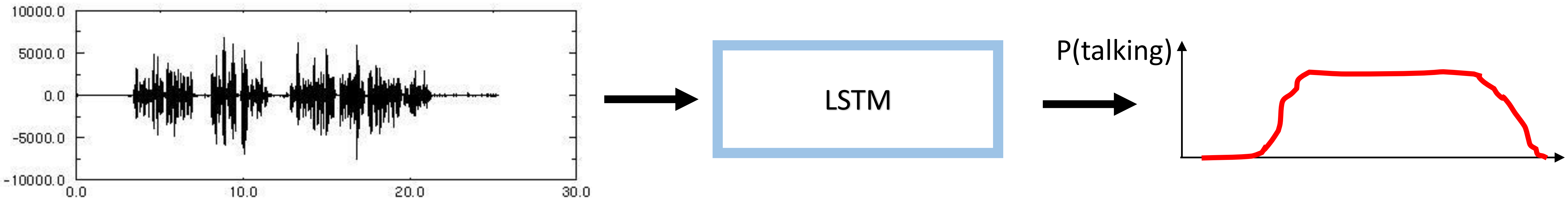
2017

ACM TOMM 17

2018

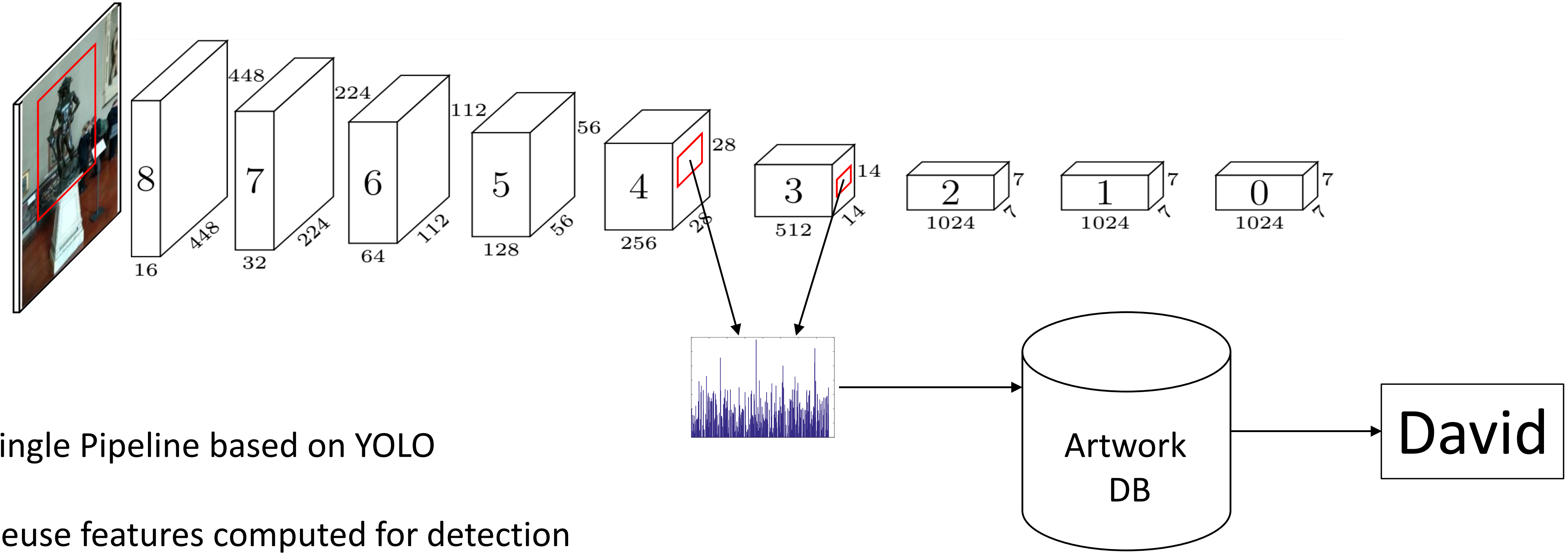
Detecting Conversations

- We detect conversations using a LSTM on the audio signal
- Audio Description fades out in case a conversation is detected



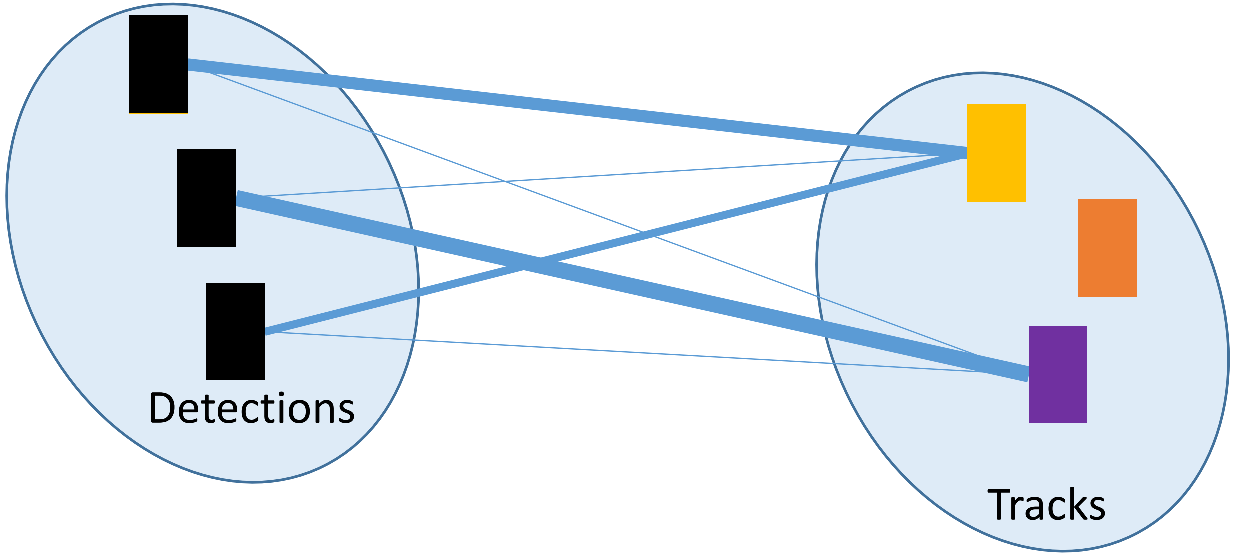
LSTM

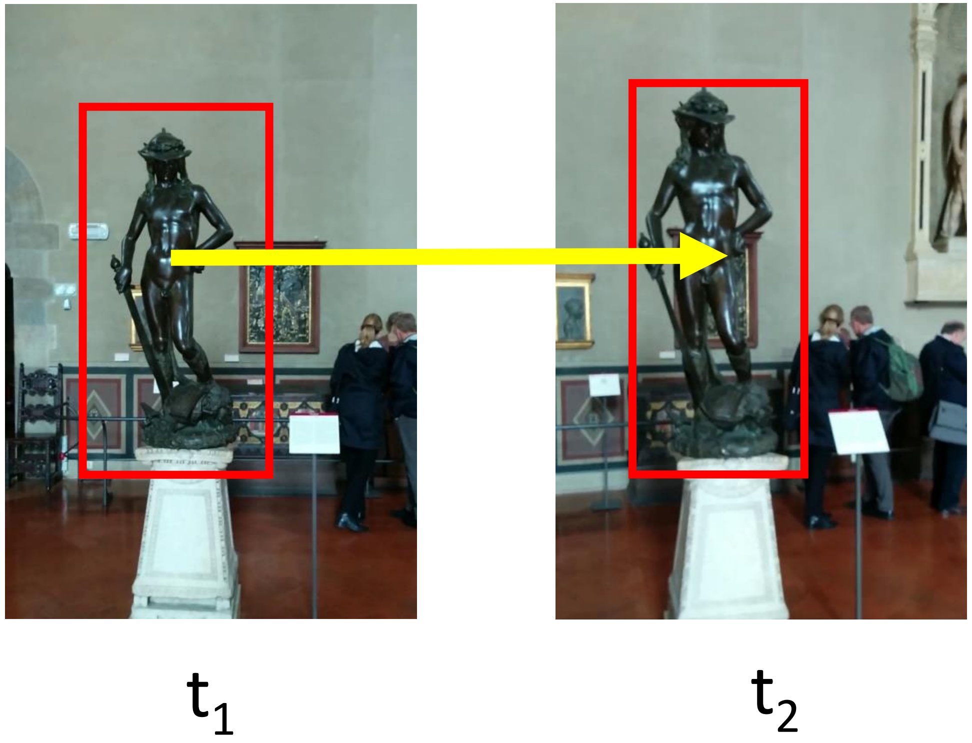
Object Detection and Recognition



- Single Pipeline based on YOLO
- Reuse features computed for detection
- Match artwork in local DB

Object Tracking



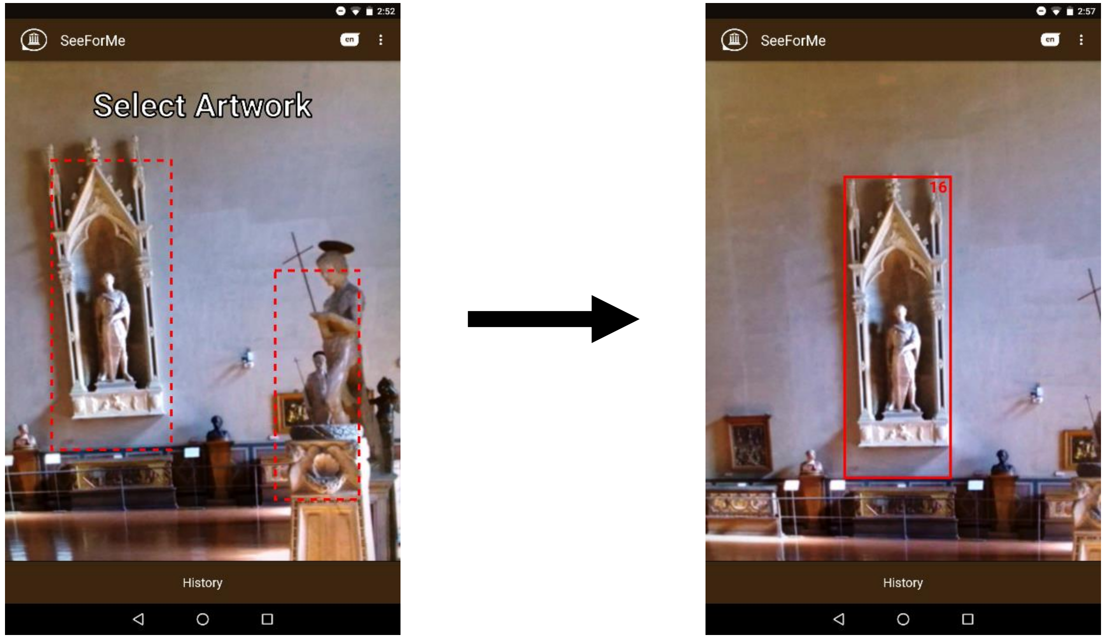


Tracking-by-detection with Greedy Data Association.

```
1. Bi-partite graph each edge is weighted by IoU
2. Associate when above a threshold
3. Unassociated detections become new tracks
4. Unassociated tracks are killed after k frames
```

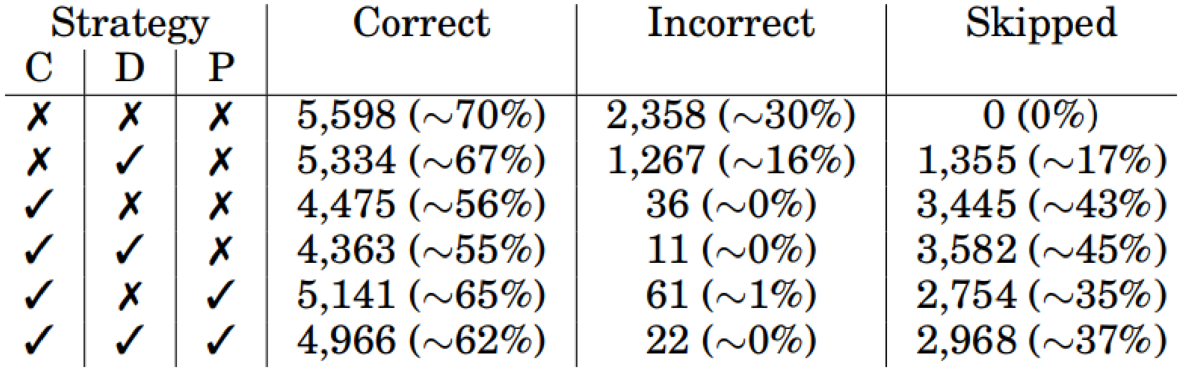
Database Bootstrap

- Extremely flexible recognition based on NN Search
- No learning required when new imagery is provided
- We bootstrap the system exploiting our tracker to annotate multiple frames.

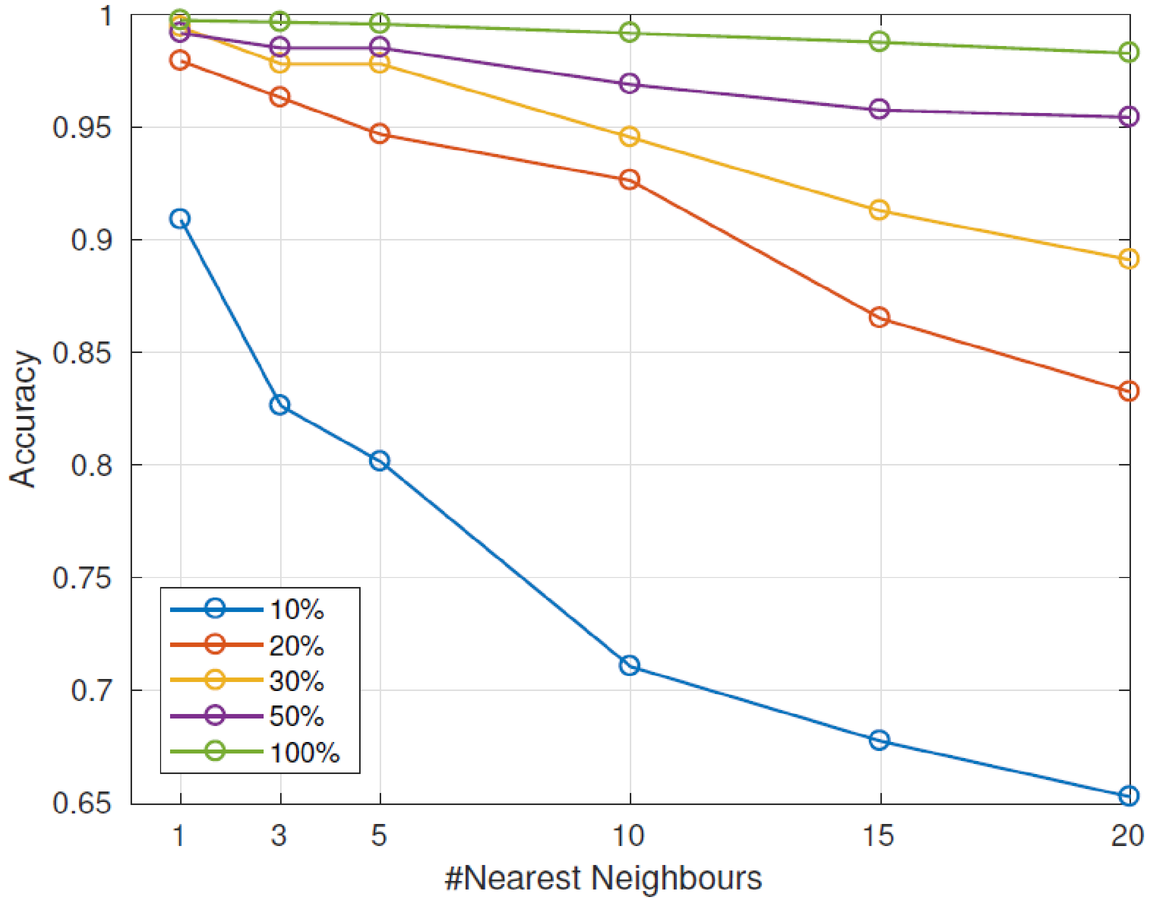


Experimental Results

- Growing dataset size allows to reach very high recognition accuracy
- Almost detrimental to use more than 1-NN for recognition
- Tracking and other filtering strategies allow a very low error rate

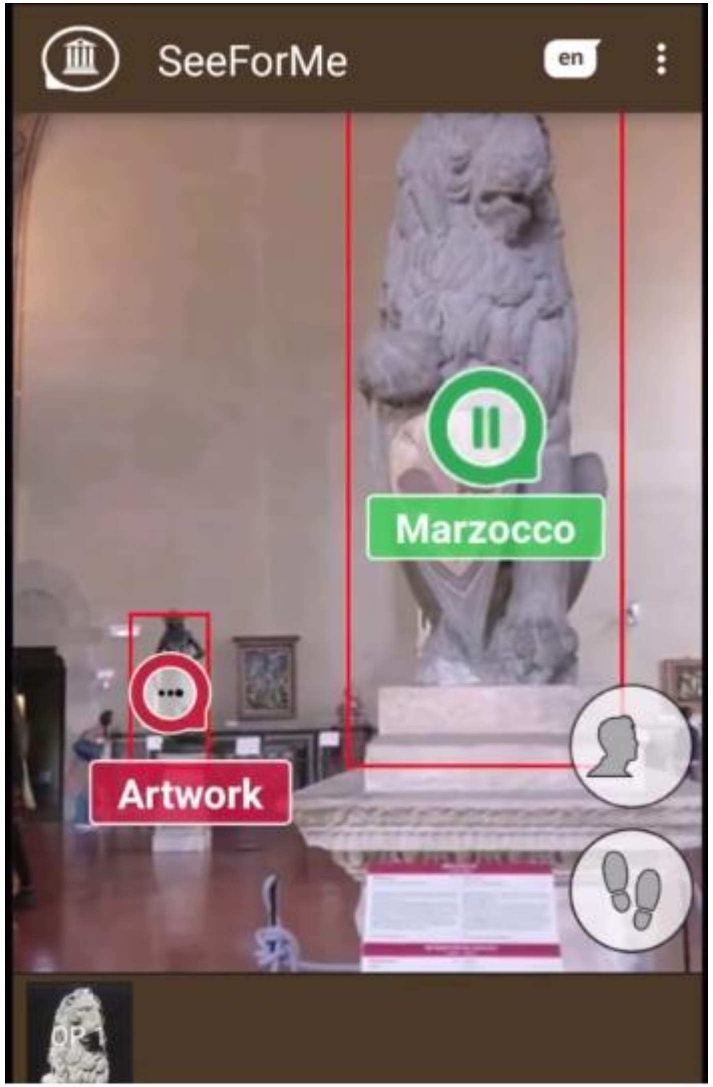


Strategy			Correct	Incorrect	Skipped
C	D	P			
✗	✗	✗	5,598 (∼70%)	2,358 (∼30%)	0 (0%)
✗	✓	✗	5,334 (∼67%)	1,267 (∼16%)	1,355 (∼17%)
✓	✗	✗	4,475 (∼56%)	36 (∼0%)	3,445 (∼43%)
✓	✓	✗	4,363 (∼55%)	11 (∼0%)	3,582 (∼45%)
✓	✗	✓	5,141 (∼65%)	61 (∼1%)	2,754 (∼35%)
✓	✓	✓	4,966 (∼62%)	22 (∼0%)	2,968 (∼37%)



Demo!

Video Demo Available at:
https://vimeo.com/187957085



Conclusion

We presented a fully automatic smart audio-guide understanding *user* attention and needs

Our method is based on an incremental library of artworks that can be grown by *curators*

Further Reading:

- Seidenari et al., ''Deep Artwork Detection and Retrieval for Automatic Context Aware Audio Guides'', ACM TOMM, 2017

Acknowledgements:

We would like to thank NVIDIA for the donation of a TITAN X Pascal. This project is partially supported by MIUR – Social Museum Smart Tourism (SMST)

MINISTERO DELL'ISTRUZIONE,
DELL'UNIVERSITÀ E DELLA RICERCA