PhD in Information Engineering - 17 October 2017, Firenze

Exploiting Digital Traces and Human Factors to Design and Improve HCI Systems for Online, Outdoor and Indoor Environments

A. Ferracani

http://www.micc.unifi.it

Report of Activities of the third year

- **Events detection** in outdoor environments using personal geo-localized Twitter Data [geolocation / frequency]
- **User profiling** in indoor for Automatic Context-Aware Audio Guides providing Obect Detection and Artwork Retrieval [user behaviour / context / interaction design]
- HCI systems for improving naturalness in virtual museums exploiting **Virtual Reality and Voice Commands** [human factors / interaction design]

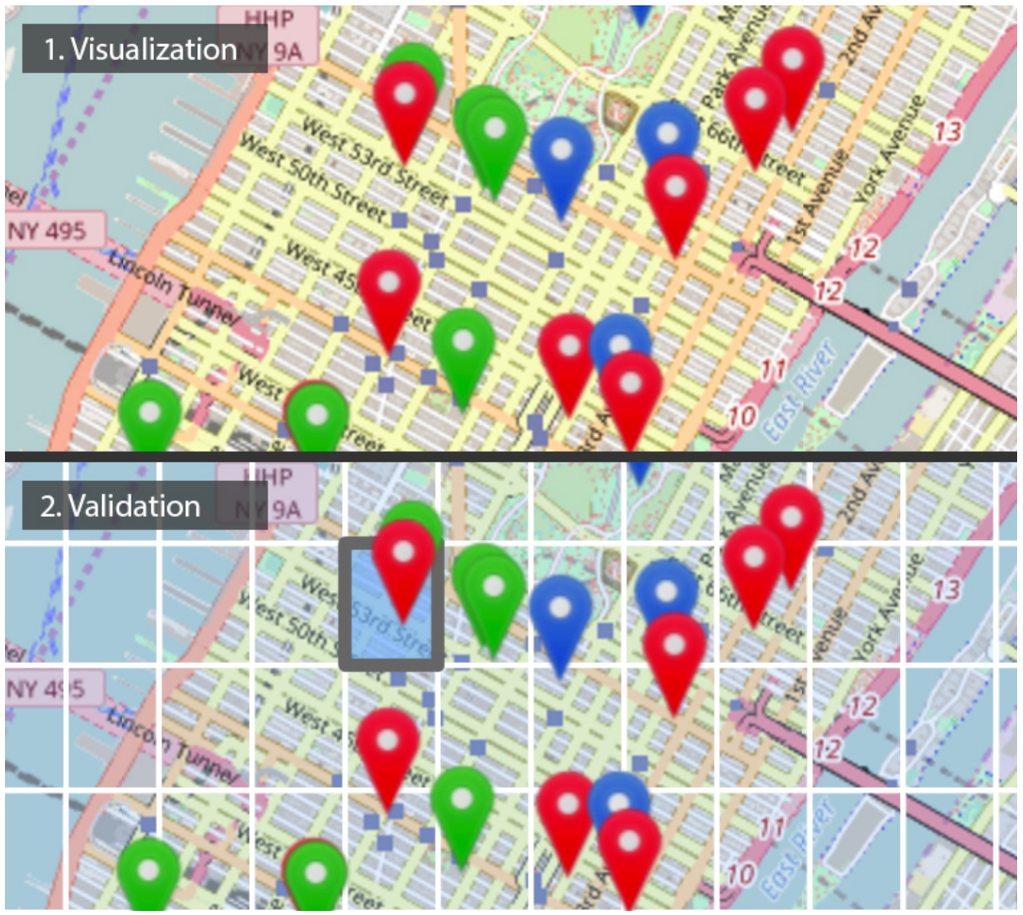
Events Detection - Summary

- A simple but effective method has been proposed and assessed for **unknown event detection** designed to alleviate computational issues in traditional approaches (amount of text, sparsity of geo-location, noisy text).

- The method is exploited by a **web interface** that exposes interactive tools to inspect, validate the data and configure the processing pipeline.

- The system can be exploited for the rapid creation of macro and micro-events **datasets** of geolocalized messages useful to improve supervised and unsupervised events classification on Twitter.

- **detection** and **classification of macro** and **micro-events** in cities through the analysis of geolocalized messages from Twitter;
- a **grid-based** approach is used



The proposed **processing pipeline** contemplates three main steps:

I. **tweets extraction**;
II. **abnormalities detection**;
III. **mining and visualization**.

The events detection is a sub-pipeline of three core algorithms based on statistical analysis of **temporal and spatial characteristics** of tweets distributions. An event is proposed as a candidate through this pipeline:

1) **DTW**: unusual volumes of tweeters in a grid cell are detected if plus/minus the average of the DTW distances plus/minus the DTW standard deviation.

$$d_i > \overline{d} + \sigma_d \quad \| \quad d_i < \overline{d} - \sigma_d$$

2) **Crest-Detection**: peaks are identified in the distribution in 1) as

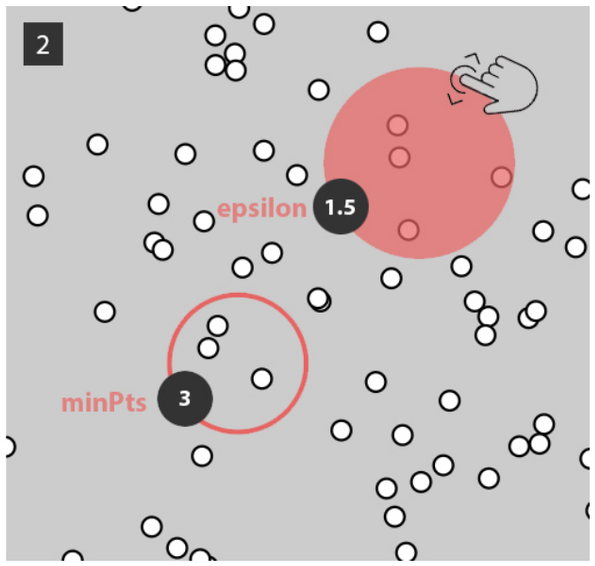
$$v(t) > \sum_{j=1,..,\Gamma_t} v(t-j) + \epsilon \quad \&\& \quad v(t) > \sum_{j=1,..,\Gamma_t} v(t+j) + \epsilon$$

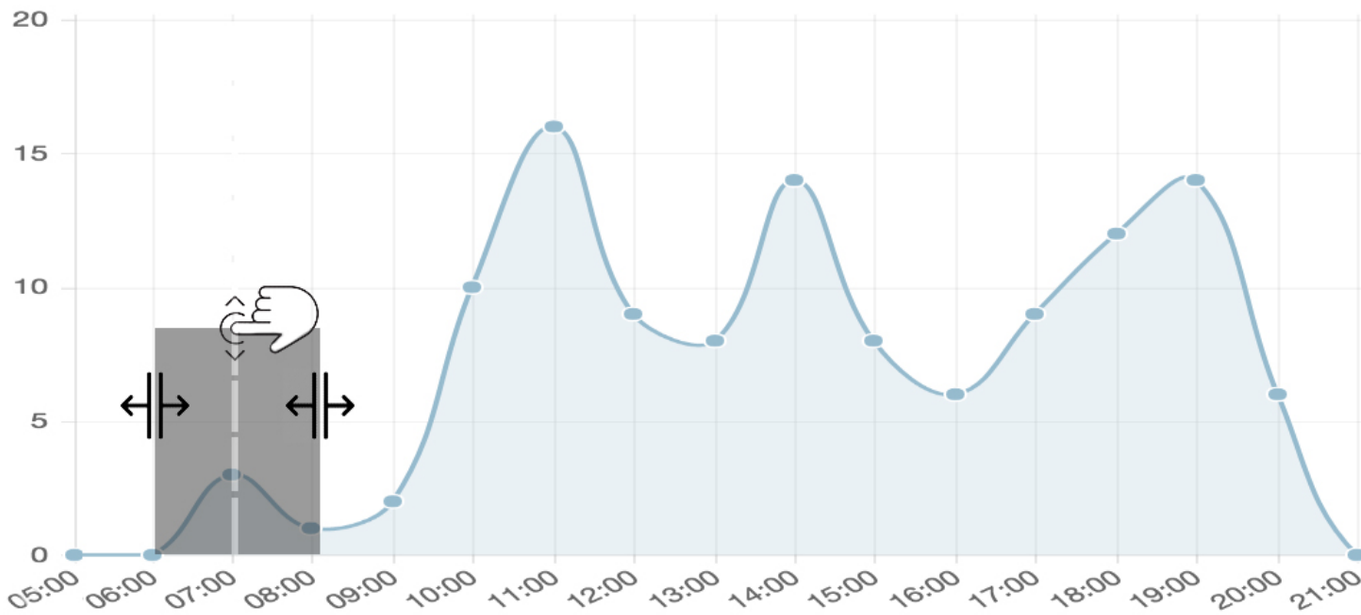
3) **DBSCAN**: density of geolocalized tweets in peaks detected in 2) is tested

The event detection method is exploited by

- a demo **web interface** that visualizes the results of the automatic computation and exposes **interactive tools** to manage the processing pipeline and validate the results;

- researchers can exploit the web application for the rapid creation of **macro** and **micro-events datasets** of geolocalized messages.





Evaluation:

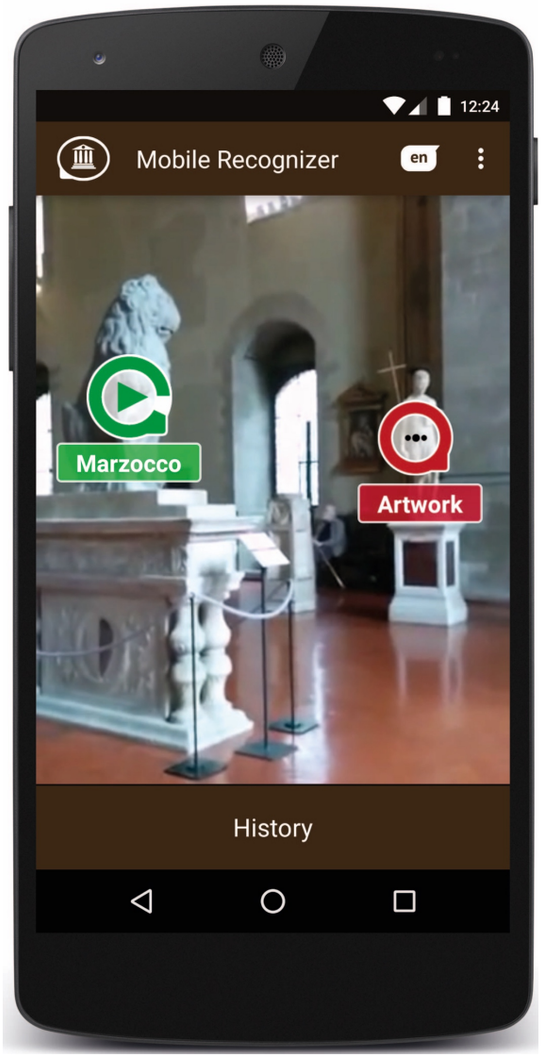
- tweets published in New York and London from March 31 to April 9, 2016
- **LND**: 17176 users / 44932 tweets > **190/340 events**
- **NY**: 17378 users / 43186 tweets > **516/900 events**
- **PRECISION: 0.57**

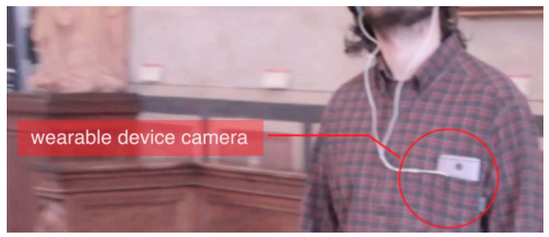
[1] **Andrea Ferracani**, Daniele Pezzatini, Lea Landucci, Giuseppe Becchi and Alberto Del Bimbo. 2017. **Separating the Wheat from the Chaff: Events Detection in Twitter Data**. In Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (CBMI '17 Florence). ACM, New York, NY, USA.

User profiling for Audio Guides - Summary

A smart audio guides featuring an embedded system which adapts to interests and behaviour of museum visitors through:

- Object / Person detection and **Artwork Recognition**
- Voice Activity Detection (**VAD**)
- **User behavioural understanding** (movements)
- **Voice Commands** enabled Playback Control
- Automatic or semi-automatic **Text-To-Speech** Synthesis



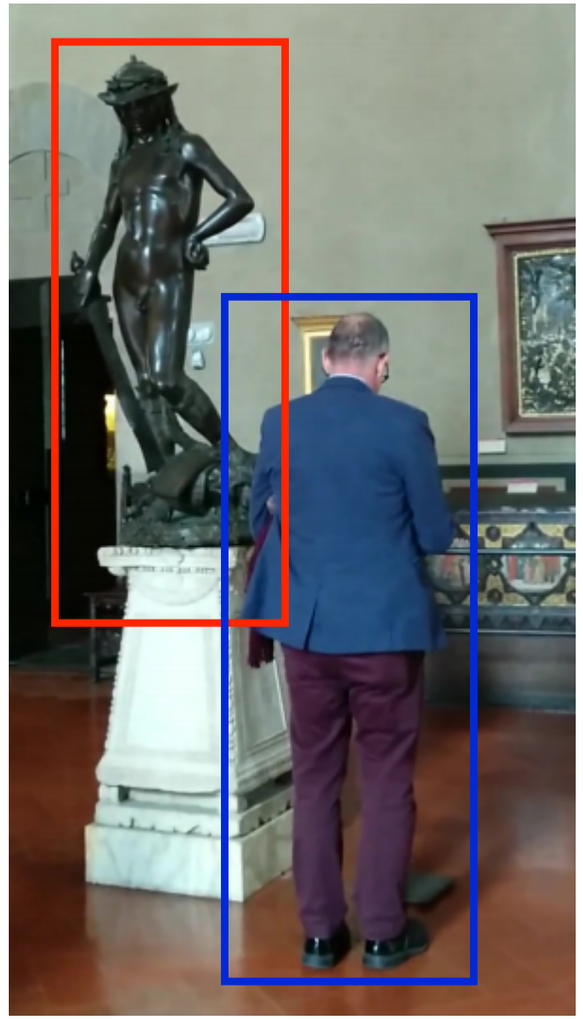


User profiling for Audio Guides - Object Detection

- Person detection allows **context understanding**, e.g. if the system stops ‘’seeing’’ an artwork it may be temporary occluded and the audio should not stop playing

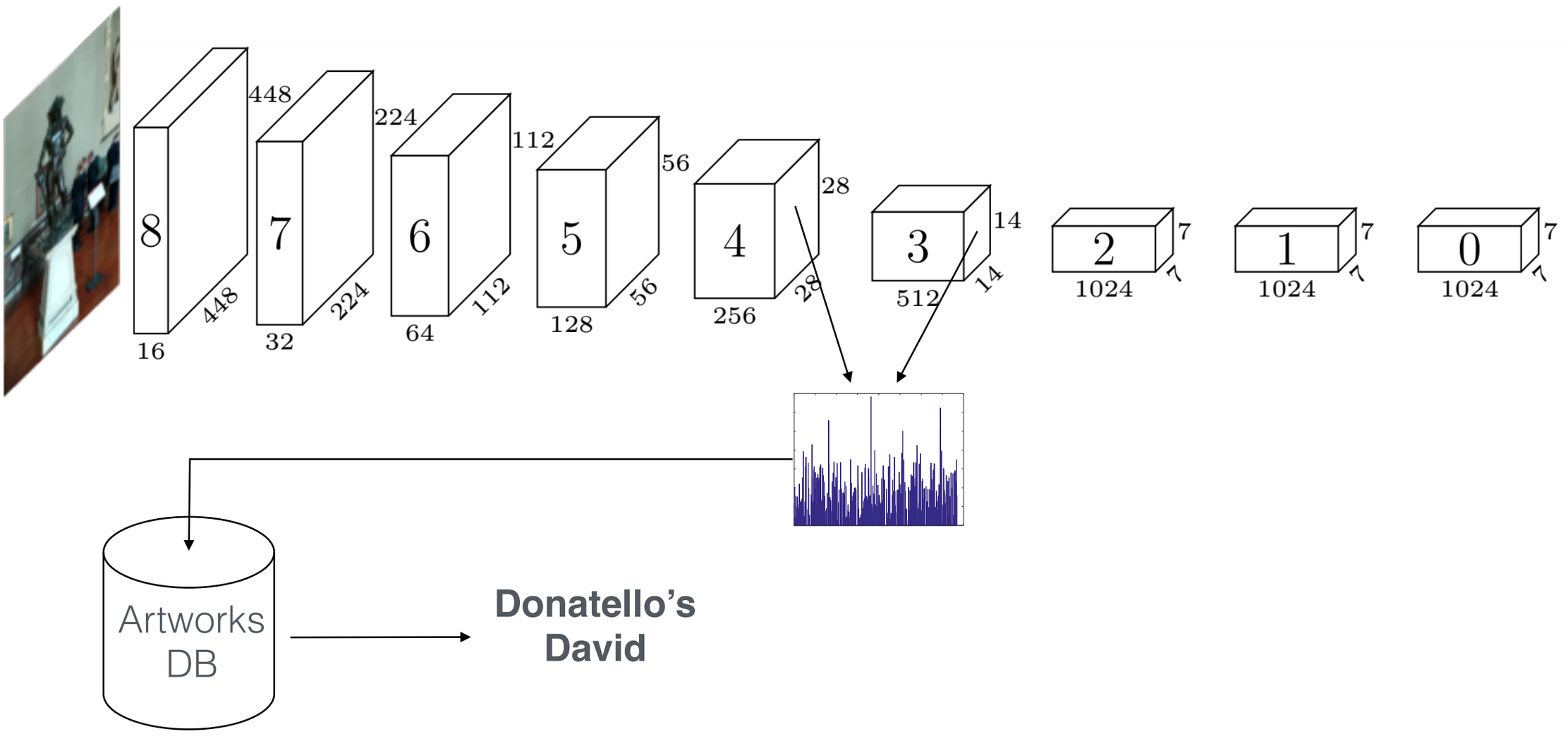
- Most artworks depict human figures, therefore we may get false artwork detections on people

- We fine-tuned the **Tiny-YOLO** CNN network to recognize person and artwork classes, using few hundreds (700+) of pictures collected from shots taken by tourist in museum, downloaded from TripAdvisor



User profiling for Audio Guides - Artwork Recognition

- We fine-tuned the network to recognize artworks and people. The CNN features computed for classification are then used to recognize specific artworks. **Artwork labels** are predicted using a **NNS** with respect to a pre-acquired dataset of artwork patches

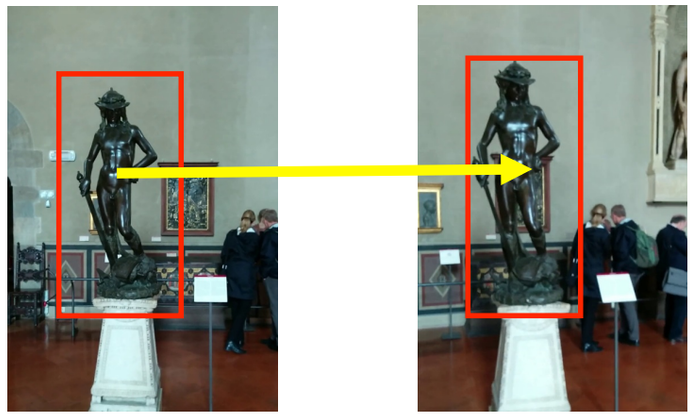


User profiling for Audio Guides - Tracking and temporal smoothing

- **Distance**: to reduce the error rate we avoid artwork recognition on distant objects with a simple heuristic

$$\frac{w_{bb} h_{bb}}{W H} > T$$

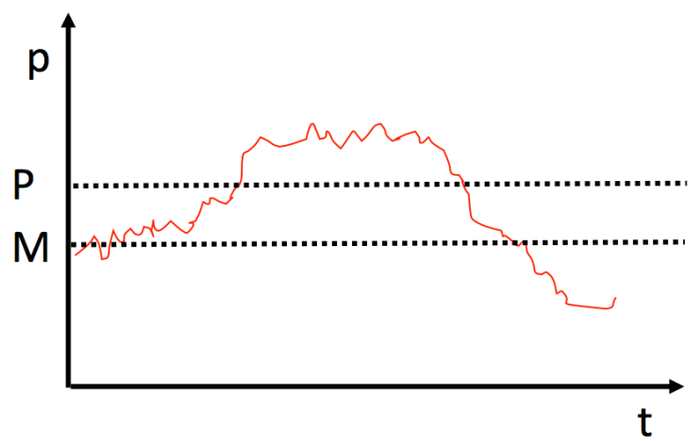
- **Consistency**: a prediction is considered only after it persists for *M* frames



t_1 t_2

- **Persistence**: a counter p is incremented every time the recognition label for a box is unchanged, keeping track of the most frequent label $\overline{y}$. Every time a label $y*$ differs from $\overline{y}$ we decrement p. We predict the artwork identity as $y*$ only if

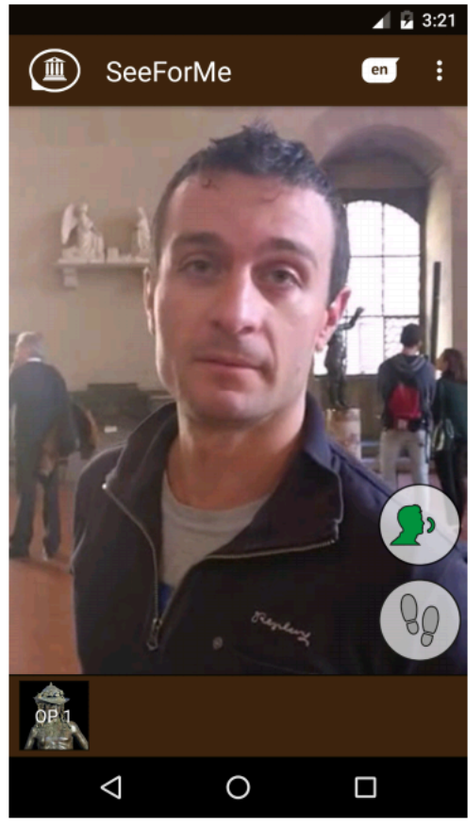
$$p > P > M$$



User profiling for Audio Guides - Context Modeling

- **Voice Activity Detection** is provided by the system exploiting the state-of-the-art Long Short Term Memory recurrent neural network by Eyben et al. [2013] implemented in the **OpenSmile** Framework.

- **Peaks in accelerometer device data** are detected to understand if visitor is walking and how fast. Gyroscope data is exploited to detect strong changes in direction and orientation.



These data are used in combination to the **Vision System** and the **Persistence** score in order to measure user attention and understand the context (i.e. occlusions, looking direction, talking with people). Audio information delivered to user is consequently managed.

User profiling for Audio Guides - Results and evaluation

- **Testing Dataset**: labeled video sequences of 8.820 frames with 7.956 detections

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%)

- **Evaluation**: we tested two different scenarios, supervised and unsupervised, in which users were asked to perform two simple tasks. Both experiments have shown a high usability score (SUS) of 74 and 79.5.

[3] Lorenzo Seidenari, Claudio Baecchi, Tiberio Uricchio, **Andrea Ferracani**, Marco Bertini, and Alberto Del Bimbo. 2017. **Deep Artwork Detection and Retrieval for Automatic Context-Aware Audio Guides**. ACM Trans. Multimedia Comput. Commun. Appl. 13, 3s, Article 35 (June 2017).

Virtual Reality and Voice Commands - Summary

- A system (a library) for immersive experiences in museums using **Voice Commands** and Virtual Reality specifically designed for use by people with motor disabilities

- The VR is visualised through an **Head-Mounted Display**. The library has been developed in Unity 3D

- Natural interaction is provided through **Automatic Speech Recognition exploiting** The Microsoft Speech API

- The system exploits **ontologies** in order to ‘augment’ the possibilities of the system in terms of voice interactions

Virtual Reality and Voice Commands - Issues

There's still **poor research** on how to exploit progress in ASR for developing effective and accessible speech controlled interfaces.

- Perceived distance between the player and the game character defined as **'identity dissonance'** in VR
- The **social context** where voice interaction takes place (e.g. the quiet environment of museums, privacy concerns);
- **Errors** in ASR (due to noise, spelling, etc.)
- Restricted freedom of speech in limited domain applications with VCs constituted by simple words or short phrases due to the **complexity of ASR in the wild**

Virtual Reality and Voice Commands

The library allows

- To insert and **position artworks** in a 3D environment
- To describe items using **triples {s, p, o}** through ontologies imported in or created by the system
- to dynamically define possible Voice Commands in a **grammar** [Microsoft Speech Recognition Grammar Specification **(SRGS)** Version 1.0]

Voice Commands have a certain degree of freedom since are automatically fed and augmented via a semantic storage provided with a **reasoner** capable of inferring concepts.

Virtual Reality and Voice Commands - Inference

Class(vc:ActionPainter complete intersectionOf(vc:Artist restriction(vc:exponentOf someValuesFrom(a:ActionPainting))))

Class(vc:AbstractPainter complete intersectionOf(vc:Artist restriction(vc:exponentOf someValuesFrom (vc:AbstractArt))))

Class(vc:ActionPainting partial vc:AbstractArt)

The following **class inference** can be derived:

- An Action Painter is an exponent of the Action Painting
- Action Painting is a type of Abstract Art
- an Action Painter is an exponent of the Abstract Art, so must be an Abstract Painter.

Virtual Reality and Voice Commands - Demo

- **Text-ToSpeech** synthesis is used by the guide to explain possible questions and to give responses.
- **E.g. “What artworks are there in the museum?”** is interpreted as a VC and mapped to SPARQL. The guide lists the artworks
- **E.g. “I’d like to see ‘The Starry Night’ by Vincent Van Gogh”**. The agent guides the visitor to the place where the artwork is located walking through the halls of the museum (the **A* algorithm** is exploited)

Virtual Reality and Voice Commands - Evaluation

- The usability of the system was preliminarily tested using the popular **Standard Usability Scale** (SUS)
- **10 users** were asked to perform the task of navigating the museum using VCs obtaining insights from the virtual guide on at least an artistic movement and an artwork.
- **Average SUS score was 71.0**. Scores are in the range [0–100] and over 68 mean that the interaction design is above average

[4] **Andrea Ferracani**, Marco Faustino, Gabriele Xavier Giannini, Lea Landucci, Alberto Del Bimbo, **Natural Experiences in Museums through Virtual Reality and Voice Commands**, Proceedings of the 25th ACM international conference on Multimedia (MM '17). ACM, New York, NY, USA, to appear.

Bibliography

[1] **Andrea Ferracani**, Daniele Pezzatini, Lea Landucci, Giuseppe Becchi and Alberto Del Bimbo. 2017. **Separating the Wheat from the Chaff: Events Detection in Twitter Data**. In Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (CBMI '17). ACM, New York, NY, USA, Article 14, 5 pages.

[2] Alberto Del Bimbo, Marco Bertini, Lorenzo Seidenari, Tiberio Uricchio, Claudio Baecchi, **Andrea Ferracani**, **Portable computer vision for new "intelligent" audio guides, EVA 2017 Florence** - Electroning Imaging & the Visual Arts, Aracne Editrice, 2017

[3] Lorenzo Seidenari, Claudio Baecchi, Tiberio Uricchio, **Andrea Ferracani**, Marco Bertini, and Alberto Del Bimbo. 2017. **Deep Artwork Detection and Retrieval for Automatic Context-Aware Audio Guides**. ACM Trans. Multimedia Comput. Commun. Appl. 13, 3s, Article 35 (June 2017)

[4] **Andrea Ferracani**, Marco Faustino, Gabriele Xavier Giannini, Lea Landucci, Alberto Del Bimbo, **Natural Experiences in Museums through Virtual Reality and Voice Commands**, Proceedings of the 25th ACM international conference on Multimedia (MM '17). ACM, New York, NY, USA, to appear.