SAPIENZA
UNIVERSITÀ DI ROMA

First International Workshop on Biometrics-as-a-Service:
Cloud-based Technology, Systems and Applications (IW-BAAS2017)

A Smart Peephole on the Cloud

Maria De Marsico
Eugenio Nemmi
Bardh Prenkaj
Gabriele Saturni

Overview

- Introduction
- Azure
- Biometric as a Service
- Detection Module
- Face Recognition Module
- Voice Verification Module
- Emotion Detection Module

Goals

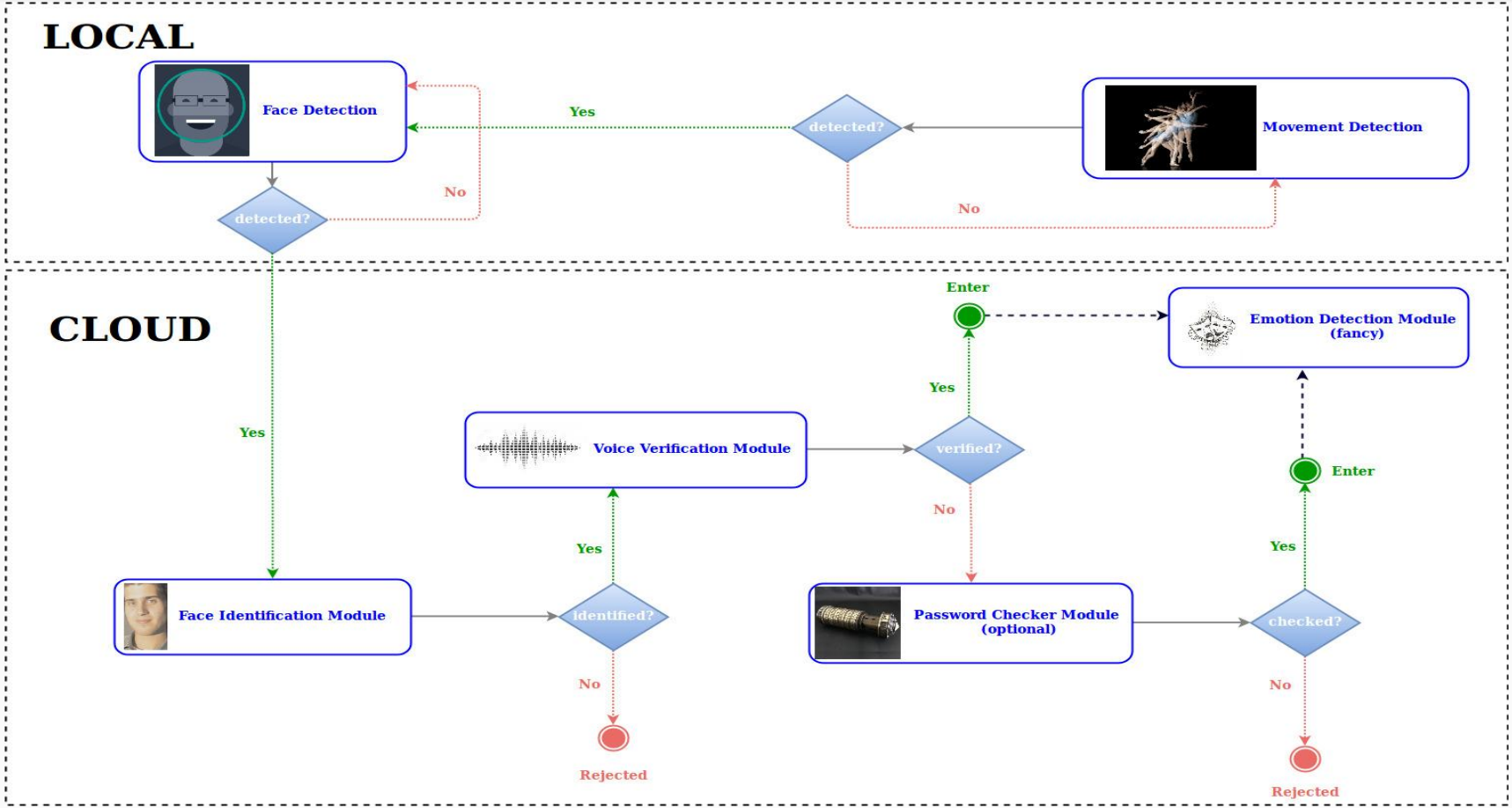
- To design a new open-set identification system for a smart home authentication: **Smart Peephole**

- To provide the following functionalities:
  - To correctly **recognize** a family member and, basically, “allow them to pass”;
    - *“Member discovery”*
  - To correctly **refuse** non-family members to enter the house;
    - *“Intruders detection”*
  - To **notify** the landlord that an intruder tried to enter his/her house

Abstract Idea for People Recognition

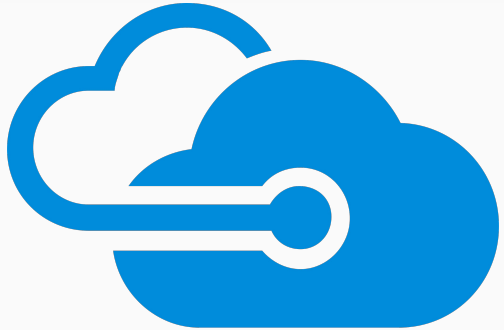
- Multibiometric System
- Enrollment:
  - User's face
  - User's voice
  - Password setup
- Operation:
  - Movement detection
  - Face detection
  - Face recognition (open set - 1:n)
  - Voice verification (text-dependent key phrase-based - 1:1)
  - Password validation (optional - requires additional hardware and/or software)
    - Backup method if voice verification fails

Abstract Idea for People Recognition



Microsoft Azure

- *Microsoft integrated cloud services*
- *Category:*
  - *IaaS - Infrastructure as a Service*
  - *PaaS - Platform as a Service*
  - *SaaS - Software as a Service*



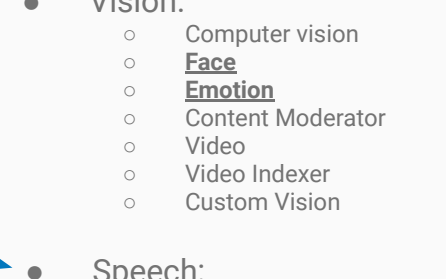
Services:

- *Networking*
- *Storage*
- *Web + Mobile*
- *Containers*
- *Databases*
- *Developer Tools*
- *Data + Analytics*
- ***AI + Cognitive Services***
- *Internet of Things*
- *Enterprise Integration*
- *Security + Identity*
- *Monitoring + Management*
- *Microsoft Azure Stack*

MCS - AI + Cognitive Services

- Microsoft Cognitive Services (MCS) let you build apps with powerful algorithms to see, hear, speak, understand and interpret our needs using natural methods of communication.

- Category:
  - Vision
  - Speech
  - Language
  - Knowledge
  - Search



- Vision:
  - Computer vision
  - **Face**
  - **Emotion**
  - Content Moderator
  - Video
  - Video Indexer
  - Custom Vision

- Speech:
  - Translator Speech
  - **Speaker Recognition**
  - Bing Speech
  - Custom Speech Service

Relevant services for Vision & Speech

- Vision:
  - Face API - among others:
    - face detection
    - face verification
    - face identification
  - Emotion API
    - recognize emotions in images and videos
  - Video - among others:
    - detect and track faces
    - detect motion
  - Custom Vision Service - among others:
    - train on uploaded labeled images
- Speech:
  - Speaker recognition API
    - speaker verification
    - speaker identification

Biometrics As A Service

- Microsoft Cognitive Services include APIs for biometric recognition
  - At present, face and voice

- It is possible to delegate complex biometric processing to remotely designed and implemented algorithms

- Microsoft Cognitive Services
  - Usage of **APIs** to **facilitate** app building using few code lines...
  - ... and support **cross-platform** development (iOS, Android, Windows)

Biometrics As A Service but ...

- **Problem**: to prevent burst-rate of requests towards MCS servers

- **Solution**: to devise a reasonable compromise between local and remote processing

Local vs. remote processing

- Detection: possibly continuous process - possibly computation-intensive
  - Natural candidate for local processing
  - Movement detection + Face detection

- Recognition: sophisticated algorithms - sporadic execution
  - Delegated to remote services
  - Face recognition + Speaker recognition + (Emotion recognition)

Movement Detection

- Movement Detection
  - Problems:
    - **To detect movement** in front of the doorstep
    - To prevent to leave the system continuously active
  - Solutions:
    - **DOPTFlow** (Dense Optical Flow Algorithm) - OpenCV implementation

DOPTFlow Algorithm

- Dense Optical Flow:
  - It computes the optical flow for **all** the points in the **frame**:
    - based on Gunner Farneback's algorithm
    - builds **motion vectors** between two consecutive frames

Face Detection

- Face detection implemented locally
  - faster if implemented locally that through API call
  - based on **haar-cascade classifiers** - OpenCV implementation

Adopted MCS APIs

- Our scenario APIs:
  - Face API
    - face identification
  - Speaker Recognition API
    - speaker verification
  - Emotion API
    - Not necessary for authentication (fancy addition)

Face Identification - Enrollment

- The users that have to be recognized correctly (genuine users) must be registered
  - The system creates a new person associated to a group, which is defined by the administrator
- Different photos with possible PIE variations per user
  - **multiple-template-based** enrollment
- The image quality plays a crucial role
- Enrollment needs to be supervised

Face Identification - Operation

- After face detection, the photo is sent to the API
- The API's response is associated to a confidence value
  - if the confidence is greater than or equal to a threshold, the person is accepted; otherwise he/she is rejected

```
[{
    "faceid" : face_id,
    "candidates" : [{"personId": person_id,
                     "confidence": confidence_value}]
}]
```

Voice Verification - Enrollment

- The API provides the user with the **list** of all possible acceptable *recognition phrases*
- Recording is **completed successfully** if the audio is at least 1s long and shorter than 15s
- The enrollment requires the user to record his/her voice **three** times with **good quality** (otherwise the API calls for repetition)
    - **multiple-template-based enrollment**

Voice Verification - Operation

- The user is required to **speak up** the phrase chosen during enrollment
- The user is either **accepted** or **rejected**
  - Levels confidence for speaker recognition : *Low, Normal, High*
- Spoken phrase attached to the response
  - Useful for debugging and logging purposes

```
[
    {
        "result" : "Accept",
        "confidence" : "Normal",
        "phrase": "My name is unknown to you"
    }
]
```

Voice Verification - Limitations

- Noise reduction used by the API might be insufficient
  - Misleading rejection of genuine user
  - Repeating voice enrollment for many times

Emotion Detection

- Feature to catch the user’s mood
- The input is the first picture taken during face detection
- The module takes some **action** accordingly with the emotion recognized, for example playing a song
- The actions that the module takes have as an objective to **improve** or **favor** the mood

Emotion Detection

- The emotions that the module is able to **detect** are:
  - happiness, surprise, fear, disgust, neutral, sadness, anger and contempt
- Limitations:
  - Users could express their emotions in an **ambiguous** way
  - Facial expressions have to be **emphasized** to get recognized correctly

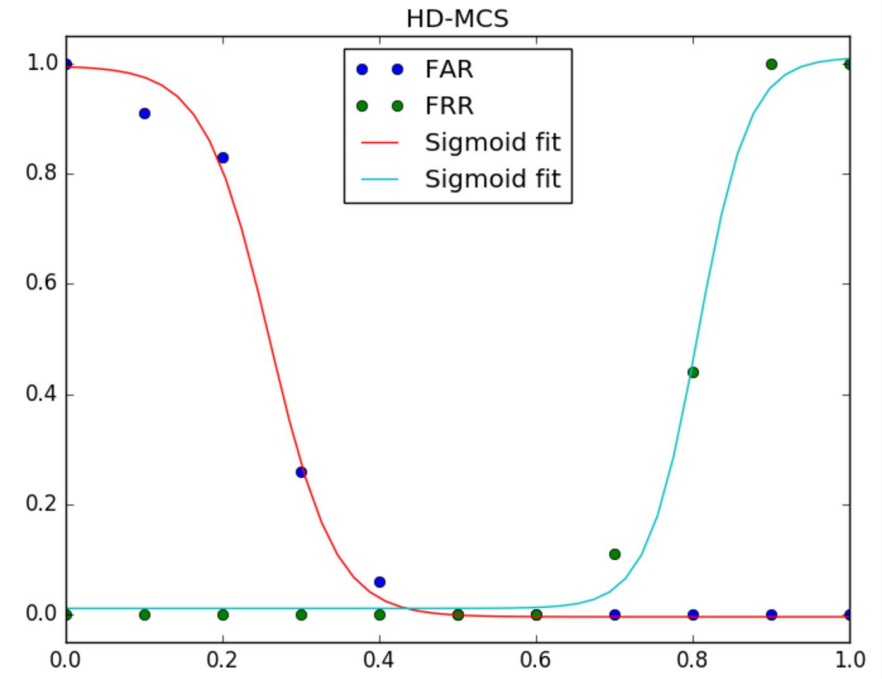

```
[
  {
    "faceRectangle": {
      "left": 68,
      "top": 97,
      "width": 64,
      "height": 97
    },
    "scores": {
      "anger": 0.00300731952,
      "contempt": 5.14648448E-08,
      "disgust": 9.180124E-06,
      "fear": 0.0001912825,
      "happiness": 0.9875571,
      "neutral": 0.0009861537,
      "sadness": 1.889955E-05,
      "surprise": 0.008229999
    }
  }
]
```

Experiments - Face Recognition

- Two experiments with two different datasets:
  - HD-Master of Computer Science (HD-MCS), an **in-house manually** constructed dataset
  - Labeled Faces in the Wild (LFW) - subset of ~1500 images
- Response Time (RT) analysis

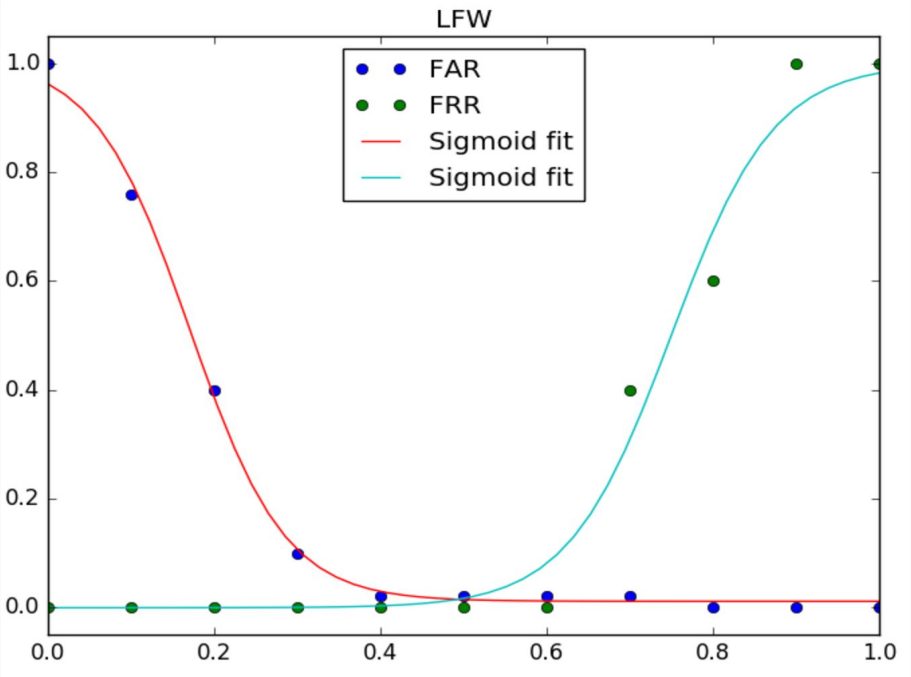
Experiments - HD-MCS Dataset

- **15** students from the Master's CS Degree at "La Sapienza" University
  - **3 different position** for each individual (Straight, Half-Left, Half-Right)
- **EER = 0** which implies optimal **discriminative** power
- The threshold adopted is, thus, equal to **0.5**
- Approximation with **sigmoid** curves



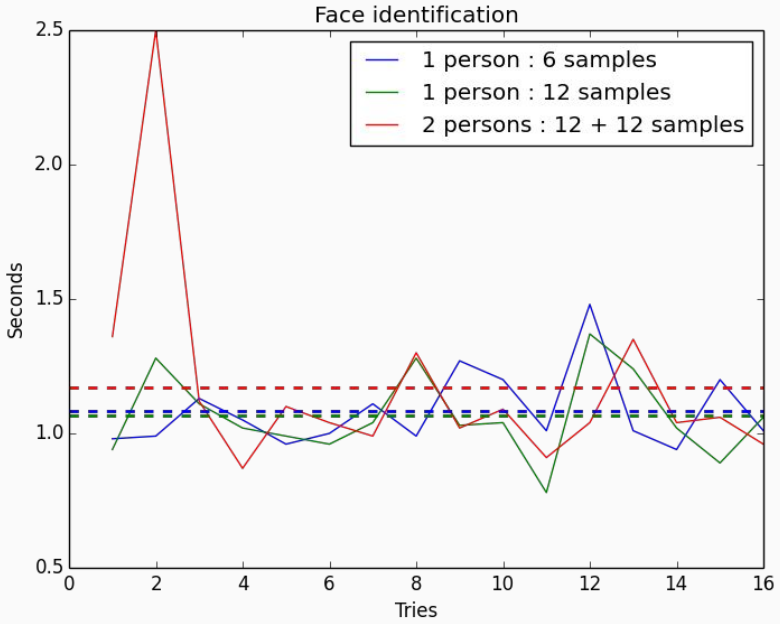
Experiments - LFW (subset)

- Subset of the LFW dataset
  - ~1500 face images
  - 200 persons - 50 genuine & 150 impostors
- **EER = 0.02**
  - The dataset consists of the Olsen **twins** that generate false alarms
- Threshold setted to **0.7** in order to **minimize** these false alarms
- Approximation with **sigmoid** curves



Experiments - Face Identification RT

- 3 trained set
  - 1 person : 6 samples
  - 1 person : 12 samples
  - 2 person : 12 + 12 samples
- Same mean RT almost
- Most important aspect is the network bandwidth
- Red peak is a network problem



Experiments - Emotion Detection

- Tests conducted using **KDEF** [1] emotion dataset:
  - Three kinds of positions (half-right, half-left, straight)
  - 7 emotion taken into account (contempt was excluded because it doesn't belong to the basic emotions, however some images are recognized as contempt by the MCS API)
  - 140 images per emotion
  - Results are represented in a confusion matrix
    - rows are the genuine emotion
    - columns the API result

[1] Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

Experiments - Emotion Detection (straight)

	Anger	Disgust	Contempt	Fear	Sadness	Happiness	Surprise	Neutral	**Error**
Anger	55	2.1429	2.8571	0.7143	2.8571	0	0.7143	35.7143	0
Disgust	7.8571	71.4286	0	0	14.2857	2.1429	0	4.2857	0
Fear	1.4286	0.7143	2.1429	17.142	20.7143	4.2857	42.8571	10.7143	0
Sadness	0	0	0	0	86.4286	0.7143	0.7143	12.1429	0
Happiness	0	0	0	0	0	100	0	0	0
Surprise	0	0	0	0	0	0.7143	95.7143	3.5714	0
Neutral	0	0	0	0	0	0	0	100	0

Experiments - Emotion Detection Notes

- The **best** results were reported by the **straight** position
- There's some **confusion** with anger and fear
- Happiness is the emotion that **outperforms** the others
- As expected, when **half-left** and **half-right** position are taken into account Neutral emotion gains, in most cases, a high (either true or wrong) recognition rate

Conclusions

- The aim of the system is to provide a cheap and easy-to implement solution for biometric control of accesses exploited “at home”
- In order to test the feasibility of adopting MCS in a challenging context, experiments on face recognition have been carried out on both a in-house collected dataset, and on LFW (at present one of the most adopted in literature)
- The results show that the system has a good capability to distinguish the faces of people registered in the system from intruders

Future work

- Speech tests are planned for the future because of the nature of the API
  - Speaker and speech recognition at the same time
- A field test will evaluate the performance in a real context
- Implementation of Telegram chatBot could be added for:
  - remote operation by the housekeeper, e.g., remote door opening and check of refused people
  - capability to extend the gallery with incorrectly rejected images
- Addition of (local) anti-spoofing algorithms could improve system security

THANK YOU!

QUESTIONS?

Citations

[1] M. De Marsico - “*Face Recognition 2D*”, Biometric System Lectures, University “La Sapienza” (a.y. 2016/2017)

[2] M. Cahoui, A. Elkefi, W. Bellil, C. Ben Amor - **“***A Survey of 2D Face Recognition Techniques”*, REGIM: Research Groups on Intelligent Machines, University of Sfax, National School of Engineers (ENIS), Sfax, Tunisia (2016)

[3] L. Muda, M. Begam, I. Elamvazuthi - “*Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”*, Journal of Computing, Volume 2, Issue 3 (2010)

Face Detection Module - Another approach

- Algorithm by Hsu, Mottaleb and Jain (2002)
  - Two principal steps:
    - Face candidates detection
    - Face candidates verification through facial features detection

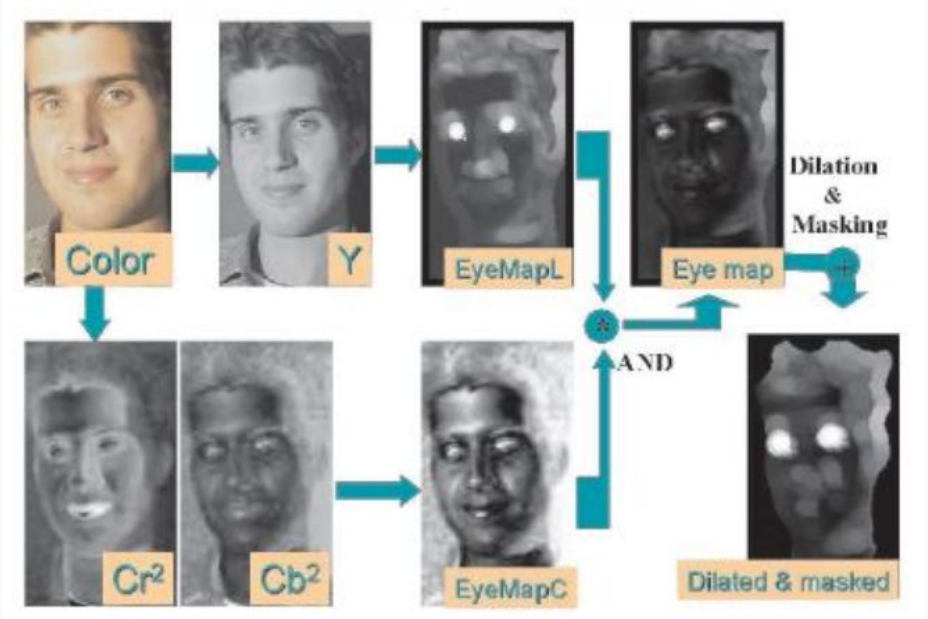
Face Detection Module - Hsu et al.
Face candidates detection

- Illumination compensation
  - Normalize the lighting because the skin tone depends on the scene overall illumination
- Colour space transformation
  - Transform RGB to YCbCr because two near colours in the RGB space doesn’t mean they have to be similar
- Localization based on skin model

Face Detection Module - Hsu et al.

Face candidates verification through facial feature detection

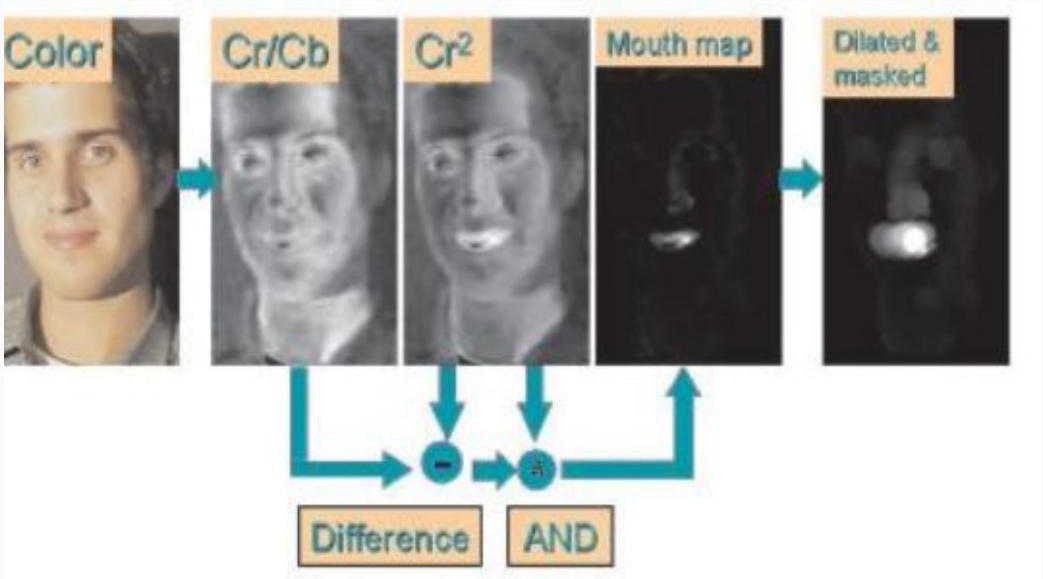
- **Eye localization**
  - Relies on the observation that the region around eyes is characterized by high values of Cb and low values of Cr
- Mouth localization
- Face contour



Face Detection Module - Hsu et al.

Face candidates verification through facial feature detection

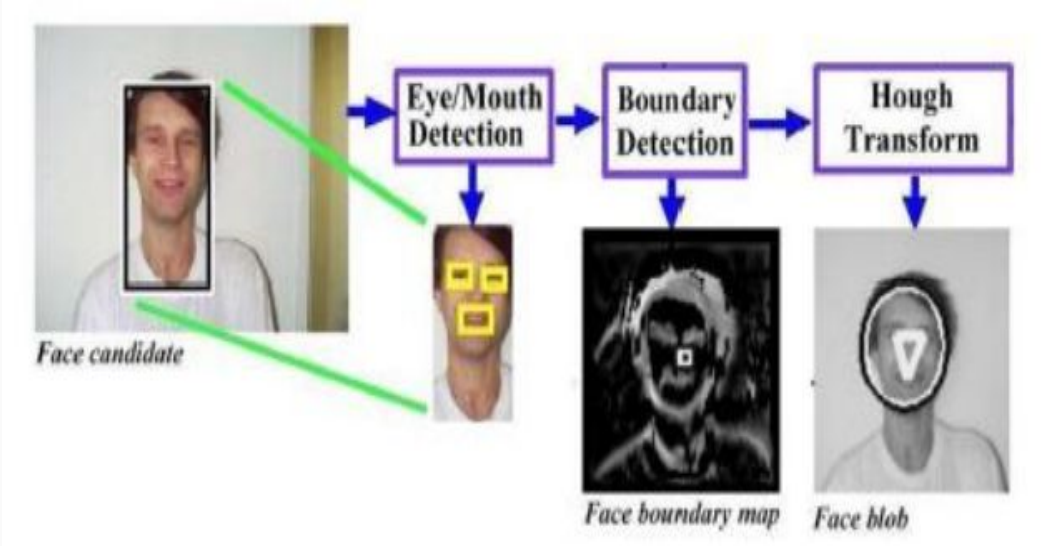
- Eye localization
- **Mouth localization**
  - The Cr component is higher than Cb
  - The response to Cr /Cb is low, while the response to Cr^2 is high
- Face contour



Face Detection Module - Hsu et al.

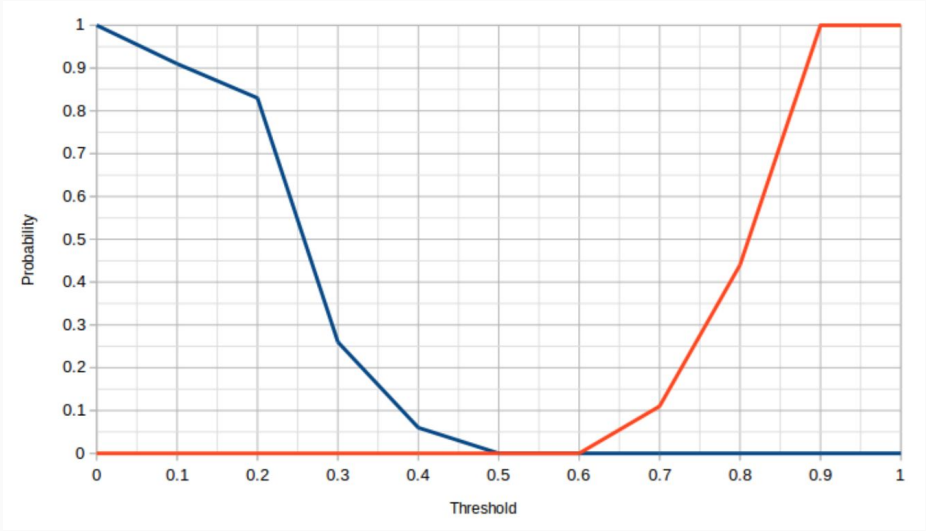
Face candidates verification through facial feature detection

- Eye localization
- Mouth localization
- **Face contour**
  - Analyze all triangles composed by two eyes and a mouth
  - For each triangles assign a specific score and return the highest scored triangle as a face contour



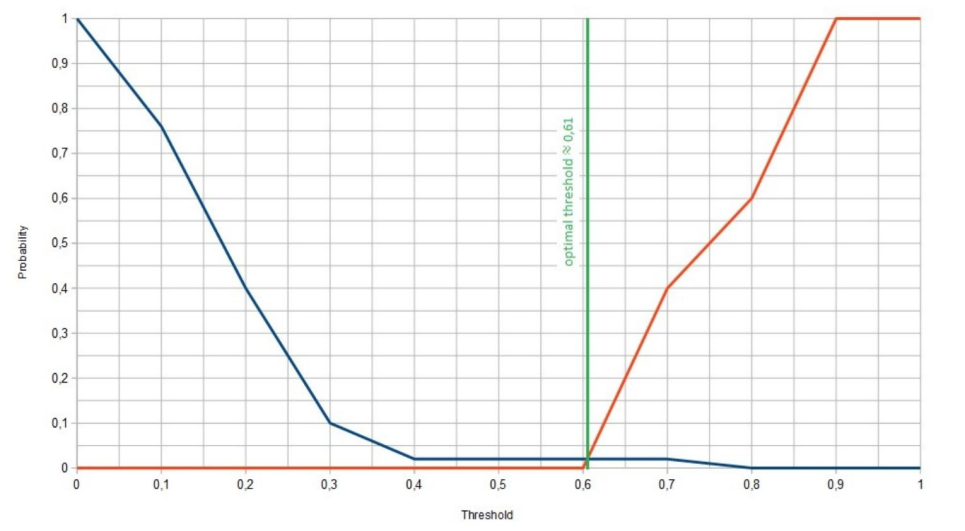
Experiments - HD-MCS Dataset

- **15** students from the Master's CS Degree at "La Sapienza" University
  - **3 different position** for each individual (Straight, Half-Left, Half-Right)
- **EER = 0** which implies optimal **discriminative** power
- The threshold adopted is, thus, equal to **0.5**



Experiments - Labeled Faces in the Wild

- **EER = 0.02**
  - The dataset consists of the Olsen **twins** that generate false alarms
- Threshold setted to **0.7** in order to **minimize** these false alarms



Smart Peephole as a Multibiometric System

- Sort of a Multibiometric Decision Level Fusion System
- The user is accepted if:
  - Face recognition has a positive outcome

    AND

  - Speak recognition has a positive outcome

    OR

  - Recovery password inserted correctly
- The face and speaker modules work in serial and not in parallel

Movement Detection - DOPFlow Algorithm

The image represents the motion vectors between frames **f** and **f-1**.

In this case, there is a head movement **from left to right**.

Haar-cascade Detection

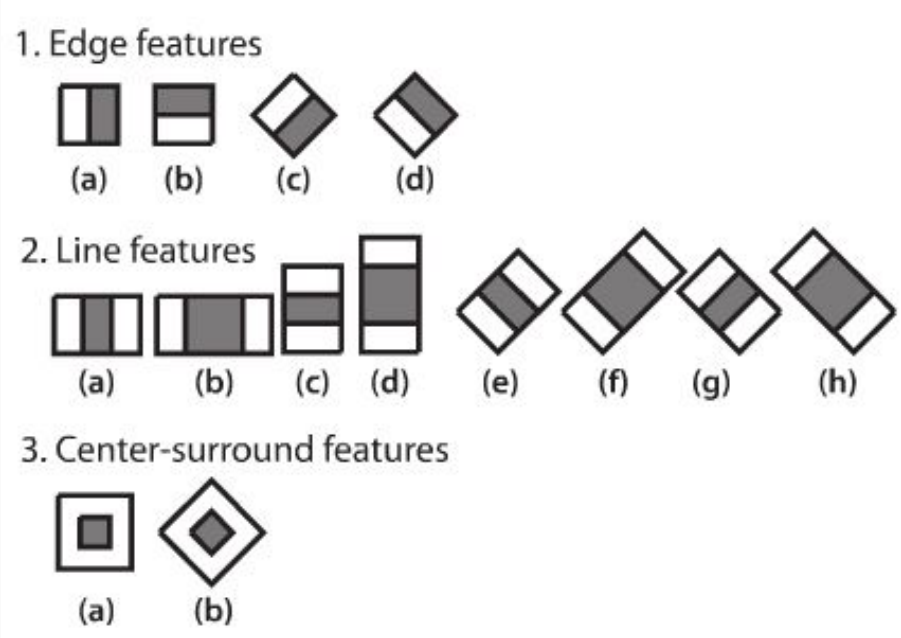
- Viola - Jones intuition in 2001
- Uses classifiers trained with positive and negative examples
  - Extracts features
  - Selects discriminative ones
  - 2 classes: face, non face
- Boosting

Haar-cascade Detection

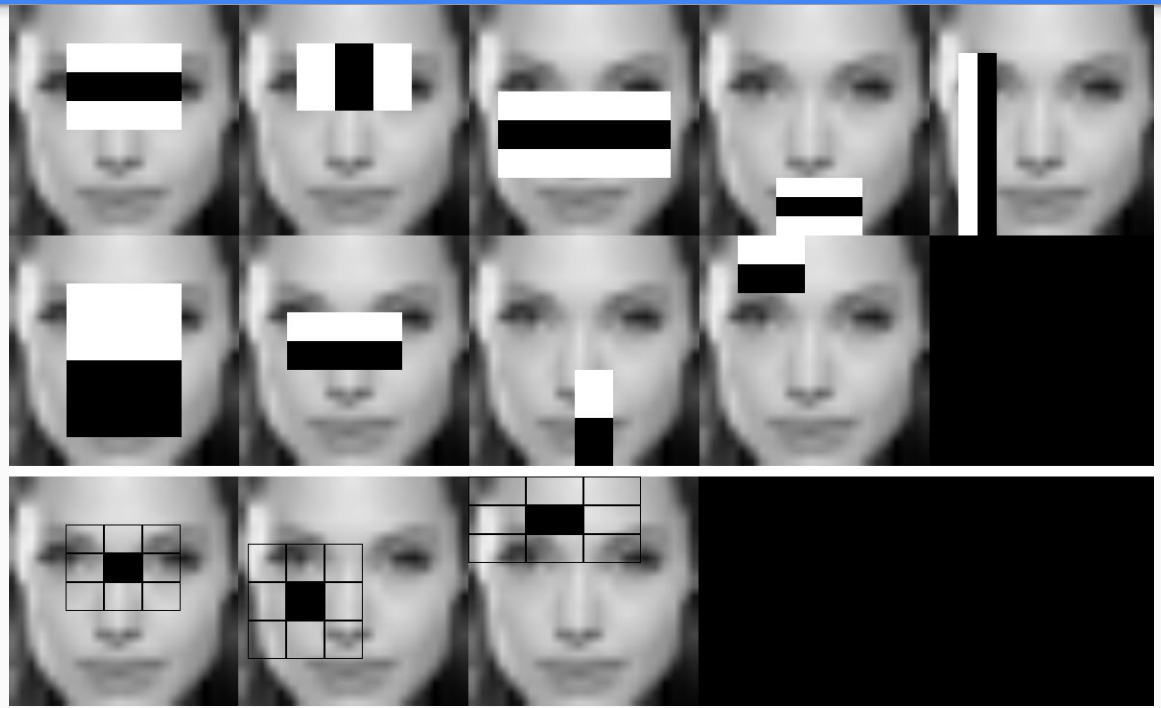
- Haar-like feature considers **adjacent** rectangular **regions** at a specific location in a detection window
- Sums up the **pixel intensities** in each region
- Calculates the difference between these sums

Haar-cascade Detection

- Haar-like feature **examples**:



Haar-cascade Detection - Applied

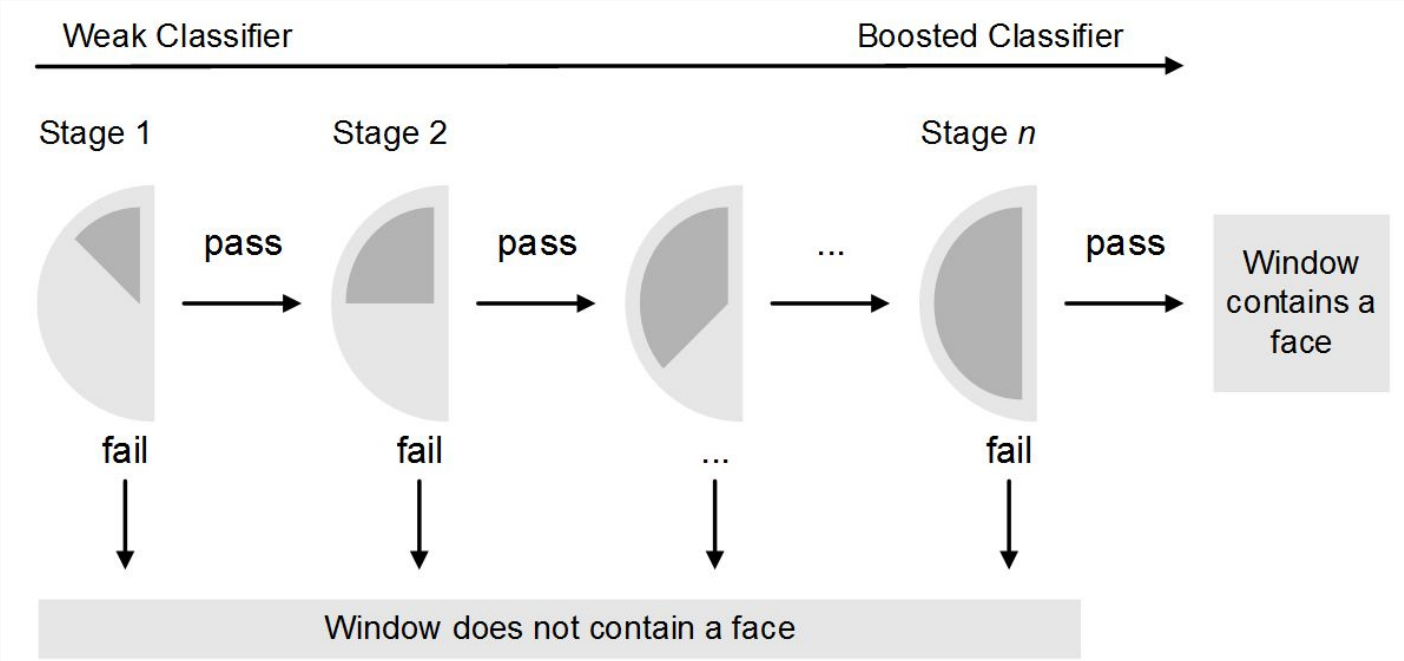


Haar-cascade Detection

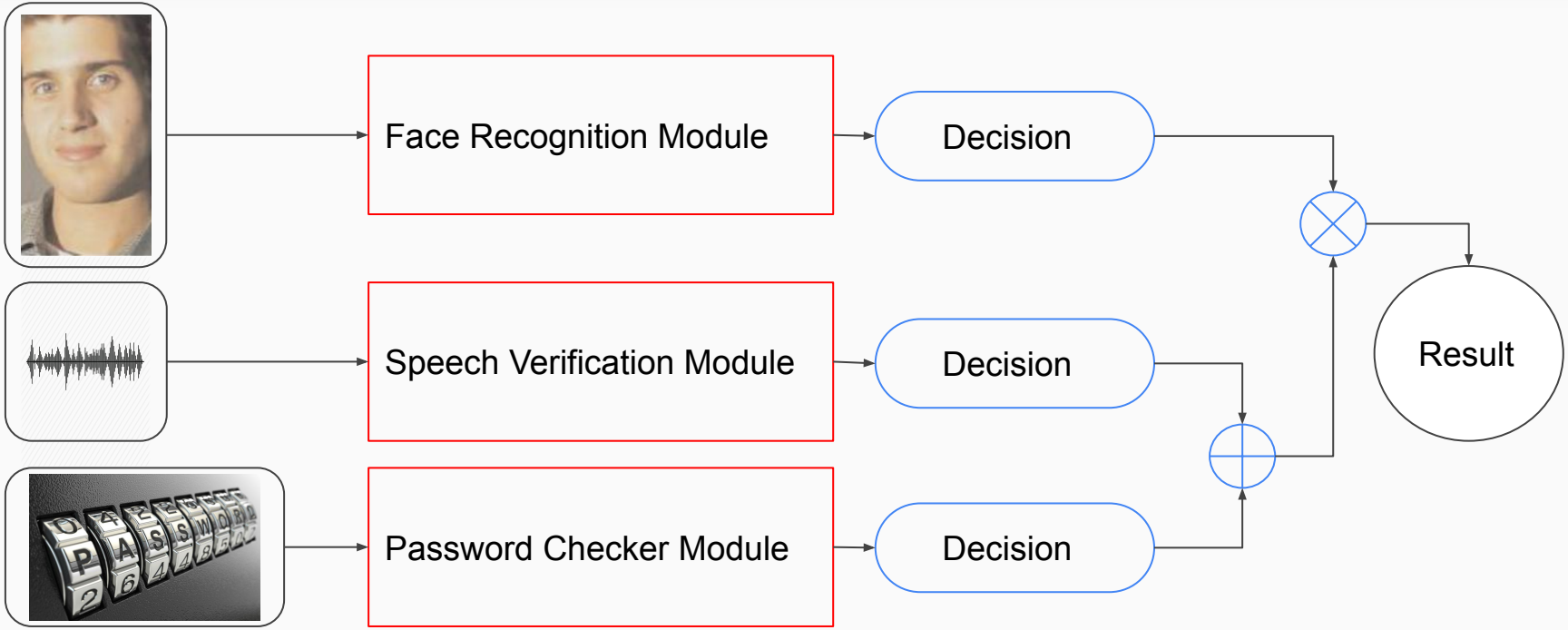
- Uses series of weak classifiers to build a strong one (Boosting)

- Why cascade?
  - The result from the previous group of weak classifiers is sent to the followers group of more complex classifiers
  - if a sub-windows is rejected, it’s not sent to the followers ones.

Haar-cascade Detection



Smart Peephole as a Multibiometric System



Face Recognition - example

1. A photo is given as input using a REST call
2. API response
3. Check the confidence and output Accept/Reject

```
[{
    "faceid" : face_id,
    "candidates" : [{"personId": person_id,
                     "confidence": confidence_value}]
}]
```

Speech Verification Module - Example

- Example of a successful call


```
[
    {
        "result" : "Accept",
        "confidence" : "Normal",
        "phrase": "My name is unknown to you"
    }
]
```

Emotion detection

- The actions that the module takes have as an objective to:
  - **Improve** the mood if the emotion recognized is negative (i.e.: fear, anger, sadness, contempt)
  - **Favor** the mood (i.e.: happiness, surprise, neutral)

Emotion detection

- Example of successful call


```
[
  {
    "faceRectangle": {
      "left": 68,
      "top": 97,
      "width": 64,
      "height": 97
    },
    "scores": {
      "anger": 0.00300731952,
      "contempt": 5.14648448E-08,
      "disgust": 9.180124E-06,
      "fear": 0.0001912825,
      "happiness": 0.9875571,
      "neutral": 0.0009861537,
      "sadness": 1.889955E-05,
      "surprise": 0.008229999
    }
  }
]
```


- There are **different** scores for each emotion type
- The emotion recognized is the one that has the **maximum** value
  - **happiness**

Emotion detection

- Limitations:
  - Users could express their emotions in an **ambiguous** way
  - Facial expressions have to be **emphasized** to get recognized correctly

Experiments - Emotion detection
Half-Left Results

	Anger	Disgust	Fear	Sadness	Happiness	Surprise	Neutral	**Error**
Anger	10.7143	4.2857	0	3.5714	0	0	81.4286	**0**
Disgust	7.1426	57.8571	0	16.4286	2.8571	0	14.2857	**1.4286**
Fear	1.4287	2.1429	3.5714	15	7.1429	36.4286	34.2857	**0**
Sadness	0	0	0	63.5714	1.4286	0	34.2857	**0**
Happiness	0	0	0	0	100	0	0	**0**
Surprise	0	0	0	0	0.7143	85	14.2857	**0**
Neutral	0	0	0	0	0	0	99.2857	**0.7143**

Experiments - Emotion detection
Half-Right Results

	Anger	Disgust	Fear	Sadness	Happiness	Surprise	Neutral	**Error**
Anger	17.1429	2.1429	0	0.7143	0.7143	0.7143	78.5714	**0**
Disgust	13.5714	53.5714	0	13.5714	2.1429	0	17.1429	**0**
Fear	1.4286	2.1429	2.8571	13.5714	4.2857	36.4286	39.2857	**0**
Sadness	0	0	0	72	1	0	67	**0**
Happiness	0	0	0	0	99.2857	0	0.7143	**0**
Surprise	0	0	0	0	0.7143	78.5714	20	**0**
Neutral	0	0	0	0	0	0	99.2857	**0.71423**

Experiments - Further Considerations (abbiamo preso 200 foto di LFW)

- The measurements taken into account aren't sufficient for a complete evaluation
- We must consider:
  - Image size and quality
  - Number of images inside the dataset
  - Number of PIE variation
- It's better to use standard datasets, but authorization for usage is required as in the following case:
  - Simulations for the FRVT 2002 experiments resulted impossible

Experiments - Voice verification RT