Online Biometric Identification With Face Analysis in Web Applications

Gerasimos Arvanitis, Konstantinos Moustakas, and Nikos Fakotakis

Artificial Intelligence Group
Wire Communications Laboratory
Dept. of Electrical & Computer Engineering
University of Patras

Why Internet Security is so Important

- Sensitive personal data (photos, videos e.g.)
- Business information
- Financial data, money transactions and payments
- Waste of money for help desk

Hacking Tools and Categories

- hacking/ pentesting categories: Application Specific Scanners, Debuggers, Encryption Tools, Firewalls, Forensics, Fuzzers, Intrusion Detection Systems, Multi Purpose Tools, Packet Crafting Tools, Packet Sniffers, Password Crackers, Port Scanners, Linux Hacking Distros, Rootkit Detectors, Traffic Monitoring Tools, Vulnerability Exploitation Tools, Vulnerability Scanners, Web Browser Related Tools, Web Proxies, Web Vulnerability Scanners and Wireless Hacking Tools.
- Some of the most known hacking tools: Nmap , Metasploit, John The Ripper, THC Hydra, OWASP Zen, Wireshark, Aircrack-ng, Maltego, Burp Suite.

How Easy a Password can be Stolen

- Brute force is the simplest method that can be applied from anyone, not necessarily a hacker
- There are lists with the most common passwords and also anyone can create dictionaries for passwords recovery.

Password length	Lower Case	Upper Case	Digits	Full ASCII
5	1 minute	7 minute	16 minute	136 minutes
6	6 minute	6 hours	16 hours	10 days
7	134 minute	12 days	41 days	29 months
8	59 hours	21 months	8 years	232 years

Some Statistics Facts

- **65%** of workers use the same password for different applications or services
- **70%** of people do not use a unique password for each Web site
- A third of Internet users have shared their log-in information with their partner
- **64%** of end users report that they have written down their password at least once
- **82%** of people have forgotten a password used on a Web site
- **44%** of people use passwords without letters and numbers
- **21%** of people use password with less than 6 characters

People who reuse passwords across sites:

%	Ages
76%	18 to 24 years old
58%	25 to 34 years old
61%	35 to 49 years old
56%	50 to 64 years old
62%	65+ years old

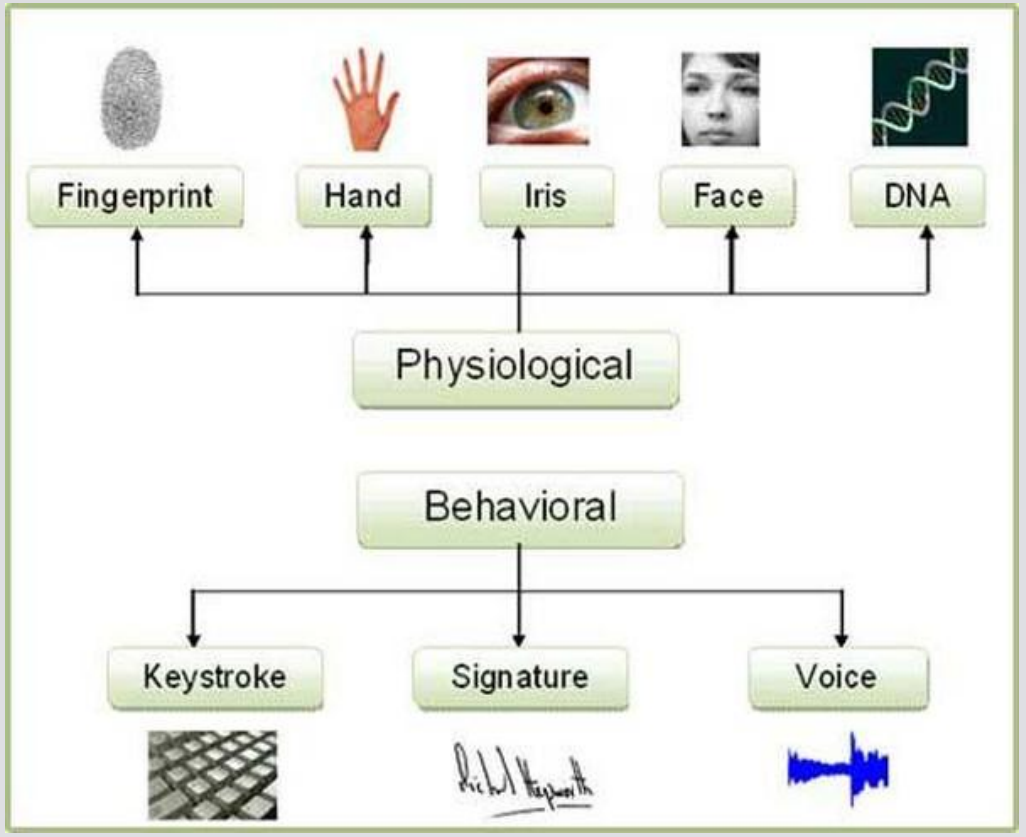
Common sources of keywords found in cracked passwords (n=441,960):

%	Keywords
8.40%	Top 2,000 Baby Names
4.98%	U.S. City Names
1.06%	Top 100 Dog Names
0.66%	Top 1,000 Word Cities (By Population)
0.12%	U.S. State Names

User-based Security

- Million of dollars are spent every year for internet security
- Companies try to make their web sites more secure with strong encryptions and other technological innovations
- It is important to give emphasis on user-based security
- Even if a company uses state of the art technologies for their users' data protection, it is useless when the user has "123456" as password.

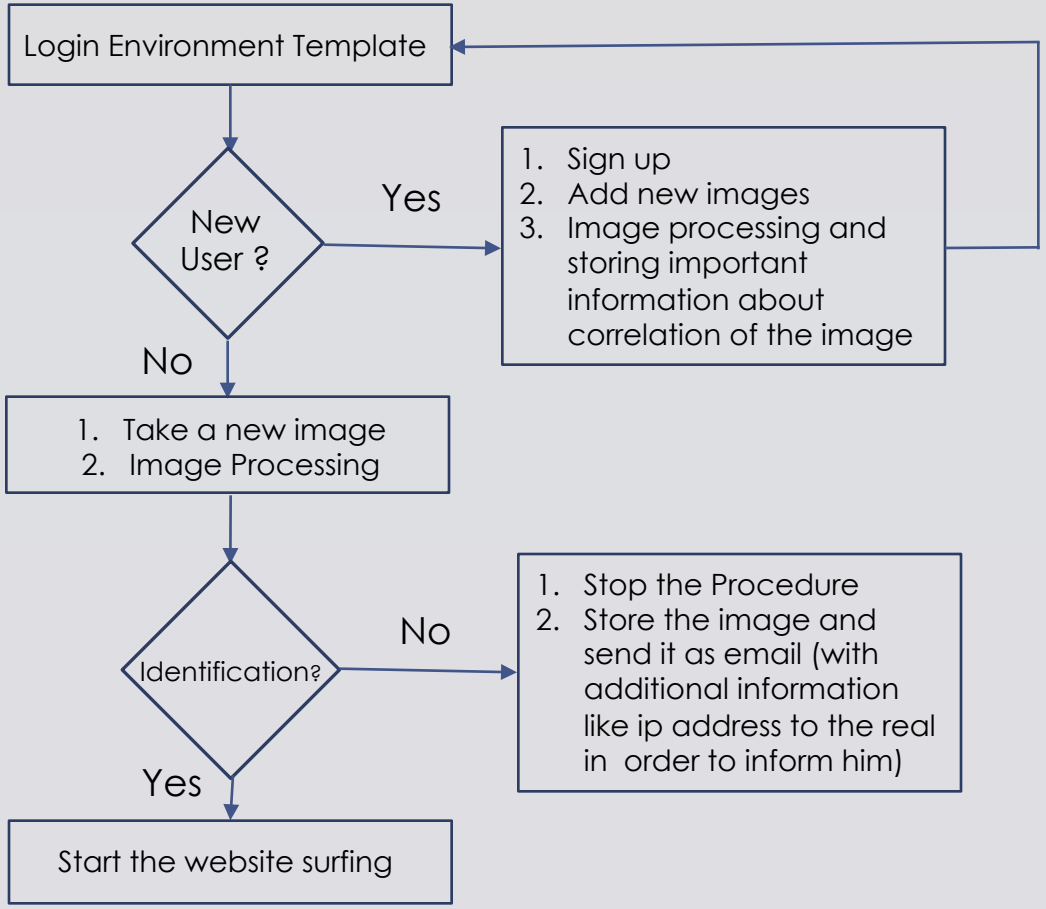
Biometric Characteristics



Biometric Identification

- The security field uses three different types of authentication:
  - Something you know — a password, PIN, or piece of personal information (such as your pet name).
  - Something you have — a card key, smart card, or token (like a SecurID card).
  - Something you are — a biometric.
- Why is safer? Because biometric characteristics are unique for any person.
- If someone has your password then has access to your application but is not easy to have your biometric characteristics.

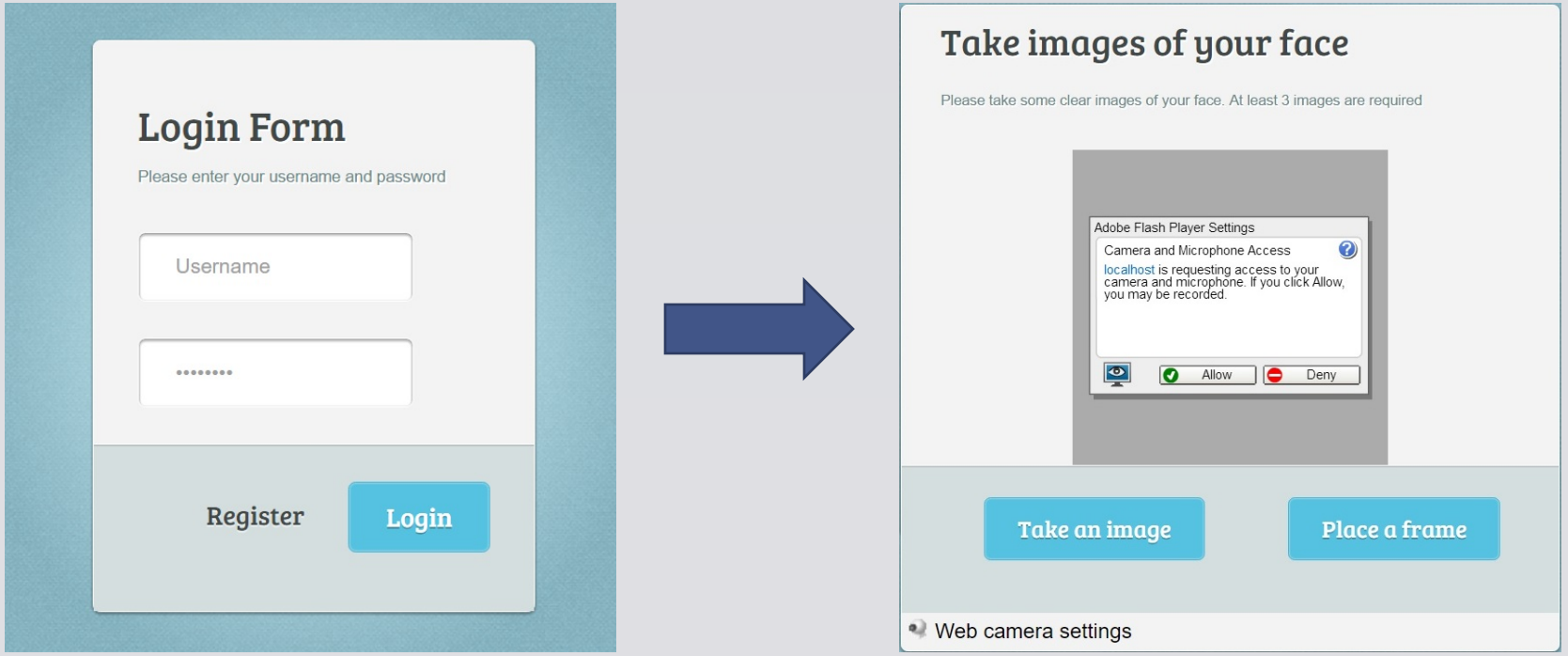
Proposed Login Procedure



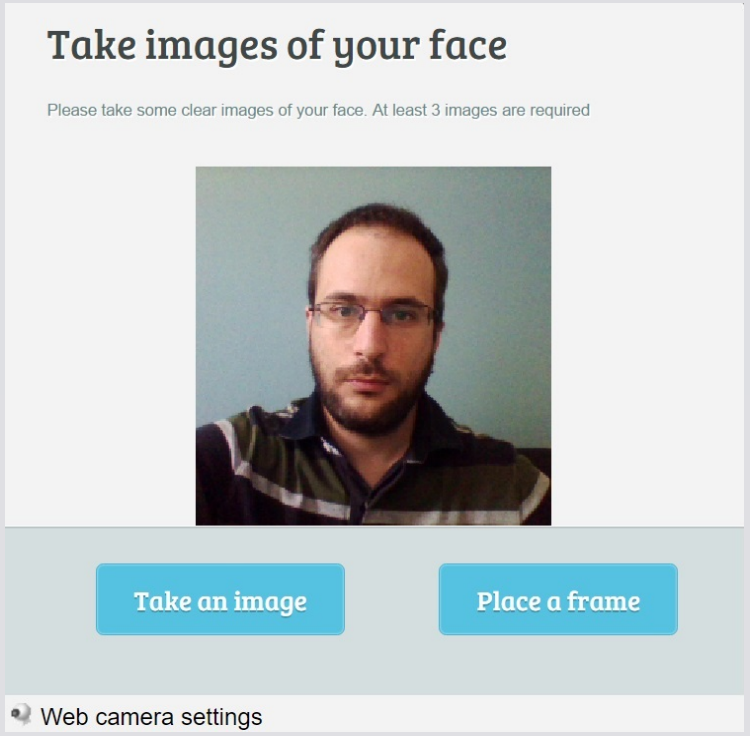
Main Advantages

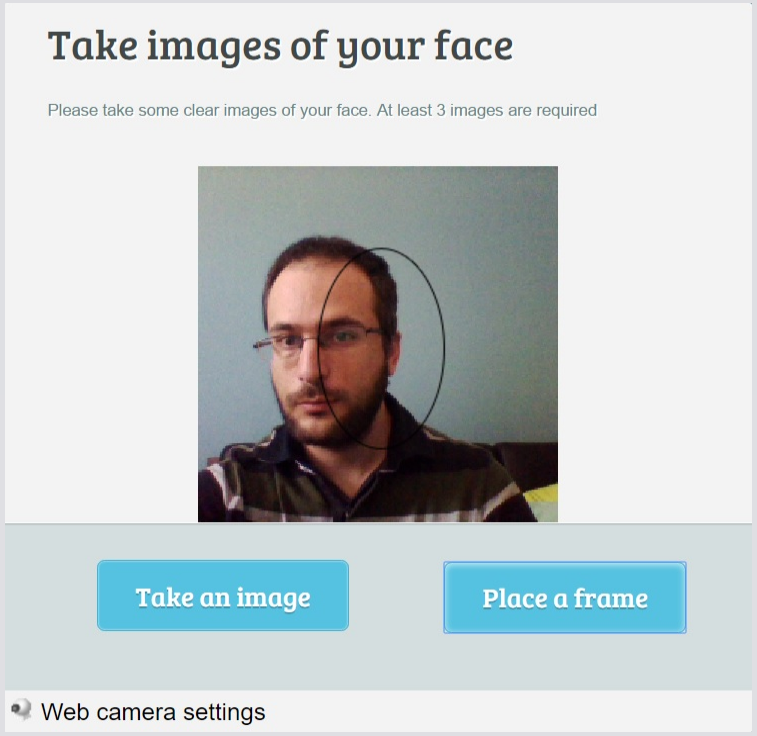
- A password is necessary too.
- It can be used at every web site that has a login frame as a second layer of security
- Because of the fast real time execution of the algorithm it can be used for online applications
- It makes the security stronger because the application adds a new layer of safety
- Each time the user's face image is compared with his own face images.

Login Frame


Login Form
Please enter your username and password
Username
Register
Login
Take images of your face
Please take some clear images of your face. At least 3 images are required
Adobe Flash Player Settings
Camera and Microphone Access
localhost is requesting access to your camera and microphone. If you click Allow, you may be recorded.
Allow
Deny
Take an image
Place a frame
Web camera settings

Take Images


Take images of your face
Please take some clear images of your face. At least 3 images are required
Take an image
Place a frame
Web camera settings


Take images of your face
Please take some clear images of your face. At least 3 images are required
Take an image
Place a frame
Web camera settings

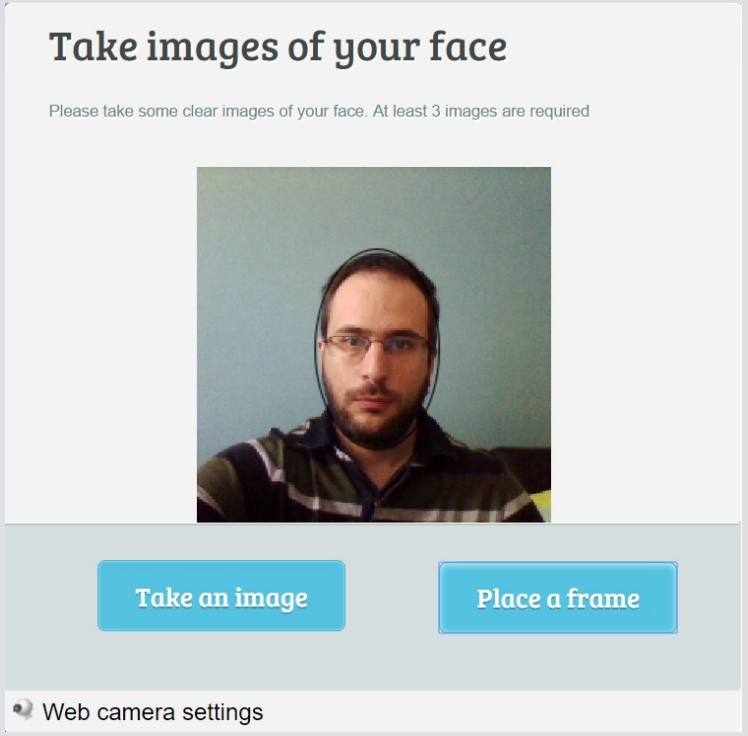
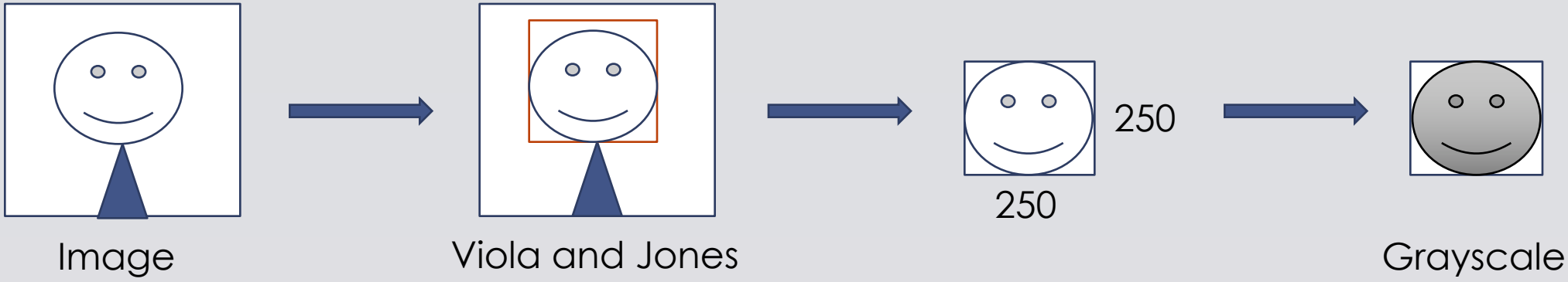

Take images of your face
Please take some clear images of your face. At least 3 images are required
Take an image
Place a frame
Web camera settings

Image Processing

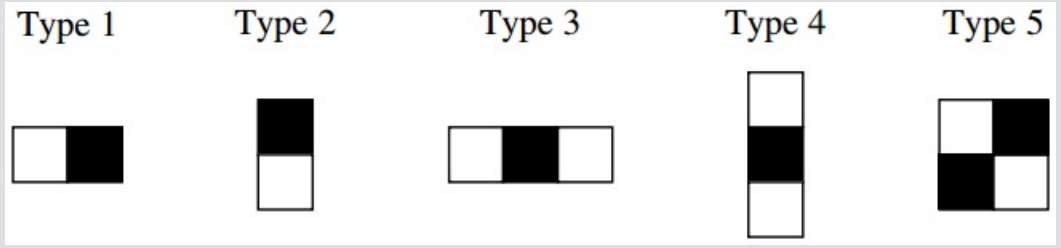
- Image taking
- Face recognition
- Background removement keeping only the face
- Grayscale
- Real time analysis and eigenfaces creation
- Comparison of the eigenvectors of the new face with the saved ones and a decision is taken.

Face Recognition and Background Removal



Haar Feature Selection

- Each feature results in a single value which is calculated by subtracting the sum of the white rectangle(s) from the sum of the black rectangle(s).

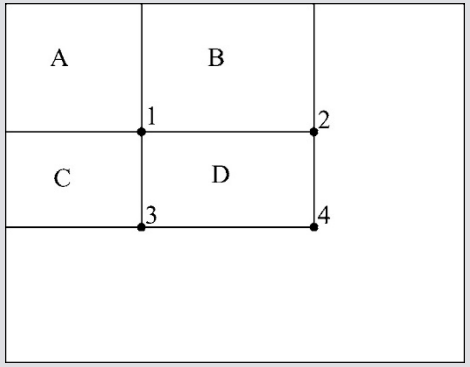


Creating an Integral Image

- The integral image at location x, y contains the sum of the pixels above and to the left of x, y, inclusive:

$$ii(x,y) = \sum_{x' \leq x, y' \leq y} i(x', y')$$

where ii(x, y) is the integral image and i(x, y) is the original image.



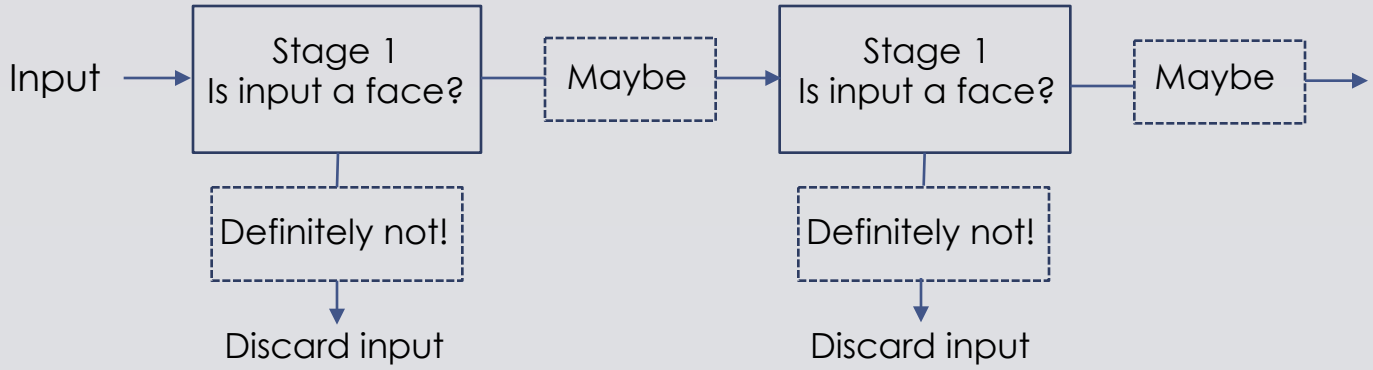
Adaboost Training

- Each feature is considered to be a potential weak classifier. A weak classifier is mathematically described as:

$$h(x, f, p, \theta) = \begin{cases} 1 & if \quad pf(x) < p\theta \\ 0 & otherwise \end{cases}$$

where f feature, threshold and p polarity that indicate the direction of the inequality.

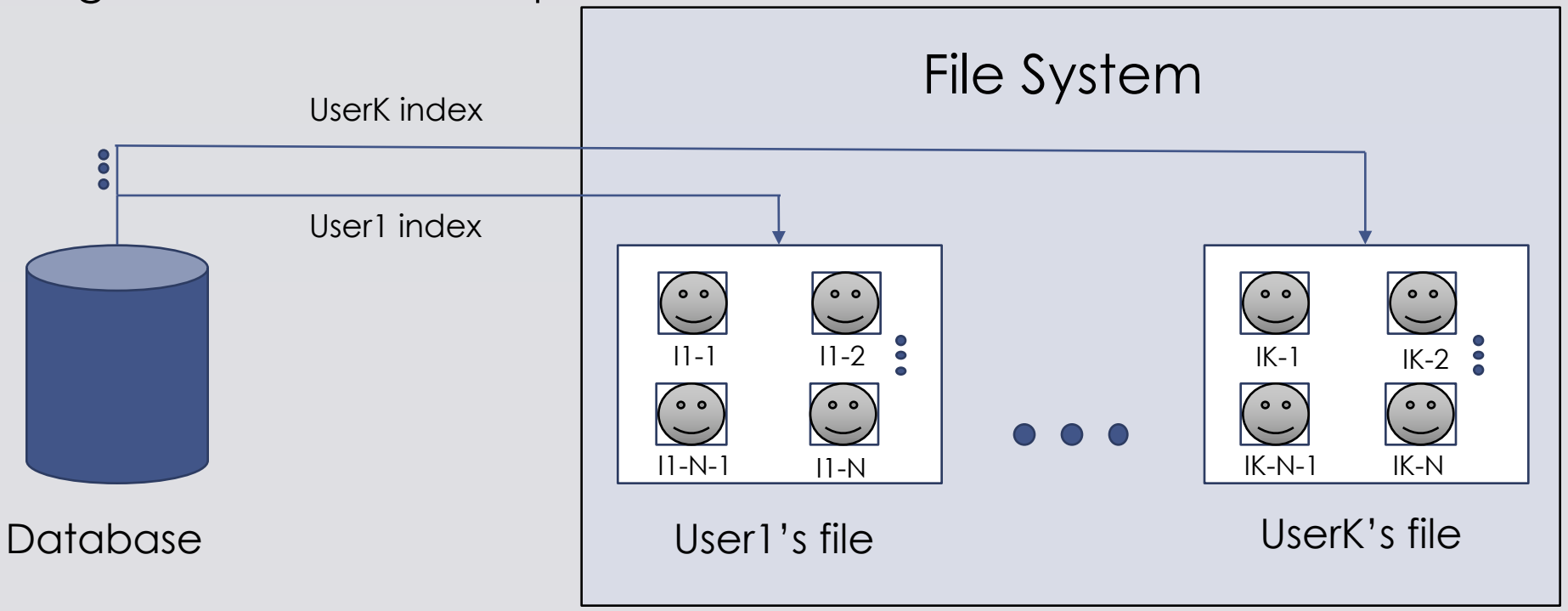
Cascading Classifiers



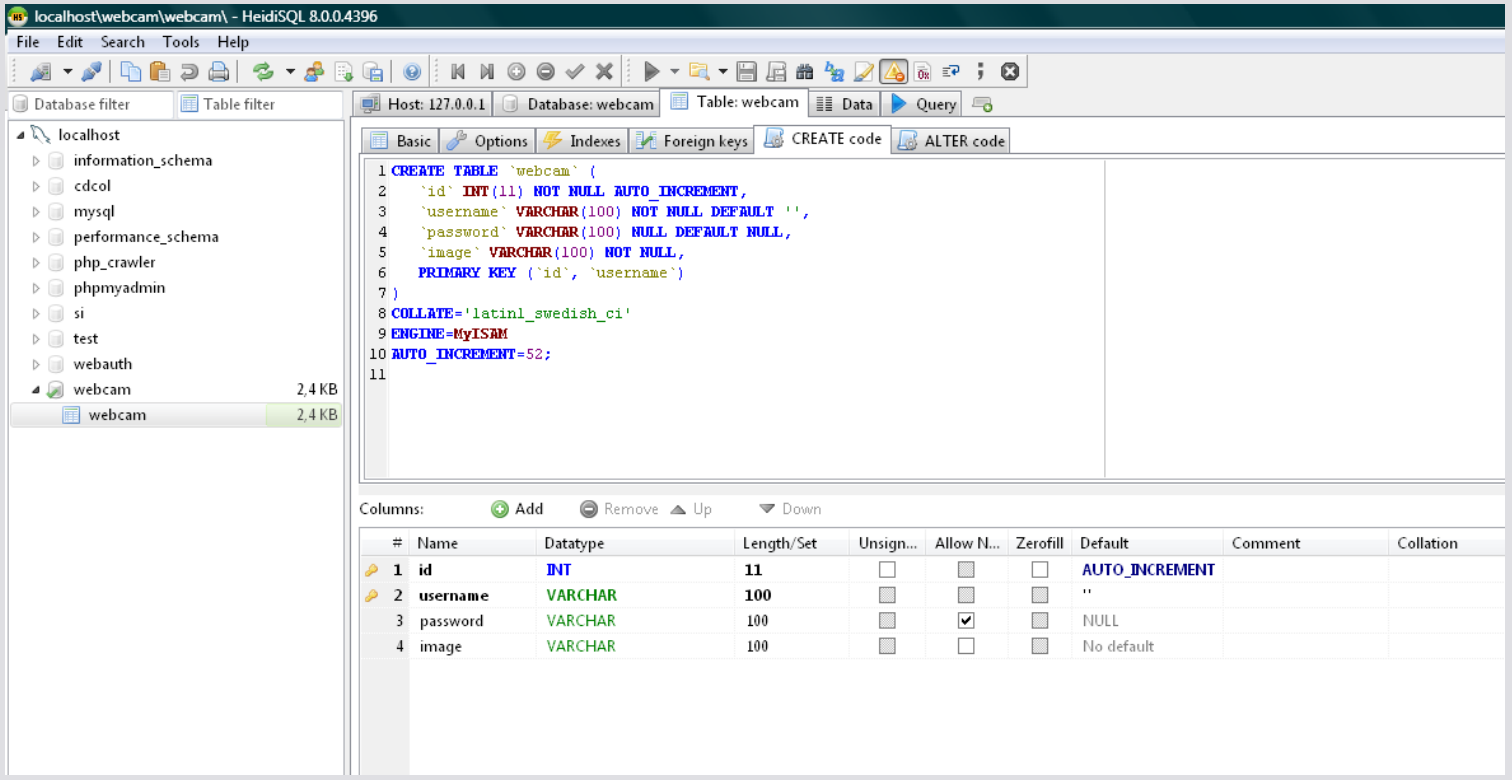
- The cascaded classifier is composed of stages each containing a strong classifier.
- The job of each stage is to determine whether a given sub-window is definitely not a face or maybe a face.

Database and File System Connection

- A database and a file system that are common for web sites with registered users are required.



Mysql Database Example



PCA Recognition

- It is used for user identification
- Firstly we assume that we have N normalized images $n\ x\ n$ that are stored at user's file system.
- These N images make up our training set $\{ I_1\ I_2\ \dots\ I_N \}$. The next step is to reshape them and create image vectors that are represented by Φ.

$$I_i = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix}_{nxn} \Rightarrow \begin{bmatrix} p_{1,1} \\ \cdots \\ p_{1,n} \\ \cdots \\ p_{2,n} \\ \cdots \\ p_{n,n} \end{bmatrix}_{n^2x1} = \Phi_i$$

PCA Recognition

- All N images of the training set are taken into account to determine the mean image vector :

$$M = \frac{1}{N}\sum_{i=1}^{N}\Phi_i$$

- We remove the common information $L_i = \Phi_i - M$, and we find the covariance matrix $C = XX^T$ where $X = [\, L_1\, L_2\, \ldots\, L_N\,]$. Next, we select the best K Eigenvectors. Each face in the training set, Φ_i can be represented as a linear combination of these Eigenvectors ui:

$$L_i = \sum_{j=1}^{K} w_j u_j$$

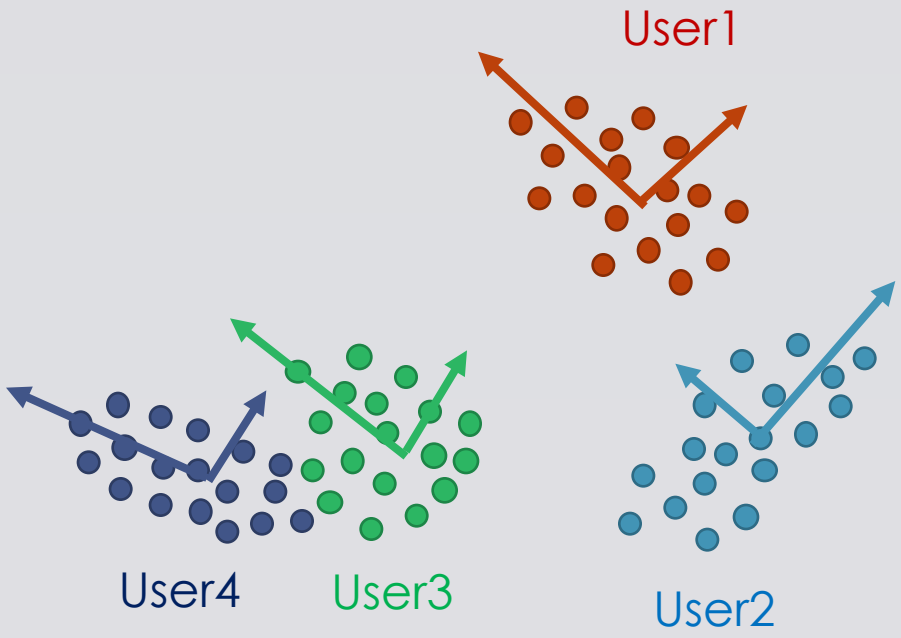
PCA Recognition

- The weights are calculated as follows, $w_j = u_j^T L_i$. Each normalized training image is represented in this basis as a vector $W = [\, w_1\ w_2\ w_3\ \ldots\ w_k\,]^T$. The vector is stored and subsequently is compared with a new vector W' when a new image is received. The same procedure is followed for the calculation of W' as well, with the difference that now $N + 1$ images exist.

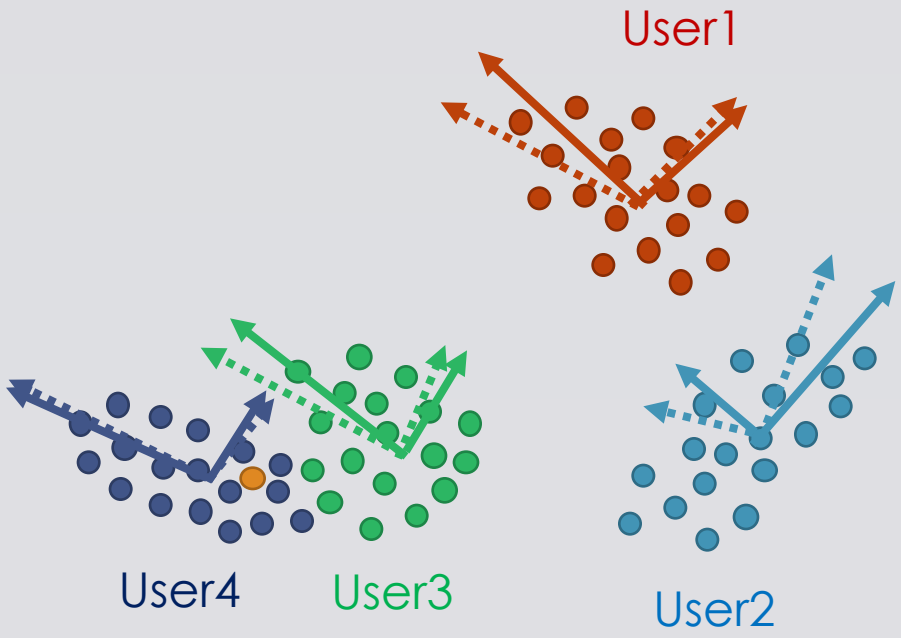
$$e_r = min \left\| W - W' \right\|$$

- If $e_r < \Theta$ the image is deemed to belong to the user. If $e_r > \Theta$ the image is deemed that it does not belong to the user.

User's Subspace



Entrance of New Image



Recognition

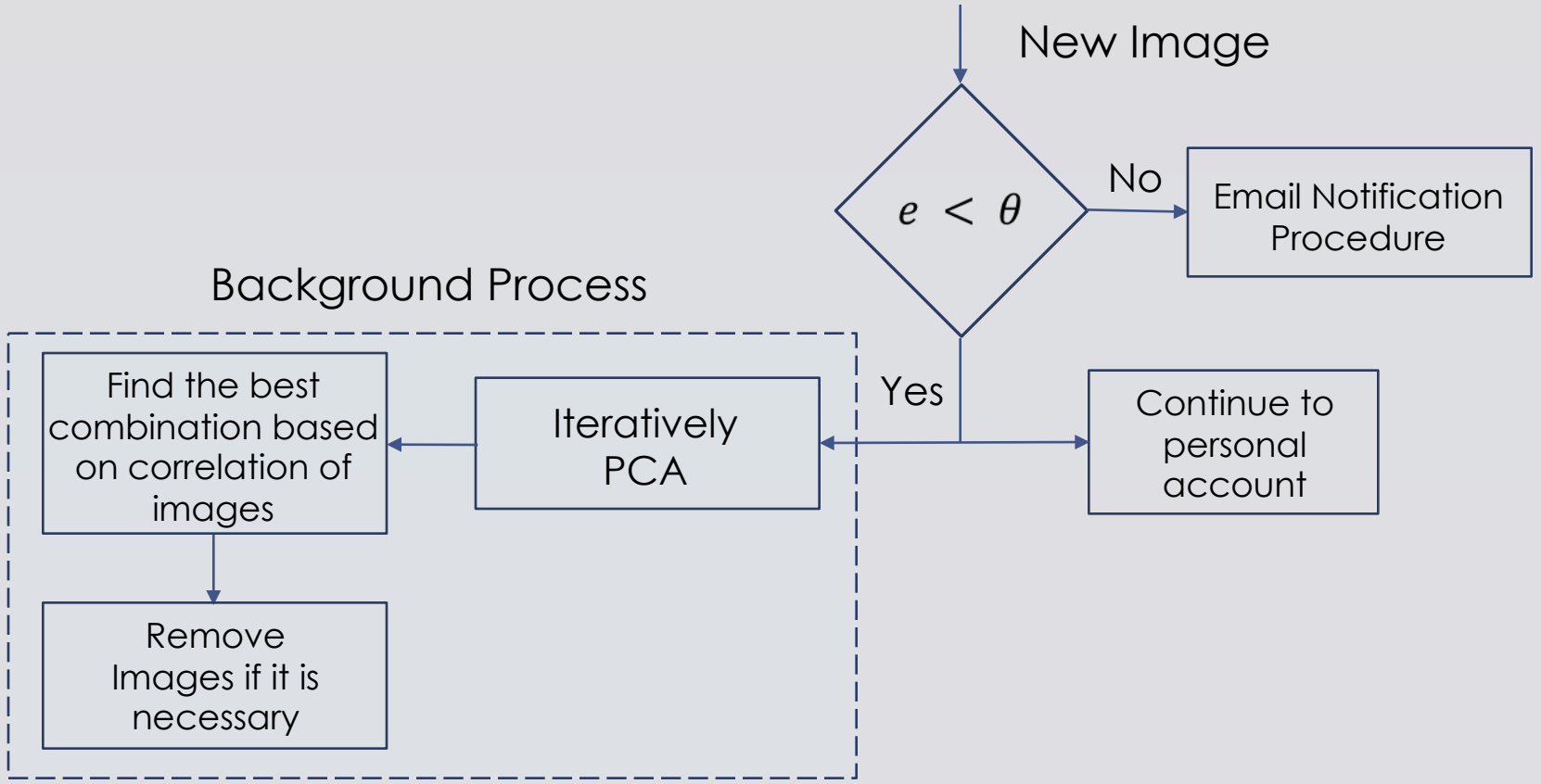
- Next, a query is made at the database to find the path of the file system where user's images are located.
- A PCA method is applied and the vector W' is calculated
- This vector compared with the value of vector W that is stored in database.
- The length of vector is 5 equally to the number of eigenvectors that we have decided to keep.
- If the Euclidean distance between two vectors is lower than a threshold $\Theta = 0.03$ then we assume that the new image is located in the subspace of user's images and the new image is marked as correct image, otherwise we assume that distance is far from the user's subspace and image belongs to someone else.

Auto-Learning and Image Database Update

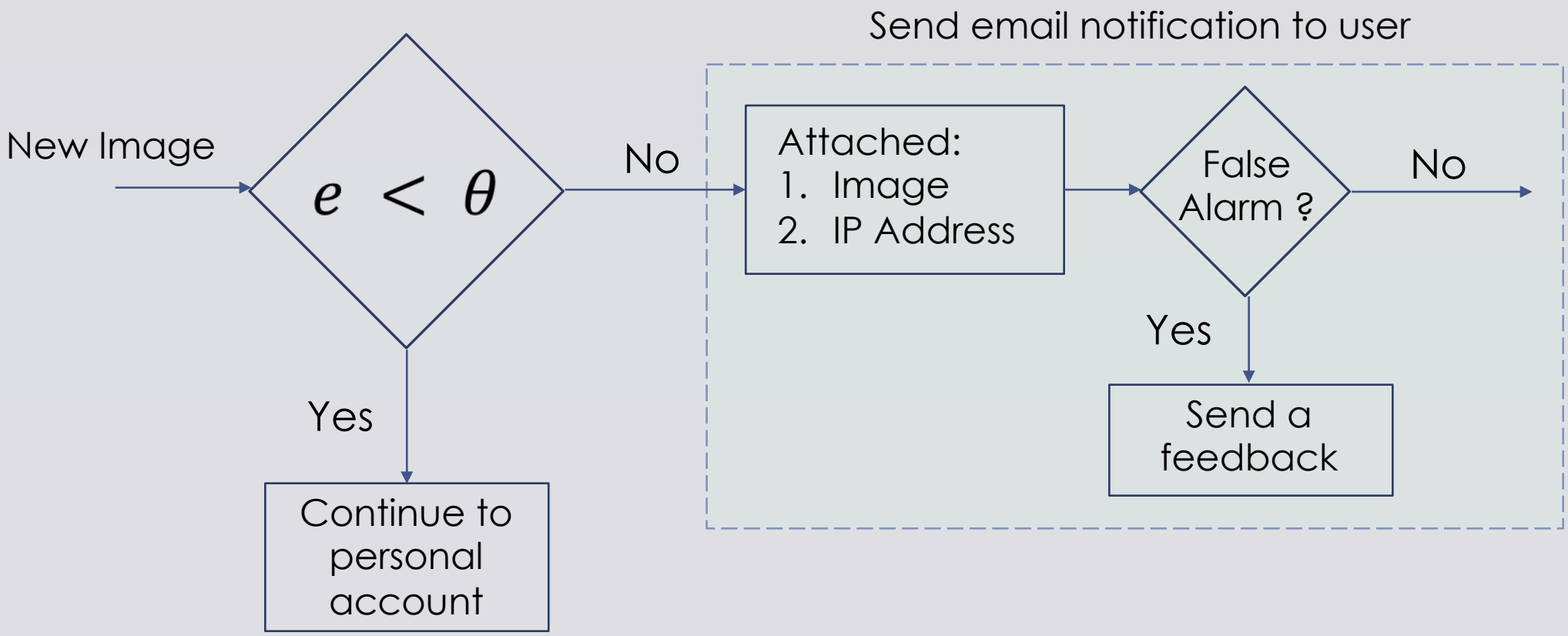
- In order to create a strongly auto correlated database, a function that finds the ideal combination of images is executed every time that a new image is verified to belong to the user. Images with low correlation factor are removed.
- More recent images have higher weights because they represent the current situation.
- Considering the fact that physical characteristics change with time, the latest images represent better the current features of the user and make more efficient a future recognition process. The new mean image is determined by the following formula taking into account the weighted average:

$$M' = \frac{1}{N'} \sum_{i=1}^{N} a_i \Phi_i \; where \; N' > N \; and \; N' = \sum_{i=1}^{N} a_i$$

Iteratively Searching for low Correlation Image



Email Notification



Web Technologies That are Used

- Xampp (Apache, Mysql Database)
- PHP
- Mysql queries
- HTML
- Javascript
- CSS
- Python for image processing and machine learning techniques

Limitations

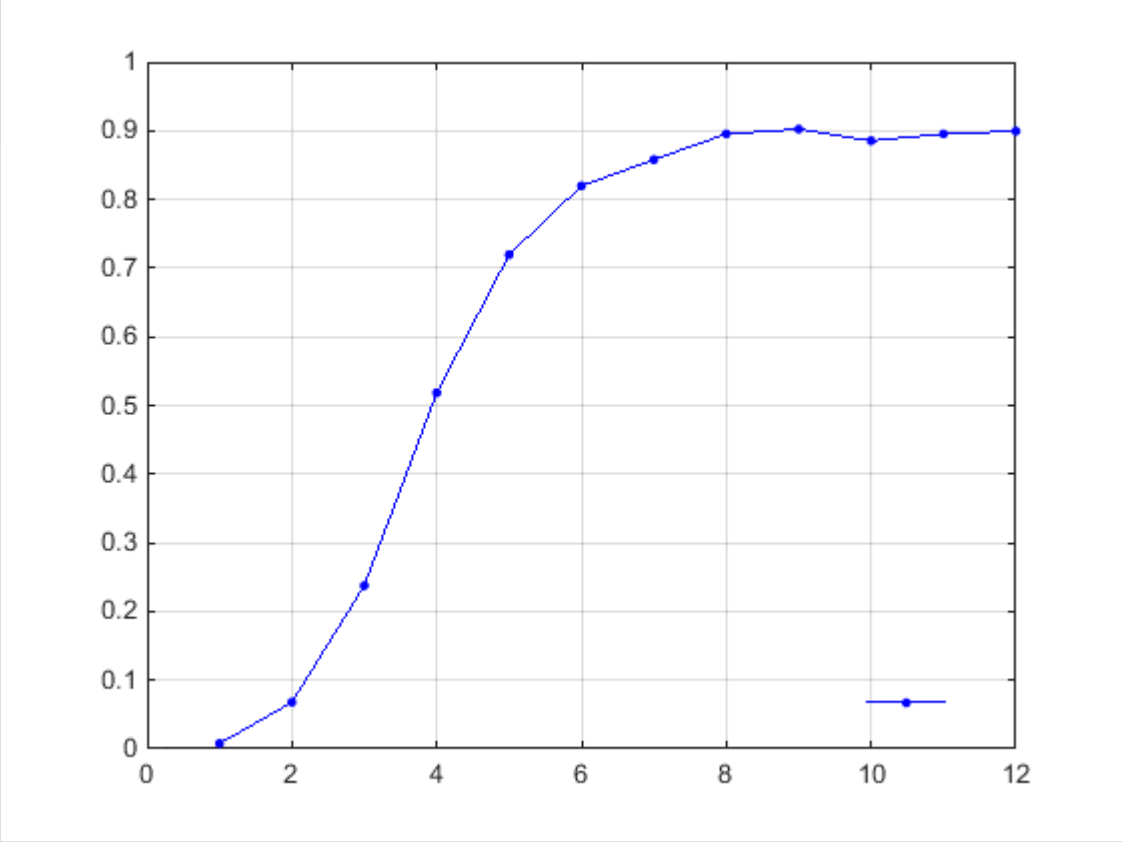
a) The variations in lighting conditions
b) The differences in pose or head orientation
c) Image quality, Web-cams
d) Expressions and partial occlusion (hats, glasses, different hair cutting, beards etc.).

Results

- 10 different Users (6 male, 4 female)
- We investigated different case combinations
- **Case1** User trying to enter at his own account
- **Case2** User trying to enter at another user's account
- **Case3** User trying to enter at his own account with sunglasses
- **Case4** User trying to enter at his own account from different computer and light conditions

	Case 1	Case 2	Case 3	Case 4
User1	10/10	0/10	7/10	10/10
User2	10/10	0/10	6/10	10/10
User3	10/10	0/10	7/10	10/10
User4	10/10	0/10	8/10	10/10
User5	10/10	0/10	5/10	9/10
User6	10/10	0/10	6/10	10/10
User7	10/10	0/10	4/10	10/10
User8	10/10	0/10	7/10	10/10
User9	10/10	0/10	6/10	10/10
User10	10/10	0/10	3/10	9/10
Sum	**10/10**	**0/10**	**5.9/10**	**9.8/10**

How the Number of Images Affects to the Results



Conclusions

- In this research we present a way to create a more secure login environment for websites.
- This extra safety level protects mainly the most careless users with weak passwords.
- The proposed method uses face detection and recognition in order to identify the user.
- The algorithms that we use are robust and trustworthy. Moreover, the real time processing is very fast and seems an ideal technique for online usage.
- A minimum number of images are required in order to achieve the best results (>7).
- Differences of light conditions, pose and web camera resolution can be managed.
- Sunglasses and other objects can affect the identification results. If a change is permanent, a new folder of images should be created.
- More accuracy with a sequence of images (short video) or voice.

Thank You

- Questions ?