Video killed the text star

-

Francisco Jesus Gomez Rodriguez and Cesar Jimenez Zapata

Contents

1 Presentation

We talk in this document about facial recognition and video processing. In this document we will try to show you which are the latest trends in facial recognition and the technology behind this type of work.

We will talk about OSINT, video analysis, facial recognition and how to find faces in images and videos, how to encode these faces so that a computer understands them, how to identify people in these videos and we will see a demonstration of all this in operation with a possible architecture hardware and software that can be used for fast and effective processing. In a second phase we will see the code that makes this possible in Java and Python, and finally we will talk about how we can classify the people that appear in the videos which have not been identified. In a final appendix we will also see some ways to defend against this type of identifications, we will see that some of them are quite daring to use on a daily basis, and thus avoid these automatic facial recognition systems.

OSINT is a series of techniques, technologies and methodologies that are used to obtain intelligence information from various open sources, understood as open to the public. This type of information is very common nowadays because social networks and the different services and web pages that people use to communicate , provide a great source of information. As you can see in 1992 "The official definition of OSINT by the U.S. Intelligence Community: By Open Source we refer to publicly available information appearing in print or electronic form. Open Source information may be transmitted through radio, television, and newspapers, or it may be distributed by commercial databases, electronic mail networks, or portable electronic media such as CDROM's. It may be disseminated to a broad public, as are the mass media, or to a more select audience, such as gray literature, which includes conference proceedings, company shareholder reports, and local telephone directories. Whatever form it takes, Open Source involves no information that is: classified at its origin; is subject to proprietary constraints (other than copyright); is the product of sensitive contacts with U.S. or foreign persons; or is acquired through clandestine or covert means."[1], There was an official definition of what was understood by OSINT given by the United States intelligence community.

A more modern definition could be that OSINT[2] is the data collection publicly available and used in an intelligence context. What is an intelligence context? Well, intelligence [3] is the development of behaviour forecasts or recommended courses of action, for the leadership of an organisation, based on a wide range of open information available.

OSINT has been used in many cases, in issues of national security, counterterrorism, cyber terrorist tracking, search for lost people, identification of people in sexual crimes robberies and other crimes, identification of people involved in crimes of identity theft. It is used in companies to monitor the activity of the competition, or in the case of hackers, to obtain information for a specific purpose.[4] There are also cases in which the analysis of videos is used to see the impact of a brand (amount of time seen on television) at sporting events and thus calculate the return on investment of sports sponsorship[5][6]. Others uses can be automated border processing system and prevent voter fraud

2 Video analysis

Video analysis, and in particular the recognition of people, is a very interesting task for different security managers, CISOs and people with responsibilities in physical security.

However, video and image are gaining importance on the Internet where content grows exponentially in this type of formats. We believe that obtaining information from these video sources should begin to take importance within the techniques and technologies used in

OSINT and we should be able to process and analyze this type of information.

Video traffic has not stopped growing on the Internet and in 2018 it is estimated that traffic circulating in video on the Internet is 90 million terabytes per month, of this traffic approximately 2% is traffic that comes from surveillance videos, which would give us 180,000 terabytes per month in content of surveillance videos[7]

One interesting article The FIVE report or "NIST Interagency Report 8173: Face In Video Evaluation (FIVE) Face Recognition of Non-Cooperative Subjects" [8] say "Face in Video Evaluation (FIVE), an independent, public test of face recognition of non-cooperating subjects who are recorded passively and are mostly oblivious to the presence of cameras" "he report enumerates accuracy and speed of face recognition algorithms applied to the identification of persons appearing in video sequences drawn from six different video datasets mostly sequestered at NIST"[9]

Governments and some private institutions have seen in this video analysis a very valuable source of information and, to mention a few cases, the Chinese government is known to have large projects of research in this field to allow the control of its citizens and always obtain useful and valuable information about their activities. We will talk about these social impacts at the end of the presentation.

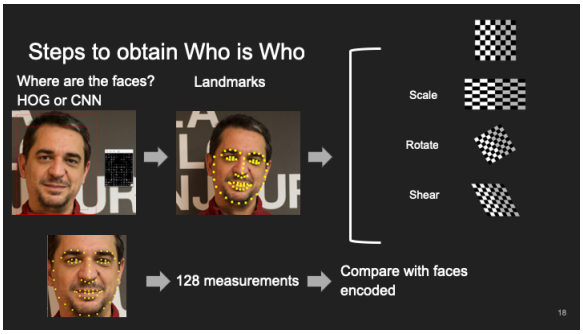
A very clear example of its usefulness is what we show below, which, we warn, can hurt the sensitivity of some people, so we apologise in advance to anyone who may feel uncomfortable. [10]

In this video, a series of terrorists kill hostages and record promotional videos to spread their ideas and beliefs. This video has been analysed and, thanks to it, you can identify different places correlating data with an application such as Google Maps. As you can see a search in Google Maps gives us a possible location based on the bushes, the buildings in the background and the dirt roads that appear in the image. Again the satellite images of Google Maps allow us to know in what approximate day these murders were made, since the different images show bloodstains that coincide with the position of the people that appear in the video.

3 Facial recognition

In this discipline there has been great advances in recent years. These advances, of which we will speak a little later, and the processing power we have in current computers, allow us to implement this type of services at a very low cost, compared to the cost that this type of systems had few years ago

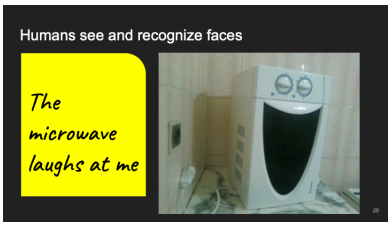
Making these facial recognitions involve a series of steps, the first of which will be to identify the positions of the faces that appear in the videos. Followed by the identification of the reference points of the faces, we will apply a series of affine transformations to later get a vector of 128 dimensions that will identify each face. Finally, by comparing the proximity of each of the vectors obtained, we will know which images correspond to the same person.



Let's explain each of these steps a little more

in detail.

The first step will be to find out in which images faces appear and what coordinates those faces have in each of the images. Humans are very good at recognising faces, and many times we see them where they do not exist as in this microwave that we can see in the image.



The two most commonly used techniques at the present time are HOG (oriented gradient histogram technique) and CNN or convolutional neural networks.

3.1 HOG

The HOG technique was developed in 1986 by Robert K. McConnell of Wayland Research Inc. [11] but in 2005 Navneet Dalal and Bill Triggs [12] use to identify human models in images.

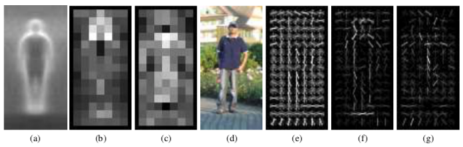
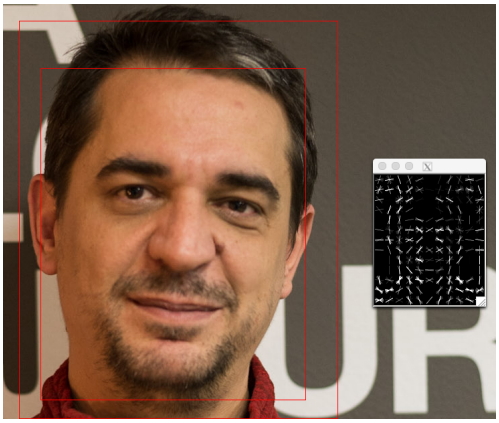
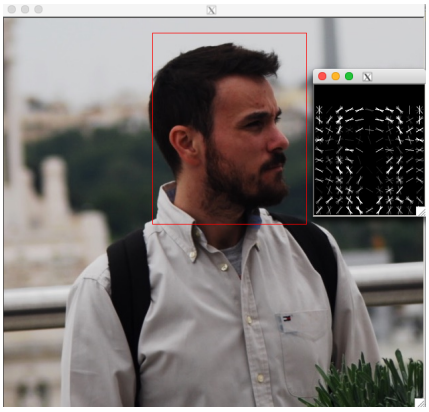


Figure 6. Our HOG detectors cue mainly on silhouette contours (especially the head, shoulders and feet). The most active blocks are centred on the image background just *outside* the contour. (a) The average gradient image over the training examples. (b) Each "pixel" shows the maximum positive SVM weight in the block centred on the pixel. (c) Likewise for the negative SVM weights. (d) A test image. (e) It's computed R-HOG descriptor. (f,g) The R-HOG descriptor weighted by respectively the positive and the negative SVM weights.

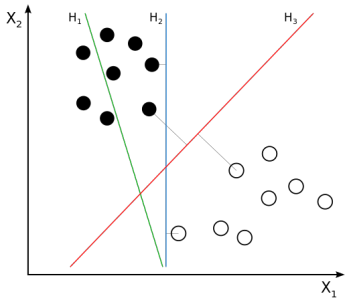
We will use it to identify only the faces that appear in the videos. This technique is based on each of the images to be treated. We must identify in which direction the darkest points of the image move and draw oriented gradients towards this direction, which allows us to have a symbolic representation of said images. When operating in local cells it is invariant to geometric transformations, but variant to the orientation of the objects. As we can see, in these two examples the representation of each face is identified with a specific histogram of oriented gradients, and all the faces have a very similar representation which allows us to identify through algorithms the positions in which faces appear in each image.





In this technique the classification of the areas of the images that have or do not have faces can be made with a Support Vector Machines (SVM) system, a discriminative classifier formally defined by a separating hyperplane. In other words, with labeled training data (supervised learning), the algorithm generates an optimal hyperplane that categorises new examples. [13]

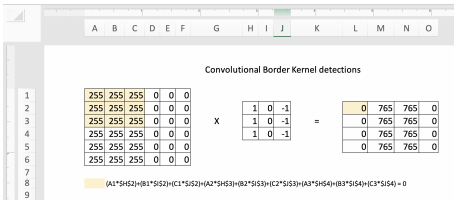
This sounds complicated, right? Let's see the simplest example, in two dimensional spaces, this hyperplane is a line that divides a plane into two parts, where each class is on each side of the line. This, carried to "n" dimensions, and the straight line being any surface in the n dimensions, is what we use to classify the examples.



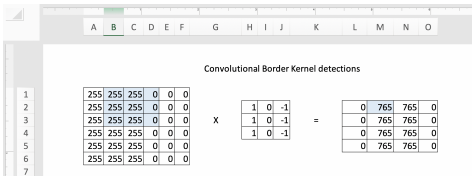
Graphic showing how a support vector machine would choose a separating hyperplane for two classes of points in 2D. H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin.[14]

3.2 CNN

The other technique that can be used to identify faces in images or videos is convolutional neural networks. These Neural Networks have been widely used to address various problems in recent years, giving very good results in these studies. Convolutional neural networks are a class of deep neural networks as a variant of multilayer sensors designed to require much lower processing than traditional deep neural networks. These networks are inspired by the biological processes of connective patterns between neurones that are found in the visual cortex of animals. This operation combines a series of convolutional layers, consisting of filters and image reductions, using different passes on the same image with what is called a kernel, and putting these layers before the different layers of a neural network. Let's explain this a little more. A convolutional layer is simple to explain. We start with a matrix with the values of an image (from 0 to 255 in each value of the matrix, in the case of a single layer of a grayscale image), in the presentation we put an image of 5x5 points. We are going to apply a 3x3 kernel with the values of 1, 0 and -1 in each of its three columns. We will move the kernel along the whole image and calculating the value of the multiplication of each of the values of the image by the value of the same cell in the kernel and adding all the values to obtain the final value.

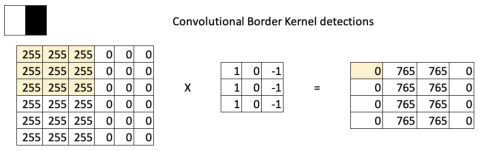


In the first step , (255 * 1) + (255 * 0) + (255 * -1) + (255 * 1) + (255 * 0) + (255 * -1) + (255 * 1) + (255 * 0) + (255 * -1) = 0

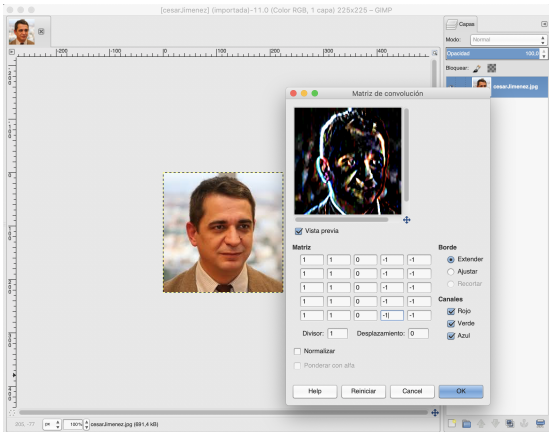


In the second step , (255 * 1) + (255 * 0) + (0 * -1) + (255 * 1) + (255 * 0) + (0 * -1) + (255 * 1) + (255 * 0) + (0 * -1) = 765 , normaliced to 255

For example, if we want to detect the vertical edges of an image, we can apply the following kernel and we will highlight those edges.

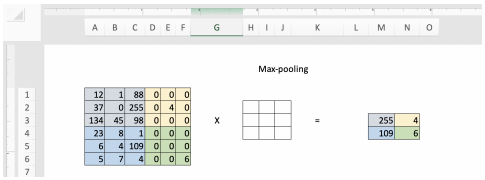


The GIMP tool allows us to play with this type of convolution operations.



The other operation on the matrices that we do in the CNN is a subsampling operation, one of the most used is the Max Pooling. These operations are also very simple, we simply choose a matrix size, for example 3x3 , and we divide the image into regions of the same size, taking from the whole region only the largest of the values that are in it. Notice that in the subsampling

we divide the image into zones and we do not move the "operator" of 3x3 as we did with the kernel.



Finally , with all the values obtained in the different operations of convolution and subsampling, flattening the matrices in a vector, we feed a traditional multi-layer perceptron of artificial neural networks, with completely connected layers, with their inputs, their activation function and their outputs.

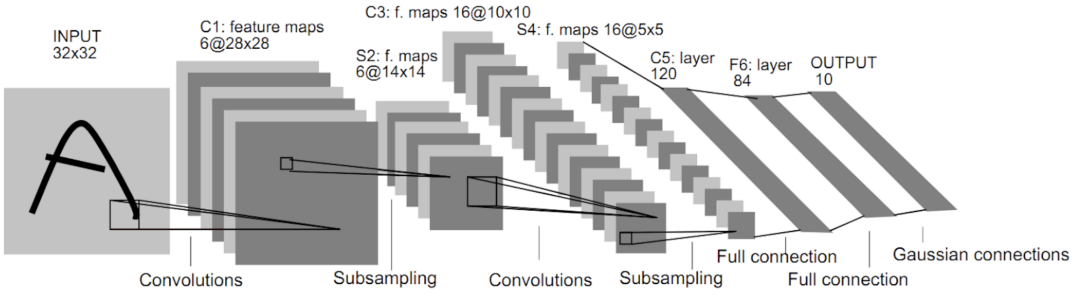


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Reviewing the image that we can found in (Gradient-Based Learning Applied to Document Recognition by LeCun et al.) [15] we will obtain from the letter A of 32x32 and "suppose" a convolution C1 of 6 layers of 28x28 (for example, two filters of 5x5 kernel on the 3 RGB channels since the formula of the size of the resulting matrix, if we start from an image of nxn with a kernel of fxf without padding is (n-f + 1) x (n-f + 1)), we will apply a subsampling S2 of 2x2 until reducing the input to 6 channels of 14x14, followed by other convolutions C3 obtaining 16 channels of 10x10, a subsampling S4 to obtain 16 channels of 5x5 and finally flattened the matrix to a vector and obtaining 120 values or 120 neurons in the input layer.

These same techniques also allow us to identify any type of objects such as flags, weapons, vehicles, buildings, and so on. Let's see some images of how these same techniques can be used to identify flags in videos.

The recognition of objects is often not a simple task, not even for humans, as we can see in this image where it is difficult to differentiate which ones are Chihuahuas and which ones are

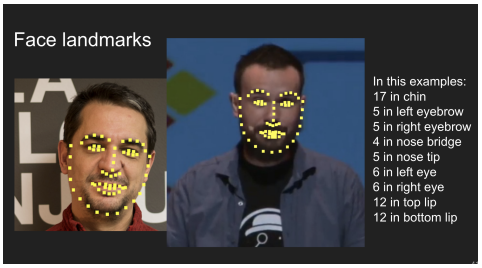
muffins .

When do we use CNN or HOG? The CNN algorithm is more precise in object recognition, by contrast, the HOG algorithm is faster in processing.

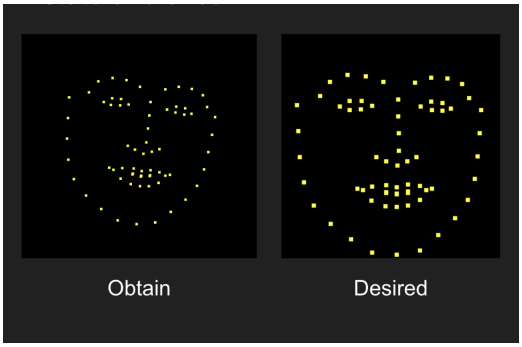
4 Face landmarks

After knowing the positions of the different faces within the images, we will identify different points in a image of a face.

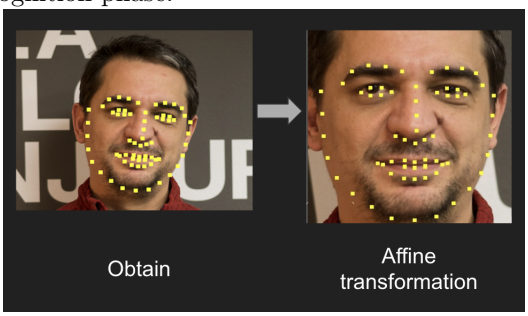
These points , landmarks , and their identification arose from a series of investigations in 2014 by Vahid Kazemi and Josephine Sullivan. In these examples we can identify 17 points for the line of the chin and the contour of the face, 5 points for each of the eyebrows, 4 points for the length of the nose, 5 points for the lower part of the nose, 6 points for each eye and 12 points for each upper and lower lip.



Why do we need to identify these points? We will use them in the next phase, so that all the faces are aligned with each other as much as possible. For this by using affine transformations of the images with operations like scaling, rotating and stretching the images we will obtain a pattern of points on the similar face in all the faces that we have detected in the images.



This prevents images of the same person in different positions from giving different results in the next phase, which will allow us to obtain a numerical code for each face and its subsequent recognition phase.



5 Face encoding

How do we get that numerical code for each of the faces we are analysing? This we will be based on a research study carried out by Google in 2015[16]

In this study researches talk about the analysis of the different images using a neural network and training this network with triplets of faces in which two faces belong to the same person and one face belongs to another person.

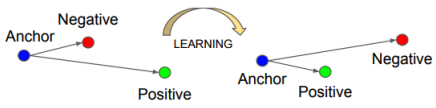
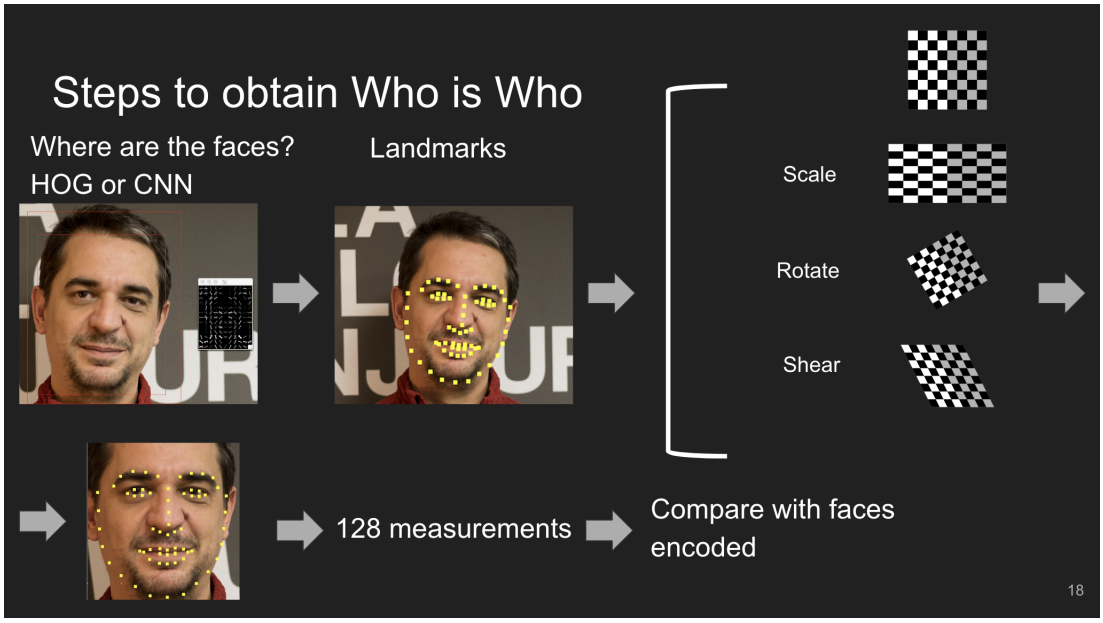


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

This allows minimising the distances between positive and negative results, finally obtaining a vector in 128 dimensions that identifies each face, something like this [-0.07996784150600433, 0.15368174016475677, 0.1399957537651062, -0.13417772948741913, -0.11255471408367157, -0.045209698379039764, -0.04773351550102234, -0.09028300642967224, 0.22902674973011017, -0.10360846668481827, 0.21877643465995789, 0.026112765073776245, -0.1701708734035492, 0.014600440859794617, -0.022751256823539734, 0.19936016201972961, -0.2355552613735199, -0.15861636400222778, -0.1386137157678604, -0.10872261971235275, 0.0037924107164144516, 0.10839948803186417, 0.100742407143116, 0.0105014368891716, -0.12778553366661072, -0.3286711871623993, -0.055458199232816696, -0.019983578473329544, 0.10712774097919464, -0.13793827593326569, 0.009068422019481659, 0.09064967930316925, -0.1343669295310974, 0.017166361212730408, 0.044025078415870667, 0.050798915326595306, -0.07990403473377228, -0.15819168090820312, 0.2576793134212494, 0.004978094715625048, -0.21836695075035095, -0.02650071680545807, 0.09678448736667633, 0.21061867475509644, 0.3103189170360565, 0.0113909300416708, 0.03714804723858833, -0.15217365324497223, 0.11215507984161377, -0.24408987164497375, 0.0763545036315918, 0.21450649201869965, 0.02021719515323639, 0.11482518911361694, 0.09251868724822998, -0.15836083889007568, -0.031153470277786255, 0.14305520057678223, -0.1970769166946411, -0.018679585307836533, 0.030500710010528564, -0.046877145767211914, -0.027389904484152794, -0.11235783994197845, 0.1841321885585785, 0.15685541927814484, -0.12544497847557068, -0.21651992201805115, 0.17368975281715393, -0.1649446338415146, -0.1660388559103012, 0.058186858892440796, -0.1616286337375641, -0.14266015589237213, -0.2833820879459381, -0.07552418112754822, 0.2916351556777954, 0.13597646355628967, -0.13732105493545532, 0.05478135496377945, -0.04149997979402542, -0.10759354382753372, -0.064942866563797, 0.15462595224380493, -0.11839039623737335, -0.03558868169784546, -0.0315987765789032, 0.029462680220603943, 0.19941174983978271, -0.011531651020050049, -0.08430354297161102, 0.21216872334480286, 0.009728733450174332, -0.017769530415534973, -0.030391577631235123, 0.07924357056617737, -0.1493520438671112, 0.022632021456956863, -0.17380347847938538, -0.12114172428846359, 0.01225726306438446, 0.0387752428650856, -0.023811623454093933, 0.19535689055919647, -0.18216219544410706, 0.24811719357967377, -0.033723384141922, -0.017802998423576355, -0.03762838989496231, -0.023586824536323547, -0.016338270157575607, -0.03003990650177002, 0.15570029616355896, -0.1804899275302887, 0.20914554595947266, 0.21273823082447052, 0.04884400963783264, 0.14714393019676208, 0.07800989598035812, 0.09556171298027039, 0.035943403840065, 0.023756101727485657, -0.13991279900074005, -0.05304291099309921, 0.015071794390678406, -0.022036418318748474, 0.07947075366973877, 0.10163198411464691]

This vector is very similar for faces that belong to the same person and different for faces of different people. These 128 characteristics do not correspond with real face measurements. They are not the width between the eyes or the length of the nose or others face measurements. They are the output of the neural network that determines them. To train these models we will need different faces of the people we want to recognise and faces of different people.

In the following image we show a summary of all the previous steps, where we can see the location of the faces, obtaining the landmarks, affine transformations, obtaining the vector of 128 dimensions and the final comparison with our know faces data model.



6 Face recognition

This training can be done with a small number of images of the same person or, if we want more precision, with a greater number of images

We obtain an optimal result with 42 images of the same person. This optimised process for its calculation with GPU can reduce in 10 times the time used in the calculation only using CPU. The recognition of the people will be based on the comparison of the 128 measurements of a face obtained from an image and our model where the different known faces will be identified. We can determine the tolerance or how close these two vectors should be within the space of 128 dimensions to consider that two faces are the same person. This comparison in its simplest way is simply to calculate the Euclidean distance between these vectors and if this distance is within our tolerance range we will consider it is the same person. This tolerance will allow us, choosing a correct value for the cases we are studying, to usefully balance the range of false positives, false negatives, true positives and true negatives that we will obtain in our analysis. Finally, we can highlight the persons that have been identified in the images with different squares in which we can include information about the person we know: name, date of birth, profession, or any other data that we have.

7 Architecture

All the work that we are going to show here can be reproduced with the code that we will provide in a link for its download and with the following tools or OpenSource libraries.

- We will use OpenCV, a software widely used in computer vision that is developed by Intel and started in 1999.[17]
- Apache storm is an Open source project that we will use for the real-time processing of different video images.[18]
- Dlib is a library written in C ++ what contains different machine learning algorithms and tools to process images[19]
- We also use Python and Java as program-

ming languages and the face_recognition library, which facilitates the use of different facial recognition functions[20]

The architecture will be based on a series of programs implemented in a stream of Storm that will allow us to process hundreds of hours of video in a distributed way.

We will use Storm for its ability to process in real time (such as processing real-time video surveillance images), because it is easy to prototype and program in any programming language (due to its rapid process of tuples or events), its ability to be scalable and fault tolerant and because it allows us to execute many of the tasks in parallel.

The general process will begin with a download module that will allow us to obtain the different videos that we obtain from the Internet. There are different tools to do this and each video source requires a special treatment to download them, although this task does not seem in principle a complex task, there are several problems that must be taken into account to obtain more video recordings that will allow us to perform a better analysis. We must bear in mind that, as we all know, more data means better analysis. At this point it should be noted that we require a large storage space of hundreds of terabytes to be able to store all this content in video and it will be important that this storage has a sufficiently high access speed in order to process the largest amount of information in the shortest time

At the time of processing the video, the task we will perform is quite simple, it consists of scaling the videos of their original size to a smaller size to speed up their process. From this video reduced in size, what we will do is obtain a series of fixed images as frames of said video. The video can usually be found in different qualities but it is quite common to find recorded videos at 60 frames per second, this means that in one hour of video we can obtain 216,000 images for processing. Here we must solve a first obstacle, do all video frames really offer us information that is just as interesting for facial recognition? The answer is no, it is not necessary to obtain the 216,000 images of an hour of recording for the task that we will do, which will be the recognition of the people that appear in this video. The number of frames that we will obtain from each second of video will determine the amount of time we have to use to the processing of that video hour. One approach that can offer good results is to make a combination of this phase of the video processing with the next phase, which we will see to identify the faces that appear in it, in such a way that we can obtain two or three frames per second, analyze and if faces appear in those frames, take that portion of the video and get more images of each second of video for analysis. This will allow us to process a much smaller number of images in an hour of video, and these images will mainly contain the faces of the people we want to identify.

You can obtain the code from public Devo github https://github.com/DevoInc and have all the steps in it.

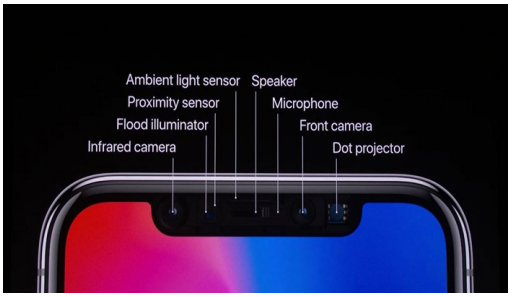
8 Three dimensional face recognition

Three dimensional facial recognition is a modality of facial recognition methods in which the three dimensional geometry of the human face is used. It has been shown that 3D facial recognition methods can achieve significantly greater accuracy than their 2D counterparts, rivalling fingerprint recognition. 3D scanners acquire a 3D mesh and the corresponding texture. This allows to combine the output of pure 3D purifiers with the more traditional 2D facial recognition algorithms, which produces better performance [21]

8.1 Face ID

Face ID and TrueDepth camera technologies, developed by Apple, works "once it confirms the presence of an attentive face, the TrueDepth camera projects and reads over 30,000 infrared dots to form a depth map of the face, along with a 2D infrared image. This data is used to create a sequence of 2D images and depth maps, which are digitally signed and sent to the Secure Enclave. To counter both digital and physical spoofs, the TrueDepth camera randomizes the

sequence of 2D images and depth map captures, and projects a device-specific random pattern. A portion of the A11 Bionic processor-s neural engine-protected within the Secure Enclave-transforms this data into a mathematical representation and compares that representation to the enrolled facial data. This enrolled facial data is itself a mathematical representation of your face captured across a variety of poses.



Facial matching is performed within the secure enclave using neural networks trained specifically for that purpose. We developed the facial matching neural networks using over a billion images, including IR and depth images collected in studies conducted with the participants- informed consent. We worked with participants from around the world to include a representative group of people accounting for gender, age, ethnicity, and other factors. We augmented the studies as needed to provide a high degree of accuracy for a diverse range of users. Face ID is designed to work with hats, scarves, glasses, contact lenses, and many sunglasses. Furthermore, it's designed to work indoors, outdoors, and even in total darkness. An additional neural network that-s trained to spot and resist spoofing defends against attempts to unlock your phone with photos or masks. Face ID data, including mathematical representations of your face, is encrypted and only available to the Secure Enclave To improve unlock performance and keep pace with the natural changes of your face and look, Face ID augments its stored mathematical representation over time. Upon successful unlock, Face ID may use the newly calculated mathematical representation-if its quality is sufficient-for a finite number of additional unlocks before that data is discarded. Conversely, if Face ID fails to recognize you, but the match quality is higher than a certain threshold and you immediately follow the failure by entering your passcode, Face ID takes another capture and augments its enrolled Face ID data with the newly calculated mathematical representation. This new Face ID data is discarded after a finite number of unlocks and if you stop matching against it. These augmentation processes allow Face ID to keep up with dramatic changes in your facial hair or makeup use, while minimizing false acceptance." [22]

8.2 RealSense

"The Intel RealSense Depth Camera [...] uses stereo vision to calculate depth [...] are USB-powered and consist of a pair of depth sensors, an RGB sensor, and an infrared projector. They are ideal for makers and developers to add depth perception capability to prototype development." [23]

"In this platform you can use OpenVINO. The release of the Open Visual Inference and Neural Network Optimization (OpenVINO) toolkit by Intel gives developers a rapid way to implement deep learning inference solutions using computer vision at the network edge. This addition to the current slate of Intel Vision Products is based on convolutional neural network (CNN) principles, making it easier to design, develop, and deploy effective computer vision solutions that leverage IoT to support business operations.

The components in the toolkit include three APIs:

A deep learning inference toolkit supporting the full range of Intel Vision Products. A deep learning deployment toolkit for streamlining distribution and use of AI-based computer vision solutions. A set of optimized functions for OpenCV and OpenVX. Currently supported frameworks include TensorFlow, Caffe, and MXNet. The toolkit helps boost solution

performance with numerous Intel based accelerators, including CPUs and integrated graphics processing units (GPUs), field-programmable gate arrays, video processing units, and image processing units.

Processing high-quality video requires the ability to rapidly analyze vast streams of data near the edge and respond in real time, moving only relevant insights to the cloud asynchronously. The OpenVINO toolkit is designed to fast-track development of high- performance computer vision and deep learning inference applications at the edge." [24]

9 Identify unknown people

We are now facing the task of how to identify unknown people in our videos. In which videos do the same people appear? This person In what other videos does it appear?

From March 8 to June 8, 2018, the Counter Extremism Project (CEP) conducted a study to better understand how ISIS content is being uploaded to YouTube, how long it is staying online, and how many views these videos receive. To accomplish this, CEP conducted a limited search for a small set of just 229 previously-identified ISIS terror-related videos from among the trove of extremist material available on the platform. [35]

Some of the conclusions of the studies are:

1. 1,348 ISIS videos were uploaded to YouTube, garnering 163,391 views.

2. 24 percent of those videos remained on YouTube for over two hours, receiving 148,590 views.

3. 76 percent of those videos remained on YouTube for less than two hours, receiving 14,801 views.

4. 278 accounts uploaded all 1,348 videos to YouTube.

5. 60 percent of accounts remained live after uploaded videos had been removed for content violations.

According to YouTube's Community Guidelines, it "do[es] not permit terrorist organizations to use YouTube for any purpose, including recruitment. YouTube also strictly prohibits content... that promotes terrorist acts, incites violence, or celebrates terrorist attacks." [36]

Since June 2017, YouTube has also taken measures towards hiding extremist content that does not explicitly violate their rules, such as putting inflammatory videos behind warning labels.

The main conclusion of these studie conducted by CEP was they findings call into question YouTube's claims of proactive content removal efforts. The fact that 91 percent of extremist videos had been reuploaded to YouTube at least once casts doubt on YouTube's stated efforts to prevent the upload or removal of known terrorist material.

Therefore, we believe that it is important to use content that is uploaded to social networks, in this case YouTube, as a source of information.

The proposed use case seeks to obtain a database where the different identities are related based on their appearance in different videos over time. We base the concept of identity on the detection and processing of a person's face.

This database being a starting point for any type of research.

9.1 pHash

A first approximation that could occur to us would be to use a hash algorithm by proximity as phash. pHash is a perceptual hashing algorithm "that produces a snippet or fingerprint of various forms of multimedia. Perceptual hash functions are analogous if features are similar, whereas cryptographic hashing relies on the avalanche effect of a small change in input value creating a drastic change in output value. Perceptual hash functions are widely used in finding cases of online copyright infringement as well as in digital forensics because of the ability to have a correlation between hashes so similar data can be found (for instance with a differ-

ing watermark). For example, Wikipedia could maintain a database of text hashes of popular online books or articles for which the authors hold copyrights to, anytime a Wikipedia user uploads an online book or article that has a copyright, the hashes will be almost exactly the same and could be flagged as plagiarism. This same flagging system can be used for any multimedia or text file."[25]

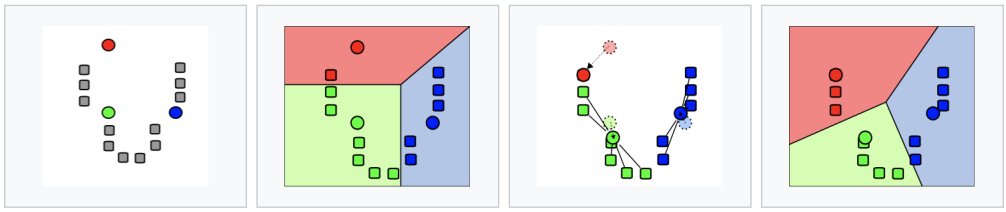
This does not work because differences in backgrounds, orientation of the faces, etc. give us quite different values for the same person, but are interesting hashing algorithms to detect logos or brands in fraudulent webs in phishing cases.

9.2 k-means

We will start then to create a dataset with the vectors of 128 dimensions of all the people that appear in all the videos that we process and we will try two unsupervised algorithms to group those vectors.

First we will use a k-means algorithm where we aim to divide n observations in k clusters in which each observation belongs to the cluster with the closest average, and serves as prototype of the cluster. This results in a partition of the data space into Voronoi cells, which are a geometric construction that allows the construction of a partition of the Euclidean plane based on the Euclidean distance around the centroids.[26]

Demonstration of the standard k-means algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).
2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.
3. The centroid of each of the *k* clusters becomes the new mean.
4. Steps 2 and 3 are repeated until convergence has been reached.

In this way each of the k-means groups will be a person. It works well if we know how many different people appear in the videos and we apply a k-means algorithm with this number of centroids, but we usually do not know that number and the estimation of the number of centroids (different people) with different methods does not work very well in our investigations.

9.3 Chinese whispers

Next we will use an algorithm called "Chinese whispers". This algorithm, also used in the processing of natural language, is based on the creation of a graph with all the faces identified (vector of 128 dimensions), joining the nodes that are less than a "tolerance" in Euclidean distance and following the following steps:

1. All nodes are assigned to a random class. The number of initial classes is equal to the number of nodes.
2. Then, all the nodes in the network are selected one by one in a random order. Each node moves to the class that the given node connects to most of the links. In the case of equality, the group is chosen randomly from the equally linked classes.
3. Step two is repeated until a predetermined iteration number or until the process converges. In the end, the classes represent the groups in the network.

[27]

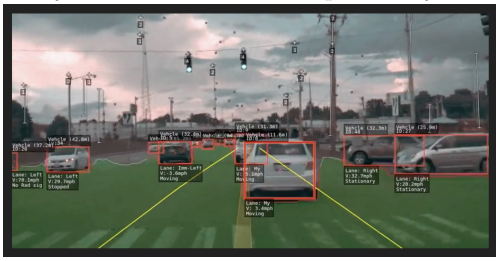
This algorithm groups the people that appear in the videos in classes quite well, where each class is a person.

Classification of 128 faces in Devo employees, 1 false positive

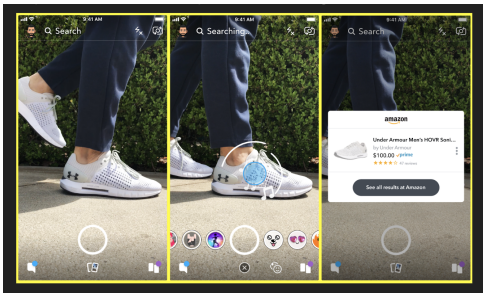
10 Social impact

We will see some examples of the social impact of this type of technology in a day-to-day basis. These techniques are used in supermarkets and stores to identify customers, observe the impact of prices and advertisements, analyze customer reactions, etc.[28]

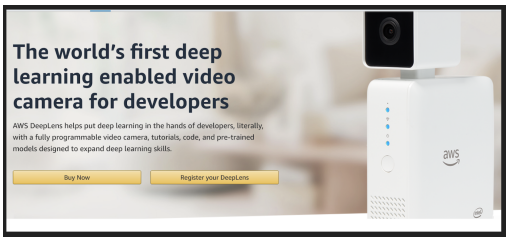
They are also used very intensively in vehicle autopilot systems.[29]



The visual search on images in snapchat allows you to make purchases of the items that appear in those images.[30]



Amazon already sells specially created cameras to develop, implement and test with these types of technologies. [31]



Perhaps the most controversial aspect is the social scoring program being developed by the Chinese government. "China's social credit system was launched in 2014 and is expected to be nationwide by 2020. In addition to tracking and

rating people, it also covers businesses and government officials." These systems are based on the video images that can be obtained from the video surveillance cameras and it is a form of massive surveillance that uses big data analysis technology. [32]

The score can be based on locations, friends, health records, insurance, private messages, financial position, hours of play, statistics of smart homes, preferred newspapers, shopping history, behaviour in appointments with other people, etc.

Some types of punishments for people with low social scoring include: prohibition of flights, exclusion of private schools, slow connection to the Internet, exclusion of highly prestigious jobs, exclusion of hotels and registration in a public blacklist. [33]

11 Defence against the dark arts

Wikipedia article says:

"Civil rights right organizations and privacy campaigners such as the Electronic Frontier Foundation, Big Brother Watch and the ACLU[71] express concern that privacy is being compromised by the use of surveillance technologies. Some fear that it could lead to a "total surveillance society," with the government and other authorities having the ability to know the whereabouts and activities of all citizens around the clock. This knowledge has been, is being, and could continue to be deployed to prevent the lawful exercise of rights of citizens to criticize those in office, specific government policies or corporate practices. Many centralized power structures with such surveillance capabilities have abused their privileged access to maintain control of the political and economic apparatus, and to curtail populist reforms.

Face recognition can be used not just to identify an individual, but also to unearth other personal data associated with an individual such as other photos featuring the individual, blog posts, social networking profiles, Internet behavior, travel patterns, etc. all through facial features alone. Concerns have been raised over who would have access to the knowledge of one's whereabouts and people with them at any given time. Moreover, individuals have limited ability to avoid or thwart face recognition tracking unless they hide their faces. This fundamentally changes the dynamic of day-to-day privacy by enabling any marketer, government agency, or random stranger to secretly collect the identities and associated personal information of any individual captured by the face recognition system. Consumers may not understand or be aware of what their data is being used for, which denies them the ability to consent to how their personal information gets shared." [34]

Let's look at some defences against these "dark arts"

Facial make-up, which, although prevents facial recognition in images, attracts a lot of attention from people

Glasses that produce reflections capable of blinding surveillance cameras

Caps with light emission that can blind the cameras

12 Conclusion

Throughout this document and the presentation that was made in "BlackHat Europe 2018" we have tried to give a general vision of the techniques and technologies that exist today to make prototypes of facial recognition. Along with the "POC" code that we have provided in our github repository you can start to try and experiment with this type of technology. The social impact of this type of technology is still to be discovered but, without doubt, they open a great moral and political debate that will be discussed in the coming years

References

[1] United States Marine Corps Comments on Joint Open Source Task Force Report and Recommendations *Working Group Draft Dated 6 January 1992*, and portions, including the definition, were subsequently OSS NOTICES, Volume 2 Issue 9, 30 November 1994

[2] Wikipedia contributors. (2018, September 23). Open-source intelligence. In Wikipedia, The Free Encyclopedia. Retrieved 18:07, October 19, 2018, from https://en.wikipedia.org/w/index.php?title=Open-source_intelligence&oldid=860809738

[3] Wikipedia contributors. (2018, October 3). Intelligence assessment. In Wikipedia, The Free Encyclopedia. Retrieved 18:07, October 19, 2018, from https://en.wikipedia.org/w/index.php?title=Intelligence_assessment&oldid=862306387

[4] https://www.cia.gov/news-information/featured-story-archive/2010-featured-story-archive/open-source-intelligence.html

[5] http://vbrandsports.com/

[6] https://vimeo.com/223217042

[7] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html

[8] https://www.nist.gov/programs-projects/face-video-evaluation-five

[9] https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8173.pdf

[10] https://foreignpolicy.com/2018/09/25/google-maps-is-a-better-spy-than-james-bond/amp/ https://www.youtube.com/watch?v=mPrxMn655lg

[11] https://patents.google.com/patent/US4567610

[12] http://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf

[13] https://en.wikipedia.org/wiki/Support_vector_machine

[14] https://en.wikipedia.org/wiki/Support_vector_machine #/media/File:Svm_separating_hyperplanes_(SVG).svg User:ZackWeinberg, based on PNG version by User:Cyc - This file was derived from: Svm separating hyperplanes.png CC BY-SA 3.0

[15] http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

[16] https://www.cv-foundation.org/openaccess/content_cvpr_2015/app/1A_089.pdf

[17] https://opencv.org/

[18] http://storm.apache.org/

[19] http://dlib.net/

[20] https://github.com/ageitgey/face_recognition

[21] Wikipedia contributors. 2018, June 29. Three-dimensional face recognition. In Wikipedia, The Free Encyclopedia. From https://en.wikipedia.org/w/index.php?title=Three-dimensional_face_recognition&oldid=848090563

[22] https://www.apple.com/ca/business-docs/FaceID_Security_Guide.pdf

[23] https://software.intel.com/en-us/realsense/d400

[24] https://software.intel.com/en-us/articles/expanding-the-possibilities-of-computer-vision-with-ai

[25] Wikipedia contributors. (2018, August 29). Perceptual hashing. In Wikipedia, The Free Encyclopedia. Retrieved 18:03, November 10, 2018, from https://en.wikipedia.org/w/index.php?title=Perceptual_hashing&oldid=857146530

[26] Wikipedia contributors. (2018, November 10). K-means clustering. In Wikipedia, The Free Encyclopedia. Retrieved 18:17, November 10, 2018, from https://en.wikipedia.org/w/index.php?title=K-means_clustering&oldid=868197050

[27] Wikipedia contributors. (2018, November 6). Chinese whispers. In Wikipedia, The Free Encyclopedia. Retrieved 18:24, November 10, 2018, from https://en.wikipedia.org/w/index.php?title=Chinese_whispers&oldid=867505208

[28] http://nymag.com/intelligencer/2018/10/retailers-are-using-facial-recognition-technology-too.html

[29] https://electrek.co/2018/10/15/tesla-new-autopilot-neural-net-v9/

[30] https://www.snap.com/en-US/news/post/introducing-visual-search/

[31] https://aws.amazon.com/deeplens/

[32] https://www.wired.co.uk/article/china-social-credit

[33] Wikipedia contributors. (2018, November 9). Social Credit System. In Wikipedia, The Free Encyclopedia. Retrieved 19:17, November 10, 2018, from https://en.wikipedia.org/w/index.php?title=Social_Credit_System&oldid=867975188

[34] Wikipedia contributors. (2018, October 31). Facial recognition system. In Wikipedia, The Free Encyclopedia. Retrieved 19:37, November 10, 2018, from https://en.wikipedia.org/w/index.php?title=Facial_recognition_system&oldid=866635017

[35] THE EGLYPH WEB CRAWLER: ISIS CONTENT ON YOUTUBE, Counter Extremism Project, from https://www.counterextremism.com/sites/default/files/eGLYPH_web_crawler_white_paper_July_2018.p

[36] Violent or graphic content policies, YouTube Help from https://support.google.com/youtube/answer/2802008?hl=en