Atas das Nonas Jornadas de Informática da Universidade de Évora

Évora, 27 de Fevereiro de 2019

UNIVERSIDADE DE ÉVORA

Author/Paper Index

Organizing Committee

- Teresa Gonçalves
- Vitor Beires Nogueira

Survey Article: Wild Forest Fire Detection and Prediction

Kashyap Raiyani

University of Évora, Portugal
kshyp@uevora.pt

Abstract. A wildfire or wildland fire is a fire in an area of combustible vegetation that occurs in rural areas. Depending on the type of vegetation where it occurs, a wildfire can also be classified more specifically as a brush fire, bushfire, desert fire, forest fire, grass fire, hill fire, peat fire, vegetation fire, and veld fire. There are several incidents of forest fire and it consumes a greater number of peoples lives every passing year. In such recent event of 2017 in Portugal, wildfires Death toll rises to 43. Due to its high social impact, many researchers with the different aspect of the problem are looking in detection and contamination of the forest fire. This survey paper shows the related work done so far in the area of forest fire detection using sensor and images. Apart from that it also covers the scope of early forest fire prediction using satellite images. The article also discusses the use of machine learning it the context of forest fire detection and prediction.

Keywords: Wildfire · Satellite Images.

1 Introduction

Fire has played a significant role in promoting the progress of human civilization. However, without proper management, it is also one of the major disasters causing huge loss of human lives and property all over the world. Therefore, it is essential to propose a reliable and effective algorithm to detect and raise the alarm for fire as soon as possible.

Traditional detection algorithms usually use heat sensors, thermocouples, or ionization sensors to detect fire, through detecting temperature and smoke particles [12]. In the past few years, Siemens launched a FDT221 fire detector[2]. It is equipped with two redundant heat sensors, which monitor rooms in which a temperature rise is expected in case of fire. Mircom also launched the MIX-200 series of intelligent sensors[3] for residential applications, equipped with photoelectric smoke detectors and electronic thermistors. Although these sensors may

The paper is submitted under the Introduction to Scientific Research course.

[2] https://www.buildingtechnologies.siemens.com/bt/global/en/products/fire-safety-(en)/fire-detection/sinteso/pages/fdt221.aspx

[3] http://www.mircom.com/media/datasheets/CAT-5904MIX-200SeriesIntelligentLowProfileSensors.pdf

work efficiently in some particular cases, they suffer from large propagation delays of smoke and temperature, resulting in the increase of fire detection latency. Other algorithms, based on beam or aspirated smoke detectors, have been used to attempt to reduce the detection latency. Still, they cannot solve the practical problem completely [14]. Moreover, all of the abovementioned sensors require being in close proximity to the fire. Optical sensors have the advantages of long detection distance and fast response. However, as point detectors, their detection area is limited. These traditional algorithms do not detect the fire itself directly, therefore they are not always reliable.

Fire image detection is a relative novel technology based on video cameras, detecting the fire through the intelligent analysis of images with advanced algorithms. Compared with traditional sensors, the advantages of video cameras are listed as follows. First of all, video cameras can avoid detection latency to a great extent. They can also monitor a larger area as volume detectors, and are adaptable to outdoor locations, where traditional sensors are difficult to place. Finally, with the increasing concern of security, more and more surveillance cameras have been installed for various security applications, providing a convenient way to embed fire image detection into existing systems.

In most cases, it is difficult to contain a forest fire beyond few minutes following ignition, and rapid detection is therefore critical. To assist human surveillance, infrared technology has been proposed to detect forest fire with thermal infrared cameras [13]. Until now, these methods do not yield good results for the main reason that the fire itself is often hidden by the trees at the start of its ignition, and the smoke plumes are too quickly cooled to be detected by infrared. More recently, a semi-automatic fire detection system uses infrared satellite images from the Very High Resolution Radiometer (AVHRR) [6, 10, 7]. Nevertheless, the satellite permits a detection service at a continental scale, and only at the moments when it passes over the same region.

1.1 Early Detection and Suppression

Early detection and suppression of forest fires are crucial to minimizing the destruction that the fires may cause due to their rapid convection propagation and long combustion cycle [18]. Massive efforts have been put into monitoring, detecting, and rapidly extinguishing forest fires before they become too large. Traditional forest fire monitoring and detection methods employ either mechanical devices or humans to monitor the surroundings, but these methods can be both dangerous and costly in terms of the required human resources [15].

Remote sensing has become one of the most frequently utilized tools for effective forest survey and management [5, 1]. Rapid advances in electronics, computer science, and digital camera technologies have allowed computer vision based remote sensing systems to provide a promising substitute for conventional forest fire monitoring and detection systems [15]. Current remote sensing approaches to forest fire monitoring and detection can be grouped into three categories: ground-based systems, manned aerial vehicle based systems, and satellite-based systems [4]. However, each of these systems presents different technological and practical problems. Ground measurement equipment may suffer from limited surveillance ranges. Satellite systems are less flexible in their path planning and technology updates, and their temporal and spatial resolution may be too low for detailed data capture and operational forest fire fighting [11]. Manned aerial vehicles are typically large and expensive. Moreover, the life of the pilot can be potentially threatened by hazardous environments and operator fatigue [2].

Unmanned aerial vehicles (UAVs) with computer vision based remote sensing systems are an increasingly realistic option, providing rapid, mobile, and low-cost alternatives for monitoring, detecting, and even fighting forest fires. The integration of UAVs with remote sensing techniques are also able to meet the critical spatial, spectral, and temporal resolution requirements, offering the potential to serve as a powerful supplement to existing methods [11]. In addition, UAVs allow the execution of long-term, monotonous, and repeated tasks beyond human capabilities. This has led to increased worldwide attention to UAV forest fire applications in recent years.

The next section will talk about the related work done in the categories of ground-based systems and satellite-based systems. Talking about sensor data and image processing.

2 Literature Review

The line of sight and the early stage of the fire process problem could be solved with the second type of sensors. A new technology called wireless sensor network (WSN) is nowadays receiving more attention and has started to be applied in forest fire detection. The wireless nodes integrate on the same printed circuit board, the sensors, the data processing, and the wireless transceiver and they all consume power from the same source batteries. Unlike cell phones, WSN do not have the capability of periodic recharging. The sensors are devices capable of sensing their environment and computing data. The sensors sense physical parameters such as the temperature, pressure and humidity, as well as chemical parameters such as carbon monoxide, carbon dioxide, and nitrogen dioxide. The sensors operate in a self-healing and self-organising wireless networking environment. One type of wireless technology is ZigBee which is a new industrial standard based on IEEE 802.15.4. This technology emphasises low cost battery powered application and small solar panels and is suited for low data rates and small range communications. Wireless sensor networks have seen rapid developments in a large number of applications. This kind of technology has the potential to be applied almost everywhere; this is why the research interest in sensor networks is becoming bigger and bigger every year.

The researchers defined more than 27 mathematical models to describe the fire behaviour where they stated that those models developed according to different countries experience of forest fire and each model is different according to the input parameters and the environments nature (fuel indexing). The researchers of forest fires manage to use some of these models in simulations or even create their own methods to create maps that can be used to analyse the fire behaviour at any time in the future so that they can help the fire fighters to determine the best method to extinguish the fire, such as BehavePlus, FlamMap, FARSITE, Geodatabase, and ArcSDE. On the contrary, researchers are trying to initiate a reliable technology that can detect the fire, localise the fire, and help in decision making in terms of requiring an immediate reaction in case of crisis possibility or a high fire risk situation. As a result, the fire can be extinguished in early stages within a short time to minimise the damage save lives, environment, fire fighter equipment, time and effort.

2.1 Related Background

The FIRESENSE (Fire Detection and Management through a Multi-sensor Network for the Protection of Cultural Heritage Areas from the Risk of Fire and Extreme Weather Conditions, FP7-ENV-2009-1-244088-FIRESENSE) is a Research Project of the European Unions 7th Framework Programme Environment (including climate change). The project aims to implement an automatic early warning system to remotely monitor areas of archaeological and cultural interest from the risk of fire and extreme weather conditions. The system consists of multi-sensors, optical, IR, and PTZ cameras in addition to temperature sensors, and weather stations. In this system, each sensor collects the data and applies pre-processing techniques and different models of data fusion algorithms in order to provide a clear understanding for the event to the local authority. The demonstrator deployments will be operated in selected sites in Greece, Turkey, Tunisia, and Italy. The project keeps track of (i)Scene model: the fire and smoke, heat flux or emitted thermal (Planck's radiation formula), the fire flickering, the reflectance, absorption emission lines, and analysis of the atoms (e.g., potassium) and the molecules(water and carbon dioxide) are characteristics to be investigated. (ii)The background emits the thermal heat, the reflectance of sunlight, the clouds (clouds shadow) the buildings and the sky polarisation. (iii)The atmosphere has a number of gases (N2 , O2 , CO, CO2 , H2O, etc.); each one has its own absorption and reflection behaviour. Water vapour concentration could vary as a result. Carbon dioxide is more uniformly distributed but its value is larger over industrial cities and vegetation fields than over oceans and deserts.

The paper [3] describes a scheme for automatic forest surveillance. A complete system for forest fire detection is firstly presented although they focus on infrared image processing. The proposed scheme based on infrared image processing performs early detection of any fire threat. With the aim of determining the presence or absence of fire, the proposed algorithms performs the fusion of different detectors which exploit different expected characteristics of a real fire, like persistence and increase. Theoretical results and practical simulations are presented to corroborate the control of the system related with probability of false alarm (PFA). Probability of detection (PD) dependence on signal to noise ration (SNR) is also evaluated.

In the light of the problem of monitoring forest fire, the design strategy and practical implementation of establishing the fire monitoring system based on digital image information are proposed [16]. The system is based of the continuous image sampling provided by CCD camera. Through this, one can obtain the configuration characteristics, dynamic characteristics and color information of interesting region with an application of the digital image processing algorithm, and then to identify the fire source according to the acquired characteristics. The experimental results show that the system can accurately identify and confirm the fire. Besides, the amount of data processed can be reduced because of the use of sampling algorithm thus shortening the execution time.

In this paper [17], an unmanned aerial vehicle (UAV) based forest fire detection and tracking method is proposed. Firstly, a brief illustration of UAV-based forest fire detection and tracking system is presented. Then, a set of forest fire detection and tracking algorithms are developed including median filtering, color space conversion, Otsuthreshold segmentation, morphological operations, and blob counter. The basic idea of the proposed method is to adopt the channel a in Lab color model to extract fire-pixels by making use of chromatic features of fire. Numerous experimental validations are carried out, and the experimental results show that the proposed methodology can effectively extract the fire pixels and track the fire zone.

This paper [9] presents an investigation of an early forest fire detection system on the basis of indoor (performed in the fire lab of the University of Duisburg-Essen) and outdoor tests. A commercial highly sensitive aspirating smoke detector, two gas sensors (H2 and CXHX), a microwave radiometer and the detection algorithms are described. Here, the main focus is early smoke detection because of the large and high-intensity forest fires are widely uncontrollable and cause very high risks. Apart from that, it also helps to reduce false alarms of video-based systems, especially in hardly accessible terrain, a remote controlled UAV can fly to the place where a fire is assumed to confirm that the origin of the smoke is most likely a fire.

In this paper [8], they have proposed a fast and practical real-time image-based fire flame detection method based on color analysis. firstly they build a fire flame color feature model based on the HSI color space by analyzing 70 training flame images. Then, based on the above fire flame color features model, regions with fire-like colors are roughly separated from each frame of the test videos. Besides segmenting fire flame regions, background objects with similar fire colors or caused by color shift resulted from the reflection of fire flames are also extracted from the image during the above color separation process. To remove these spurious fire-like regions, the image difference method and the invented color masking technique were applied. Finally, the fire flame burning degree is estimated so that users could be informed with a proper fire warning alarm.

2.2 Flame Detection

In the scientific literature there are a lot of methods and approaches for the fire and/or smoke determination which are based on image segmentation procedures. Refering to some surveys on this topic, method for the detection of fire and smoke proposed is based on the usage of color spaces RGB and YCbCr. For the fire area pixel consistent pattern Y > CRr > Cb is discovered. For the smoke detection and the feasibility of the system, this two inequalities |R-G| < Th and |R-B| < Th are checked. Mostly, time derivative of luminance component Y is used to declare the candidate fire pixels, then depending on chrominance components U and V, the candidate pixels are classified into fire and non-fire sections. As mentioned above, areas of three types are analyzed: containing fire, containing smoke and areas without smoke and/or fire.

The next section will talk about the proposed methodology that should be considered for early forest fire prediction.

3 Conclusion and Proposed Method

Forest fires represent a constant threat to ecological systems, infrastructure and human lives. Past has witnessed multiple instances of forest and wild land fires. Traditional fire protection methods use mechanical devices or humans to monitor the surroundings. The most frequently used fire detection techniques are usually based on particle sampling, temperature sampling, and air transparency testing. An alarm is not raised unless the particles reach the sensors and activate them. So we are going to capture the images through satellite and will give the captured image as a input to the time series system. This system will give the output as whether the fire is present or not. Further, it will detect important features like, object detection, different type of vegetation, gradient, magnitude and angle. The Segmentation of image would also be done. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze.

This time series system will analyze the all the images and build the story line of particular region. This will help in seeking the change of Climate, Vegetation, Civilization and, Eco System resulting into wildfire. This way it would be huge leap in area of prediction of forest fire.

References

1. R. A. Chisholm, J. Cui, S. Lum, and B. Chen. Uav lidar for below-canopy forest surveys. *Journal of Unmanned Vehicle Systems*, 01:61–68, 12 2013.
2. B. Abdalhaq, A. Cortés, T. Margalef, and E. Luque. Enhancing wildland fire prediction on cluster systems applying evolutionary optimization techniques. *Future Generation Comp. Syst.*, 21:61–67, 01 2005.
3. I. Bosch, S. Gomez, L. Vergara, and J. Moragues. Infrared image processing and its application to forest fire surveillance. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 283–288, Sep. 2007.

4. J. R. M. de Dios, L. Merino, and A. Ollero. Fire detection using autonomous aerial vehicles with infrared and visual cameras. *IFAC Proceedings Volumes*, 38(1):660 – 665, 2005. 16th IFAC World Congress.
5. D. G. Leckie. Advances in remote sensing technologies for forest survey and management. *Canadian Journal of Forest Research-revue Canadienne De Recherche Forestiere - CAN J FOREST RES*, 20:464–483, 04 1990.
6. L. Giglio, J. D. Kendall, and C. O. Justice. Evaluation of global fire detection algorithms using simulated avhrr infrared data. *International Journal of Remote Sensing*, 20(10):1947–1985, 1999.
7. T. Hame and Y. Rauste. Multitemporal satellite data in forest mapping and fire monitoring. *EARSeL Advances in Remote Sensing*, 4:93–101, 01 1995.
8. W. Horng and J. Peng. A fast image-based fire flame detection method using color analysis. *Tamkang Journal of Science and Engineering*, 11:273–285, 09 2008.
9. W. Krull, R. Tobera, I. Willms, H. Essen, and N. von Wahl. Early forest fire detection and verification using optical smoke, gas and microwave sensors. *Procedia Engineering*, 45:584–594, 12 2012.
10. T. Lillesand and R. Kiefer. *Remote Sensing and Image Interpretation.* John Wiley Sons, 01 2000.
11. H. Olsson, M. Egberth, J. Engberg, J. E S Fransson, T. Granqvist Pahlén, O. Hagner, J. Holmgren, S. Joyce, M. Magnusson, B. Nilsson, M. Nilsson, K. Olofsson, H. Reese, and J. Wallerman. Current and emerging operational uses of remote sensing in swedish forestry. *Proceedings of the Seventh Annual Forest Inventory and Analysis Symposium*, 01 2005.
12. S. Verstockt, P. Lambert, R. Van de Walle, B. Merci, and B. Sette. State of the art in vision-based fire and smoke dectection. In H. Luck and I. Willms, editors, *International Conference on Automatic Fire Detection, 14th, Proceedings*, volume 2, pages 285–292. University of Duisburg-Essen. Department of Communication Systems, 2009.
13. J. Vicente and P. Guillemant. An image processing technique for automatically detecting forest fire. *International Journal of Thermal Sciences*, 41(12):1113 – 1120, 2002.
14. A. Vincitore, H. Wang, A. Finn, and O. Erdinc. Spatial-temporal structural and dynamics features for video fire detection. In *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV)*, WACV '13, pages 513–519, Washington, DC, USA, 2013. IEEE Computer Society.
15. V. Vipin. Image processing based forest fire detection. *International Journal of Emerging Technology and Advanced Engineering*, 2:87–95, 01 2012.
16. J. Xiao, J. Li, and J. Zhang. The identification of forest fire based on digital image processing. In *2009 2nd International Congress on Image and Signal Processing*, pages 1–5, Oct 2009.
17. C. Yuan, Z. Liu, and Y. Zhang. Uav-based forest fire detection and tracking using image processing techniques. pages 639–643, 07 2015.
18. C. Yuan, Y. Zhang, and Z. Liu. A survey on technologies for automatic forest fire monitoring, detection and fighting using uavs and remote sensing techniques. *Canadian Journal of Forest Research*, 45:150312143318009, 03 2015.

Biometric Facial Recognition[*]

Margarida Ventura

Universidade de Evora, Departamento de Informatica,
Colegio Luis Antonio Verney, Evora, Portugal
d43427@alunos.uevora.pt

Abstract. Facial recognition is a biometrics technique, based on human facial traits, that is being used more frequently in our society, especially in the areas of security and research. Over the years, several methods of recognition, based on biometric information, have been developed. As such, several facial recognition algorithms that use Principal Component Analysis (PCA), Artificial Neural Networks (ANN) or Support Vector Machines (SVM) have been proposed. Even though facial recognition is simple for most people, the computational cost of implementing such a task on a machine might be tremendous. The struggle resides in the computation of a model to extract the characteristics that can different between each face, especially because these have features that are similar among each other, like a nose, a mouth or two eyes, and very few considerable differences. The process of facial recognition as numerous problems that can prevent a correct recognition and lower the recognition accuracy, especially because of the non-rigid structure of the human face. The most significant problems to be considered are the variations in angle and illumination of the face, the different facial expressions as well as the aging process. The new EU General Data Protection Regulation (GDPR) has recognized the importance of protecting personal data, such as facial images, introducing several requirements organizations must comply with to keep up with the legislation.

Keywords: Biometrics · Facial Recognition · Facial Recognition Methods · GDPR.

1 Introduction

Human beings possess the natural ability of quickly identifying any individual, and easily recognize familiar faces even after a long period of time [1], [2], [3]. Facial recognition is an essential and important skill of the perception system that is hardly affected even after changes like aging, various expressions, the use of glasses, and conditions such as beards or different hair styles [2]. Building an intelligent system like the human perception system is an active area of research [1], [2]. Even though facial recognition is a simple task for human beings, since they are capable of detecting facial features and other components of an image

[*] This paper is submitted for the Introduction to Scientific Research course.

instantly, its not trivial to implement this process on a computer [2], [3]. The great difficulty is in having a model that isolates the characteristics that differentiate a specific face from other faces, since, although different, they present few substantial differences between them, for all faces have similar characteristics like a mouth, two eyes and a nose [3].

The advent of computer systems and their capacity to store large amounts of information has led to the emergence of biometric recognition systems, that use physiological attributes to measure and analyze the unique characteristics of a person in order to distinguish between different people [4], [5]. These systems can be built through various techniques, like fingerprints, palm prints, voice recognition and iris methods, that require an individuals participation or involvement to access the system [1], [6], [5]. New systems, for human identification scenarios, that provide participants access without their direct intervention or physical contact are being created [1] [4] [2]. Among such biometric systems, facial recognition is a popular and more naturally accessible technique than many other biometrics, because individuals can be efficiently captured and monitored, at a distance, through these systems [1] [4], [7], [8]. By utilizing biometrics, a person can be recognized considering ”who she/he is” as opposed to ”what she/he has” (e.g. ID card) or ”what she/he knows” (e.g. secret key, PIN), which can be particularly important in defeating the problems of credential-based authentication, like IDs being stolen and faked or passwords being forgotten or cracked, likely resulting in identity theft [6], [9]. Since biometric information is difficult to forge or spoof, it is widely considered to be more secure and advantageous than traditional credential-based authentication mechanisms [9].

The field of biometrics, in specific facial recognition, has been, for a long time, an important and challenging research area that has attracted the interest of many researchers, and a reliable option for recognition because of its wide range of sophisticated techniques that help secure organizations, information, assets and people [4], [6], [2], [5], [10], [11], [8]. Nevertheless, despite significant recent advances in the field of facial recognition, implementing facial recognition efficiently presents serious challenges to current approaches [12]. As such, efforts are still being made to develop more user-friendly systems, that improve the requirements of security systems and yield more accurate results in protecting assets and ensuring privacy [2], [9].

Facial recognition is a necessity of the modern age, because the need for the identification of individuals has increased with the world globalization, especially since people move between countries and continents in a nearly unrestricted way [11]. As such, researchers have enhanced numerous algorithms and methodologies in order to recognize a face in an effective and efficient manner [1], [11]. For this purpose, they have focused on the detection and recognition of traits and features for individuals, such as the nose, eyes, mouth, face shape, face position and size, and the relationship among these traits and features. Furthermore, ongoing research in facial recognition as tried to develop systems that could work well in multitude of real-world applications [1].

The remainder of this paper is as follows. Section 2 describes the main steps of the facial recognition process. Section 3 presents several advantages of using facial recognition, while Section 4 explains some of the limitations of this process. In Section 5, different methods used in the facial recognition process are discussed, and in Section 6 some of the applications for facial recognition systems are presented. Section 7 pertains to the laws which bound local governments, agencies and organizations in the use of facial recognition systems. Finally, the main conclusions of this paper are drawn in Section 8.

2 The Facial Recognition Process

Facial recognition is a difficult task that has been an interesting field of research, for many decades, attracting significant attention from many scholars, in the fields of biometrics, computer vision, image processing and analysis, pattern recognition, and network and multimedia information access [1], [5], [10], [13], [14], [15], [16], [17]. It is a very useful, but complicated and challenging process, through which faces can be detected, the biometric facial features can be extracted and then used as distinguishable evidence in the automatic identification of a specific individual, against a given set of database face images, by using computational filters, methods and algorithms [6], [5], [10], [11], [8].

Over the past decade, the facial recognition systems, that have appeared, have become increasingly useful, faster and accurate tools, capable of identifying a person from a digital image or a video frame [13], [14]. These systems, automatically detect faces present in images and videos, while ignoring the background, recognize specific facial features through their inherent traits, and relate a query face image against all the face images in a database, to determine the identity of the query face image [1], [10], [14].

The different methods employed for facial recognition can use the entire face as input data for the recognition system, may not consider the whole face, but only some features or areas of the face, or could combine the previous methods simultaneously [5]. However, all of the facial features, that are extracted from the face, must represent an algorithmic connection of distances and sizes, such as the exact distance between nose and ears, skull size, among other details [7].

The facial recognition problem can be categorized into three main steps: facial detection, feature extraction and facial matching [1], [6], [2], [5], [11], [8]. In the first step, an attempt is made to discover faces in an input image [1], [5]. If a face is detected then the specific features of that face are captured and record in the form of mathematical templates [1], [5], [10]. Before performing the facial matching step, the face is normalized as to line-up the eyes and mouth [1]. Finally, the face is compared with several other database face images, to find the identity of the face among several possibilities, thus returning a possible output match [1], [5], [10].

3 Facial Recognition Advantages

Facial recognition is a process with several advantages, against other biometrics, that range from the possibility of mass scanning at greater distance, easy access and use, non-intrusive and hands-free requirements, user friendliness and the fact that it is inexpensive [4], [6], [5], [10], [8], [9], [14].

Facial recognition is one of the fastest, most reliable and consistent biometric methods [5], [15]. It is an easy to install real-time application, where the user must go through only once, making it less invasive [5], [15], [18]. It generally requires little resources to capture the face, a simple webcam will suffice, a minimum amount of training to get the system operational and there is no need for expensive password administrators [15], [18].

In organizations, this type o biometric system allows not only employees to verify their presence, thus deterring fraud, but since any visitor can be added to the system, it does not provide access to individuals not included in such system [5], [15]. In this way, it increases the security level of organizations, making it possible for any person, present in the system, to be identified and tracked, or rejected in a matter of seconds [15].

These systems work automatically, without being controlled by a person, and as a result organizations wont need to worry about having someone there to monitor the systems [5], [15].

Biometric facial systems are easy to integrate and use in organizations, since they only require the installation of capturing equipment, like a camera, and they may work with existing software [5], [15].

Facial biometrics technology today has achieved high recognition rates, especially with the emergence of three-dimensional technology, which makes it very difficult to deceive, and as such, users can feel secure and confident with these systems [5], [15].

Its considered to be a convenient security solution because users dont have to remember passwords, or carry extra badges, documents, or ID cards [15]. This is especially important, since people tend to forget passwords, codes or PINs, and ID cards or keys might be damaged, lost, robbed and duplicated, which can be a big problem in traditional security methods [1], [2], [15].

4 Facial Recognition Problems

Despite all of the previously mentioned advantages of facial recognition and the fact that there are many approaches to carry out this task, in images and videos, none can accomplish it with 100 percent accuracy, because there are numerous limitations that could prevent recognition, consequently leading to a strong decrease in the recognition efficiency, especially since algorithms dealing with complex environments can be computationally expensive [1], [4], [2], [5], [8].

The face is not a rigid object, as a person becomes older significant alterations in the facial appearance (e.g. wrinkles) and face shape of an individual occur [2], [5], [8], [13]. Indeed, the shape of the skull and the skin texture change from childhood to adolescence, which represents a problem in facial recognition, because the images used in passports and identity cards are not frequently updated [5]. As such, aging is an inevitable natural process, during the lifetime of a person, that influences facial recognition techniques, as well as the performance and accuracy of such systems [1], [2], [8].

The presence of natural or artificial obstacles (e.g. glasses, scarves, hats, masks, hands, nose rings, facial-hair or hair style) blocking a face in an image [1], [2], [5], [8], [13], [14], [15], also known as occlusion, can be a problem for recognition systems. These objects can severely affect the accuracy of the facial detection process and the performance of such systems [1], [4], [2], [8], [15].

The rotation of the face, in different images, is another challenge in achieving a successful facial recognition system, since pose variation can severely degrade the performance of such systems, by reducing the recognition accuracy [1], [2], [5], [11], [8], [15]. In fact, people pose differently every time they take a picture, causing a rotation of the face, and if the facial poses are not properly aligned in the camera view, the recognition process will suffer [1], [4], [2], [5].

Variation in lightening conditions (e.g. background light, brightness, contrast, shadows, dim lights) and inappropriate illumination are factors that can greatly challenge the accuracy of facial recognition, in images or videos, especially because the appearance of the face changes with variations in illumination [1], [4], [2], [5], [11], [8], [14], [15], [19].

Different facial expressions modify the geometry of the face, due to contractions of facial muscles and, therefore, challenge the accuracy of the facial recognition process [4], [2], [5], [11], [8], [13].

There are several specific characteristics of the input images that can prove problematic in the facial recognition process. For instance, images with different dimensions make the facial recognition process difficult, especially in the phases of facial features extraction and matching, because the larger the image dimensions, the larger the vector matrix and, therefore, the greater the computational cost and the smaller the recognition accuracy [3]. Also, blurred images, usually caused by peoples movement, can impede the correct recognition process [5], [8].

Another of the challenges faced by recognition systems has to do with the fact that they can be easily fooled by a picture of a persons face, which can be available on the Internet (e.g. social networks) [5], [14].

There are numerous distracting effects that an image, taken in an uncontrolled environment, can have and that may cause problems in the recognition process. For example, an image of the face can have a complex background, causing the recognition system to mistake some areas of the background as a face [2], [8]. There is also the fact that the use of makeup can constitute a distraction in the recognition process, especially because a person can have different visuals in uncontrolled situations [13], [20].

5 Facial Recognition Methods

Within the last several years, researchers have suggested numerous algorithms, techniques and methodologies for facial recognition, but so far no method has yielded satisfactory results under all unrestricted conditions [2], [5], [14]. In this section, some of the many methods that are used in the facial recognition process are discussed.

5.1 Principle Component Analysis (PCA)

PCA is a very popular approach that can be used for facial recognition, as long as the input image and the database images are the same size, and are normalized to line up the eyes and mouth of the subjects [1], [2], [5], [11], [14]. Its fundamental idea is the following: given an input face image, it essentially aims at retrieving unique features (eigenfaces) that can describe the face [1], [2], [5]. These eigenfaces are sets of orthogonal vectors (eigenvectors) of the covariance matrix, that can describe the image through their linear combination [2], [5], [11]. The eigenvectors are computed by measuring the distance between key features like the nose tip, the mouth and the eye corners and chin edges [2]. To construct the covariance matrix, the face image is transformed into a vector, where each element of that vector corresponds to the pixel intensity [5]. Finally, in order to find a match, the input image is compared against a set of database images, by calculating the distance between their respective eigenvectors [1], [2], [7], [14].

The main advantage of the PCA method is that it decreases the data needed to identify a specific match, because it reduces the dimensionality of the face and only the important parts for facial recognition are left [2], [14].

5.2 Artificial Neural Networks (ANN)

Among the several techniques for facial recognition, one of the most widely used is ANN [1], [18]. In simple terms, they consists of networks of many simple processing units (neurons), inspired by the human nervous system, which includes neurons and their synaptic transmissions, as well as the properties of plasticity and adaptability [1], [18].

These networks are processors capable of learning from previous experiences (training) and then using the knowledge they gained in new situations, of the same scope of their learning [1], [2], [18]. This way, ANN can be trained to recognize faces [2].

The accuracy of the facial recognition process has been increased with the help of this method, especially because they reduce the complexity of the recognition process [1], [2]. However, they present a main disadvantage when it comes to the large amount of time required for their training [1], [2].

5.3 Support Vector Machines (SVM)

SVM are a supervised learning approach that can be used in the facial matching step (or classification step) after the facial features extraction phase [1], [2], [5]. This technique takes out the discriminatory information, from the training data, and when given a set of points that belong to two different data sets, SVM find the hyperplane that separates the maximum possible distance between them [2], [5], [14].

SVM have the advantage of offering fast computational speed along with an effective high performance rate, reducing the risk of misclassification not only for the learning set, but also for the test set [1], [5]. Nevertheless, SVM cannot be applied directly when some of the features are occluded, because the values for those dimensions are unknown and represent missing entries in the feature vectors [14].

5.4 Gabor Wavelets

The gabor wavelets technique, also known as gabor filters, detects local properties for position estimation, in order to match faces in the facial recognition process [1], [2]. They have the capacity to extract the properties of spatial relations, spatial localization, spatial frequency structure and orientation information of a face image, [1], [2], [5]. In order to model the relationship between these properties, a topological graph is built for each face [5]. Gabor wavelets also work well over the extraction of edge and shape information [1], [2].

The main advantages of gabor wavelets are facial feature reduction, its global feature representation in facial recognition, the ability to represent faces in a compact way, as well as the fact that it is a fast recognition method that requires a small training set [1], [2].

5.5 Hidden Markov Models (HMM)

HMM are a powerful statistical modelling technique that have proven to be efficient, in facial recognition systems, since their invention [1], [2], [5]. It is a model composed of states and transitions, where significant facial regions, in a face image, are placed in a natural order, from top to bottom [5]. These facial regions can be grouped into five facial features (e.g. mouth, eyes, nose, chin, forehead), or seven facial regions (e.g. hair, forehead, eyebrows, eyes, nose, mouth and chin) [1], [2], [5]. For each of these regions, a state from left to right is affected [5]. However, the number of states can be increased or decreased depending upon the systems requirements [1], [2].

5.6 Linear Discriminant Analysis (LDA)

LDA, also known as fisherfaces, is an appearance-based technique, used for class-specific dimensionality reduction and feature extraction, that is commonly applied in facial recognition with good performance [5], [11], [7], [14]. This method constructs a discriminant projection subspace, to provide a small set of features that carry the most relevant information, with the purpose of distinguishing between the faces of different people [5], [14]. With this technique data is projected in such a way that each feature can be easily separable [11].

6 Facial Recognition and Its Applications

In recent years, the demand for biometric systems, that use facial recognition, has risen due to a wide range of commercial, governmental and law enforcement applications that use this type of technology, like security systems [1], [4], [2], [5], [13], video surveillance [1], [4], [2], [10], [7], [8], [14], [20], access control [1], [4], [2], [8], [14], [20], human-computer interfaces [4], [10], [14], [18], [20], identification systems [2], [20], and attendance technologies [1].

Facial recognition can be applied to control the access of people to buildings, offices, computer systems, ATM machines, airports, email authentication, among others [2]. For example, these automatic facial recognition systems could control who is using a PC or an ATM machine [1], [2]. This way, if a user left a PC, for a specific amount of time, the system could halt until the authorized user came back and was recognized, while denying any unauthorized access from other users [1], [2]. At ATM machines, instead of using ATM cards or PINs, the machine could take a picture of the user and compare it with the images in the database, to legitimate a persons access to an account [1], [2].

Facial recognition can be used for identification purposes, for instance, to search for missing people, to verify someones identity in an electoral voting system, or in banks, airports, schools, mobile devices, and even for criminal identification [1], [2], [5], [10], [18], [21].

Surveillance systems offer several benefits to different organizations, since they can be used for intelligence gathering, crime control, protecting people, crowd and people monitoring, border control, etc. [1], [4], [2]. For example, this can be achieved by using security cameras, to monitor well-known criminals [1], [2], [18], [21].

Biometric attendance technologies are among the latest solutions in facial recognition systems, that allow attendance marking from facial recognition of a specific individual, after their face is captured by the systems cameras [1].

7 Facial Recognition and Data Protection

The widespread adoption of facial recognition technology poses an increasing threat to privacy violations [1], [13]. It is often criticized by the civil society, groups or activists, because it can be used for far more than to identify individuals or to track their whereabouts [1], [13]. They also criticize the silent nature of the facial recognition technology because current applications lack the adequate mechanisms for informing users of the presence of cameras and the usage given to the collection, processing, dissemination and storing of potentially sensitive data [13]. As part of this data, face images are considered critical data and when processed by facial recognition systems serious privacy concerns might be raised, because they are particularly prone to misuse, which signals the end of the basic right of public anonymity [21]. As such, laws which bound local governments and agencies in the use of surveillance, restricting it to circumstances where public safety is compromised, have been imposed [1].

The European Union (EU) General Data Protection Regulation (GDPR), which came into force in May 2018, is applicable to all organizations inside the EU, the European Economic Area (EEA) and to all organizations from other countries, as long as they process data from European citizens, thus producing effects worldwide [21], [22], [23]. This data protection law introduced strict requirements for personal data protection, setting a new global standard for privacy rights and regulating the way organizations worldwide collect, store and process such data [21], [22]. This way, the GDPR, recognized these concerns and risks, highlighting the protection of private personal data, by considering images with identifiable people as sensitive biometric personal data [13], [21], [22], [23]. The GDPR clearly states that the processing of biometric personal data for identification purposes is prohibited [21]. Processing is only allowed if explicit and informed consent of the subject exists (if the subject is not capable of giving a valid consent, consent must be given by someone on behalf of the subject), if specific laws or regulations apply, or if the processing is required for reasons of public safety [13], [21], [22], [23]. Additionally, the consent must be limited to a specific purpose of data processing [22].

8 Conclusions

The study of facial recognition and facial recognition systems has remained an active area of research, especially since the security of people, information or assets is becoming more difficult and criminal activity keeps increasing day by day [1], [5], [8], [11]. They are particularly important, since traditional identification methods can suffer from lack of reliability [1]. On the other side, physiological characteristics and traits of an individual cannot be stolen, forgotten or misplaced [1], [2]. As such, continuous efforts are being made to develop facial recognition methods with the best accuracy and efficiency possible, especially in unconstrained environments, because of its relevance in so many areas [1], [2]. Among these facial recognition methods, some of the most relevant are Artificial Neural Networks, Support Vector Machines and Principal Component Analysis [1]. However, legitimate concerns over the privacy and misuse of sensitive data, generated by facial recognition technology, poses a significant obstacle to the widespread use of such a technology [13].

References

1. Lal, M., Kumar, K., Arain, R. H., Maitlo, A., Ruk, S. A., Shaikh, H.: Study of Face Recognition Techniques: A Survey. International Journal of Advanced Computer Science and Applications 9(6), 4249 (2018)

2. Sharif, M., Naz, F., Yasmin, M., Shahid, M. A., Rehman, A.: Face Recognition: A Survey. Journal of Engineering Science and Technology Review 10(2), 166177 (2017)
3. Diniz, F. A., Silva, T. R., Alencar, F. E. S.: Um estudo emprico de um sistema de reconhecimento facial utilizando o classificador KNN. Revista Brasileira de Computao Aplicada 8(1), 5063 (2016)
4. Zhou, S., Xiao, S.: 3D face recognition: a survey. Human-centric Computing and Information Sciences 8, (2018)
5. Chihaoui, M., Elkefi, A., Bellil, W., Amar, C. B.: A Survey of 2D Face Recognition Techniques. Computers 5(21), 4168 (2016)
6. Ranjani, R., Priya, C.: A Survey On Face Recognition Techniques: A Review. International Journal of Pure and Applied Mathematics 118(5), 253274 (2018)
7. Okabe, R. K., Carro, S. A.: Reconhecimento Facial Em Imagens Capturadas Por Cmeras Digitais De Rede. Colloquium Exactarum 7(1), 106119 (2015)
8. Ranganatha, S., Gowramma, Y. P.: Face Recognition Techniques: A Survey. International Journal for Research in Applied Science and Engineering Technology 3(4), 630635 (2015)
9. Chen, S., Pande, A., Mohapatra, P.: Sensor-Assisted Facial Recognition: An Enhanced Bio-metric Authentication System for Smartphones. In: 12th annual international conference on Mobile systems, applications, and services, pp. 109122. Bretton Woods (2014)
10. Kesharwani, S., Shantaiya, S.: A Survey on Face Recognition Techniques in Video. European Journal of Advances in Engineering and Technology 2(5), 112116 (2015)
11. Naeem, M., Qureshi, I., Azam, F.: Face Recognition Techniques and Approaches: a Survey. Science International (Lahore) 27(1), 301305 (2015)
12. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: Conference on Computer Vision and Pattern Recognition, pp. 815823. Boston (2015)
13. Das, A., Degeling, M., Wang, X., Wang, J., Sadeh, N., Satyanarayanan, M.: Assisting Users in a World Full of Cameras. In: Conference on Computer Vision and Pattern Recognition Workshops, pp. 8392. Honolulu (2017)
14. Reddy, A. M., Kishore, M. R., Sreenivasulu, P., Jyothi, V.: A Survey paper for Face Recognition System. International Journal of Engineering Research in Computer Science and Engineering 5(4), 4550 (2018)
15. Khade, B. S., Gaikwad, H. M., Aher, A. S., Patil, K. K.: Face Recognition Techniques: A Survey. International Journal of Computer Science and Mobile Computing 5(11), 6572 (2016)
16. Liu, J., Deng, Y., Bai, T., Wei, Z., Huang, C.: Targeting Ultimate Accuracy: Face Recognition via Deep Embedding. In: Arxiv, (2015). https://arxiv.org/abs/1506.07310. Last accessed 21 Dec 2018
17. Parkhi, O. M., Vedaldi, A., Zisserman, A.: Deep Face Recognition. In: British Machine Vision Conference, pp. 41.1-41.12. Swansea (2015)
18. Maia, H. L. F.: Deteco and Reconhecimento Facial por meio de Aprendizado de Mquina. Universidade de Braslia, Braslia (2016)
19. Gurton, K. P., Yuffa, A. J., Videen, G. W.: Enhanced facial recognition for thermal imagery using polarimetric imaging. Optics Letters 39(13), 38573859 (2014)
20. Echeagaray-Patron, B. A., Miramontes-Jaramillo, D., Kober, V.: Conformal parameterization and curvature analysis for 3D facial recognition. In: International Conference on Computational Science and Computational Intelligence, pp. 843844. Las Vegas (2015)

21. The EU General Data Protection Regulation (GDPR) and Face Images. In: D-ID, (2018). https://www.deidentification.co/wp-content/uploads/2018/09/White-Paper-GDPR-and-D-ID.pdf. Last accessed 20 Dec 2018
22. Gruschka, N., Mavroeidis, V., Vishi, K., Jensen, M.: Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR. In: Arxiv, (2018). https://arxiv.org/pdf/1811.08531.pdf. Last accessed 31 Dec 2018
23. Morrison, M., Bell, J., George, C., Harmon, S., Munsie, M., Kaye, J.: The European General Data Protection Regulation: Challenges and considerations for iPSC researchers and biobanks. Regenerative Medicine 12(6), 693703 (2017)

A Survey of Time Series Classification for Device Monitoring

Pedro Pessoa

Universidade de Évora, Évora, Portugal,
pessoa@angulosolido.pt, d36719@alunos.uevora.pt

Abstract. In the field of device monitoring, the context of this survey, metrics are collected over time and in most cases in a fixed interval. These characteristics make it very convenient to store this data on a Time-Series Database (TSDB).
However, given the typical amount of hundreds of metrics that devices running just a few simple services may report, together with the current trend to have multiple of these devices in most of the simpler services architecture, we can easily see thousands of metrics being reported and stored on such Time-Series Databases.
Eventually all of these metrics will be useful. They will need human analysis, correlation or threshold setting of alerts which allow for comprehensive service monitoring and alerting.
There has been a pressing requirement to provide the human systems operator with distilled insights into the monitoring data - all the available metrics. This distilled data is comprised of system changing events as well as thresholds and alerts on those metrics.
This survey aims to present the state of the art of tools and methodologies for pattern detection or event correlation on time-series data stored in one or more Time-Series Databases. As well as identifying behaviours of related metrics with the ultimate goals of being able to detect anomalies in single and related metrics, predict their behaviour or directing the human operator to the most relevant events or in defining alerts thresholds for such metrics.

Keywords: time-series, time-series database, monitoring, pattern analysis, temporal analysis, pattern detection, prediction

1 Introduction

Measurements that are performed over time, eventually without limit, lead to a collection of organized data called time series.

The goal of time-series pattern detection and data mining is to try to extract knowledge and events from the shape of this data. In the monitoring field, automating this extraction will allow to bring a automatically curated subset of patterns and events for further analysis of the human operator. Although humans have a natural ability to identify these patterns and events, they cannot review all the potential metrics systems nowadays produce.

Even though time series has been matter of study for decades, effective techniques that provide satisfactory pattern identification in short compute times, or even real-time, are still not generally available.

This paper surveys existing techniques such as artificial intelligence, machine learning for time series pattern detection, data mining, modelling and analysis. Focus will be on highly dimensional time series stored on a Time-Series Database (TSDB) covering device monitoring data.

A TSDB is a database optimized for time-stamped or time series data. Time series are simply measurements or events that are tracked, monitored, down sampled, or aggregated over time.

In the field of device monitoring, multidimensional time series, those that measure more than one related variable over time, are common. Some typical examples are CPU usage and device power usage; memory usage and number of running processes; network IO and disk IO. However these time series are typically stored on the TSDB as multiple, uncorrelated, single dimension time series adding to the correlation and time series class identification problem.

2 Definitions

This section lists the definitions that will be used throughout this paper.

Definition 1. A ***device*** is a computer system that, for the scope of this paper, is source of multiple time series.

Definition 2. A ***time series*** T is a time ordered sequence of n real values.

$$T = (t_1, \dots, t_n) \in \mathbb{R}$$

In the case of device monitoring, time series are semi-infinite as the devices continuously feed the series with new data. This semi-infinite nature of device time series metrics forces the next definition of *subsequence* of time series.

Definition 3. Given a time series $T = (t_1, \dots, t_n)$ of length n, a ***subsequence*** S of T is a time series of length $m \leq n$ built from time consecutive data points of T.

$$S = (t_k, \dots, t_{k+m})$$

where $1 \leq k \leq n - m + 1$.

Definition 4. Given a time series $T = (t_1, \dots, t_n)$ of length n, a ***representation*** time series of T is a model T' of reduced dimensionality d ($d \leq n$) where T' closely resembles T.

Definition 5. Time series ***clustering*** results from partitioning a time series $T = (t_1, \dots, t_n)$ of length n into $P = (p_1, \dots, p_n)$, so that homogeneous time series are classified together based on a chosen similarity measure.

2.1 Representation

Given the fundamental characteristic of time series, their high dimensionality, it is often recommended to represent the time series in a reduced dimensionality [19] also referred as segmentation [9].

Time series representation approaches are [13] classified according to the type of applied transformation into following categories:

1. **Non-data adaptative:** non-data adaptive representation are used when the parameters of the transformation are fixed. That is, they do not depend on the nature of the time series data. Examples of non-data adaptive representation are:
   - Discrete Fourier Transform (DFT) [2],
   - Discrete Wavelet Transform (DWT) [5]
   - Piecewise Linear Approximation (PLA) [22]
   - Piecewise Aggregate Approximation (PAA) [12]
   - Symbolic Aggregate ApproXimation (SAX) [15]
2. **Data adaptive:** these transformations use the available time series data to adjust the parameters of the transformation. When an adaptation stage, consisting of a data sensitive parameter, is added to a non-data adaptative algorithm, an adaptive representation algorithm is considered:
   - Adaptive Piecewise Constant Approximation (APCA) Adaptive Piecewise Constant Approximation (APCA) [22]
   - Singular Value Decomposition (SVD) [14]
3. **Model based:** These algorithms assume the time series is generated from a model and that model parameters are discoverable and represent the time series:
   - Statistical modelling by feature extraction [18]
   - Autoregressive Moving Average (ARMA) models [8]
   - Markov Chains (MCs) [20]
   - Hidden Markov Models (HMM) [10]
4. **Data dictated:** in this approach, unlike the previous non-data adaptive, data adaptive, and model based approaches where the user can define the compression ratio taking into account their application, data dictated approaches automatically define the compression ratio. Clipped [17] is one of such approaches.

Time series representation considerations for device monitoring: the previously presented methods to represent time series have been considered effective to significantly reduce:

1. the time series cardinality,
2. the time series storage needs,
3. the data access rates and compute time on subsequent processing tasks

however, by reducing the cardinality in systems that enable visualization of device metrics time series, changing the representation from T to T', would prevent the users of those systems from inspect every single data point. This is not desired and would require to maintain the data series T even after the transformation had produced T'. This would then lead to increased storage requirements, to keep T and T', instead of reducing them [23].

In most of the experiments with time series data data mining [27], [4], has been demonstrated that transforming the time series leads to lower amount of data being analysed and also lower subsequent processing compute time.

However, more recent research, in particular after [7], we can find research [28] that confirms the constant performance of such systems with growing data sets. So in this case future work may prove that a TSDB backed by HBase (Bigtable) supports time series processing with constant latency regardless of data volume.

HBase (Bigtable): A Bigtable can be described as a sparse, distributed multi-dimensional sorted map [7]. It is designed to scale into the petabyte range across "hundreds or thousands of machines, and to make it easy to add more machines [to] the system and automatically start taking advantage of those resources without any reconfiguration". There are two implementations of [7], Google's own Compute Platform *Bigtable* and Apache's HBase.

MapReduce: The *MapReduce* framework methodology aims to distribute the processing of data. Data needs to be divided into, ideally independent chunks. Hence the importance of Time Series clustering, see next. The *MapReduce* technique is mainly used for parallel processing of data sets across various clusters. [24]

2.2 Clustering

This section covers the most studied time series clustering methods. As justified by the previous section, focus will be on clustering methods that work on the multidimensional time series resulting from device monitoring. Methods that require dimensionality reduction, and less relevant for device monitoring time series will be mentioned for completeness.

1. **Partitioning:** This clustering method makes k groups from n, with $(k \leq n)$, data points so that each group contains at least one data point. The mostly referenced algorithm for partitioning is *k-Means* [16] which uses a prototype to each cluster from the mean of its data points. This prototype becomes the center of the cluster and the algorithm will group data points that minimize the distance to this center. Determining this prototype on a semi-infinite time series is non-trivial.
2. **Hierarchical:** [21] introduces hierarchical clustering which makes a hierarchy of clusters by using agglomerative or divisive algorithms. An agglomerative algorithm builds the cluster in a bottom-up approach by starting to consider each data point as a cluster and gradually merging the remaining data points to the same or new clusters. On the other hand, divisive algorithms considers all the objects as a single cluster and splits them to the limit of clusters with one data point in a top down approach. This clustering method does not effectively support large time series [26] as it is of quadratic computational complexity.

3. **Grid based:** Thesed methods classify space into a finite number of the cells in a grid. Afterwards, clustering is applied to the grid cells. [25] is oneal example grid based clustering algorithms. This survey was unable to uncover works in the literature applying grid clustering on time-series.
4. **Model based:** This method uncovers a model from the data set. When in the device monitoring context, there are references to at least two drawbacks: first the algorithm needs to be parametrized and is slow processing on large data sets [3].
5. **Density based:** In this method, clusters are formed by subspaces of dense data points which are disjoint by subspaces in which data points have low density. In [6] a density based method in kernel feature space for clustering multivariate time-series data of varying length is proposed. With it, a heuristic method of finding the initial values of the parameters is also proposed. During this survey it became apparent that there is very little research in using density based clustering on time series data.
6. **Multi-step Clustering:** Newer research directions are proposing a combination of multiple methods to overcome the previous limitations with data size or parametrization. Two stand out in the context of device monitoring data:
   In [1], the author addresses very large time series data sets on movement of the stock market using a 3 phase method (3PTC) (1) pre-clustering of the time series, (2) cleaning and summarizing, (3) merging.
   Fast Shapelet Selection (FSS) is proposed in [11]. FSS works by first sampling the time series using a subclass splitting method. Then the FSS identifies the Local Farthest Deviation Points (LFDPs) for the time series and selects the subsequences between two nonadjacent LFDPs as shapelet candidates.

3 Conclusion

There is abundant research on time series data mining tasks. However two of the characteristics of device monitoring metrics time series, their high dimensionality together with the requirement of accessing individual data points and their semi-infinite nature, are difficult barriers for most of the surveyed methods. Those have focused mostly on:

- reducing the dimensionality of the time series,
- finding distance between time series for classification and clustering,

and balance between high quality, accurate but computational expensive and low quality but fast and computational inexpensive methods.

Judging by the promising results of hybrid methods that combine the capacity of working on large data sets and computation methods such as hbase/bigtable and map reduce, it is one conclusion of this survey that future work should focus on testing and development of these methods.

References

1. Aghabozorgi, S., Teh, Y.W.: Stock market co-movement assessment using a three-phase clustering method. Expert Systems with Applications **41**(4 PART 1), 1301–1314 (2014). https://doi.org/10.1016/j.eswa.2013.08.028, http://dx.doi.org/10.1016/j.eswa.2013.08.028
2. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Lomet, D.B. (ed.) Foundations of Data Organization and Algorithms. pp. 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg (1993)
3. Andreopoulos, B., An, A., Wang, X., Schroeder, M.: A roadmap of clustering algorithms: finding a match for a biomedical application. (2009). https://doi.org/10.1093/bib/bbn058
4. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery **31**(3), 606–660 (2017). https://doi.org/10.1007/s10618-016-0483-9
5. Chan, K.P., Fu, A.W.C.: Efficient time series matching by wavelets (1999). https://doi.org/10.1109/ICDE.1999.754915, http://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.754915&site=eds-live
6. Chandrakala, S., Sekhar, C.C.: A density based method for multivariate time series clustering in kernel feature space. Proceedings of the International Joint Conference on Neural Networks pp. 1885–1890 (2008). https://doi.org/10.1109/IJCNN.2008.4634055
7. CHANG, F.A.Y., DEAN, J., GHEMAWAT, S., HSIEH, W.C., WALLACH, D.A., BURROWS, M., CHANDRA, T., FIKES, A., GRUBER, R.E.: Bigtable: A Distributed Storage System for Structured Data. ACM Transactions on Computer Systems **26**(2), 4:2 (2008), http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=40080119&site=eds-live
8. Dilling, S., MacVicar, B.J.: Cleaning high-frequency velocity profile data with autoregressive moving average (ARMA) models. Flow Measurement and Instrumentation **54**(August 2016), 68–81 (2017). https://doi.org/10.1016/j.flowmeasinst.2016.12.005, http://dx.doi.org/10.1016/j.flowmeasinst.2016.12.005
9. Esling, P., Agon, C.: Time-series data mining. ACM Computing Surveys **45**(1), 1–34 (2012). https://doi.org/10.1145/2379776.2379788, http://dl.acm.org/citation.cfm?doid=2379776.2379788
10. Hewahi, N.M.: Hidden Markov Model Representation Using Probabilistic Neural Network pp. 50–63
11. Ji, C., Liu, S., Yang, C., Pan, L., Wu, L., Meng, X.: A Shapelet Selection Algorithm for Time Series Classification: New Directions. Procedia Computer Science **129**, 461–467 (2018). https://doi.org/10.1016/j.procs.2018.03.025, https://doi.org/10.1016/j.procs.2018.03.025
12. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Knowledge & Information Systems **3**(3), 263–286 (2001). https://doi.org/10.1007/PL00011669, http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=50006266&site=eds-live

13. Keogh, E., Lonardi, S., Ratanamahatana, C.A.: Towards parameter-free data mining. Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04 p. 206 (2004). https://doi.org/10.1145/1014052.1014077, http://portal.acm.org/citation.cfm?doid=1014052.1014077
14. Korn, F., Jagadish, H.V., Faloutsos, C.: Efficiently supporting ad hoc queries in large datasets of time sequences. In: Proceedings of the 1997 ACM SIGMOD International Conference: Management of Data. pp. 289–300 (1997). https://doi.org/10.1145/253260.253332, http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=101198774&site=eds-live
15. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop: Research Issues in Data Mining & Knowledge Discovery. pp. 2–11 (2003). https://doi.org/10.1145/882082.882086, http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=83827006&site=eds-live
16. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. pp. 281–297. University of California Press, Berkeley, Calif. (1967), https://projecteuclid.org/euclid.bsmsp/1200512992
17. Ratanamahatana, C., Keogh, E.J., Bagnall, A.J., Lonardi, S.: A Novel Bit Level Time Series Representation with Implication of Similarity Search and Clustering. In: PAKDD (2005)
18. Ravikumar, P., Devi, V.S.: Weighted feature-based classification of timeseries data. IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CIDM 2014: 2014 IEEE Symposium on Computational Intelligence and Data Mining, Proceedings pp. 222–228 (2015). https://doi.org/10.1109/CIDM.2014.7008671
19. Ren, H., Liu, M., Li, Z., Pedrycz, W.: A Piecewise Aggregate pattern representation approach for anomaly detection in time series. Knowledge-Based Systems **135**, 29–39 (2017). https://doi.org/10.1016/j.knosys.2017.07.021, http://dx.doi.org/10.1016/j.knosys.2017.07.021
20. Santilli, M., Gasparri, A., Oliva, G.: Optimal Redesign of Markov Chains with Prescribed Repulsive Distribution. 2018 IEEE Conference on Decision and Control (CDC) (Cdc), 1610–1615 (2018)
21. Sarle, W.S.: Finding Groups in Data: An Introduction to Cluster Analysis. (1991). https://doi.org/10.2307/2290430, http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=9705075457&site=eds-live
22. Shatkay, H., Zdonik, S.B.: Approximate queries and representations for large data sequences (1996). https://doi.org/10.1109/ICDE.1996.492204, http://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.492204&site=eds-live
23. Shurkhovetskyy, G., Andrienko, N., Andrienko, G., Fuchs, G.: Data abstraction for visualizing large time series. Computer Graphics Forum **37**(1), 125–144 (2018). https://doi.org/10.1111/cgf.13237
24. Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., Indragandhi, V.: Unstructured data analysis on big data using map reduce. Procedia Computer Science **50**, 456–465 (2015). https://doi.org/10.1016/j.procs.2015.04.015, http://dx.doi.org/10.1016/j.procs.2015.04.015

25. Wang, W., Yang, J., Muntz, R.: STING : A statistical information grid approach to spatial data mining. In: Proceedings of the 23rd International Conference on Very Large Databases, VLDB 1997. pp. 186–195 (1997)
26. Wang, X., Smith, K., Hyndman, R.: Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery **13**(3), 335–364 (2006). https://doi.org/10.1007/s10618-005-0039-x
27. Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E.: Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery **26**(2), 275–309 (2013). https://doi.org/10.1007/s10618-012-0250-5
28. Yu, B., Cuzzocrea, A., Jeong, D., Maydebura, S.: On managing very large sensor-network data using bigtable. Proceedings - 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGrid 2012 pp. 918–922 (2012). https://doi.org/10.1109/CCGrid.2012.150

Blockchain: Data Protection and Privacy*

Nuno Miquelina[1,2][0000−0002−3202−2242]

[1] Universidade de Évora, Évora, Portugal
d37384@alunos.uevora.pt
[2] Compta Business Solutions, S.A., Maia, Portugal
nuno.miquelina@compta.pt

Abstract. Blockchain is a rising technology, born as the base of the cryptocurrency implementation. Blockchain is used to register and validate the transactions in a way that all virtual money owners agree and trust. Soon investigators saw the potential of using blockchain as a public and trusted way to record information of all kinds. This technology, as in any other technology, needs also to have a great concern about data privacy and data protection on the user personal data. Any solution built over blockchain, used in European space, must also respect the General Data Protection Regulation (GDPR), to avoid European's Authorities sanctions. This work intends to raise relevant challenges in data protection and privacy on blockchain implementations and also the challenges and restrictions that GDPR enforces in such solutions.

Keywords: Blockchain · Data Protection · GDPR · Privacy

1 Introduction

The single idea of having a trusted payment system that does not rely on any financial institution, for sure drawn attention from everyone, especially from people (buyers and sellers) that make online transactions. This idea came from Satoshi Nakamoto, an anonymous person that is only identified by this name, that published a work called: "Bitcoin: A Peer-to-Peer Electronic Cash System". He or she stated that "what is needed is an electronic payment system based on cryptographic proof instead of trust, allowing any two willing parties to transact directly with each other without the need for a trusted third party [9]".

In this paper, Satoshi Nakamoto, made the first reference to blockchain technology by describing it "as an ongoing chain of hash-based proof-of-work, forming a record that cannot be changed without redoing the proof-of-work [9]". This new technology created the trust necessary to avoid having a third party that validates the transactions and avoids the double-spending problem. This ongoing chain acts as a ledger system, that registers the transactions of the electronic coin. Satoshi Nakamoto defined the electronic coin as a chain of digital signatures.

* This paper is submitted for the Introduction to Scientific Research course

Bitcoin was the first cryptocurrency, but others have emerged like: Ethereum, Dash, Monero, Ripple, and Litecoin. These ones were the most significant cryptocurrencies after Bitcoin, as of April 2008 [5].

The blockchain is the base of the cryptocurrency systems and was created for it, but the properties and the trust created by the blockchain technology soon captured the attention in the context of other uses.

In this work, there is a brief introduction to blockchain and some applications of the blockchain. These applications were selected, among others, because of the nature of the information that is registered in the blockchain: financial transactions and personal data. Both are sensitive data and must be protected. The Rights for data protection and privacy exist and are legislated. We can see European Union legislation regarding data protection and privacy in the following sections.

2 Blockchain

Blockchain can be considered as a distributed storage, distributed over a network of peers that guarantee the consistency of the chain. This distributed storage records blocks of information, that the network of peers validates as trustworthy. Each block holds a list of transactions functioning like a public ledger, and a reference (hash) to the previous or parent block. If there is any attempt to change data in the chain, the change is detected by the network - invalidates the hashes - and the attacked chain is replaced by a valid one. Is almost impossible to attack all the network, so the chain remains valid. Figure 1 illustrates an example of a blockchain architecture.

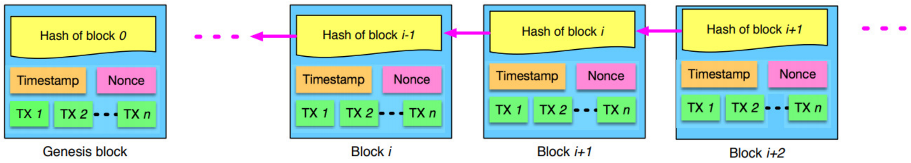


Fig. 1. Blockchain architecture [11]

The example structure of the block [11] is composed by the block header and by the block body. The block header contains the following information:

- **Block version** : indicates which set of block validation rules to follow.
- **Parent block hash** : 256-bit hash value that points to the previous block.
- **Merkle tree root hash** : the hash value of all the transactions in the block.
- **Timestamp** : current timestamp as seconds since 1970-01-01T00:00 UTC.
- **nBits** : current hashing target in a compact format.
- **Nonce** : a 4-byte field, which usually starts with 0 and increases for every hash calculation.

The block body is composed of a transaction counter and the transactions. The block has limited size, so the number of transactions that a block can hold is limited by the block size.

3 Data Protection and Privacy on Blockchain Applications

Blockchain as born as support to cryptocurrency implementation, recording the transactions performed by the users. This new environment created a trusted way to perform transactions between users and the chain works as a public ledger. But, quickly, other sectors started to see the potential of having such a system to record information other than financial information. Figure 2 shows some of the applications of blockchain that are been tested and used.

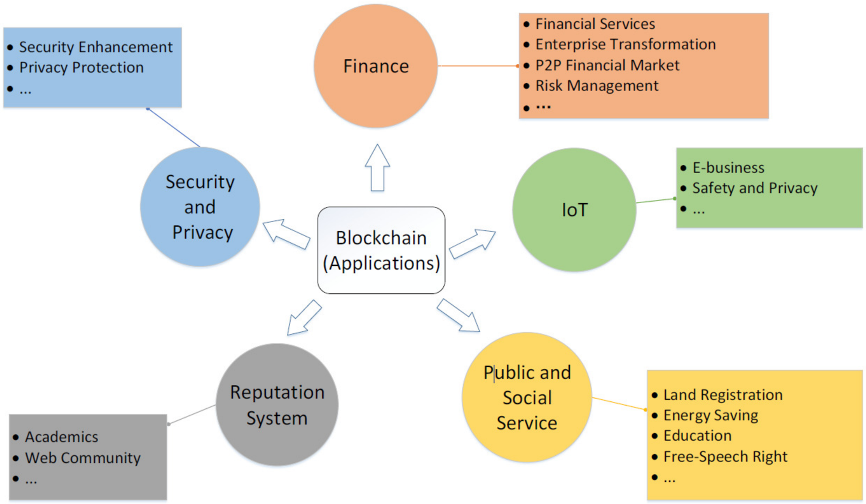


Fig. 2. Applications of Blockchain [11]

Each of these applications has challenges of data protection and privacy. The following sections will be focused on only a set of applications of blockchain: Financial services and Healthcare services.

3.1 Data Protection and Privacy in Financial Services

Bitcoin and other alternatives are playing an important role in financial services, becoming more mature and growing in market value [4]. This cryptocurrency facilitates fast and inexpensive transactions.

"Just how private are today's blockchains? The ephemeral nature of users' pseudonymous identities in Bitcoin played a key role in its early success. However, eight years of intense scrutiny by privacy researchers has brought to bear an arsenal of powerful heuristics using which attackers can effectively link disparate Bitcoin transactions to a common user and, in many cases, to that user's real-world identity [4]."

This breach of user privacy and data protection can break the trust in the system because traditional financial transactions are well regulated by governments and the financial institutions play an important rule in creating this trust and privacy.

This privacy problem has been studied by cryptography and privacy research communities, and they have proposed and implemented several protocols, trying to resolve this privacy problem. One solution is using the Tor network (https://torproject.org), but this does not resolve everything. "However, running complex protocols over general-purpose, low-latency anonymity networks such as Tor is fraught with risks and can expose users to subtle-yet-devastating deanonymization attacks, thereby undermining the privacy guarantees of the entire blockchain system [4]".

Some cryptocurrency applications, like Zcash (https://z.cash) and Monero (https://getmonero.org), use cryptography to hide personal data from unauthorized access and to prevent identifying the transaction users. But even so, with network-level information and access patterns for specific blocks can reveal the transactions users.

The use of blockchain in financial services, in spite of the value that is recognized to it, needs to get mature enough to ensure privacy and protection of the transaction users. There must be real confidence that the data is well protected and continues to be private. This confidence must be built and tested in real life scenarios.

3.2 Data Protection and Privacy in Healthcare Services

It is common sense that a decentralized database of medical records can help patients when they need to have their information available at any place and need to share information with many different parties. All medical stakeholders involved with the treatment can benefit from this decentralized database and common protocol to access the data [8].

There are already some blockchain applications or networks like Gem Health Network, Estonia Government, healthbank (Swiss digital health startup) that prove that a health infrastructure can be operated using Blockchain [8].

Healthcare is a data-intensive domain where a large amount of data is created, disseminated, stored, and accessed daily [2]. Even a simple act of taking a computerized tomography generates data that needs to be shared by the radiographer and the physician. Later, the patient can go to another hospital and access the information immediately.

The healhcare environment has well defined protocols and data types [2]:

- **EMR** : Electronic Medical Records, contain medical and clinical data related to a given patient and stored by the responsible healthcare provider.
- **EHR** : Electronic Health Records, designed to allow patient medical history to move with the patient or be made available to multiple healthcare providers.
- **PHR** : Personal Health Records, where patients are more involved in their data collection, monitoring of their health conditions, etc, using their smart phones or wearable devices.

Figure 3 shows an ecosystem in healthcare and the different uses of the medical records.

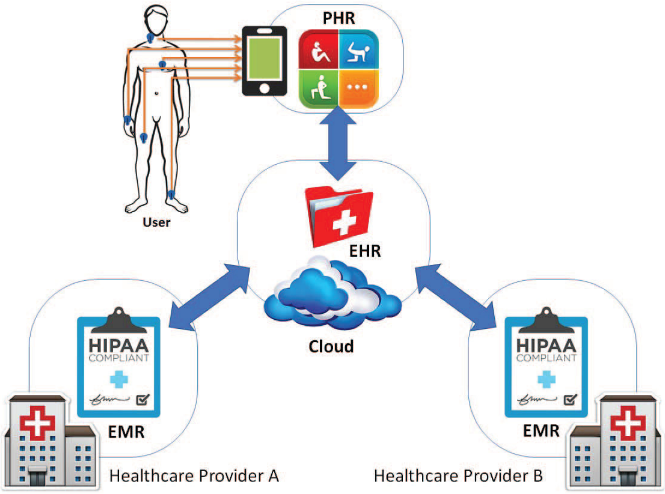


Fig. 3. A conceptual cloud-based EMR/EHR/PHR ecosystem [2]

.

This sharing of information on a public network, information that is personal and sensitive, may be attractive to cybercriminals. Cybercriminals can take advantages if they access the data and identify the owner of the data, by blackmailing persons or entities using the privileged information. Blockchain can be a solution to avoid attacks on personal information, but “there are limitations associated with a blockchain-based approach that need to be carefully studied. For example, blockchain technology can be somewhat disruptive and requires a radical rethink and significant investment in the entire ecosystem (e.g. replacement of existing systems and redesigning of business processes) [2].”

There is work in progress regarding the challenge of using blockchain in healthcare, like ModelChain [6], MedRec [1] or ProvChain [7]. Modelchain is a proposed framework to share patient medical data for research proposes in the field of predictive modeling based machine learning algorithms. Medrec is a decentralized record management system to handle Electronic Health Records, giving the possibility to share the records of patients across providers and treatment sites. ProvChain is a cloud data provenance solution, that records the creation and modifications on a cloud data object. These examples have different purposes for the use of medical data, but all try to understand the usability of blockchain technology and all have great concern about privacy and data protection.

4 Data Protection and Privacy: Rights and Legislation

Data protection and privacy are a deep concern in the European Union space. There is legislation to protect individual privacy and to protect the data. These rights are fundamental and the European Union makes efforts to guarantee these fundamental rights.

These two rights, data protection, and privacy are not the same as stated by Giakoumopoulos et al., "The right to respect for private life and the right to personal data protection, although closely related, are distinct rights. The right to privacy – referred to in European law as the right to respect for private life – emerged in international human rights law in the Universal Declaration of Human Rights (UDHR), adopted in 1948, as one of the fundamental protected human rights. Soon after adoption of the UDHR, Europe too affirmed this right – in the European Convention on Human Rights (ECHR), a treaty that is legally binding on its Contracting Parties and that was drafted in 1950. The ECHR provides that everyone has the right to respect for his or her private and family life, home and correspondence. Interference with this right by a public authority is prohibited, except where the interference is in accordance with the law, pursues important and legitimate public interests and is necessary in a democratic society [3]."

If blockchain acts like a public distributed database it is necessary to protect user privacy. If a user buys a good or service, he has the right that this information does not become public or, by tracking the transactions, reach the user identity. Again, this is not only applied to transactions but to all information that the block could register.

"Data protection in Europe began in the 1970s, with the adoption of legislation – by some states – to control the processing of personal information by public authorities and large companies. Data protection instruments were then established at European level and, over the years, data protection developed into a distinct value that is not subsumed by the right to respect for private life. In the EU legal order, data protection is recognized as a fundamental right, separate to the fundamental right to respect for private life. This separation raises the question of the relationship and differences between these two rights [3]."

If a company (or a public authority) adopts blockchain, it has to be careful with the protection of the data generated or collected from the user. They have to create the security mechanisms to prevent access to this data from users or entities which do not have that right granted.

“Article 8 of the EU Charter of Fundamental Rights (the Charter) not only affirms the right to personal data protection but also spells out the core values associated with this right. It provides that the processing of personal data must be fair, for specified purposes, and based on either the consent of the person concerned or a legitimate basis laid down by law. Individuals must have the right to access their personal data and to have it rectified, and compliance with this right must be subject to control by an independent authority [3].”

4.1 European Union: Regulation (EU) 2016/679

On April 2016, the European Union created a regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation). The European Data Protection Regulation became applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe (https://gdpr-info.eu/).

This regulation is applicable to the European Union members and the countries that have relations with the state members must also follow this regulation. This information can be found in article 3, Territorial scope [10].

Some articles from the regulation, related to a person’s rights, have a direct impact in blockchain fundamental aspects. The next table enumerates some of the most important ones.

Table 1. Table 1 - GDPR [10] vs Blockchain

Article	Subject	Blockchain impact
15	Right of access by the data subject	The blockchain application must control if personal information is getting registered and at any time should report to the user if and what is registered.
16	Right to rectification	This is a major drawback for blockchain applications, because if the data/transactions in the chain are immutable, then they cannot be changed. If a change is performed, the hashes in the chain become invalid. As result, the blocks should not be changed but a new block is added with the updated information.
17	Right to erasure (‘right to be forgotten’)	As Article 16, this is also a drawback. But in this case, if personal data is recorded, they should be removed from the chain. If not possible to remove, the blockchain application should have a way to apply data anonymization methods.

Blockchain's immutable property could lead to a non GDPR-compliant implementation. This situation could lead to penalties applied by government authorities.

"The responsibility and liability of the controller for any processing of personal data carried out by the controller or on the controller's behalf should be established. In particular, the controller should be obliged to implement appropriate and effective measures and be able to demonstrate the compliance of processing activities with this Regulation, including the effectiveness of the measures. Those measures should take into account the nature, scope, context and purposes of the processing and the risk to the rights and freedoms of natural persons [10]."

5 Conclusion

Blockchain technology can be applied in other applications and not only to cryptocurrency applications. This trusted way to register information (and not only financial transactions) and to store it in a decentralized way can create new services that depend on flows of validated data. In any case, exposing public information needs to be treated carefully because it could expose also personal data and violate user privacy rights. The European Union regulated the way that personal data and data protection should be handled. Blockchain applications must also follow these regulations to avoid penalties. Protecting data and privacy, having decentralized storage, being trusted, makes blockchain a disruptive technology to create new services at national and global level.

References

1. Ekblaw, A., Azaria, A., Halamka, J.D., Lippman, A.: A case study for blockchain in healthcare:"medrec" prototype for electronic health records and medical research data. In: Proceedings of IEEE open & big data conference. vol. 13, p. 13 (2016)
2. Esposito, C., De Santis, A., Tortora, G., Chang, H., Choo, K.K.R.: Blockchain: A panacea for healthcare cloud-based data security and privacy? IEEE Cloud Computing **5**(1), 31–37 (2018)
3. Giakoumopoulos, C., Buttarelli, G., O'Flaherty, M.: Handbook on European data protection law. European Union Agency for Fundamental Rights and Council of Europe (2018)
4. Henry, R., Herzberg, A., Kate, A.: Blockchain access privacy: challenges and directions. IEEE Security & Privacy **16**(4), 38–45 (2018)
5. Hileman, G., Rauchs, M.: Global cryptocurrency benchmarking study. Cambridge Centre for Alternative Finance **33** (2017)
6. Kuo, T.T., Ohno-Machado, L.: Modelchain: Decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks. arXiv preprint arXiv:1802.01746 (2018)
7. Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., Njilla, L.: Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In: Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. pp. 468–477. IEEE Press (2017)

8. Mettler, M.: Blockchain technology in healthcare: The revolution starts here. In: e-Health Networking, Applications and Services (Healthcom), 2016 IEEE 18th International Conference on. pp. 1–3. IEEE (2016)
9. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008)
10. Regulation, P.: Regulation (eu) 2016/679 of the european parliament and of the council. REGULATION (EU) p. 679 (2016)
11. Zheng, Z., Xie, S., Dai, H.N., Wang, H.: Blockchain challenges and opportunities: A survey. Work Pap.–2016 (2016)

Algoritmos de segmentação para identificação de defeitos na indústria da pedra[1]

Marco Tereso[0000−0003−4644−5227]

UEVORA,Évora, Portugal
d41655@alunos.uevora.pt

1

Resumo A evolução tecnológica permite à indústria modernizar e otimizar os seus processos de trabalho, bem como aumentar a sua rentabilidade. Num mercado de concorrência, é crucial inovar e tirar o máximo proveito da tecnologia e aplicá-la. A concorrência de mercados obriga o tecido empresarial a investir no controlo de qualidade dos seus produtos, e neste contexto a automatização é essencial. A deteção de defeitos na matéria-prima é um fator essencial para a qualidade do produto final. Este estudo baseia-se na análise e enumeração de um conjunto de algoritmos que aplicados na indústria, permitirão uma melhoria na diminuição da produção de resíduos e um aumento da rentabilidade da matéria-prima. A aplicação de técnicas de Machine Learning trazem vantagens e benefícios na otimização de processos e consequente lucratividade de recursos. O foco deste trabalho é a enumeração de algoritmos de segmentação de imagens, que possam ser aplicados na indústria na deteção de defeitos. Neste caso específico, concentramos a nossa pesquisa nas indústrias de transformação de pedra, cujo volume de exportação tem uma taxa muito significativa. A qualidade dos produtos rochosos, reconhecida no exterior, permitiu que as empresas se solidificassem e criassem mais empregos. Sendo a pedra um recurso natural em extrema abundância em Portugal, e com uma gama de diferentes tonalidades, este é um estudo com alguma dimensão. Os algoritmos aqui descritos, podem ser aplicados noutros tipos de indústrias, como cerâmica, madeira, metalúrgica, calçados e têxteis.

Keywords: Visão Computacional · Segmentação de imagens · Canny · Chan-Vese.

1 Introdução

Os avanços significativos de áreas como a Inteligência Artificial(IA) e Machine Learning(ML), permitem uma melhoria e renovação da indústria. Face à concorrência empresarial, a indústria necessita de se modernizar cada vez mais, de modo a tornar-se competitiva. A otimizando dos seus recursos, a capacidade de produzir mais com menos custos e ao mesmo tempo garantir a qualidade do

[1] Trabalho submetido no âmbito da Disciplina de Introdução à Investigação

produto final é uma meta. O foco deste trabalho prende-se essencialmente com o levantamento do estado da arte de técnicas e algoritmos de Visão Computacional, com o intuito de os podermos aplicar na deteção de defeitos em produtos acabados. Neste caso especifico, este trabalho, foca essencialmente em algoritmos que na literatura apresentem características capazes de detetar defeitos na indústria da pedra ou similares. Este trabalho encontra-se estruturado com o presente capítulo de introdução; de seguida surge o capítulo da descrição do problema, onde é identificado o problema que se pretende solucionar; posteriormente surge a secção do estado da arte, onde serão enumerados os principais métodos de segmentação de imagens; logo após ao estado da arte surge a secção reservada aos algoritmos, na qual são abordados os que são considerados viáveis para a implementação prática; após a descrição dos algoritmos surge uma secção reservada às ferramentas, onde serão identificadas ferramentas capazes de facilitar na implementação dos métodos enumerados; por fim, a secção de conclusão e trabalho futuro, onde são analisadas as conclusões e definido o ponto de partida para um trabalho futuro.

2 Descrição do Problema

Face à necessidade de acompanhar o desenvolvimento tecnológico e de apresentar produtos de qualidade num mercado competitivo, a indústria necessita de se modernizar. A indústria de extração e transformação de pedra, é uma das principais atividades económicas do nosso país. Com uma taxa de exportação significativa, a optimização e automatização de processos de manufatura são uma necessidade. A maioria do produto transformado desta indústria, destina-se à exportação, portanto é essencial manter a qualidade do produto final, de modo a manter a procura dos consumidores internacionais[1]. Numa indústria cada vez mais exigente, em que o rigor e a qualidade fazem a diferença, este trabalho de pesquisa tem por base apresentar métodos de Visão Computacional que possam ser soluções para a automatização na deteção de defeitos. Ainda que nesta fase o foco seja a Visão Computacional do problema, no futuro será interessante aplicar estes métodos tendo em conta o conceito de Machine Learning, fazendo com que as próprias máquinas tenham capacidade de selecionar defeitos de forma automática. Mediante o estudo [2] o autor faz a relação entre o processamento de imagens e o conceito de Visão Computacional. Na mesma fonte, o autor escreve que o processamento de imagens pode ser interpretado pela entrada de uma imagem e a saída de um conjunto de valores numéricos, que podem representar uma outra imagem. Enquanto que a Visão Computacional tende a aproximar a sua ação como se de uma emulação da visão humana se tratasse. Deste modo, este processo recebe uma imagem que tem como saída uma interpretação dessa mesma imagem de forma parcial ou total.

Tendo em conta o foco do nosso trabalho ser as indústrias de transformação de pedra, convém referir que as imagens digitalizadas neste tipo de indústria, não têm um formato linear, e também não são homogéneas, nem ao nível de textura nem ao nível de dimensão e formato. A digitalização de cada chapa de pedra, acontece à entrada para uma máquina CNC (Computer Numeric Control) de corte, que ao digitalizar a pedra reconhece os limites da mesma. O reconhecimento dos limites da chapa cortada é feito através de um processo de diferença de escala de cores, entre a pedra e o tapete de cor verde[2]. Após esta digitalização, é realizado o processo de identificação dos defeitos na superfície da mesma, processo este que é realizado manualmente por um operador fabril. Posteriormente é realizado o processo de nesting[3], de seguida passa-se ao corte da chapa em peças. Após a descrição do processo, é possível perceber que a automatização da deteção de defeitos permitirá um melhor aproveitamento dos recursos, diminuindo desta forma o trabalho do operador e retirando-lhe a responsabilidade de algumas decisões mais complexas. Este processo exige que o operador tenha uma boa saúde ocular, para garantir total fiabilidade na deteção de defeitos. Deste modo, a automatização contribuirá para uma melhoria do processo e otimização de recursos, esperando-se que consequentemente permita diminuir a taxa de erro na anotação e classificação dos defeitos. Os principais defeitos que se podem encontrar numa chapa são: quebras, existência de fósseis, existência de zonas profundas, e alterações de tonalidades. A Figura1 que se segue representa alguns desses defeitos. A linha de cor violeta faz a fronteira entre duas tonalida-

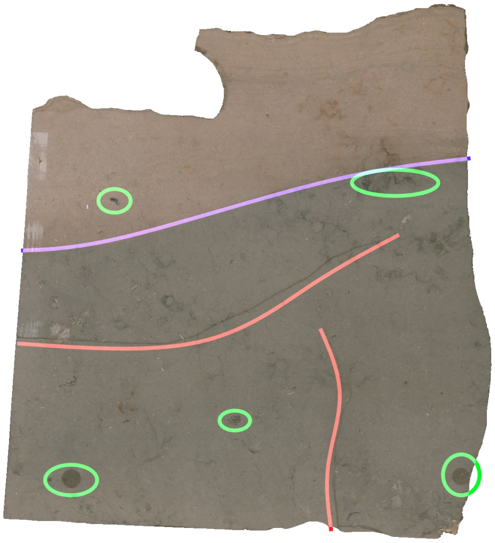


Figura 1. Chapa digitalizada com defeitos

[2] O tapete das CNC's de corte, é de cor verde tendo em conta que, é uma tonalidade invulgar de cor e praticamente inexistente na variedade de rochas em Portugal

[3] Nesting - processo de disposição das peças de corte sobre a chapa, de forma virtual, evitando as zonas defeituosas assinaladas

des diferentes, as linhas de cor vermelha realçam quebras existentes na chapa, e as elipses desenhadas a verde identificam existência de fósseis na pedra. Estes são alguns defeitos que podem ser considerados.

3 Estado da Arte

Apesar da escassez de investigação de algoritmos de deteção de defeitos aplicados na indústria de transformação de pedra, vários são os estudos similares, ainda que em áreas diferentes. Este artigo tem como foco, o levantamento de um conjunto de algoritmos que podem ser aplicados na indústria da pedra, como já foram aplicados a indústrias similares.

Em [4] e [5], foram realizados estudos similares aplicados à indústria das madeiras. Em [6] é feita uma abordagem sobre a deteção de defeitos em peças metálicas com ranhuras afinadas. No estudo de [7], são apresentadas técnicas de deteção de defeitos ao nível da textura e da tonalidade em peças cerâmicas, mais concretamente em pavimentos. Em [8], o autor procurou solucionar o problema de defeitos resultantes do processo de polimento de chapas de pedra. Um outro estudo [9], o autor apresenta um método de deteção de porosidade em superfícies rochosas. Tendo em conta os estudos anteriores, ainda que não sejam correlacionados diretamente com este estudo, apresentam contextos similares e com características muito semelhantes. Com base na pesquisa desenvolvida, foi possível identificar algumas etapas essenciais para a segmentação e deteção de defeitos a partir de imagens. No processamento de imagens, as principais técnicas utilizadas são[3]:

- Pré-processamento de imagens
- Melhoria da imagem
- Segmentação da imagem
- Extração de recursos
- Classificação da imagem

As próximas sub-secções descrevem em detalhe cada uma das técnicas enumeradas, e apresentam algoritmos especializados para cada processo.

3.1 Pré-processamento de imagens (Thresholding)

Thresholding é segundo[10] o método mais simples de segmentação de imagens. A base de funcionamento deste método, é a recriação de uma imagem a cores, para uma imagem em tons de cinza. A imagem resultante posteriormente pode ser utilizada para construir uma imagem binária.

Segundo[11], o Thresholding é um dos métodos mais antigos, mas ainda hoje dos mais utilizados. Um dos processos mais simples de conversão da imagem a cores numa imagem em escala de cinza, resulta da definição de uma constante, T, que através da análise de cor pixel-a-pixel, verifica se os valores RGB(Red, Green, Blue) do pixel é inferior a T. Caso a condição seja verdadeira, esse pixel passa a ter a cor preta, caso contrário passa a ter a cor branca.

A constante T tem o nome de limiar (ou *threshold*), e pode ser aplicada globalmente a toda a imagem, definindo-se um único valor constante de intensidade, ou localmente, exigindo a aplicação de um outro tipo de limiar, denominado de limiar dinâmico ou adaptativo. Segundo[12] geralmente, as abordagens automáticas aplicam análise estatística ao histograma, para determinarem o melhor valor quer seja global, quer seja local.
A cor é um recurso poderoso para a análise de imagens [13]. Em imagens coloridas cada pixel corresponde a uma intensidade das três cores primárias, RGB. O histograma é a representação dos valores RGB da cor de cada pixel, de forma gráfica, de modo a facilitar uma análise fundamentada sobre o número de vezes que determinado valor surge representado na nossa distribuição. Quando aplicado a superfícies pouco homogéneas, a imagens cujos píxeis possuem um nível de intensidade consideravelmente diferente ou que não se diferenciem muito do plano de fundo, este poderá ser um processo extremamente difícil. Nestes casos considera-se que o método não é suficiente para obter os resultados esperados. Na maioria dos casos ela surge apenas como primeira ou última etapa no processo de segmentação.

3.2 Melhoria da imagem

A qualidade da imagem original é bastante importante e nem sempre estão reunidas as melhores condições para a captura das mesmas. As imagens devem de ser captadas com luminosidade a uma escala de equilíbrio. Devem ser testadas as melhores condições para garantir que as capturas de imagem aconteçam sempre nas mesmas condições[3]. Após a aquisição das imagens digitalizadas também é possível realizar melhorias, aplicando uma das seguintes técnicas:

- Contraste Alongamento
- Filtragem de ruído
- Modificação do histograma

Contraste de alongamento Algumas imagens têm um contexto muito homogéneo (como por exemplo imagens com o céu, o mar, deserto ou neve como plano de fundo), neste tipo de imagens praticamente não existe diferenciação de tonalidades. Para estes contextos torna-se difícil fazer a identificação de eventuais defeitos.

Filtro de ruído A técnica de filtragem de ruído de uma imagem, consiste na identificação e eliminação de ruído dela mesma, este processo melhora a qualidade da imagem para análise posterior.

Modificação do histograma O histograma detém extrema importância na caracterização de uma imagem, a alteração do histograma permite alterar também a equalização da imagem[14].

3.3 Deteção de descontinuidade

O processamento da imagem original, origina uma imagem segmentada com base nas alterações bruscas de intensidade, nos níveis de cinza, originando desta forma uma imagem com contornos. Ideal na deteção de pontos isolados, linhas e bordas de imagens[15].

3.4 Segmentação de bordas

Para além da análise da intensidade dos píxeis, também é possível extrair outras características derivadas da intensidade, tais como, a deteção de bordas e contornos de objetos presentes numa imagem. Para a realização deste processo procede-se à aplicação de um conjunto de filtros aplicados à imagem.

4 Algoritmos

Segundo[16] algoritmo é um conjunto de passos sequenciais a seguir, para a realização de uma tarefa. A grandeza de um algoritmo não está na sua execução para uma finalidade, mas na possibilidade da sua adaptação a problemas distintos. Esta secção enumera alguns dos principais algoritmos de segmentação de imagens, tendo em conta os métodos analisados na secção anterior.

4.1 Algoritmos de limite (threshold)

Tendo em conta o conceito de threshold anteriormente enumerado, os algoritmos desenvolvidos para a aplicação desta estratégia, têm por base a classificação de uma imagem tendo em conta o valot 'T' definido para o *threshold*. Suponhamos que para uma imagem representada pela função f(x,y) com o valor de *threshold* 'T' temos que:

$$f(x,y) = 1, se(x,y) > Touf(x,y) = 0, se(x,y) \leq T \tag{1}$$

Dos valores obtidos, 1 representa a estrutura e 0 representa o fundo da imagem[17].

4.2 Algoritmos baseados em regiões

Este tipo de algoritmos, segue uma ideia de que numa imagem existem regiões que tendem a ser homogéneas. A base do seu funcionamento é fazer uma análise numa imagem, pegando num pixel ou num conjunto de píxeis e fundindo píxeis que tenham tonalidades semelhantes, criando desta forma regiões. Se regiões adjacentes, forem consideradas semelhantes, o algoritmo agrupa estas regiões em apenas uma região[17].

4.3 Algoritmos baseados em contornos

O paradigma de contorno, assenta sobre a determinação de um limite que serve de fronteira entre duas regiões com propriedades distintas numa imagem. Estes contornos são o limite entre diferentes gradientes de intensidade. A base de deteção de contornos de uma imagem é calculada através da derivação da função da imagem f(x,y). Sendo que:

$$|g| = \sqrt{[Gx^2 + Gy^2]} \tag{2}$$

Em que Gx e Gy representam os valores de gradiente nas direções x e y respetivamente. Alguns dos algoritmos mais conhecidos que implementam este paradigma, segundo[17], são o de Prewitt[18], de Roberts[19], de Sobel[20], o Laplaciano[21] e o de Canny[22]. Estes métodos são de rápido processamento e dispensam informações de base sobre as imagens, limitando-se a processar cada imagem que a eles seja associada. Por outro lado a existência de ruído em imagens dificulta a delimitação da imagem, apresentando por vezes, linhas descontinuas e com falhas. Estes são algoritmos bons para utilizar numa primeira fase de análise, os quais devem ser contemplados com técnicas de segmentação mais complexas[17]. O algoritmo de Canny é um dos mais utilizados e mais completos na deteção de contornos[23], segundo[24] é mesmo um dos melhores na deteção de bordas.

Algoritmo Canny edge detect O algoritmo de Canny, é um algoritmo para deteção de bordas que foi desenvolvido por John Canny. Ainda que a deteção de limites não seja de todo ideal, John Canny procurou desenvolver um algoritmo que fosse o mais eficaz possível. Ainda que seja um algoritmo que tenha surgido nos primórdios da Visão Computacional, em 1986, segundo[25] este detetor de bordas, é ainda nos dias de hoje considerado um excelente detetor e sem concorrência à altura. Segundo[26] este algoritmo, para um melhor desempenho, segue um conjunto de fases, sendo elas:

1. Uniformização da imagem - O primeiro passo, passa por reduzir o ruído da imagem original. Este método utiliza um filtro gaussiano para suavizar a imagem.
2. Diferenciação - A segunda etapa, passa por calcular as imagens de ângulo e magnitude do gradiente. Para isso Canny utiliza os métodos de Roberts, Prewit ou Sobel. Estes métodos passam por fazer o reconhecimento de linhas ou pontos através de cálculos matriciais.
3. Omissão de pontos de mínima intensidade - O terceiro passo, tem como objetivo tornar a borda o mais fina possível. Nesta fase, aplica-se a supressão de não-máxima no plano de magnitude do gradiente. Esta técnica, procura em cada ângulo encontrar a direção do ponto (horizontal, vertical e as duas direções diagonais). Caso o valor neste ponto seja menor que o mínimo de dois dos seus vizinhos na mesma direção, então o ponto é suprimido, passando a fazer parte do fundo em vez de fazer parte da borda.

4. Limiarização da borda (threshold) - Por fim, o último passo, utiliza o método "*histerese*", consiste em analisar limiares e conectividades duplas para descobrir e vincular arestas. Este processo utiliza dois limites, um baixo e um alto, em que todos os pontos acima do limite alto são adoptados e os pontos entre o limite inferior e superior apenas são incluídos se estiverem ligados aos pontos fortes calculados a partir do limite alto.

A Figura2 mostra uma aplicação prática do algoritmo de Canny, este exemplo aplica-se em bordas de imagens que necessitem de ser operadas por convolução.

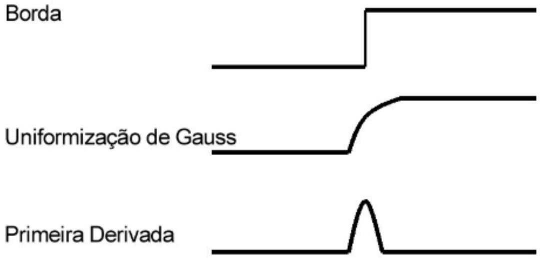


Figura 2. Exemplo da aplicação do algoritmo de deteção de bordas de Canny[27]

A função Gaussiana para uma dimensão é dada pela expressão[27]:

$$G(x) = \frac{1}{\sqrt{2\Pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} \tag{3}$$

Sendo a sua primeira derivada, dada pela fórmula[27]:

$$G'(x) = \frac{-x}{\sqrt{2\Pi\sigma^3}} e^{\frac{-x^2}{2\sigma^2}} \tag{4}$$

A ideia do algoritmo de Canny, para detetar bordas usando G'(x), é fazer a convolução desta derivada, para uma nova imagem *I*, mostrando bordas mesmo na presença de ruído. Segundo[27] esta operação é complexa, e torna-se mais exigente a nível de processador se for aplicada em contextos de duas dimensões. A expressão a aplicar em contextos bidimensionais é dada por[27]:

$$G(x,y) = \frac{1}{\sqrt{2\Pi\sigma_x\sigma_y}} e^{\frac{-x^2}{2\sigma_x^2} + \frac{-y^2}{2\sigma_y^2}} \tag{5}$$

Mediante esta popularidade e a qualidade reconhecida deste algoritmo, num trabalho futuro, será testado na identificação de bordas das chapas na indústria da pedra. As bordas a identificar nas chapas de pedra, são essencialmente as margens da chapa, e possíveis cortes, quebras, furos ou existência de fósseis. De seguida surgem duas imagens, a Figura3 que representa a fotografia original proveniente da digitalização de uma chapa, e a Figura4 que representa os limites da chapa, aplicando o algoritmo de Canny.

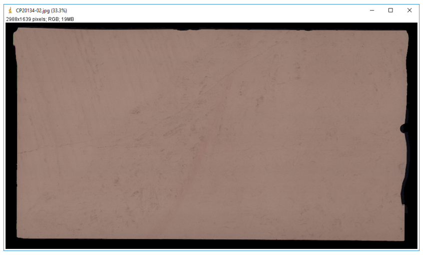


Figura 3. Imagem original

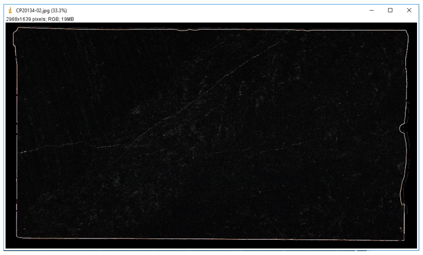


Figura 4. Imagem final (Canny)

A imagem da Figura4 foi trabalhada através de um aplicação JAVA, com o nome ImageJ[4].

4.4 Algoritmos Híbridos

Os algoritmos híbridos fazem uma combinação de técnicas utilizadas pelos algoritmos baseados em regiões e pelos algoritmos de contorno. Este tipo de algoritmos combina diferentes propriedades da imagem original, utilizando-as no processo de segmentação[17]. Um dos algoritmos deste tipo é o *Watershed*.

Algoritmo Watershed O paradigma de *Watershed* (método do divisor de água), consiste na recriação da imagem num plano tridimensional, que baseado na técnica de regiões utiliza a morfologia da imagem. Neste algoritmo é necessário definir pelo menos um marcador (semente), interior a cada objeto da imagem incluindo também o fundo da imagem. Estes marcadores são definidos ou pelo utilizador ou de forma automática[28]. A representação da imagem tridimensional, permite visualizar através de montanhas, a diferenciação de cores. O watershed representa a imagem através de gradientes que representam as variações locais de intensidade na imagem. Se imaginarmos uma imagem topográfica num recipiente, e se simularmos a queda da chuva, à medida que vai chovendo vão enchendo os socalcos mais baixos, e subindo, desta forma as regiões vão sendo fundidas à medida que a água vai subindo, desta forma o algoritmo vai fazendo a limitação da imagem[29,15]. O algoritmo interpreta os pontos de fusão, como sendo as bordas presentes na imagem.

4.5 Modelos Deformáveis Geométricos

A segmentação de imagens, baseada nas técnicas de deformação de objetos, é considerada na Visão Computacional como uma técnica de sucesso[30]. Segundo

[4] disponível para download na internet em https://imagej.nih.gov/ij/download.html

a mesma fonte, a sua utilização na área da medicina, tem se revelado bastante eficaz. Os modelos geométricos, também conhecidos na literatura internacional por *level set*, "são uma alternativa aos modelos de otimização de uma função objetivo, sendo a deformação do contorno formulada como uma frente de onda que se propaga e que é considerada como um level set de valor zero de uma função envolvente"[29]. Esta função envolvente se traduzida para uma equação diferencial parcial, que ao analisar a constante de velocidade, ou seja, valor e direção do vetor, força a paragem da propagação do limite tendo em conta a informação proveniente da imagem a segmentar[29]. Os modelos deformáveis dividem-se em dois tipos:

- Contornos ativos (ou snakes)
- Modelos geométricos (ou level sets)

De um modo resumido o método snake, indica que é possível seguir bordas em imagens através da definição de uma curva sobre a mesma, deixando essa mesma curva mover-se a uma forma e estrutura desejáveis. Esta curva contém propriedades físicas, tais como, elasticidade e rigidez e deve também ela ser atraída pelas bordas da imagem. Geometricamente uma snake é um contorno explícito diretamente sobre uma imagem.
Quanto ao *level sets*, segundo[31] surgiu para suprimir algumas lacunas no método *snake*, nomeadamente ao nível de problemas de topologia e cálculo numérico. O autor[31] remete para um exemplo prático da utilização deste método, suponhamos que existe uma fronteira que separa duas regiões (uma curva fechada no espaço bidimensional ou uma superfície num espaço tridimensional) e uma velocidade *F*, que define a forma de movimento de cada ponto da curva fronteira. Em que *F*, pode depender de qualquer fenómeno físico, como a dissipação de calor numa superfície. Assumindo-se que a função *F* é conhecida e nos dá a velocidade na direção perpendicular da interface. Um exemplo é o algoritmo Chan-Vese que aplica os conceitos descritos no modelo *level set*.

Algoritmo de Chan-vese O algoritmo de Chan e Vese[32], baseia-se num método de contorno ativo sem bordas, para a segmentação baseada em regiões e deteção de objetos numa imagem. Este modelo tem como base a técnica de segmentação *Mumford-Shah* (método que estabelece um critério otimizado para segmentar uma imagem em sub-regiões[33]) bem como no método *level set* para representar a curva da imagem[34]. Segundo a mesma fonte, as vantagens deste método são:

1. Permitir definir a posição da curva inicial em qualquer parte da imagem.
2. Detetar automaticamente os contornos interiores, sem a necessidade de especificar uma nova curva, esta é uma vantagem face ao método *level set*.
3. Tornar possível a deteção de diferentes objetos, de intensidades diferentes e com boa deteção de fronteiras com ruído.
4. Fazer uma mudança topológica automática da curva.
5. Conseguir detetar objetos mesmo onde o contorno não possua gradiente, graças ao critério de paragem da evolução da curva até que a fronteira desejada não dependa do gradiente da imagem.

6. Ter bons resultados na deteção de objetos em imagens com ruído.

Em[35] o autor exemplifica a forma como o método se comporta. Definindo C, como a curva inicial que pode obter qualquer forma, que não nula, e estar localizada em qualquer zona da imagem, obtendo o mesmo resultado final. Desta forma independentemente da forma que o objeto a segmentar tenha, esta curva irá convergir de modo a ajustar-se ao objeto que se pretende segmentar. Também não importa se a curva está dentro, fora ou a sobrepor as zonas internas e externas do objeto, tendo em conta que a função atingirá sempre o valor ótimo[35]. Consideremos:

$$F_1(C)+F_2C = \int_{Dentro(c)} |u_0(x,y)-c_1|^2 dxdy + \int_{Fora(c)} |u_0(x,y)-c_2|^2 dxdy, \quad (6)$$

Onde c_1, c_2 são constantes que representam os valores de medida da imagem dentro e fora da curva.
Para simplificar a expressão, supondo que o objeto e o fundo sejam uniformes temos que: Se a curva está fora do objeto, $F_1(C) > 0$ e $F_2(C) \approx 0$; Se a curva estiver dentro do objeto, $F_1(C) \approx 0$ para $F_2(C) > 0$; caso a curva C esteja dentro e fora do objeto, $F_1(C) > 0$ e $F_2(C) > 0$.
As imagens da Figura5 mostram que os contornos se intersetam se minimizarmos $F_1(C) + F_2(C)$.[35]

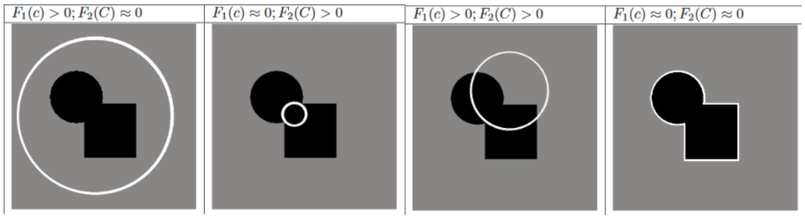

$F_1(c) > 0; F_2(C) \approx 0$
$F_1(c) \approx 0; F_2(C) > 0$
$F_1(c) > 0; F_2(C) > 0$
$F_1(c) \approx 0; F_2(C) \approx 0$

Figura 5. Aplicação do método Chan-Vese[35]

A Figura5 apresenta os resultados obtidos através do método Chan-Vese, neste caso específico o método faz uma boa representação dos limites da imagem. Este modelo tem um problema, não nos permite fazer o controlo da curva, então podem ser adicionados os parâmetros de distância e área. Segundo[35], basta adicionar os seguintes parâmetros:

$$F(c_1, c_2, C) = \mu.long(C) + v.area(dentro(C)) + \lambda_1 F_1(c_1, C) + \lambda_2 F_2(c_2, C); \quad (7)$$

$$F(c_1, C) = \int_{Dentro(C)} |u_0(x,y) - c_1|^2 dxdy, \quad (8)$$

$$F(c_2, C) = \int_{Dentro(C)} |u_0(x,y) - c_2|^2 dxdy. \quad (9)$$

$Onde \mu, v \geq 0 e \lambda_1, \lambda_2 > 0$, são parâmetros predefinidos da imagem. Tendo em conta os resultados ilustrados e a capacidade de ajuste do método Chan-Vese às formas dos objetos a selecionar, este algoritmo tem potencial para ser utilizado na segmentação dos defeitos em rochas ornamentais.

5 Ferramentas

Existem várias soluções para a implementação prática dos algoritmos aqui enumerados, soluções comerciais e outras gratuitas. Tendo em conta que a maioria das indústrias de transformação de rochas ornamentais são PMEs (Pequenas e Médias Empresas), a minha investigação, a este nível, foca-se em soluções open source. O objetivo é apresentar soluções a preços apelativos para este tipo de indústrias. Existe um conjunto de ferramentas open source para a segmentação de imagens, tais como, OpenCV[36], Scikit-Image[37], Mahotas[38] entre outras. Tendo em conta que as bibliotecas Scikit-Image e Mahotas são bibliotecas da linguagem Python, enquanto que OpenCV disponibiliza interfaces para várias linguagens de programação, vou apresentar o OpenCV em detrimento das outras, apenas por abranger mais linguagens de programação.

5.1 OpenCV

O OpenCV (Open Source Computer Vision) trata-se de uma biblioteca desenvolvida pela Intel que disponibiliza acima de 500 funções. Esta biblioteca foi desenvolvida com a finalidade de tornar a Visão Computacional acessível a utilizadores e programadores que têm as áreas de interação homem-máquina e robótica, como as suas áreas de investigação e interesse. Esta biblioteca tem uma licença BSD (Berkeley Software Distribution), sendo gratuita para uso académico e comercial[39]. Segundo a mesma fonte, oficial, é uma biblioteca multiplataforma, suportando os sistemas operativos mais comuns, Windows, Linux, Mac OS, iOS e Android. Ao nível de linguagens de programação, possui interfaces para as linguagens C++, Python e JAVA. Sendo uma ferramenta open source, o código encontra-se disponibilizado no site oficial da ferramenta para download[5]. Na sua página surge ainda indicação de que esta biblioteca possui mais de 47 mil utilizadores em todo o mundo, sendo as suas áreas de utilização bastante abrangentes. Segundo[2] a biblioteca está dividida em cinco grupos de funções:

- Processamento de imagens
- Análise estrutural
- Análise de movimento e rastreio de objetos
- Reconhecimento de padrões
- Calibração de camera e reconstrução 3D

Tendo em conta que é uma biblioteca de funções open source, e com várias funções implementadas, esta é uma ferramenta viável e interessante de analisar

[5] Toda a documentação e download do OpenCV em https://opencv.org/

em contextos práticos na obtenção de resultados, quando utilizada na indústria da pedra. No futuro pretendo fazer a implementação desta biblioteca e apresentar resultados práticos.

6 Conclusões e trabalho futuro

Este trabalho teve como foco a investigação sobre a área de Visão Computacional e o aprofundar de conhecimentos sobre métodos e técnicas de identificação de bordas e objetos em imagens. Neste estudo foram contemplados um conjunto de conceitos diferentes mas que podemos considerar que se possam relacionar. Como foi referido ao longo deste artigo, na maioria dos casos, a utilização de um único método não apresenta por si só resultados ótimos, melhorando esses resultados aquando da associação de diversos métodos. Foram abordados diferentes conceitos de deteção de bordas, mas neste caso específico e tendo como objetivo a aplicação de algoritmos que possam ser utilizados na indústria da transformação de rochas ornamentais, o algoritmo de Canny, algoritmo de deteção de bordas aqui apresentado, possui capacidade para ser aplicado nesta indústria. Para além da demonstração experimental na Figura4, o algoritmo apresenta também alguma capacidade na deteção de alguns defeitos apresentados na chapa. Tendo em conta a necessidade de utilização de métodos mais sofisticados na deteção dos defeitos, o algoritmo Chan-Vese, tem uma boa capacidade de responder às necessidades de identificação de objetos nas chapas de pedra digitalizadas. Sendo um algoritmo com bons indicadores, quando utilizado em imagens com ruído, é aparentemente uma boa solução a testar e implementar. Como trabalho futuro, proponho-me a utilizar os algoritmos aqui enumerados e a fazer uma implementação prática dos mesmos. Neste contexto irei utilizar a biblioteca OpenCV, tendo em conta que é uma biblioteca gratuita, com suporte para diferentes linguagens de programação e com bastantes funções implementadas.

Referências

1. Carvalho, J. M., Lisboa, J. V., Casal Moura, A., Carvalho, C., Sousa, L. M., Leite, M. M.: Evaluation of the Portuguese ornamental stone resources. In Key Engineering Materials (Vol. 548, pp. 3-9). Trans Tech Publications. (2013).
2. Marengoni, M., Stringhini, S.: Tutorial: Introdução à visão computacional usando opencv. Revista de Informática Teórica e Aplicada, 16(1), 125-160. (2009)
3. de Queiroz, J. E. R., Gomes, H. M.: Introdução ao processamento digital de imagens. RITA, 13(2), 11-42. (2006)
4. Funck, J. W., Zhong, Y., Butler, D. A., Brunner, C. C., Forrer, J. B.: Image segmentation algorithms applied to wood defect detection. Computers and electronics in agriculture, 41(1-3), 157-179. (2003).
5. Cavalin, P., Oliveira, L. S., Koerich, A. L., Britto, A. S.: Wood defect detection using grayscale images and an optimized feature set. In IEEE Industrial Electronics, IECON 2006-32nd Annual Conference on (pp. 3408-3412). IEEE. (2006, November).
6. Beyerer, J., León, F. P.: Detection of defects in groove textures of honed surfaces. International Journal of Machine Tools and Manufacture, 37(3), 371-89. (1997).

7. Xie, X.: A review of recent advances in surface defect detection using texture analysis techniques. Publisher, ELCVIA Electronic Letters on Computer Vision and Image Analysis, 7(3), 1-22. (2008).
8. Lee, J. R. J., Smith, M. L., Smith, L. N., Midha, P. S.: Robust and efficient automated detection of tooling defects in polished stone. Computers in Industry, 56(8-9), 787-801. (2005).
9. Tajeripour, F., Fekri-Ershad, S.: Developing a novel approach for stone porosity computing using modified local binary patterns and single scale retinex. Arabian Journal for Science and engineering, 39(2), 875-889. (2014).
10. Beltzac, A., Hess, R. C., Yamaguchi, S. Y.: DNA shot-reconhecimento de DNA através de imagens.
11. Pinheiro, C. I. C.: Algoritmos de Segmentação para Aplicações Biomédicas (Dissertação de Mestrado). (2017)
12. Filho, O. M., Neto, H. V.: Processamento Digital de Imagens. Rio de Janeiro: Brasport. (1999)
13. Chamorro-Martinez, J., Sánchez, D., Soto-Hidalgo, J. M.: A novel histogram definition for fuzzy color spaces. Publisher, In Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference on (pp. 2149-2156). IEEE. (2008, June).
14. Cordeiro, M. B., Tinôco, I. F., Roque Filho, M., Sousa, F. C. D.: Digital image analysis for young chicken's behavior evaluation. Engenharia Agrícola, 31(3), 418-426. (2011)
15. Qahwaji, R., Green, R.: Detection of closed regions in digital images. Publisher, The International Journal of Computers and Their Applications, 8(4), 202-207. (2001),
16. Campos, E. A. V., Ascencio, A. F. G.: Fundamentos da Programação de Computadores. Editora Prentice Hall (2003)
17. Silva, P. F., Zhen Ma, Tavares, J. M. R. S.: Segmentação de imagem médica: algoritmos para aplicação à cavidade pélvica feminina. in CIBEM 10, Porto, Portugal, 2011
18. Yang, L., Wu, X., Zhao, D., Li, H., Zhai, J.: An improved Prewitt algorithm for edge detection based on noised image. Publisher, In 2011 4th International Congress on Image and Signal Processing (Vol. 3, pp. 1197-1200). IEEE. (2011, October).
19. Galimberti, R.: An algorithm for hidden line elimination. Communications of the ACM, 12(4), 206-211. (1969).
20. Abbasi, T. A., Abbasi, M. U. (2007). A novel FPGA-based architecture for Sobel edge detection operator. Publisher, International Journal of Electronics, 94(9), 889-896.
21. Vollmer, J., Mencl, R., Mueller, H.: Improved laplacian smoothing of noisy surface meshes. Publisher, In Computer graphics forum (Vol. 18, No. 3, pp. 131-138). Oxford, UK and Boston, USA: Blackwell Publishers Ltd. (1999, September).
22. Olmos, A., Kingdom, F. A. (2004). A biologically inspired algorithm for the recovery of shading and reflectance images. Perception, 33(12), 1463-1473.
23. Baştürk, A., Günay, E.: Efficient edge detection in digital images using a cellular neural network optimized by differential evolution algorithm. Publisher, Expert Systems with Applications, 36(2), 2645-2650. (2009).
24. Shrivakshan, G. T., Chandrasekar, C.: A comparison of various edge detection techniques used in image processing. International Journal of Computer Science Issues (IJCSI), 9(5), 269. (2012).
25. Gonzalez, R. C., Woods, R. E.: Digital image processing third edition. Beijing: Publishing House of Electronics Industry, 719. (2008)

26. Haugsdal, K.: Edge and line detection of complicated and blurred objects (Master's thesis, Institutt for datateknikk og informasjonsvitenskap). (2010)
27. Canny, http://www2.ic.uff.br/∼aconci/canny.pdf. Last accessed 03 Jan 2019
28. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. Publisher, IEEE Transactions on Pattern Analysis & Machine Intelligence, (6), 583-598. (1991)
29. Silva, T. D., Tavares, J. M. R.: Algoritmos de segmentação de imagem e sua aplicação em imagens do sistema cardiovascular. Publisher, In Actas do 10^{o} Congresso Iberoamericano de Engenharia Mecânica (CIBEM 10). (2011)
30. Silva, J. S., Santos, B. S., Silva, A., Madeira, J.: Modelos deformáveis na segmentação de imagens médicas: uma introdução. Electrónica e Telecomunicações, 4(3), 360-367. (2004)
31. Bertuol, G.: Análise e aplicaçao da técnica de contornos ativos com métodos de segmentaçao tradicionais em imagens médicas. (2007)
32. Esedog, S., Tsai, Y. H. R.: Threshold dynamics for the piecewise constant Mumford–Shah functional. Journal of Computational Physics, 211(1), 367-384. (2006).
33. Tsai, A., Yezzi, A., Willsky, A. S.: Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. (2001).
34. Oliveira, R. B.: Método de detecção e classificação de lesões de pele em imagens digitais a partir do modelo chan-vese e máquina de vetor de suporte. Publisher, Universidade Estadual Paulista(UNESP) (2012)
35. Vidal Lloret, P.: Método de detección de contornos: implementación dinámica del modelo de Chan-Vese y detección de contornos basado en multirresolución. Publisher, Universitat Jaume I (2014)
36. Sobral, A.: BGSLibrary: An opencv c++ background subtraction library. Publisher, In IX Workshop de Visao Computacional (Vol. 2, No. 6, p. 7). (2013, June).
37. Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Yu, T.: scikit-image: image processing in Python. PeerJ, 2, e453. (2014).
38. Coelho, L. P.: Mahotas: Open source software for scriptable computer vision. Publisher, arXiv preprint arXiv:1211.4907. (2012).
39. Bradski, G., Kaehler, A.: OpenCV. Dr. Dobb's journal of software tools, 3. (2000)

Simple Event Model Ontology Population using an Information Extraction System: A Preliminary Approach

Gonçalo Carnaz[1,2](orcid.org/0000-0001-8285-7005)

[1] Informatics Departament, University of Évora, Portugal
d34707@alunos.uevora.pt
[2] LISP - Laboratory of Informatics, Systems and Parallelism, Portugal

Abstract. The present study proposes an Information Extraction (IE) system combined with the Simple Event Model (SEM) Ontology, aiming to obtain instances of semantic relations extracted from unstructured documents (crime police reports). An analysis of past works regarding IE and Ontologies applied to the criminal domain was performed. This contribution aims to develop an IE system to obtain instances of semantic relations regarding textual documents, like named-entities (NEs) and relations between them. Additionally, the same system includes a module that evaluates the possibility of SEM population with NEs and relations extracted.

Keywords: Information Extraction, Ontologies, Crime Police Reports.

1 Introduction and Motivation

The field of IE aims to obtain instances of semantic relations in textual documents, such as in crime police reports. Computer science methods applied to criminal investigations are a necessity, as argues [1], to face the deluge of structured and unstructured data obtained from heterogeneous sources like forensic reports or wiretap transcriptions. Therefore, crime police reports record the crimes committed or suspicions of it, and several associated NEs, such as Dates, Organizations or Locations. Extracting those NEs and relations from crime police reports becomes a vital task to support police investigators in crime investigation. Our main motivations for this paper are: retrieve relevant information, such as NEs and relations, and populate the SEM ontology.

The remainder of this paper [3] is organized as follows. First, section 2 details the works related to RE and ontologies applied to criminal domain. Next, in section 3 we proposed a system to RE from crime police reports and to populated the selected ontology. In section 4 we describe the experimental results. Finally, in section 5 we present the conclusions and future work.

[3] This paper is for the assessment of Doctoral Seminar IV.

2 Literature Review

In this section, we discussed a two-fold approach: the Relation Extraction (RE) systems and knowledge representation using Ontologies applied to criminal or legal domain.

2.1 Information Extraction Related Works

Information Extraction (IE) could be defined as the *"automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources"* [2]. A subtask related to IE is Relation Extraction (RE), defined by Culotta et al. [3] as *"the task of discovering semantic connections between entities. In the text, this usually amounts to examining pairs of entities in a document and determining (from local language cues) whether a relation exists between them."*, where three elements are retrieved with such task, a triple $(Subject, Predicate, Object)$. RE is applied to different applications, like question answering (QA) systems, information retrieval, summarization, semantic web annotation, construction of lexical resources and ontologies.

There are several methods for the development of RE systems: from Traditional Relation Extraction [4] [5] [6] [7] to the Open Information Extraction (OIE) [8] approaches, where several methods could be used, such as rule-based, supervised, semi-supervised, distantly supervised and others.

Within recent research projects and initiatives, RE has been proposed to extract relevant information from unstructured data. In 2008, the SEI-Geo System [4] recognized *part-of* relationships between geographic entities, using handcrafted patterns based on linguistic features to detect geographic entities in text documents. Bruckschen et al. [5] proposed a system named by *SeRELeP* to recognize three types of relations: *occurred*, *part-of*, and *identity*, using heuristic rules applied over linguistic and syntactic features. Nuno Cardoso [6] proposed a system called $REMBRANDT$, to identify 24 different relations using handcrafted rules and supported by two knowledge bases: DBpedia [4] and Wikipedia.

Garcia et al. [7] proposed in 2011, a system to extract occupation relationship instances over Portuguese texts. Training sentences to detect relations, using Support Vector Machines (SVM) classifier, evaluated over the extracted word, lemma, and PoS-tag, to compute the syntactic dependencies between words, using for that a syntactic parser.

Souza et al. [8] in 2014, developed a supervised OIE approach for extracting relational triples from Portuguese texts. Using annotated sentences from corpus CETENFolha [5], where positive and negative examples of relationships are labeled.

In 2015, Ricardo Rodrigues [9] proposed the RAPPort, a Portuguese Question-Answering System that uses a natural language processing (NLP) pipeline with a Named-Entity Recognition (NER) system and a fact extractor.

[4] See https://wiki.dbpedia.org/ [Accessed: January 2019].

[5] See https://www.linguateca.pt/cetenfolha/ [Accessed: January 2019].

Collovini et al. [10] in 2016, aims to evaluate the Conditional Random Fields (CRF) classifier to extract relations between named-entities, such as Organizations, Locations, and Persons, from Portuguese texts. David Batista [11] changed the original ReVerb algorithm, instead of looking for noun-phrases, tagged all the NEs, such as Persons or Organizations in a document collection. Then tried to find relational phrases, according to the pattern based on PoS, which is connected the NEs in a relationship [6].

In 2017, Sena et al. [12] aim to extract facts in Portuguese without predetermining the types of facts. Furthermore, the authors used an inference approach (identification of transitive and symmetric issues) to increase the quantity of extracted facts using open IE methods.

2.2 Ontologies Related Works

In artificial intelligence (AI), knowledge in computer systems is thought as an inference process over an explicitly and well-represented data. For understanding the *World*, we have to represent the entities, properties, relations or even attitudes, such as hypothesize, believe, expect, hope, desire, or fear [13]. Ontology as defined by Tom Gruber *"An ontology is a specification of a conceptualization."* [14] or *"...defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members)..."* [15]. We can formalize an ontology with the following concepts:

- Classes: defines how entities are organized. These classes, are unary predicates, containing different types, attributes or features that could be defining the application domain, such as *Person*, *Criminal* or *Time*.
- Entities: represent the instances or objects that exist in the application domain. That is, instances such as persons, chairs, or cars could be associated with classes, by *Axioms*. We can define those with a *InstanceOf* relation, like *instanceOf(AlCapone,Criminal)* or *instanceOf(SherlockHolmes,Detective)*;
- Relations: define a set of relations between classes or entities. These relation are dependent on the application domain. Therefore, we can have relations such as *hasCommitedBy(Person,Crime)*.
- Rules: statements in the form of an *if-then* (antecedent-consequent) sentence that describe the logical inferences, and the Axioms, or assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application.

Within recent research projects and initiatives, ontologies have been proposed to represent criminal or legal related data used by different police institutions. In 2003, Asaro et al. [16] found some issues regarding some ongoing projects, they aimed to develop a set of supported tools for Judges activities regarding the

[6] See https://github.com/davidsbatista/information-extraction-PT [Accessed: January 2019].

criminal domain. Despres et al. [17] applied in 2004, the *TERMINAE* method for alignment purposes of legal domain terms and a core ontology. Resulting in an ontology supported by the method below and the reuse of LRI-Core [18] and DOLCE [7].

Ortiz-Rodríguez et al., [19] aimed in 2006, to present an e-Government ontology, called *EGO*, to development of tools that allow legal documents to be model in the electronic support, regarding the Spanish and Mexican legal cases.

Casanovas et al. [20] aimed to develop an Ontology of Professional Judicial Knowledge, named *OPJK*, based on manual selection of relevant terms from legal questions and modeled according to the to *DILIGENT* methodology. Thus, the *OPJK* ontology has 700 terms (relations and instances), to reduce the minimization of concepts at the class level, *PROTON* [21] has been selected as an upper-level ontology. Daniela Tiscornia [22] has proposed the *LOIS* project supported in semantic metadata notions, i.e., what the resource is about. To achieved these results, she used as a support, the search engines, to retrieve legal information for the enrichment of the legal knowledge into their searching strategies. Additionally, a methodology was presented to build a multilingual semantic lexicon for the law. In the conceptual model, a multilingual lexicon is proposed, composed by 35000 concepts in five languages, the *JukWordNet* database, and used the WordNet and EuroWord Net resources. At the domain model, authors used DOLCE ontology for knowledge representation. Additionally, the Eurovoc thesaurus is integrated for project lexicon enrichment purposes. Also in 2007, Francescon et al. aimed to ensure that legal drafters and decision makers lead to control over a legal language, specified by *DALOS* Knowledge System. The *DALOS* project is divided into ontological and lexical layers. A domain ontology supports the Ontological Layer, and *LOIS* database [23] supports lexical Layer.

In 2009, Hoekstra et al. [24] proposed a legal core ontology that was part of the Legal Knowledge Interchange Format, know as *LKIF Core* Ontology, as a core in a legal knowledge system.

Rajput et al. [25] tried in 2014, to find suspicious financial transactions through an expert system, based on an ontology and a set of rules. Using a set of classes, objects, and properties that represent the transactions to be processed by the expert system. Markovi et al. [26] proposed a structural and semantic annotation approach for complaints, using a Serbian Judiciary use case for validation.

In 2016, Ghosh et al. [27] proposed an approach to building legal ontologies applied to the Lebanese Legal System, based on two processes: Conceptual Modeling and Ontology Learning. A middle-out approach was used for domain ontology development.

Rodrigues et al. [28] proposed in 2017, reuse of the *UFO-B* and LKIF ontology for property crimes applied to the Brazilian Criminal Code, called *Onto-PropertyCrime*. Mezghanni el al. [29] proposed *CrimAr* ontology is defined by a handcrafted approach, for the Arabic legal domain, supported by LRI-Core as top-level ontology. McDaniel et al. [30] proposed a framework, based on an

[7] See http://www.loa.istc.cnr.it/old/DOLCE.html [Accessed: January 2019].

ontology for physical evidence from a crime scene. The ontology includes a situation ontology, focus on physical evidence. Additionally, a physical bio-metrics, FOAF ontology was added. For development, the authors used Semantic Web standards (Resource Description Framework (RDF) and the Web Ontology Language (OWL)).

3 Proposed System

Figure 1 shows the first steps of our proposal to extract relations from criminal police reports, defining each module and outcomes that will be used on RE between NEs and other concepts.

The designed system was divided into several modules, following a standard NLP approach with some tweaks, such as trained models for NER module and others, regarding the domain applied (crime police reports). Therefore, the modules are:

- Pre-Processing (1): using stop-words detection, sentence detection and tokenization that generates defined tokens;
- Named-Entity Recognition (2): extract the NEs, such as Persons, Organizations or Locations. We presented this module and discussed the obtained results in [31];
- Relation Extraction and Semantic Role Labeling (SRL) module (3): that will generate the triples relations extracted from criminal police reports. Also a SRL sub-task was added to identify the roles present in each sentence extracted. The output results, will be used to populate an ontology, representing the relevant knowledge extracted from the criminal police reports. A review related to relation extraction was performed in [32];
- POS-Tagging (A): for syntax identification, specifying parts of speech to each word, such as noun, verb, adjective;
- Lemmatization (B): to remove the inflectional ends and to return the base or dictionary form of a word, known as the lemma;
- Dependency Parser (C): analyzes the grammatical structure of a sentence, using Maltparser [8] system to generate a trained file using a Portuguese treebank [9].

3.1 Ontology Module

The ontology module allows the population of the concepts, relations, and instances present in the chosen ontology to represent the knowledge extracted from the crime police reports. Using Simple Event Model (SEM) [10] ontology for mapping concepts and relations, see figure 2. As we can see in figure 3, we selected

[8] See http://www.maltparser.org [Accessed: January 2019].

[9] See https://github.com/UniversalDependencies/UD_Portuguese-Bosque [Accessed: January 2019].

[10] See https://semanticweb.cs.vu.nl/2009/11/sem/ [Accessed: January 2019].

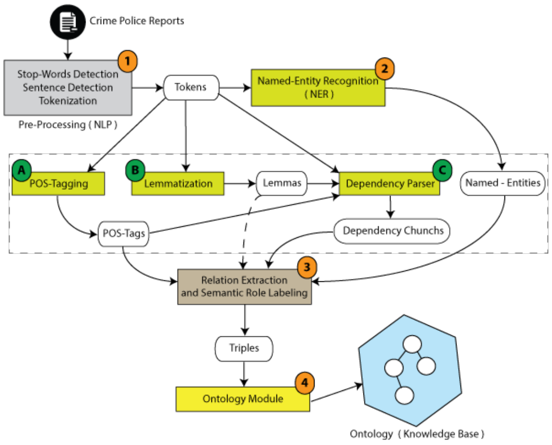


Fig. 1: Proposed System.

the SEM ontology, that has been created to represent model events in various domains, without making assumptions about the domain-specific vocabularies used [33], and for an introductory work, their simplicity will benefit our first steps regarding ontology population.

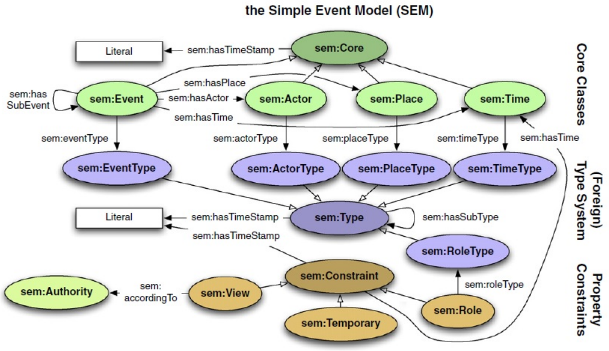


Fig. 2: Simple Event Model (SEM) design [33].

Figure 3 shows the module design, where annotated triples are analyzed and retrieved by the Extraction Tool in conjunction with Initial Ontology that originated the Concept/Relation Instances matched. After this the Ontology Population Tool populates the Extraction Tool ontology, that originates the Populated Ontology.

Notice that the ontology population is the task performed for adding new concepts and relations instances into an existing ontology. A concept/relation

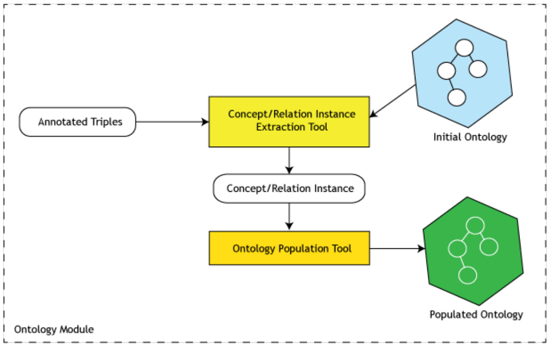


Fig. 3: Ontology Population Tool proposal.

instance is a realization of the concept in the domain, such as the instantiating of the concept as a phrase in a textual corpus. The proposed process does not change the structure of an ontology, like the concept hierarchy and relations, is not modified. Of course, during the analysis of textual corpus, the process may need another task, called ontology enrichment, extending an existing ontology with additional concepts and semantic relations and placing them at the correct position in the ontology.

4 Experimental Results

The details in Table 1 are merely illustrative of the Natural Language Processing (NLP) pipeline, using a sentence as an example, from sentence detection to relation extraction in the form of triples. For ontology mapping purposes, we used the results obtained regarding the NER module (marked as 2 in fig. 3). The recognition of named-entities, such as Person and Time that could be mapped into SEM ontology.

Figure 4 describes an example based on a sentence, that represents a typical situation from the criminal domain, where suspects or other actors are described, the name and date used on sentence is fictional but follows the form, syntax, and semantics of a crime police report.

Therefore, the possibility of mapping NEs extracted from a sentence (an example, also used on fig. 1), in this use case we mapped the named-entities: Persons and Places, found in (2) into the SEM ontology (1). In this study phase, some of the relations found on crime police reports cannot be extracted automatically, such as *nascido-em* or *natural-de*. Therefore, taking a manual approach, we can prove the feasibility and applicability of SEM ontology to support NEs or events and relations mapping and populating tasks, we need to extend the ontology with concepts or relations to facilitate such tasks.

Figure 5 shows a "snapshot" of SEM ontology, that uses a template-based approach to map NEs, events or relations to map as instances into SEM ontology,

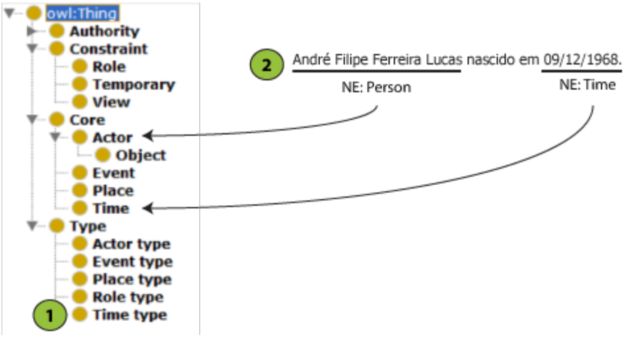
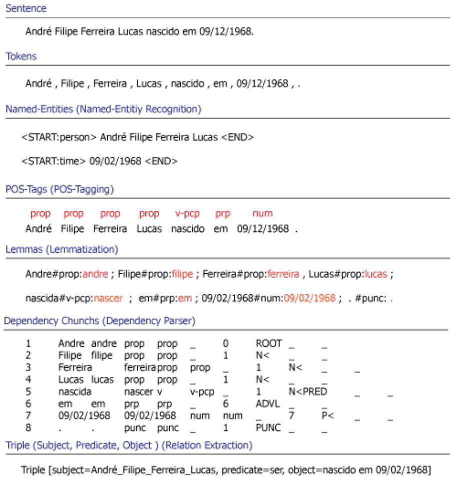


Fig. 4: Sentence example.



Sentence

André Filipe Ferreira Lucas nascido em 09/12/1968.

Tokens

André , Filipe , Ferreira , Lucas , nascido , em , 09/12/1968 , .

Named-Entities (Named-Entitiy Recognition)

<START:person> André Filipe Ferreira Lucas <END>

<START:time> 09/02/1968 <END>

POS-Tags (POS-Tagging)

```
prop   prop   prop     prop   v-pcp    prp   num
André  Filipe Ferreira Lucas  nascido  em    09/12/1968 .
```

Lemmas (Lemmatization)

Andre#prop:andre ; Filipe#prop:filipe ; Ferreira#prop:ferreira , Lucas#prop:lucas ;

nascida#v-pcp:nascer ; em#prp:em ; 09/02/1968#num:09/02/1968 ; . #punc: .

Dependency Chunchs (Dependency Parser)

```
1   Andre  andre   prop   prop   _      0     ROOT   _      _
2   Filipe  filipe prop   prop   _      1     N<     _      _
3   Ferreira       ferreiraprop  prop   _     1      N<     _      _
4   Lucas  lucas   prop   prop   _      1     N<     _      _
5   nascida        nascer v      v-pcp  _     1      N<PRED        _      _
6   em     em      prp    prp    _      6     ADVL   _      _
7   09/02/1968     09/02/1968    num    num   _      7      P<     _      _
8   .      .       punc   punc   _      1     PUNC   _      _
```

Triple (Subject, Predicate, Object) (Relation Extraction)

Triple [subject=André_Filipe_Ferreira_Lucas, predicate=ser, object=nascido em 09/02/1968]

Table 1: System output for a sentence (as example).

using Jena [11] framework to populate ontology and mapped into class, relations or other properties. In this case, we mapped only NEs, such as Persons, Places or Time.

5 Conclusion and Future Work

The work we described in this paper tries to achieve to achieve an analysis of RE systems for ontology population with relevant information that exists in the crime police reports retrieved from heterogeneous data sources. Analyzing the work developed until now, we achieved the possibility to extract relations from reports and the feasibility and applicability of SEM ontology to populate NEs, relations or events, and properties.

[11] See http://jena.apache.org [Accessed: January 2019].

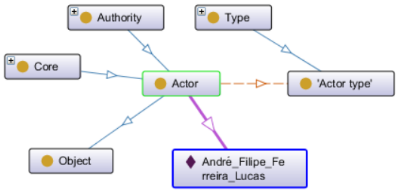


Fig. 5: Example of a Person named-entity mapped into SEM Ontology (Actor class) as instance.

For future work, our primary goals are: (1) to improve the automatic population of the ontology; (2) to improve the relation extraction module and SRL, to give more accurate results.

References

1. José Alberto Campos Braz. *Criminal Investigation*. Leya, 2013.
2. Sunita Sarawagi et al. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377, 2008.
3. Aron Culotta, Andrew McCallum, and Jonathan Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303. Association for Computational Linguistics, 2006.
4. Cristina Mota and Diana Santos. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter : Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM, page 436. 2008.
5. José Guilherme Mírian Bruckschen, Re-nata Vieira Souza, and Sandro Rigo. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter 14, page 436. 2008.
6. Nuno Cardoso. Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto, 2008.
7. Marcos Garcia and Pablo Gamallo. Evaluating various linguistic features on semantic relation extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 721–726, 2011.
8. Erick Nilsen Pereira Souza and Daniela Barreiro Claro. Extração de relações utilizando features diferenciadas para português. *Linguamática*, 6(2):57–65, 2014.
9. Ricardo Rodrigues and Paulo Gomes. Rapport—a portuguese question-answering system. In *Portuguese Conference on Artificial Intelligence*, pages 771–782. Springer, 2015.
10. Sandra Collovini, Gabriel Machado, and Renata Vieira. A sequence model approach to relation extraction in portuguese. In *LREC*, 2016.

11. David Soares Batista. *Large-Scale Semantic Relationship Extraction for Information Discovery.* PhD thesis, Instituto Superior Técnico, 2016.

12. Cleiton Fernando Lima Sena, Rafael Glauber, and Daniela Barreiro Claro. Inference approach to enhance a portuguese open information extraction. In *Proceedings of the 19th International Conference on Enterprise Information Systems*, volume 1, pages 442–451, 2017.

13. B Chandrasekaran, John R Josephson, and V Richard Benjamins. What Are Ontologies , and Why Do We Need Them ? *IEEE Intell. Syst.*, pages 20–26, 1999.

14. Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993.

15. Mauricio B. Almeida and Marcello P. Bax. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Ciência da Informação*, 32(3):7–20, 2003.

16. Carmelo Asaro, Maria Angela Biasiotti, Paolo Guidotti, Maurizio Papini, Maria-Teresa Sagri, Daniela Tiscornia, and Lucca Court. A Domain Ontology: Italian Crime Ontology. *Proceedings of the ICAIL 2003 Workshop on Legal Ontologies and Web based legal information management*, 1(January):1–7, 2003.

17. Sylvie Despres and Sylvie Szulman. Construction of a Legal Ontology from a European Community Legislative Text. *Leg. Knowl. Inf. Syst. Jurix 2004 Seventeenth Annu. Conf.*, pages 79–88, 2004.

18. Joost Breuker and Rinke Hoekstra. Epistemology and ontology in core ontologies: Folaw and lri-core, two core ontologies for law. In *In Proceedings of the EKAW04 Workshop on Core Ontologies in Ontology Engineering*, pages 15–27. Northamptonshire, UK, 2004.

19. F Ortiz-Rodriguez and B Villazón-Terrazas. EGO Ontology Model: law and regulation approach for E-Government. *3rd Eur. Semant. Web Conf. ESWC 2006*, 1, 2006.

20. Pompeu Casanovas, Núria Casellas, Christoph Tempich, Denny Vrandečić, and Richard Benjamins. OPJK and DILIGENT: Ontology modeling in a distributed environment. *Artificial Intelligence and Law*, 15(2):171–186, 2007.

21. John Davies. Lightweight ontologies. In *Theory and Applications of Ontology: Computer Applications*, pages 197–229. Springer, 2010.

22. Daniela Tiscornia. The Lois Project: Lexical Ontologies for Legal Information Sharing. In *Proceedings of of the V Legislative XML Workshop, European Press Academic Publishing*, pages 189–204, 2007.

23. Enrico Francesconi, Pierluigi Spinosa, and Daniela Tiscornia. A linguistic-ontological support for multilingual legislative drafting: the DALOS Project. *CEUR Workshop Proceedings*, 321:103–111, 2007.

24. Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. Lkif core: Principled ontology development for the legal domain. In *Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, pages 21–52, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press.

25. Quratulain Rajput, Nida Sadaf Khan, Asma Larik, and Sajjad Haider. Ontology Based Expert-System for Suspicious Transactions Detection. *Comput. Inf. Sci.*, 7(1):103–114, 2014.

26. Marko Markovi, Stevan Gostoji, and Zora Konjovi. Structural and Semantic Markup of Complaints : Case Study of Serbian Judiciary. In *12th Int. Symp. Intell. Syst. Informatics*, pages 15–20, 2014.

27. M. El Ghosh, H. Naja, H. Abdulrab, and M. Khalil. Towards a Middle-out Approach for Building Legal Domain Reference Ontology. *Int. J. Knowl. Eng.*, 2(3):109–114, 2016.
28. Cleyton Mario De Oliveira Rodrigues, Frederico Luiz Goncalves De Freitas, and Ryan Ribeiro De Azevedo. An Ontology for Property Crime Based on Events from UFO-B Foundational Ontology. *Proceedings - 2016 5th Brazilian Conference on Intelligent Systems, BRACIS 2016*, pages 331–336, 2017.
29. Imen Bouaziz Mezghanni and Faiez Gargouri. Crimar. *Procedia Comput. Sci.*, 112(C):653–662, September 2017.
30. Marguerite McDaniel, Emma Sloan, William Nick, James Mayes, and Albert Esterline. Ontologies for situation-based crime scene identities. *Conf. Proc. - IEEE SOUTHEASTCON*, 2017.
31. Gonçalo Carnaz, Vitor Nogueira, Mário Antunes, and Nuno Ferreira. An automated system for criminal police reports analysis. In *14th International Conference on Information Assurance and Security (IAS'18)*, 2018.
32. Gonçalo Carnaz, Paulo Quaresma, Vitor Nogueira, Mário Antunes, and Nuno Ferreira. A review on relations extraction in police reports. In *WorldCist'19 - 7th World Conference on Information Systems and Technologies*, 2019.
33. Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136, 2011.