Fall 9-2019

Image based recognition of the monuments in Prizren

Annea Futko
University for Business and Technology - UBT

Mechatronics and Management Program

IMAGE BASED RECOGNITION OF THE MONUMENTS IN PRIZREN

Bachelor Degree

Annea Futko

September / 2019
Prishtinë

Mechatronics and Management Program

Diploma Thesis
Academic year 2018-2019

Annea Futko

IMAGE BASED DESCRIPTION AND RECOGNITION OF THE MONUMENTS IN PRIZREN

Mentor: Dr. Sc. Bertan Karahoda

September / 2019

This paper has been compiled and submitted to meet the partial requirements for the Bachelor Degree

ABSTRACT

Image classification application has recently been covering a high number of research fields. In the other hand as the performance of the mobile devices is being updated day by day, the implementation of image recognition algorithms in them, is not only being trendy but very helpful in everyday tasks. With the automatic monument recognition, visiting a city is easy and fun. This application recognizes the captured monument, gives useful information and describes that particular landmark.

In this thesis there are used four historical monuments of the city of Prizren, Kosovo. The images where taken specially for the research from the different angles of the city and the dataset for the training and testing process has been created. Although these monuments differ from one another in the archaeological structure, the classification process is not an easy approach. Here will be presented an approach for the recognition of these particular monuments by using computer vision and machine learning methods on images. The image processing classification techniques and algorithms used in the literatures not only for the landmark recognition but overall the methods, will be described.

ACKNOWLEDGEMENT

Foremost, I would like to express the deepest appreciation to my mentor prof. Dr Bertan Karahoda for the continuous support, motivation and immense knowledge during my Bachelor study. He has been a real inspiration with his guidance and by continually giving me the courage throughout the process of writing this thesis.

Besides my mentor, I would also like to thank the rest of my professors of the Mechatronics department in the UBT for always giving me the opportunity to express myself and for always supporting my steps.

In addition, a very special thank you goes to my parents who have given me the opportunity of an education from the best institutions, who have been supporting me throughout my whole life and never had a doubt on the steps I made,

This thesis is dedicated to you, Nana and Baba.

September, 2019

Prishtinë

LIST OF FIGURES

LIST OF ABREVIATIONS

CNN - Convolutional Neural Network

CV - Computer Vision

SVM - Support Vector Machine

RGB - Red Green Blue

NN - Neural Network

kNN- k nearest neighbor

ID3 - Iterative Dichtomiser

CONTENTS

1 INTRODUCTION

In this section there will be a small introduction about the main concept and purpose of this thesis. In other terms, there will be described how the importance and the reason of this research.

The application of Machine Learning algorithms has expanded in almost every field of our lives. They are helping us deal with the problems in a quicker, easier and more efficient way. From the personal software assistants to learning from medical records for finding patient diagnosis, this field has started to become a part of our everyday life.

This thesis proposes an application of these so called intelligent systems in the tourism industry. Automatically recognizing the monuments and describing the history and features of them can help a tourist learn more about them, just by taking a picture.

By using Machine Learning algorithms and Image Processing methods is created a system which describes the monuments in Prizren. Or else we can call it as - an intelligent tour guide. The idea of this intelligent tour guide is that the tourists who visit Prizren, by only photographing the monument will be able learn about the history, the time of construction and the purpose of that particular monument.

Prizren is a historic city located in Kosovo. Surviving since the ancient times, Prizren holds a mixture of cultural and religious heritage. For this application are used only four famous monuments in Prizren with a different architecture and cultural meaning for the city. The collected images formed a dataset which were used for the training process using deep learning.

The main objective of this thesis is to test with how much accuracy can the created program classify or so to say recognize these monuments by the dataset of the images created specifically for this.

2 LITERATURE REVIEW

The automatic monument recognition approach is still a new area of interest. The literature and the work done for this particular field is only been found in a very small number of papers. There are some of the working groups an authors who have proposed and applied different methods for the landmark recognition. Some of the work and used techniques by this authors are also represented in this thesis.

This chapter is divided in different parts to express the theoretical work of this thesis so that the readers could have a clarified understanding about it. Firstly, it will talk about the captured images and the digital image processing. And the other part will deal with the machine learning algorithms used for image classification, recognition and object detection.

2.1 Digital Images

A digital image is a discrete representation of data as a 2D array which possess both layout and intensity information (Solomon & Breckon, 2011). It can be defined as a function f (x, y) where x and y are spatial coordinates, and the amplitude of at any pair of coordinates (x, y) is called the intensity of that image at that point. When x, y, and amplitude values of F are finite, we call it a digital image.

A digital Image is a composition of a lot of different elements which have a particular value at a very particular location. This element is called pixels. Each pixel of an image refers to a physical object in the 3D world and is formed from the combination of light and R, G and B sensors. Part of the reflected light reaches the array of sensors used to image the scene and is responsible for the values recorded by these sensors. One of these sensors makes up a pixel and its field of view corresponds to a small patch in the imaged scene as it is seen an image is a complicated concept and it can be measured by its quality. An image is considered of a

good quality if it is not noisy, blurred, has high resolution and good contrast (Petrou & Petrou, 2010).

Digital images play a very important role in computer vision and artificial intelligence systems because this type of input data is very helpful for the algorithms of these systems. From the trained digital image inputs, processes such as object detection, classification, recognition and so on, can be performed.

Another input data could be also the signal, but in this thesis there are only used the digital images because of the nature of the problem.

2.2 Computer Vision

Computer Vision is a combination of a lot of processes and representations used for vision perception. It follows different tasks and techniques such as image processing and classification methods to gain a high level understanding from digital images, videos and so, to achieve specific goals (Ballard & Brown, 1982)

From the engineering point, a computer vision system tries to create an autonomous system which imitates the human visual system so that this system can sense the environment, understand the sensed data, take appropriate actions, and learn from this experience in order to enhance future performance just as humanlike perception capabilities (Sebe, Cohen, Garg, & Huang, 2005).

Computer vision deals with images which must follow certain steps of techniques to give the desired results. The techniques involved in CV are image collection, image preprocessing, image processing and finally analyzing of the image to achieve the main goal which can be the classification for the detection or recognition of the image. For this, Machine Learning offers effective methods for computer vision for automating the model/concept acquisition and updating processes, adapting task parameters and representations, and using experience for generating, verifying, and modifying hypotheses.

The base line of this thesis is that by using the steps of the so called computer vision system, to achieve the goal of recognizing and describing of the specific monuments, the data set of whom was created from the images taken specially for this project.

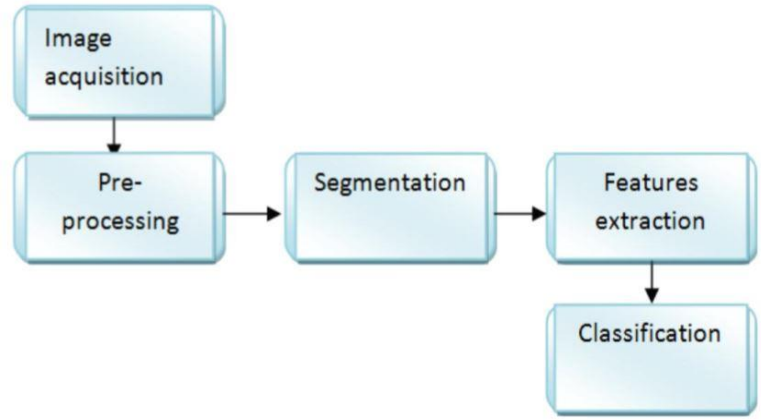


Figure 1. Process of a Computer Vision System

Image Acquisition

Image Acquisition is the primary step of the process. A physical scene or a structure of an object is represented digitally in the form of a digital image. Tools such as digital cameras, smart phone cameras, scanners or video recorders are used to collect the images as the input data from the environment one wants to use.

Andrew Crudge, Will Thomas and Kaiyuan Zhu in their research with the title Landmark Recognition using Machine learning used the input data of 193 images, all collected from the Google Images. Another proposed approach for the Indian Monument Recognition has used 100 different images of temples in India where the images where used from Google and where from different angular views.

There are also other projects who use different platforms for building their dataset with different landmark image. But however in this research the input images of the monuments were taken directly at various angles, horizon shifts, zoom etc.

Also one of the main factors of collecting good data in this process is the illumination and the focus to the main object. The performance of the illumination system can greatly affect the image quality as it plays an important role in the accuracy of the system.

When all the image data is collected or photographed, it is uploaded to the computer where the other processes will be performed. Usually the images are in RGB format, but may be captured in different dimensions so they have to follow other steps. In our case, the images are also in RGB and in different dimensions since the main idea is that all tourists do not have the same camera, lightning or photographing perspective.

Image Processing

In computer vision and image data analyzing, Image processing is the main process that provides us with better results for coming closer to the main goal.

It is a method to convert an image into digital form and perform different Image processing operations on it, so a lot of useful information can be extracted from the image.

Some of the purposes of image processing is to visualize the objects that are not visible, create a better image, seek for the image of interest, measure various objects in an image and so on. And as we can see, by applying the image processing tools we get better results for the next steps of the process. These tools are offered form almost every language that is used for image processing, computer vision or machine learning.

To start with the image processing, first the image must be imported from the image acquisition tools. Since the image can be represented in different types or formats such as binary images, 8-bit color images, RGB images, all of them need to be converted in the same or easier format for the computer to read it. After all the data is read, image processing techniques can be implemented on them.

There are a lot of techniques which are used for better portray the images. Some of them are represented in the sections below.

– Image Enhancement

This image processing technique is one of the simplest ones. Image enhancement is used to improve the quality of the image in subjective way.

The main concept of Image Enhancement is to bring out hidden details or highlight certain features. And one does not know what an image should look like when it is enhanced but, he can tell whether it improved or not if the contrast is better or if the details are more obvious. Some of the methods used for enhancement are:

Smoothing and low pass filtering, Sharpening or high pass filtering, Mean, median or Gaussian filtering, Generic deblurring algorithm and Histogram manipulation which can be seen in Figure 2.

a, b, c

Figure 2. Histogram maniputlation (a) original image, (b) result of histogram equalization (c) histogram of the image

(Gonzalez & Woods, 2002)

– Image Restoration

Image restoration is also used to improve the image but unlike enhancement it uses an objective criterion and is based on mathematical, probabilistic models of image degradation.

An image is degraded when the grey values of pixels are altered or are far away shifted from their old positions and that is when restoration enters the process. It is applied by using different methods for eliminating the degradation which often comes with a noise too.

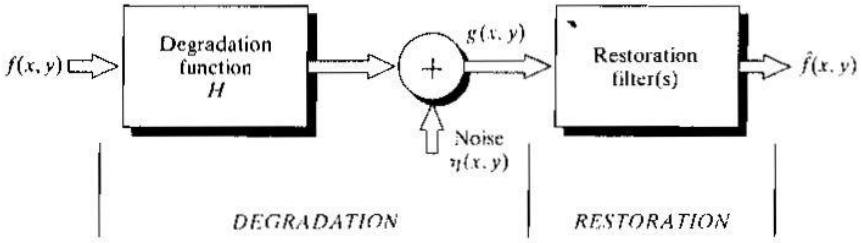


Figure 3. Image restoration process

For the restoration process there are usually used various types of restoration filters. Some of the techniques are: Mean filtering, Adaptive filtering, Inverse filtering, Weiner filtering, MAP estimation, Constrained matrix inversion (Petrou & Petrou, 2010).

There are other also other image processing techniques that are very helpful for image improvement such as:

- *Color image processing* – This method is divided into two major areas: full-color and pseudo color processing. Color image processing model and process colors of the images in a digital domain, simplifying object identification and helping for further extraction from an image.

- *Wavelets and multi resolution processing* -This theory represents and analysis signals in more than one resolution using wavelet transforms. Images are divided into smaller regions for better data compression and for pyramidal representation with which incorporated the multiresolution processing.

- Morphological Processing – This method is used to identify and extract important components of the image based on the form or shape of them (Solomon & Breckon, 2011).

- *Image compression* – Compression deals with techniques which use as few bits as possible to represent the image, without any loss of image information (Petrou & Petrou, 2010).

Image Segmentation

Image segmentation has a very important role in image processing because it is the very first step to understand an image. Usually we are not interested in all parts of the image, but only some certain areas or one object specifically. Image Segmentation uses different techniques to extract the outlines of different regions in the image and to divide the image into regions which are made up of pixels which have something in common in order to take out the area which we are interested in. The level to which these areas are divided depends on the problem being solved. Segmentation should stop when the objects or regions of interest are detected. For example, in this project the object of interest are the monuments that must be segmented from the image and after the algorithms identify these elements it is no need to continue the segmentation anymore.

How good the image is segmented is measured by the accuracy of segmentation. It determines the eventual success or failure of computerized procedures. For this reason, considerable care should be taken to improve the accuracy of the segmentation algorithms. (Gonzalez & Woods, 2002). And also another thing to keep a mind on is that there is no universally applicable segmentation technique that will work for all images but the right one depends strongly on the types of object or region we are interested in identifying (Young, I.T., Gerbrands, J.J., van Vliet, L.J., 1998).

So Image segmentation is one of the hotspots in image processing and computer vision. It helps to keep going on the further steps such as feature extraction, classification, description etc.

The techniques that are used to find the objects of interest are usually referred to as segmentation techniques – segmenting the foreground from background. Some of the segmentation techniques are:

Edge detection, Thresholding, Region based segmentation, Clustering based segmentation, Watershed method, Semantic segmentation.

- Edge Detection

Edges are important characteristics of images. They are significant local changes which typically occur on the boundary between two different regions in an image (Sebe, Cohen, Garg, & Huang, 2005). With the edge detection technique, first all the edges are detected and then are connected together to form the boundaries of the object and to segment the required regions (Kaur & Kaur, 2014).

This segmentation technique is highly used in different computer vision processes. Some of the most used edge detection algorithms that can be used for the segmentation with edge detection are: Laplacian of Gaussian filtering, Canny edge detector, Sobel operator. The result of these operators are usually binary images. And in the Figure 4. shown below in the left can be seen the original image and in the right the binary image with Canny edge detector applied in Matlab.

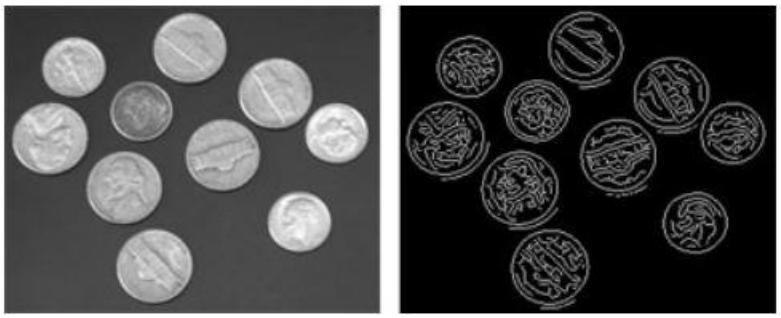


Figure 4. Canny Edge Detector

- Thresholding

Thresholding is one of the most basic techniques used for the segmentation. It produces binary images from a grey-scale or color image by setting values to 1 or 0 depending on their threshold value.

One of the main thresholding methods are global and local thresholding.

Global thresholding – for choosing a threshold value can be used the histogram of the image. If we have an image f (x, y) formed from light objects and dark background, to separate these two modes of the image from one another is to select a Threshold T. When f (x, y) ≥T the result is the object and when f (x, y) ≤T the result shows the background (Gonzalez & Woods, 2002).

In the figure below is presented the original image and the image after the global thresholding method is applied. The experiment is taken from the webpage of Matlab where it can clearly be seen how the image is separated as the background and the objects in it.

Figure 5. Global thresholding

But there can be several limitations to simple global thresholding because it is only applicable to simple images where the foreground and background are already in very different intensities.

Local or Adaptive Thresholding - designed to overcome the limitations of conventional, global thresholding and chooses different threshold values from the global one.

This method presumes that illumination may differ over the images so is generally determined from the values of the pixels in the neighborhood of the pixel (Solomon & Breckon, 2011).

- Region Based Segmentation

The region based segmentation technique segments the image into different regions having similar characteristics. The two main methods of the technique are the region growing method and region splitting and merging method.

Region growing method – this method groups pixels or sub regions into larger regions. Region growing needs a set of starting pixels called seeds, the process starts by picking a seed from the set and find connected neighbor pixels which similar characteristics as this seed and then merge a suitable neighbor to the seed. The seed is then removed from the seed set, and all merged neighbors are added to the seed set. The region growing process continues until the seed set is empty.

Region splitting and merging method- The region splitting and merging based segmentation methods uses two basic techniques i.e. splitting and merging for segmenting an image into various regions. Splitting stands for iteratively dividing an image into regions having similar characteristics and merging contributes to combining the adjacent similar regions. Following diagram shows the division based on quad tree (Kaur & Kaur, 2014).

- Watershed Algorithm

The concept of watershed is based on visualizing an image in three dimensions: two spatial coordinated versus intensity. This technique is based on topological interpretation (Gonzalez & Woods, 2002). A watershed is formed by ‘flooding’ an image from its local minima, and forming ’dams’ where waterfronts meet. When the

image is fully flooded, all dams together form the watershed of an image. The watershed of an edginess image (or, in fact, the watershed of the original image) can be used for segmentation. The idea is that when visualizing the edginess image as a three-dimensional landscape, the catchment basins of the watershed correspond to objects (Kaur & Kaur, 2014). Segmentation using watershed methods has some advantages over the other methods. A notable advantage is that watershed methods, unlike edge detection-based methods, generally yield closed contours to delineate the boundary of the objects. A number of different approaches can be taken to watershed segmentation, but a central idea is that we try to transform the initial image (the one we wish to segment) into some other image such that the catchment basins correspond to the object we are trying to segment (Solomon & Breckon, 2011).

For the monument recognition applications there has not been any study yet. The researches that has been made before in this area, have only used the features extraction, noise removal or other simpler methods since they used the convolutional neural networks as the learning algorithm.

Feature Extraction

One of the other steps one should consider on the way of having a good classification is the feature extraction. The classification can be good only if features which are extracted from images are enough to get a better accuracy. The model can be trained on any features provided it gives better results. After all of these image restoration, enhancement and segmentation we can finally extract some features that best describe our inputs.

The feature extraction methodology addresses the problem of finding the most compact and informative set of features, to improve the efficiency or data storage and processing. Defining feature vectors remains the most common and convenient means of data representation for classification and regression problems. Data can then be stored in simple tables (lines representing “entries”, “data points, “samples”, or “patterns”, and columns representing

“features”). Each feature results from a quantitative or qualitative measurement, it is an “attribute” or a “variable” (Guyon, Gunn, Nikravesh, & Zadeh, 2006).

So finding good features for classification and recognition of the monuments can sometimes be not an easy approach. On “A Survey on Mobile Landmark Recognition for Information Retrieval “from Tao Chen, Kui Wu, Kim-Hui Yap, Zhen Li, and Flora S. Tsai for the landmark/monument recognition categorized features as local and global ones.

They said that global features characterized the overall properties of the images and the local features aim to represent the contents of images using local information extracted from the salient regions or patches within the images. (Chen, Wu, Yap, Li , & Tsai, 2009).

- What are the global features?

Global features are color, edge or text based features. For the landmark and monument recognition the global feature is the earliest and simplest feature, due to its low computation cost and simplicity. But still it can only help for the global properties of the image and not with the regions of interest. Therefore, this type of feature extraction method is used together with the local features method.

- *Local Features*

Local features and their descriptors describe selected individual points or areas in an image. The extraction is executed in two steps. First, a set of key points in the image is detected. Second, the area around the selected key points is analyzed to extract a visual description. Local feature descriptors contain information that allow local feature matching, i.e. deciding that two local features from two different images represent the same point. Standard information on the position in the image, the orientation and size of the region are typically associated with the visual information that depends on the particular local features (Amato, Gennaro, & Falchi, 2015). Their applications include image registration, object detection and classification, tracking, and motion estimation. This is the reason why extracting local features is very helpful for the accuracy of the classification.

Local features are widelt used in monument recognition too, due to the capability of describing the properties of regions of interest.

On ‘A Survey on Mobile Landmark Recognition for Information Retrieval’, the authors talk about three categories of local features for this specific topic. These features are: *SIFT – based, generative model-based and patches –based.* But there are also other proposed methods such as *HOG, SURF, ORB, LBP* which can be used for extracting the local features. The usage of these local feature based algorithms depends on the image and what local characteristics of these image we want to extract.

Figure 6. Application of SIFT method for landmark recognition

2.3 Machine Learning

Machine learning is a sub-domain of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It also draws on concepts and results from many other fields, including

statistics, philosophy, information theory, biology, cognitive science, computational complexity, and control theory.

In his book titled 'Machine Learning', Tom Mitchell describes learning with the definition that says: *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

For example, a task for the learning process of a self- driving robot could be driving on public four-lane highways using vision sensor. Performance P, the average distance traveled before an error. And training experience E could be a sequence of images and steering commands recorded while observing a human drive (Mitchell, 1997).

For these so called learning problems to be better understood by a computer, it is important to have a good representation of the input data. It is helpful to think of the data as a table. Each data point that you want to reason about is a row, and each property that describes that data point is a column.

Each entity or row here is known as data point or *sample* in machine learning, while the columns, the properties that describe these entities, are called *features* (Muller & Guido, 2016).

Inside of this world called Machine learning there is a variety of algorithms and different types of learning methods. From the simplest to the most complex one, depending on the problem itself machine learning offers a lot of different ways to find an easier and better solution for the problem. So one can see that by creating systems which need a minimal human intervention, machine learning is here to make life tasks easier for us all.

- Types of Learning

Learning of an algorithm can be *Supervised, Unsupervised* and *Reinforcement Learning*. This depends on the nature of the problem we want to use the algorithm on. So we do not choose randomly the algorithm or the learning type but the problem itself does.

Reinforcement Learning

Reinforcement learning addresses the question of how an autonomous agent that senses and acts in its environment can learn to choose optimal actions to achieve its goals. This very generic problem covers tasks such as learning to control a mobile robot, learning to optimize operations in factories, and learning to play board games. Each time the agent performs an action in its environment, a trainer may provide a reward or penalty to indicate the desirability of the resulting state. For example, when training an agent to play a game the trainer might provide a positive reward when the game is won, negative reward when it is lost, and zero reward in all other states. The task of the agent is to learn from this indirect, delayed reward, to choose sequences of actions that produce the greatest cumulative reward (Muller & Guido, 2016).

Unsupervised Learning

In this type of algorithm learning method we do not tell the system about the right outputs but just shown the input data, it is asked to give us right answers about this data.

The data usually does not contaien labeled information so the computer is trained with the unlabeled input data and these algorithms can be practical when the experts don't know what to look for in the data and what the actual output should be.

Another common application for unsupervised algorithms is as a preprocessing step for supervised algorithms. Learning a new representation of the data can sometimes improve the accuracy of supervised algorithms, or can lead to reduced memory and time consumption (Muller & Guido, 2016).

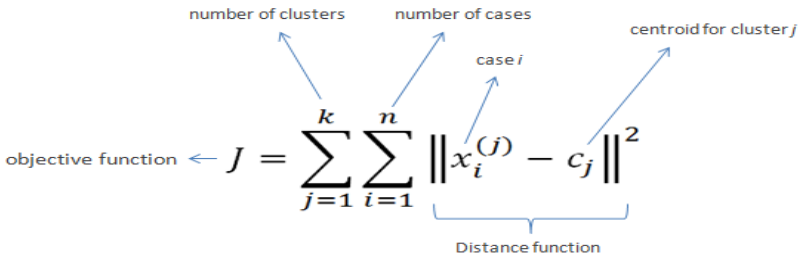
The main type of algorithms are the clustering algorithms, such as K-means clustering algorithm.

K-means clustering

K- means clustering is an iterative algorithm which is one of the simplest and commonly used as a clustering algorithm.

The algorithm requires as input data a matrix of M points in N dimensions and a matrix of K initial cluster centers in N dimensions. We keep iterating the position of the K centers until there is no change between the points and the center and the clusters are formed. The number of points in cluster L is denoted by the Euclidean distance point I and cluster L. The general procedure is to search for a K-partition with locally optimal within-cluster sum of squares by moving points from one cluster to another (Hartigan & Wong, 2012).

The formula of it is:



Supervised Learning

Supervised learning is one of the most commonly used types of machine learning. In this type of learning we have our data labeled. We provide the algorithm with inputs and tell it the desired outputs. This way it learns from the data and finds a way to produce the desired output given a new set of examples of the new input.

Supervised learning has two major types, called classification and regression.

In classification, the goal is to predict a class label, which is a choice from a predefined list of possibilities. It can be a binary classification e.g. 'Yes' or 'No', or can have more than two classes for classification.

In regression problems the output is a real number, or a floating point number in programming terms (Muller & Guido, 2016).

Since we said the supervised learning is the most common one, there is a big number of algorithms used for classification and regression problems. Below will be talked more widely about some of the most important supervised machine learning algorithms.

k-Nearest Neighbor

The k-nearest neighbor algorithm is one of the simplest and easiest algorithms to implement, which can be used for both classification and regression.

Contrary to other learning algorithms that allow discarding the training data after the model is built, kNN keeps all training examples in memory. Once a new, previously unseen example x comes in, the kNN algorithm finds k training examples closest to x and returns the majority label (in case of classification) or the average label (in case of regression). The closeness of two points is given by a distance function which can be calculated with Euclidean distance (Burkov, 2019).

In (Amato, Gennaro, & Falchi, 2015)a paper for the monument recognition is given. There are two new proposed approaches based on kNN classification for the reported problem. The first approach exploits kNN classification to classify images and relies on a relaxed definition of the local feature based image to image similarity definition, which allows efficient index for similarity search to be used. The second approach that they propose, called Local Features Based Image Classifier, uses kNN classification to classify individual local features of an image, rather than the entire image. The results of this experiment conducted in a cultural heritage scenario revealed that the approaches of this paper had better performance than other state of art approaches who used kNN algorithm.

Naïve Bayes Classifier

A Naive Bayes classifier is a probabilistic machine learning algorithm that is used for classification tasks.

This learning method involves a learning step in which the various P(vj) and P (ai l Jv) terms are estimated, based on their frequencies over the training data. The set of these estimates corresponds to the learned hypothesis. This hypothesis is then used to classify each new instance by applying the rule in Equation below.

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

There are different types of naïve Bayes Classifiers which can be used for a variety of classification problems. These types are the Multinomial Naïve Bayes mostly used for document classification, Bernoulli Naive Bayes where the predictors are Boolean variables, and Gaussian Naïve Bayes, when the predictors take up a continuous value and are not discrete.

Support Vector Machines

Support Vector Machines is another common machine learning algorithm which can be used for both regression and classification problems.

The input data is mapped to a very high dimension feature space (Cortes & Vapnik, 1995). The aim of the SVM algorithm is to find a hyperplane – decision boundary- in this feature space, which helps to classify the labeled into data.

There is a big number of hyperplane positions that can be chosen but the goal is to choose the most optimal one. The one that is with the maximum margin (maximum distance between data points of the classes) fits the best and is the most optimal because is robust to outliners and has strong generalization ability. Formula of the hyperplane is :

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + \mathbf{b} = 0, \text{ where } \mathbf{w} \in \mathbf{R}^{\mathrm{d}}, b \in \mathbf{R}$$

.

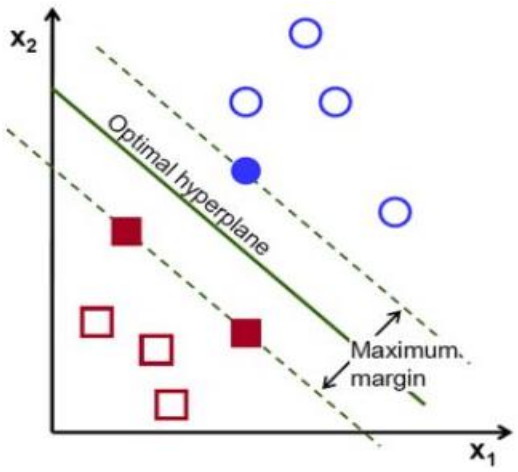


Figure 7. SVM hyperplane classification

SVM algorithm can also be used when the input data is not linearly separable in high dimensional feature space.

In this case Kernel function or else called Kernel trick which is defined as a function that corresponds to a dot product of two feature vectors in some expanded feature space. Turning the non-separable data to linearly separable.

Depending on the problem there are different types of Kernel functions, because different types of Kernel functions can lead to different results.

SVM has been used recently for landmark recognition too. In their paper Crudge, Thomas and Zhu describe that they used SVM because it was very efficient when dealing with the high dimensional feature space.

They training data was built of 193 images collected from Google and was put into the SVM as a vector that contained all labels and a matrix whose rows where examples and whose columns where features (Crudge, Thomas, & Zhu , 2014)

The experiment was tested also in three other classifiers but the highest accuracy of 92% the system had with the support vector machines.

Decision Trees

Decision trees are a widely used models for classification and regression tasks. This type of algorithm is used for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Learned trees can also be re-represented as sets of if-then rules to improve human readability.

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute in the given example. This process is then repeated for the subtree rooted at the new node.

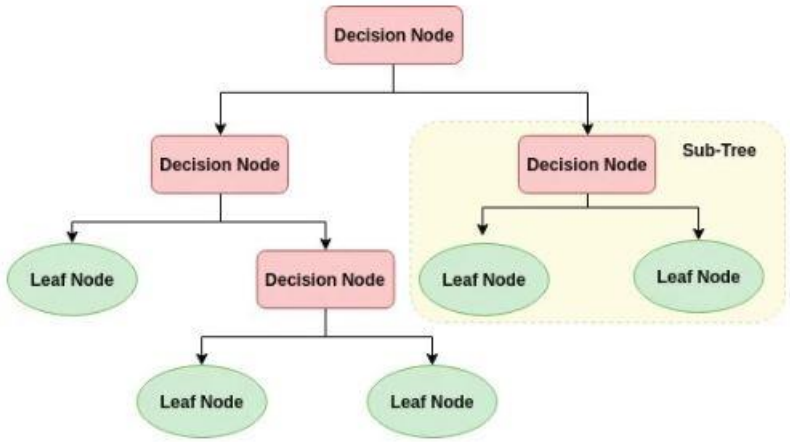


Figure 8. Decision Tree structure

The core algorithm for building decision trees called ID3. to answer the question which attribute is the best classifier the ID3 defines a statistical property, called information gain, that measures how well a given attribute separates the training examples according to their target classification. The algorithm uses this information gain measure to select among the

candidate attributes at each step while growing the tree. But to define the information gain precisely we should calculate the entropy which measures the purity or impurity of samples.

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Decision trees are easy to interpret and visualize. It requires fewer data preprocessing from the user but is sensitive to noisy data and can overfit it. It finds application in different problems but there has not been any research or paper about monument recognition which used decision tree as the algorithm for it.

Artificial Neural Networks

Artificial Neural Networks called also just Neural Networks provide a general, practical method for learning real-valued, discrete-valued, and vector-valued functions from examples (Mitchell, 1997).

The structure and functionality of this algorithm was been inspired by the biological learning systems, more specifically from the human brain.

The connection between the neurons, information sharing from one neuron to another and the structure of them has led to create such an algorithm which is motivated and imitates the work of this biological neural system.

To explain the neural networks one must start with the artificial neuron called a perceptron.

A perceptron takes a several binary inputs and produces a single binary output. As shown in the figure abow the perceptron has three inputs x1, x2, x3 (Nielson, 2015).

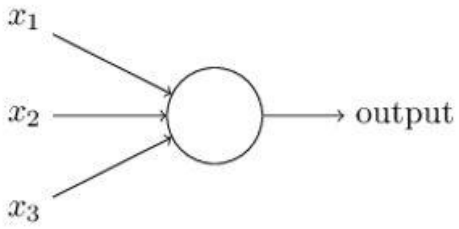


Figure 9. Perceptron structure

In general, it could have more or fewer inputs. In each input values there are also weights, real numbers expressing the importance of the respective inputs to the output. The neuron's output, 0 or 1, is determined by whether the weighted sum ∑j wj*xj is less than or greater than some threshold value. Just like the weights, the threshold is a real number which is a parameter of the neuron. In algebraic terms the results of the outputs depending on the threshold value can be seen in the below picture (Nielson, 2015).

$$\text{output} = \begin{cases} 0 & \text{if} \quad \sum_j w_j x_j \leq \text{threshold} \\ 1 & \text{if} \quad \sum_j w_j x_j > \text{threshold} \end{cases}$$

To get the approximately desired outputs the perceptron learning rule must be applied. Starting with the random weights, the perceptron is iteratively applied to each training example, modifying the perceptron weights whenever it misclassifies an example. This process is repeated, iterating through the training examples as many times as needed until the perceptron classifies all training examples correctly. Weights are modified at each step according to the perceptron training rule (Mitchell, 1997).

$$w_i \leftarrow w_i + \Delta w_i$$

where

$$\Delta w_i = \eta(t - o)x_i$$

As it can be seen perceptron is a very basic algorithm. But when there is a problem with multiple inputs of huge datasets, a perceptron can not deal with them. Because of these problems, not a single perceptron but the neural network is applied.

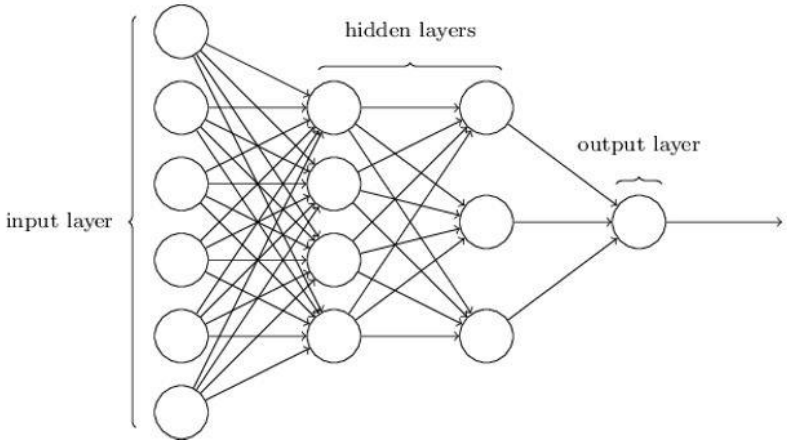


Figure 10 . Neural Network structure

NN as shown in the Figure 10 abow has an architecture built from the input layer, hidden layer and the output layer and this is why they are sometimes called Multi Layer Perceptrons too.

MLP are usually learned by the Backpropagation algorithm and are capable of expressing a rich variety of nonlinear decision surfaces.

The Backpropagation algorithm is such a widely used iterative algorithm which learns the weights for a multilayer network, given a network with a fixed set of units and interconnections. It employs gradient descent to attempt to minimize the squared error between the network output values and the target values for these outputs (Mitchell, 1997). The weights are not changed all at once but rather incrementally. A weight is changed depending on the influence it has on the error always trying to minimize it (Makin, 2006).

The error is the difference between the actual result of the output y and the desired output tj activation.

$$E := \frac{1}{2}\sum_{j=1}^{J}(t_j - y_j)^2.$$

In (Rojas, 1996)is claimed that this algorithm can be decomposed in four steps. Feedforward computation, backpropagation to the output layer, backpropagation to the hidden layer and weight updates. And the algorithm stops when the value of an error function becomes small.

Neural Networks are one of the algorithms which have yielded with the best results when we have to deal with classification problems. Although in experiments about the landmark recognition problems this algorithm has not been seen a lot. Since a large data set contained there has been used more advanced types of Neural Networks which will be mentioned below.

2.4 Deep Learning

Deep Learning is a branch of machine learning family that teaches the computer to learn from the experience. When we have a large and complex data set it gets harder and harder to train it with the shallow neural networks. Because of this the need for the better and more powerful nets arose.

It is called deep learning because it refers to training neural networks with more than three hidden layers.

Deep Learning is used almost as other machine learning algorithms in fields such as computer vision, finances, signal processing, online services and so on. But the only difference is that we use deep learning when we have a lot of data because it can be very expensive from a computational point of view to always apply it for every small task (Health, 2018).

There are different types of deep learning neural network architectures as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks for which will be talked next

Convolutional Neural Networks

Convolutional Neural Networks is a deep learning algorithm mostly used for face recognition, image classification. Basically it is heavily used in computer vision systems.

These networks use a special architecture which is particularly well-adapted to classify images. Using this architecture makes convolutional networks fast to train. This, in turn, helps us train deep, many-layer networks, which are very good at classifying images (Nielson, 2015) .

ConvNet is built from a set of layers. The neurons of these layers are in three dimensions and this is one of the great advantages of convolutional neural networks.

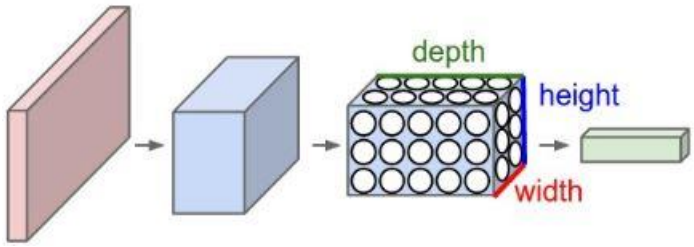


Figure 11. Three dimensionality of CNN layers

In the CNN models the input data passes through a lot of layers so technically when we train and test with CNN we do not need the preprocessing techniques because the system has already the ability to learn from the existing features when the image goes from one layer to another. And in the end the classification is fulfilled.

There are different layers with different functionalities in convolutional neural networks. The input layer, convolution layer, pooling layer, fully connected layer softmax layer and output layer.

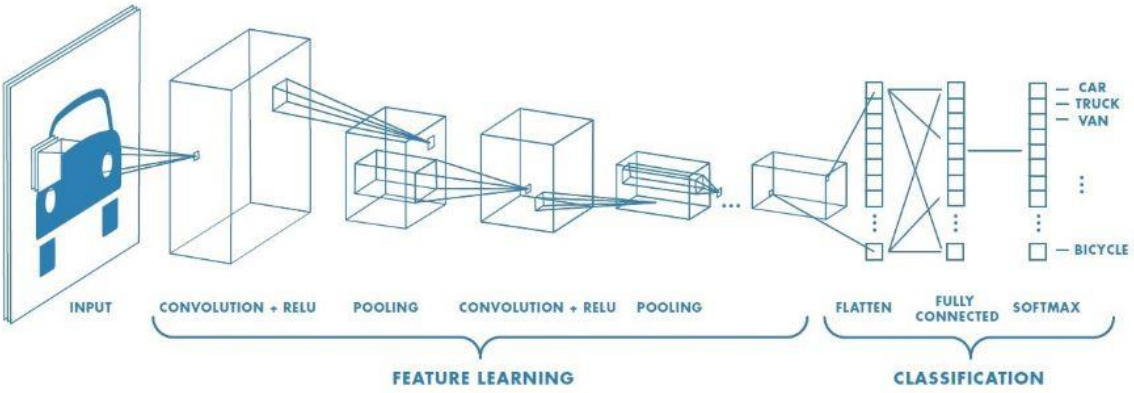


Figure 12. Convolutional Neural Network with its layers

Convolutional Layer – is the first layer used for extracting features from the input image. It is a mathematical operation which uses kernel as a filter matrix and moves to the right through all of the pixels until it completes the width of input. The output of this layer is the image reduced without losing any feature so it could be later easier processed.

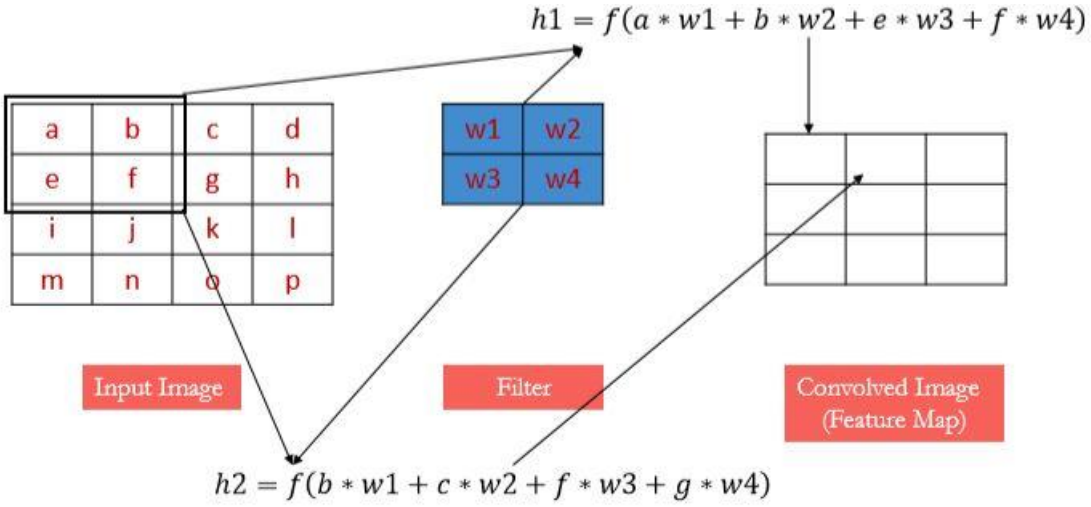


Figure 13. Process of Convolutional Layer

Pooling Layer – This layer is connected after the convolutional layer and reduces the computational cost by reducing the number of parameters.It uses a 2x2 filter and has two types , the max pooling and average pooling . The max pooling reports the maximum output within a rectangular neighborhood of the input and the average reports the average of it .
In the Figure below can be seen the filtered output result with max pooling applied .

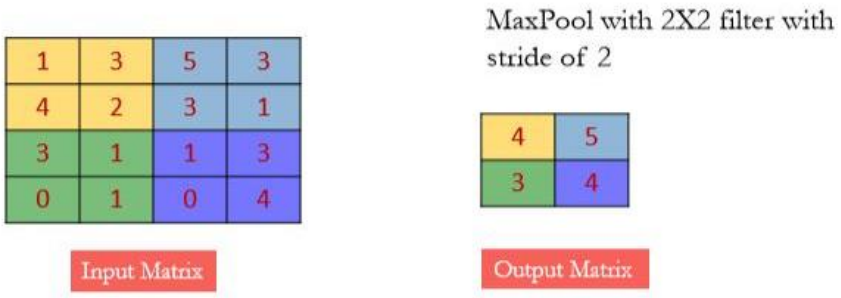


Figure 14. MaxPooling Layer

In the Fully Connected Layer and the Softmax Layer happens the classification process . Combined with the connection of weights and neurons they classify if the input belongs to class a or class b .
The structure represented below is the basic architecture or the sceleton of the convlutional neural networks . Deppending on the layers there are also different types of CNN such as the AlexNet , GoogLeNet , SqueezeNet , NU-LiteNet- A etc which are used a lot for different types of problems.

The convolutional neural network has also been applicable for the landmark recognition systems.
In (Termritthikun, Kanprachar, & Muneesawang, 2018)the NU-LiteNet A and B architectures of ConvNet were used to classify 50 landmarks of Singapore . There has been used only two modules and the application was succesful approached with the accuracy of 92.75 for NU-LiteNet A and 93.96 for the NU-LiteNetB.

Another approach for the recognition of temples in India has used convolutional neural network's Alex-Net architecture. It was able to achieve the accuracy of 92.7%. In the picture below can be seen the architecture of the CNN created for this experiment.

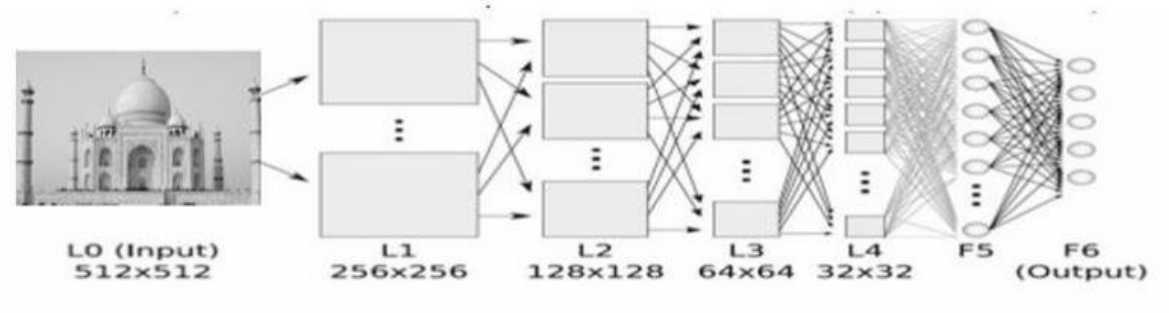


Figure 15. CNN applied for Indian tamples recognition

There are also a lot of other papers which propose the deep convolutional neural networks for the landmark recognition system as the most suitable algorithm since from this algorithm later can be created mobile applications for different types of monuments.

3 PROBLEM STATEMENT

During the process of the research it has been seen that there a lot of areas where the computer vision processes are being applied. Internationally and also in Kosovo data science is attending the market day by day with a relatively increased number of projects in sectors as medicine, finances, security, automation and so on. But when we speak about tourism, also as seen in the literature review chapter there are not many work related to this problem, neither in the world nor in Kosovo.

Kosovo, as one of the newest countries of the world owns a treasure of history, tradition and nature.

Every part of it has a different story to tell. So is the city of Prizren, or else the capital of culture. It has a multicultural and multiethnical history which makes it special from other cities.

But the world is still not well informed about this small city in the heart of the Balkans or the ones that visit Prizren sometimes have a problem to find a tour guide or look for the real story of the monuments of it.

There are barely any touristic applications in app stores which one could use as a tourist when he/she visits the city of Prizren even though the mobile device usage for different activities is growing day by day.

A mobile application which could help the tourists of Prizren to identify the monuments they see while they walk through the city, would be a good solution.

By just taking a picture of these monuments from any angle the person is standing, will not jut be helpful and more fun to better discober the city, but also an innovative and efficient project for the municipality of Prizren.

4 METHODOLOGY

This part of the thesis presents the research approach, strategy and the method of classification used to analyze and accomplish this thesis.

Because of the main purpose of this thesis is to classify monuments of the city of Prizren, there is formed a data set with images that were photographed with a camera of a mobile phone.
There are used four different monuments such as two ottoman mosques, one catholic cathedral and one orthodox church. The monuments are chosen this way to form a small variety of classes.
For training of these image it has been used the Matlab R2019a together with its apps and toolboxes.
Before training all the images have been resized with Matlabs Image Batch Processor app.
And for the classification algorithm the Convolutional Neural Networks has been used as being more practical to implement.

5 RESULTS

In this chapter will be discussed about the final results of the project. More precisely here will be described about the accuracy and outcome of classification which is made from our input dataset.

- *Data Set*

The data set of this program is created from the images photographed from the mobile phone. It includes four classes as four different monuments. The Gazi Mehmet Pasha Mosque, Sinan Pasha Mosque, Cathedral of Our Lady of Perpetual Succour and St. George Church. Each class has 15 images taken from different angles and different times of the day.

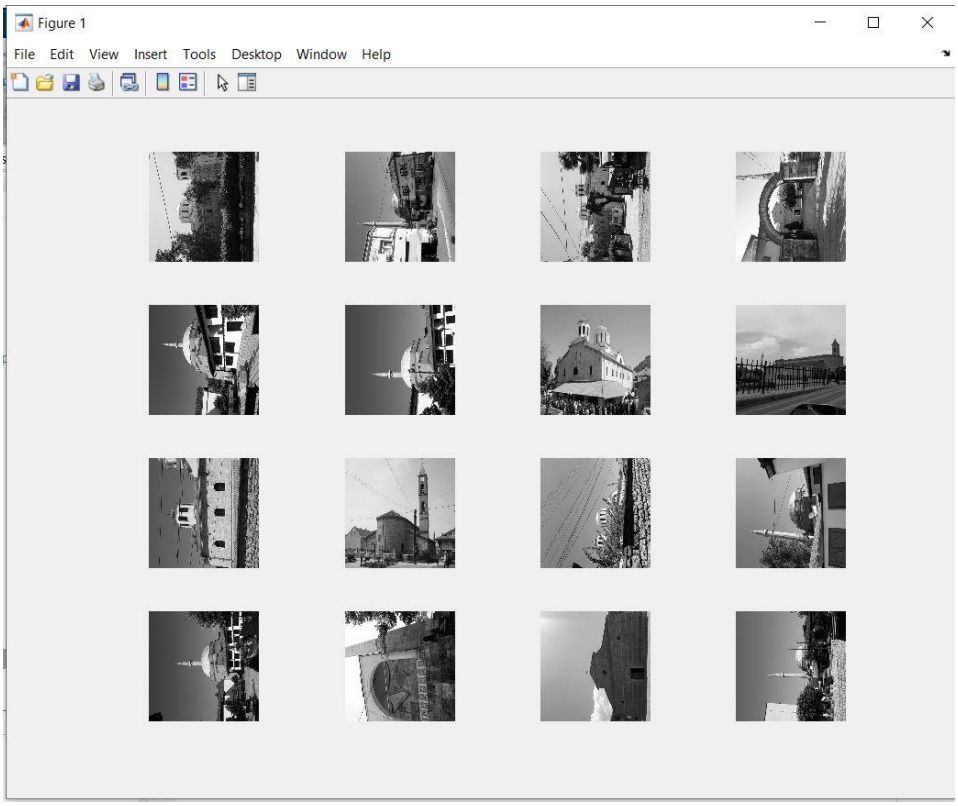


Figure 16. Image Dataset

- *Pre processing of the input data*

The input images were all in different sizes and in RGB format.

Since the size of the input images for the input layer must be the same with the size of inputs for the outputlayer the images are resized all in the same size and converted into grayscale images.

For resizing and color converting, Image Batch Processor App of Matlab has been used.

The images were uploaded into the app and function for the mentioned processes is been created. All the images were converted int grayscale and resized with 500x500 dimensions.

```
function results = myimfcn(im)
%Image Processing Function
%
% IM      - Input image.
% RESULTS - A scalar structure with the processing results.
%

%--------------------------------------------------------------------------
% Auto-generated by imageBatchProcessor App.
%
% When used by the App, this function will be called for every input image
% file automatically. IM contains the input image as a matrix. RESULTS is a
% scalar structure containing the results of this processing function.
%
%--------------------------------------------------------------------------



% Replace the sample below with your code---------------------------------

imgray=rgb2gray(im);
results=imresize(imgray,[500 ,500]);


%--------------------------------------------------------------------------
```

Figure 17. Function for the Image Batch Processor

- *Data Labeling*

For the classification process it has been used the Convolutional Neural Networks from the Deep Learning Toolbox of Matlab.
From the folder 'classification_dataset' all the subfolders with images have been called and stored in a datastore. In the figure below can be seen that all the images are labled and in the same size.

```
ans =

  4×2 table

                                Label                                Count
    ____________________________________________________________    _____

    Cathedral of Our Lady of Perpetual Succour                       15
    Gazi Mehmet Pasha Mosque                                         15
    Sinan Pasha Mosque                                               15
    St George Church                                                 15


ans =

   500   500
```

Figura 18. Label and size results of images

- *Spliting of the dataset*

Each label was randomly splited into training and validation set but the number of training files is given as 11.

```
numTrainFiles = 11;
[imdsTrain,imdsValidation] = splitEachLabel(imds,numTrainFiles,'randomize');
```

- *Architecture of the CNN*

The architecture of the convolutional network is shown in Figure 18.

In the *imageInputLayer* is specified the size of image 500x500 and 1 for the grayscale image.

There are used three convultional layers as *convlution2dLayer* with the filter size equal to 3, Padding is used to add padding to the input feature map and 'same' to insure that the spatial output size is the same as the input size.

Together with convolutional layers are used *batchNormalizationLayer*, makng network training an easier for optimization and *reluLayer a*ctivation function.

As pooling layer is used the maxPooling2dLayer with the rectangular shape of 2x2 and 'Stride' for specifying the step size that the training function takes as it scans along the input.

FullyConnectedLayer is equal to 4 because as input there are only four classes and has a *softmaxLayer* to normalize the output of it. And the final layer is the classificationLayer.

```
layers = [
    imageInputLayer([500 500 1])

    convolution2dLayer(3,8,'Padding','same')
    batchNormalizationLayer
    reluLayer

    maxPooling2dLayer(2,'Stride',2)

    convolution2dLayer(3,16,'Padding','same')
    batchNormalizationLayer
    reluLayer

    maxPooling2dLayer(2,'Stride',2)

    convolution2dLayer(3,32,'Padding','same')
    batchNormalizationLayer
    reluLayer

    fullyConnectedLayer(4)

    softmaxLayer
    classificationLayer];
```

Figure 19. CNN architecture

- *Training of the network*

For the training process is used the Learning rate which is equal to 0.0002 and a maximum number of Epoch equal to 10.

```
options = trainingOptions('sgdm', ...
    'InitialLearnRate',0.00002, ...
    'LearnRateSchedule', 'piecewise', ...
    'MaxEpochs',10, ...
    'Shuffle','every-epoch', ...
    'ValidationData',imdsValidation, ...
    'ValidationFrequency',30, ...
    'Verbose',false, ...
    'Plots','training-progress');
```

Figure 20. Training option

- *Accuracy*

The labels of the validation data in the trained network are used for prediction. And to find out how good the program predicted the accuracy is calculated.

After running the program for sometimes, the highest prediction the program has showed a is 62.50 % with accuracy = 0.6250 for the validation set.

The training time with 10 iterations is 53 secs and the accuracy of the training set is equal to 1 in 10^{th} iteration which means that the program predicts 100% for the training set. This can be seen in the first graph of the Figure 21 together with other specific details of the training process. Also in Figure 22 can be seen the values of the test set and the prediction that the system made after training in the network.

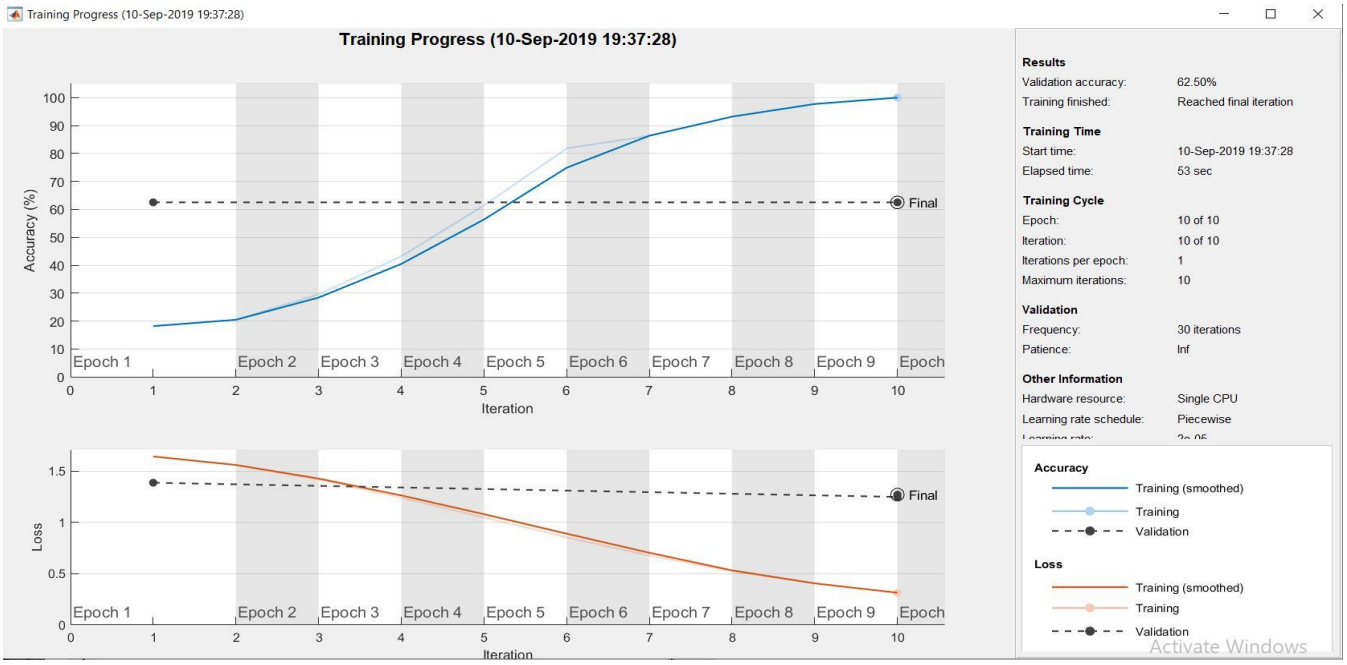


Figure 21. Training Progress

```
>> imdsTest.Labels

ans =

  21×1 categorical array

     Cathedral of Our Lady of Perpetual Succour
     Cathedral of Our Lady of Perpetual Succour
     Cathedral of Our Lady of Perpetual Succour
     Cathedral of Our Lady of Perpetual Succour
     Cathedral of Our Lady of Perpetual Succour
     Cathedral of Our Lady of Perpetual Succour
     Gazi Mehmet Pasha Mosque
     Gazi Mehmet Pasha Mosque
     Gazi Mehmet Pasha Mosque
     Gazi Mehmet Pasha Mosque
     Gazi Mehmet Pasha Mosque
     Sinan Pasha Mosque
     Sinan Pasha Mosque
     Sinan Pasha Mosque
     Sinan Pasha Mosque
     Sinan Pasha Mosque
     St George Church
     St George Church
     St George Church
     St George Church
     St George Church
```

```
>> YPred = classify(net,imdsTest)

YPred =

  21×1 categorical array

     Cathedral of Our Lady of Perpetual Succour
     Cathedral of Our Lady of Perpetual Succour
     Cathedral of Our Lady of Perpetual Succour
     Gazi Mehmet Pasha Mosque
     Cathedral of Our Lady of Perpetual Succour
     Cathedral of Our Lady of Perpetual Succour
     Gazi Mehmet Pasha Mosque
     Gazi Mehmet Pasha Mosque
     Gazi Mehmet Pasha Mosque
     Cathedral of Our Lady of Perpetual Succour
     Gazi Mehmet Pasha Mosque
     Sinan Pasha Mosque
     Sinan Pasha Mosque
     Sinan Pasha Mosque
     Sinan Pasha Mosque
     Sinan Pasha Mosque
     St George Church
     St George Church
     St George Church
     St George Church
     Cathedral of Our Lady of Perpetual Succour
```

Figure 22. Test Set and the Predicted Test Set values

6 DISCUSSIONS AND CONCLUSIONS

With the obtained results we can come to the conclusion that the created system with this data set of images is only 62.50% correct.

Even though the prediction was higher than 50%, when we first created the network we gave a learining rate of 0.01. This value was too big and after a numerous time of changing the number of epochs and running the program several times, it could always give us the accuracy of 0.25.

The learning rate of 0.0002 shows us that the program this number and type of collected images does not need a high value of the learning rate.

To increase the value of the accuracy for a better classification of the monuments pre processing of the images can be proposed.

Even though the input images are trained with Convolutional Neural Networks and this deep learning algorithm exctracts the features of the images in its layers, it has resulted that this is not enough for our dataset.

One of the methods for preprocessing before the CNN could supposably be extracting the SURF points from the images. By giving a constant rate for the points this feature detection algorithm could extract the most useful features from the monument images and better prepare for the network.

Another good option could be the segementation of the images.

With segmentation algorithms we could segment the monuments in the image, separate the background from it so this way the classification could predict which monument it is with a higher rate.

7 REFERENCES

Amato, G., Gennaro, C., & Falchi, F. (2015). Fast Image Classification for Monument Recognition. *Journal on Computing and Cultural Heritage*.

Ballard, D. H., & Brown, C. M. (1982). *Computer Vision.* New Jersey: Prentice Hall.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book.* Andriy Burkov.

Chen, T., Wu, K., Yap, K.-H., Li , Z., & Tsai, f. S. (2009). A Survey on Mobile Landmark Recognition for Information Retrieval . *Tenth International Conference on Mobile Data Management: Systems, Services and Middleware.* Singapore: Nanyang Technological University.

Cortes, C., & Vapnik, V. (1995, September). Support-Vector Networks. *volume 20*, pp. 273-297.

Crudge, A., Thomas, W., & Zhu , K. (2014). *Landmark Recognition Using Machine Learning.*

Gonzalez, R. C., & Woods, R. E. (2002). *Digital Image Processing 2nd Edition.* Prentice Hall.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. (2006). *Feature Extraction: Foundations and Applications.* The Netherlands: Springer.

Hartigan, J. A., & Wong, A. M. (2012, 01 18). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28,*, pp. 100-108.

Health, N. (2018, August 7). *ZDNet*. Retrieved from https://www.zdnet.com/article/what-is-deep-learning-everything-you-need-to-know/

Kaur, D., & Kaur, Y. (2014). Various Image Segmentation Techniques : A review , vol.3. *International Journal of Computer Science and Mobile Computing*, 809-814.

Makin, J. G. (2006, 02 15). *Backpropagation*. Retrieved from http://www.cs.cornell.edu/courses/cs5740/2016sp/resources/backprop.pdf

Mitchell, T. (1997). *Machine Learning.* USA: McGraw-Hill .

Muller, A. C., & Guido, S. (2016). *Indtroduction to Machine Learning with Python.* Sebastopol: O'Reilly Media, Inc. .

Nielson, M. A. (2015). *Neural Networks and Deep Learning.* Determination Press.

Petrou, M., & Petrou, C. (2010). *Image Processing : The Fundamentals 2nd edn.* West Sussex: John Wiley&Sons Ltd.

Rojas, R. (1996). The Backpropagation Algorithm. In *Neural Networks.* Berlin : Springer-Verlag.

Sebe, N., Cohen, I., Garg, A., & Huang, T. S. (2005). *Machine Learning in Computer Vision.* Netherlands: Springer.

Solomon, C., & Breckon, T. (2011). *Fundamentals of Digital Image Processing , A practical approach with examples in Matlab.* West Sussex: John Wiley & Sons Ltd.

Termritthikun, C., Kanprachar, S., & Muneesawang, P. (2018). *NU-LiteNet: Mobile Landmark Recognition using Convolutional Neural Networks.*

Young, I.T., Gerbrands, J.J., van Vliet, L.J. (1998). *Fundamentals of Image Processing.* Delft: Delft University of Technology.