Museum City

**Authors**: Bogdan Nicoara and Waldmilson Vilhete

**Subject** : Multimedia Computing

**Emails**: [bogdancristi012@gmail.com](mailto:bogdancristi012@gmail.com) and [w.vilhete@campus.fct.unl.pt](mailto:w.vilhete@campus.fct.unl.pt)

**Abstract**

We live in an era with access to the internet in almost every corner of our world and with devices in our pockets that have more computational power than the ones used to land on the moon in 1969. With this easy access to technology we can ease our life and improve it in every aspect. One of the main components that emerged in the last decade and has now continued to improve is our ability to make programs that perceive the world in some way, and facilitate access to it for us. From AI assisted fighter jets like the F35, to self driving cars such as Tesla, or now, the emergence of Chat GPT , a bot that has the potential to improve enormously our day to day life we move in the direction of something a bit more artistic and we pose the question: So why not use it for something fun? Why not make something that could help us navigate our cities in a new interactive way?

**Introduction**

Our aim with this project is to build, through openframeworks, an application that provides in a more interactive way, features that allows users to discover and search for contents concerning monuments and buildings that take important roles in different parts of the world.

With the support of some algorithms and image recognition systems, the application allows users to upload different pictures, whose contents will then be displayed in an interactive way, allowing them to consult relevant details and history behind the images. As a simple use case, this is a short description of the system flows:

* A user uploads a picture of a relevant monument to the application
* The image is processed and recognized
* Different relevant contents concercing the story of the image are displayed
* Users can play with the application by navigating over contents, videos, images…

Looking from a technical perspective, once the image is uploaded by the user, computer vision is used to interpret and categorize the image through certain traits, which allows us to find out if it’s similar to what we already have stored. Once we were able to compare the image and do the matching, we are then able to grab the image corresponding information and display them to the user.

**Overview of related work**

Computer vision has been shaping the way we use technology to interact with the external world. Many projects have been built throughout the world, using different techniques, such as, image segmentation, Machine/Deep learning, Feature detection and matching, and image recognition to extract from images useful information about the physical world, including meaningful description of physical objects.

One technology that started emerging as state-of-the-art in the image classification is ViT, or Visual Transformers. Transformers are a type of neural network that were at first focused on language modeling by creating a way in which to evaluate how each word interacts with each other word in a sentence or phrase. The way it achieves this is a complicated process,[1] but the main idea is that each word has a key, query and value vectors associated which are used to store how important and connected are all the other words relative to the current one. This is obtained also using weighted sums with the values of all the other elements (in this case words).

Say now that instead of words we use fragments of images. Then, as before, all the fragments have somekind of corelation between them that makes our model understand if a group of pixels. This is what a Visual Transformer is. Testing this on a series of objects and comparing to the results a CNN gives, we get the conclusion that it has the potential of behaving better, also with a better accuracy.[2] The next step in this process was to analyse what the results of self-supervised ViT’s could do for us, and this proves to be a break-trough compared to the CNN’s.

The way this works is to have like before a ViT, but this time, instead of tring directly to get a label from it, put it to recognise an image, we make it solve different problems like filling out cropped parts of the image. For this task, a method called DIDO is used. This means that at first, a “teacher” model creates certain descriptors for the objects in the cropped images that are then passed to a “student” model in a “distilled” way (basicly it sortes out the outstanding most caractheristics). What we end up with is a model that gets a very good understanding of the relation between parts of an image. We train this model on many images and then pass it’s data down to an image recognition or another kind of assignment which has a way easier time to assign a certain category to the objects [3]. This leads to one of the most advanced ways of categorizing and labeling an image there is on the market today.

One other of the articles we found that relates to computer vision is a blog posted by “Bolarinwa Oreoluwa”, a Machine Learning Engineer and Backend Developer. In the blog, they use Object Detection to identify and locate objects in an image or video. Object Detection can be considered as image recognition with some advanced features. Not only does the algorithm recognize/identify the objects in images/videos, but it also localizes them. In other words, a bounding box is created around the object in the image or video frames [5].

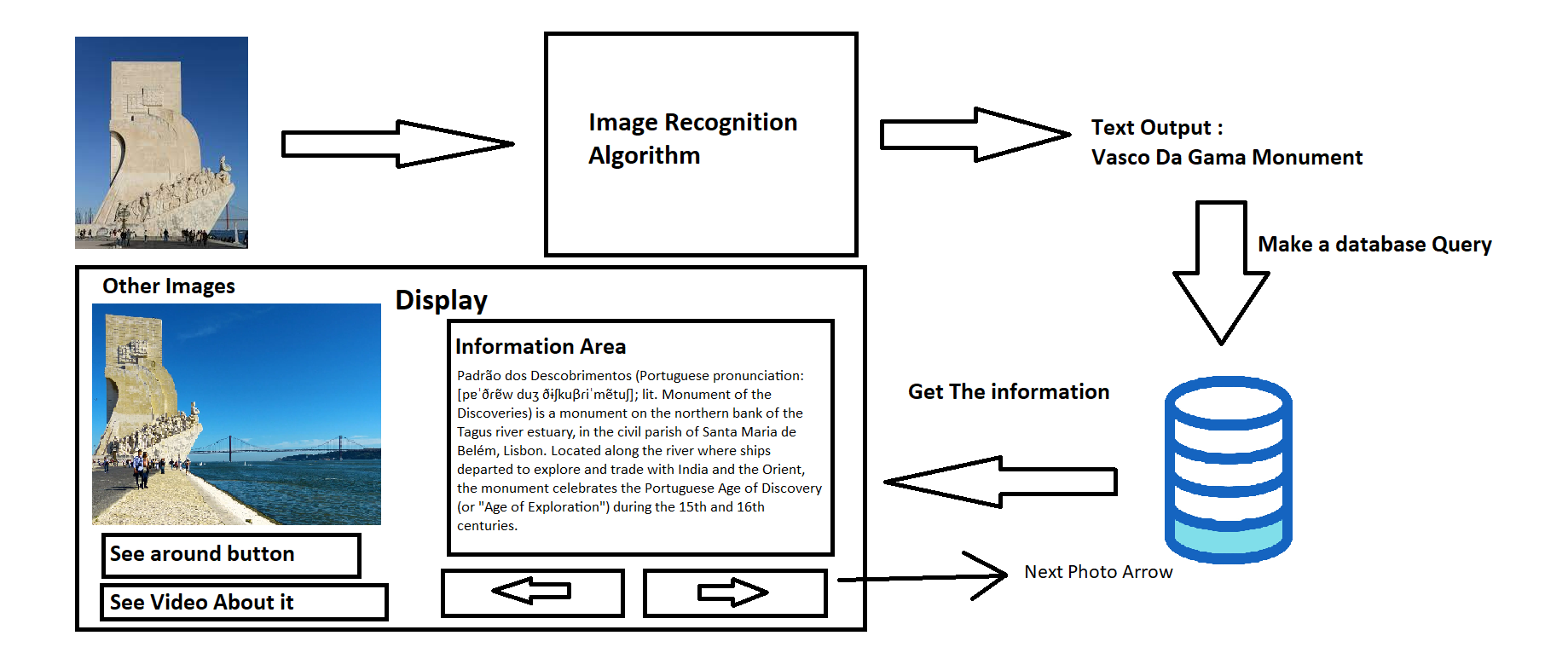
An object Detection Algorithm called YOLO (You Only Look Once) was used to produce real-time object detection. The algorithm is best known for its speed and accuracy. By locating and identifying objects, the algorithm can be used to count objects in a scene and to determine and track their precise locations, all while accurately labeling them. Furthermore, as a demo of the article, they show a crowded space with people moving/dancing and different kinds of objects moving, such as umbrellas, cars, and other vehicles, and it’s possible to recognize, and locate them in real time [6].

Another article that we found interesting is entitled “Sign Language Recognition Based on Computer Vision”. The aim is to elaborate on a project that uses computer vision domain technologies to implement sign language recognition to solve the problems of deaf people in daily communication. Sign language recognition is currently a hot topic in the field of machine learning and computer vision. In this project, two algorithm models, CNN+LSTM network structure and YOLOv5 target detection, are studied, and the functions of sign language recognition are implemented by the two algorithms respectively, and their implementation effects are compared, and the advantages and disadvantages of the two algorithms in sign language recognition are derived. Among them, YOLOv5 completes the task of sign language recognition by detecting hand movements, and its fast detection speed is more in line with the needs of life scenes and satisfies the demand for real-time sign language translation, which is more promising for applications [4].

In the project, they initially use CNN+LSTM classifier to implement the method of American Sign Language and Chinese. They first used the improved CNN algorithm for image feature extraction, and after that the optimal decision result is given by LSTM classifier to achieve sign language recognition. Since the recognition speed of this model is slow and does not match the application scenarios in life, they improved the detection of hand movements using the YOLOv5 algorithm.

The project uses the YOLOv5 target detection algorithm for hand motion detection, YOLO is a target detection method that is characterized by achieving fast detection while also achieving high accuracy. The project was pre-trained using the coco128 dataset, which contains the first 128 images from COCO Train 2017. The Chinese Sign Language dataset was collected by seven of our team members by shooting 30 Chinese sign languages in the Chinese Sign Language Finger Alphabet in different environments, and then labelling the data into YOLO format by themselves to produce more than a thousand datasets, and then partitioning the datasets into test and validation sets in proportion [4].

**Project Specification and sketch**

****

**Figure 1. Prototype Sketch**

**Application Implementation**

We used openCV to develop our project. The implementation for the interface is pretty basic. As a first version, we added some buttons to see the next and previous related image. There is a text part that appears with an image too. We have an area where you can drag the image. Then we cycle through our database and we first compare images on luminance. If they are not very different, then we compare them by key points. The way luminance comparison works is that it basically does an average over the entire image and compares it to the other one.

Comparing by key points is a bit more delicate as first we choose some key points in each image, then we see what are the characteristics of the pixels around them, and then we try to match them between images. If enough matches are made, we conclude that it is the same image. From there we get the name of the image and then we import the data from the file with the same name and we show similar pictures of the same building.

**Conclusion**

In conclusion, our paper presents the development of a multimedia computing application called Museum City, aimed at providing users with an interactive and engaging way to explore and discover information about monuments and buildings around the world. The application uses openframeworks and incorporates algorithms and image recognition systems to process and analyze user-uploaded images, presenting relevant details and historical information related to the depicted subjects.

From a technical standpoint, computer vision plays a crucial role in the functioning of the application. Through the use of computer vision techniques such as image segmentation, and image recognition, the application is able to extract valuable information from images, enabling users to gain insights into the physical world.

Museum City showcases the potential of multimedia computing and computer vision in providing users with an interactive and immersive experience in exploring the world's cultural heritage. The application opens up new possibilities for engaging with and understanding our surroundings. The ongoing advancements in technology and the growing availability of powerful computing devices continue to pave the way for innovative applications that enhance our daily lives and offer new ways to interact with the world around us.

With respect to papers that we studied, we cannot leave without mentioning the concept of Visual Transformers (ViT), a type of neural network that has shown promise in image classification tasks. By adapting the principles of transformers used in language modeling to fragments of images, ViT demonstrates improved accuracy and performance compared to traditional convolutional neural networks (CNNs).

We also studied a paper about the relevance of object detection algorithms, particularly the You Only Look Once (YOLO) algorithm, in computer vision applications. YOLO enables real-time object detection and localization, allowing for the identification and tracking of objects within images or video frames. This algorithm proves useful in scenarios where precise object labeling and counting are required, as demonstrated in the example of recognizing objects in crowded spaces.

Our project does not use all technologies, but we could in certain measure, apply the knowledge of what we’ve learnt on the computing multimedia subject and we hope to learn much more about openframework technology in another given time.

**Refferences**

[1] Attention Is All You Need - Ashish Vaswani, Llion Jones, Noam Shazeer, Niki Parmar, Aidan N. Gomez, Jakob Uszkoreit, Łukasz Kaiser, Illia Polosukhin. arXiv:1706.03762v5 [cs.CL] 6 Dec 2017

[2] AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE - Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. arXiv:2010.11929v2 [cs.CV] 3 Jun 2021

[3] Emerging Properties in Self-Supervised Vision Transformers - Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, Armand Joulin. arXiv:2104.14294v2 [cs.CV] 24 May 2021

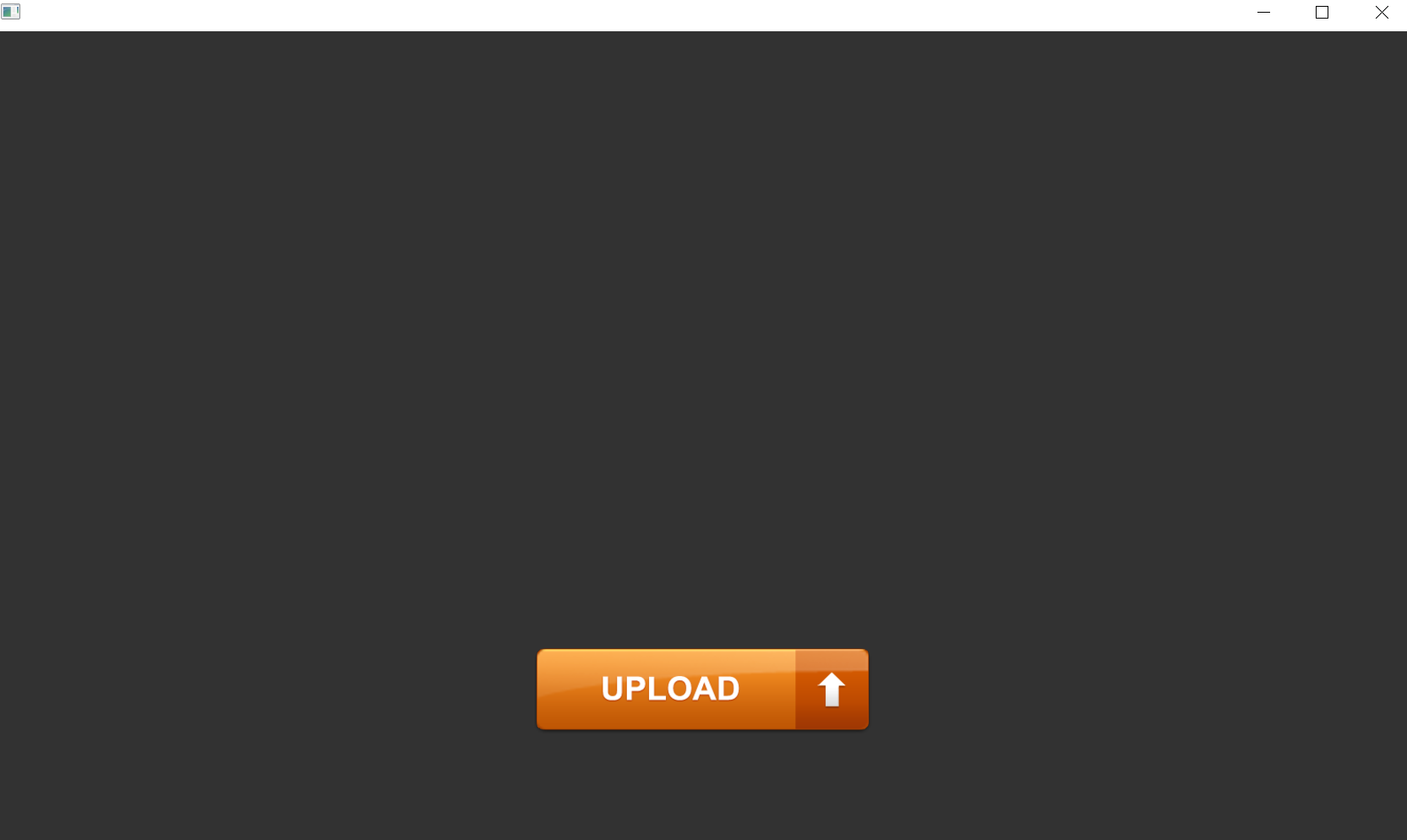
[4] Tengfei Lia, Yongmeng Yanb, Wenqing Duc. “Sign Language Recognition Based on Computer Vision- Zhengzhou University, Zhengzhou 450000, China” 2022, pp. 223-227 https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9498024&tag=1

[5] Gupta, Utkarsh. "A Simple Object Detection App Built Using Streamlit and OpenCV." Dev Genius, Feb. 10, 2021, last access date: 20 April 2023 https://blog.devgenius.io/a-simple-object-detection-app-built-using-streamlit-and-opencv-4365c90f293c

[6] zhoroh Bolarinwa Oreoluwa Seun. “Object Detection” 13 Feb 2022, last access date: 20 April 2023. https://github.com/zhoroh/ObjectDetection

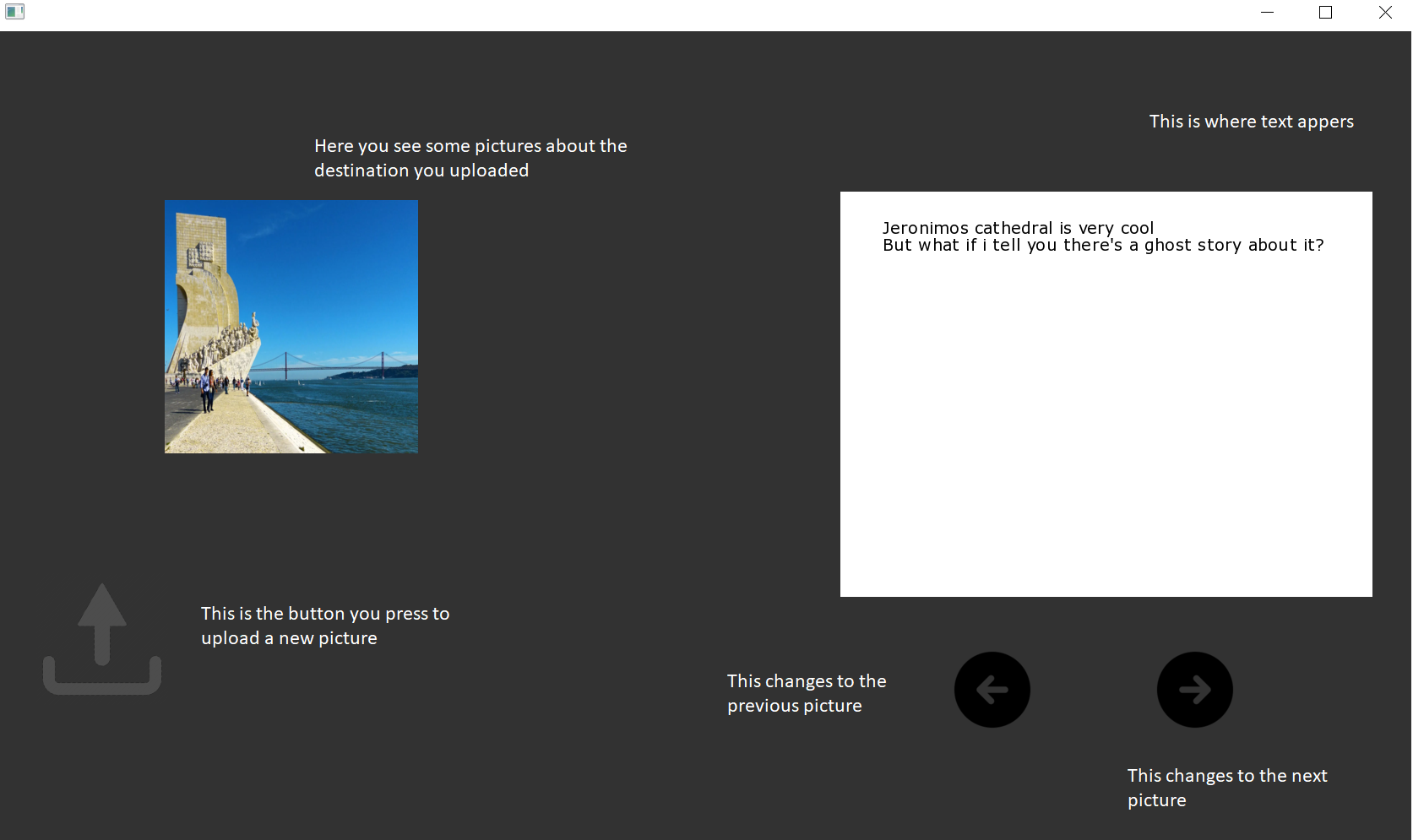
**Appendix A: User manual**

You simply drag and drop the photo you want to see data about in the up part of the screen and then you press upload.



**Figure 2. Upload Screen**

This loads the next part :

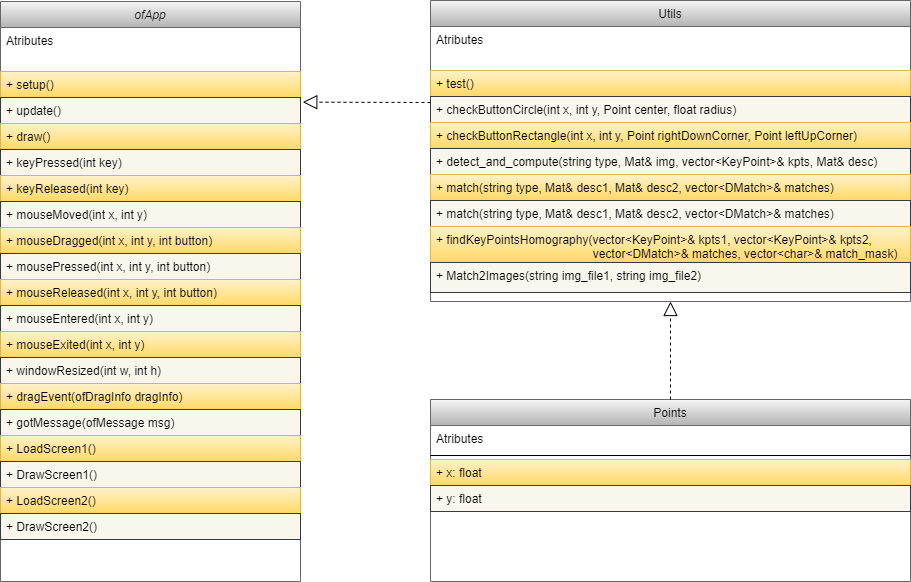


**Figure 3. Contents Screen**

Here , if we want to upload a new photo, we just press the upload sign.

The previous and next pictures are already in our database. If we don’t find the picture in our database the screen will clean and you will be back to the first drag and drop screen.

**Appendix B: Class Diagram**



**Figure 4. Class Diagram**