UNIVERSITAT POLITÈCNICA DE CATALUNYA

ESEIAAT

EgoMon Gaze and Video Dataset for Visual Saliency Prediction

Mònica Chertó Sarret

June 2016

UNDERGRADUATE PROJECT REPORT

Supervised by Xavier Giró and Cathal Gurrin

Acknowledgements

I would like to thank my project advisors Xavier Giró and Cathal Gurrin for their guidance and dedication to this project. Also I would like to thank Albert Gil and Junting Pan for their technical support and for attend all my queries about this project. Thanks to my Erasmus mates Marc Carné and Cristian Reyes for all their help to construct the dataset for my project.

Abstract

This project focuses on the creation of a new type of egocentric (first person) vision dataset. For that purpose, the EgoMon Gaze & Video Dataset is presented. This EgoMon dataset was recorded using the eye gaze tracking technology that studies the movement and position of the eyes. The Tobii glasses (wearable, eye tracker and head-mounted device) were the main tool used to record and extract the gaze data for this dataset. The dataset consists in 7 videos of 34 minutes each one of average, 13428 frames extracted from each video (with a frequency of 1 fps), and 7 files with the gaze data (fixations points of the wearer of the glasses) for each frame and video. The videos were recorded in the city of Dublin (Ireland) both indoor and outdoor.

The generated dataset has been used to evaluate the performance of a state of art model for visual saliency prediction on egocentric video.

Table of Contents

Table of Figures

Table of Tables

Chapter 1 – Introduction

The latest advances in electronics and communications have taken a high interest in wearable devices to recording and monitoring personal activities. Nowadays, electronic personal assistants typically combine information stored in the cloud with geolocation data captured by a smart phone, but the next generation will exploit information captured with wearable sensors. This project focuses in video cameras mounted on glasses.

However, a camera mounted on a glass is not enough to predict what is the interest of the user in the captured image. The human field of vision typically covers multiple objects at different scales and distances, while our vision system focuses only in one particular region of the full field of vision. This project presents a novel egocentric video and eye gaze dataset and uses it to test the performance of a state of the art system for visual saliency prediction.

This project was developed in the framework of an Erasmus+ program for mobility. It extends the research line started with the Bachelor Thesis by Sergi Imedio [1]. In that work, Imedio explored the problem of object detection with the same wearable device used in our work.

1.1. Project Planning

This project was divided in two different parts, one of them submitted in the Dublin City University on April of 2016 and its extension for the "Universitat Politècnica de Catalunya" (UPC) on June of 2016. The first part of this project is appended to this document.

While the first part focuses on the design, recording and post-processing of the dataset, its extension aimed at assessing the performance of a state of the art visual saliency predictor using the dataset.

These two main goals can be further detailed in the following tasks.

1. Construct a new and innovative egocentric dataset with a device with wearable video camera and sensor.
2. Run a state of the art visual saliency predictor with a single image.
3. Extract frames of each video that forms the dataset. Decide the rate of the frames extraction.
4. Run the saliency predictor with the frames extracted from the dataset.
5. Compare the results of the saliency estimator with the ground truth (data extracted from the videos of the dataset).

The scheduling of these tasks is described in the Gantt Chart of *Figure 1*.

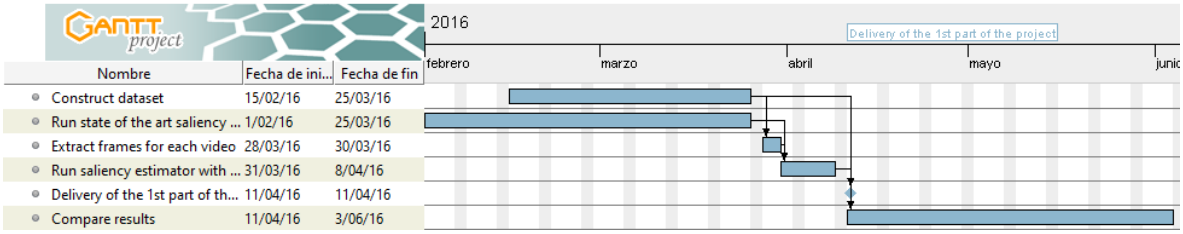

Figure 1. *Gantt Chart of the project*

1.2. Equipment and Software

This section explains the equipment used to create and record the dataset and the tools used for its post-processing.

1.2.1. Eye tracker: Tobii glasses

The Eye Gaze Tracking (EGT) is the technology that captures the movements and the position of the eyes (where the eyes are focused). This technology can be implemented using a head-mounted device, an eye tracker (as the device used to record the dataset for this project). An eye tracker extracts the point of gaze (eye position) in the scene where the user is looking at. The position of direction of the head and the direction of the eyes are used to determinate the gaze position.

The Tobii glasses *(figure 2)* were the main tool for construct the dataset for this project. This is a head-mounted eye tracking system. The eye-tracker is monocular, that means that the glasses only consider the position of only one eye, the right eye in this case. This device shows the exact point of gaze in real time and is complemented with a

recording assistant and an IR Markers. The recording assistant is used to handle the data and the processes from the glasses together with the video of the scene camera and IR Marker position to a memory card. The IR Makers are used to the required calibration process of these glasses.

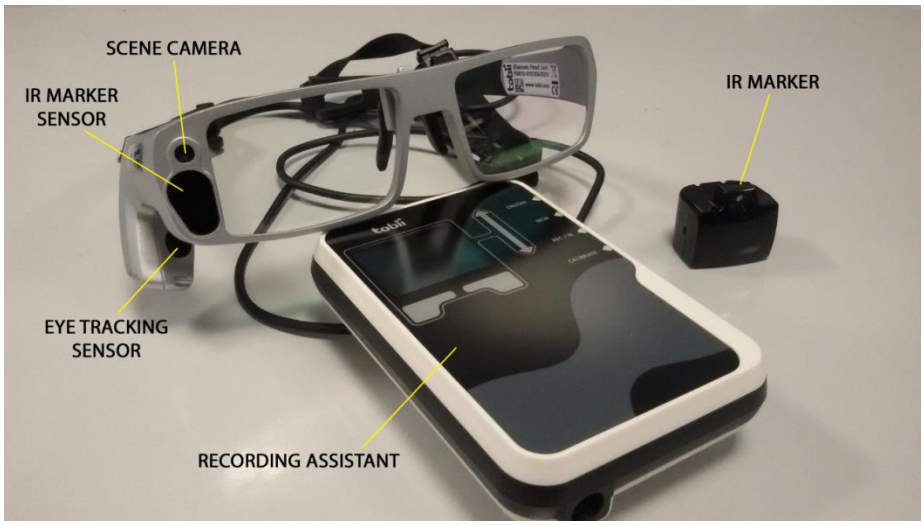

Figure 2. *Tobii Glasses*

The main features of the Tobii glasses are the following:[2]:

Frequency of the recording of the video	30fps
Tracking	Monocular sampling, 30 Hz
Visual angle	56ºx40º
Resolution of the resultant video	640 x 480
Maximum recording time	60 – 70 minutes
Calibration Procedure	System guided (9 points)
Tolerated angles of IR Markers	90 – 150 degrees (depending on viewing distance)
Supporting Tool	**Tobii Studio**

Table 1. *Main features of the Tobii glasses*

1.2.2. Tobii Studio Software

This software *(figure 3)* is used to support the extraction of the gaze data. This environment allows exporting the video with or without the point of gaze plotted, together with a text file with the gaze data in it.

The text file contains the necessary data to process the eye gaze tracking data. An example of this parameters are the resolution of the video the date and times, and the coordinates for each point of gaze and their corresponding timestamp. The timestamp is in milliseconds, and the coordinates are in pixels and between 0 and 640 for the X axis and between 0 and 480 for the Y axis of the image.

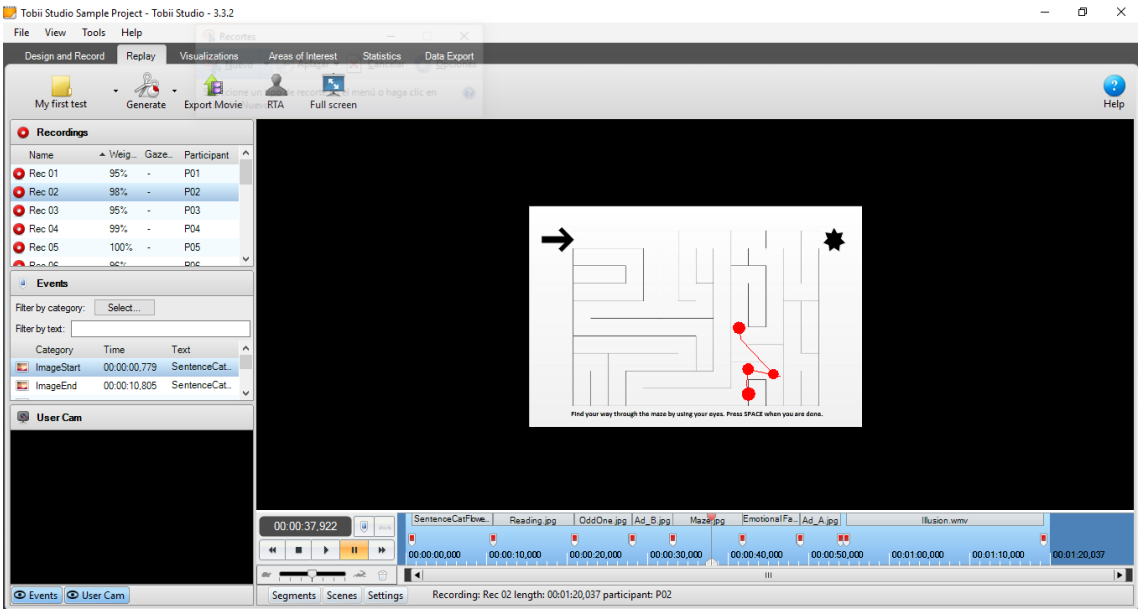

Figure 3. *Tobii Studio Software*

1.2.3. Narrative Clip

The Narrative Clip [31] is Small wearable camera that can be put through a clip on clothes. This camera is always on and takes images automatically each 30 seconds.

1.2.4. Data Analytics

FFmpeg

The FFmpeg tool was used for the frames extraction of the videos exported to the working computer. This tool is available on the basic Linux package and offers different usage options. One of these options is to extract frames from a video. It allows deciding some parameters before the frames extraction. One of these parameters is the rate of frames extracted.

Matlab

Matlab is a mathematical software tool and programming language developed by MathWorks. The scripts created for this project *(will be explained in the chapter 5)* are programmed.

GPI Servers

This project was developed in Dublin City University, but data processing was remotely performed on the computational servers provided by the Image Processing group (GPI) [6] of the "Universitat Politècnica de Catalunya" in Barcelona.

1.2.5. Publication: GitHub repository and project page

The code used on this project and the website of the dataset created in this project is allocated in a [1]repository of GitHub. GitHub is a platform that offers a distributed revision control and source code management that can be used by a command-line or a graphical interface. It also provides access control, collaboration features, task management, etc.

1. Repositori of Egocentric-saliency in GitHub [online] Available: https://github.com/imatge-upc/egocentric-saliency

Chapter 2 – State of the art

This section contains a study of the state of the work field of this project. This brief description of the technical background allows understanding the goals of this project.

2.1. Datasets

The main goal of this project is to create a new video dataset for egocentric vision. This section contains existing datasets in this domain.

2.1.1. GTEA (Georgia Tech Egocentric Activities) Gaze Dataset [8, 9]

The GTEA Gaze dataset was collected using Tobii eye-tracker glasses [2]. It consists of 17 sequences that were recorded while the wearer of the glasses was seating and preparing different types of meals. For example, pouring milk into cup. The dataset consists of the frames extracted at a sampling rate of 15 fps and the annotation is based on the number of extracted frames.

2.1.2. UT Ego Dataset (University of Texas) [10, 11]

This dataset was collected using the Looxcie wearable (head-mounted) camera [12] and contains four videos. Each video is 3-5 hours long, captured in a natural, uncontrolled setting. The videos capture a variety of activities such as eating, shopping, attending a lecture, driving, and cooking.

2.1.3. DogCentric Activity Dataset [13]

This video dataset focuses on dog activity, captured with a Go-Pro camera [14] attached to the back of a dog. The dataset contains 10 different types of activities such as walking in the street, drinking water, playing with a ball, etc. The videos were recorded with a different type of camera (a non-head-mounted camera) and in all videos appear the dog's head because the camera was located on it back.

2.2. Applications

There exist previous works that have explored the role of saliency (bottom-up) or attention (top-down) in the domain of egocentric vision. This investigations focus in the validity of using conventional saliency maps for every type of egocentric images or videos.

2.2.1. Yamada et al (ACCV 2010) [15]

This experiment tries to calculate the best option for predict the human egocentric visual attention. The dataset they used was recorded in the same environment: in a room *(figure 4)*. Four different persons (one at a time) sit on a chair and another different person walking randomly around the first person. The two subjects were looking all the room moving their head freely for one minute. This dataset was not studied for the creation of the dataset of this project because is a non-published dataset very similar than others studied.

Figure 4: *Example of a frame extracted from the dataset*

The saliency model they used for compute saliency maps for egocentric videos (based on Itti et al's model [16] and Harrel et al's model [17]) follows a hierarchical architecture. The first level makes a feature decomposition of the input image using simple linear filters. These filters decompose the image in static features (intensity, colour and orientation) and dynamic features (motion and flicker). In the second level features maps are computes from Gaussian pyramid. In the last stage, the final saliency map is obtained by combining the normalized features maps.

The results of the experiment show that the saliency maps are the best way to predict the visual human attention. But the conventional saliency map models detect better the static features than the dynamic features. For this reason they conclude that the motion introduced by the wearer (egomotion) is the responsible of the performance in the

dynamic case, and that a specific model should be trained to handle the egocentric saliency prediction in videos.

2.2.2. Matsuo (CVPR 2014). [18]

In this paper, the authors use visual saliency to improve the activity recognition from egocentric videos. A recent work [19] presented a method of predict key objects by hand manipulation. But this can make problems because not all the objects can be manipulated by hands and no all the objects are being manipulated are important objects for the visual human attention.

In this method, the visual attention is also calculated not only from static saliency, but also from dynamic ego motions. They made changes in the block diagram *(figure 5)* of the conventional method allowing focusing the analysis in the most informative part of the region [20].

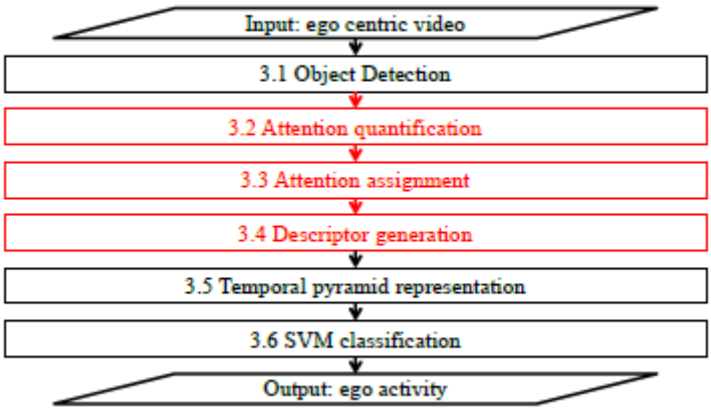

Figure 5. *Block diagram of the proposed method. The red parts are additional steps to achieve attention-based activity recognition.*

The dataset of this work consisted on 20 different persons recording egocentric videos in their own homes. This dataset was not studied for the creation of the dataset of this project because is a non-published dataset very similar than others studied.

They conclude that their method can predict saliency regions better than the hand-based approach.

2.2.3. Yu and Grauman (arXiv 2016). [11]

This work explores how a camera wearer is engaged with the environment. The authors claim that the eye-tracker methods allow to know what a person chose to look at but it does not immediately reveal when this person is engaged with the environment. This study proposes a learning-based approach to discover the connection between first-person motion and engagement *(figure 6).*

Figure 6. *Ego-engagement detection*

This paper uses the UT Ego Dataset *(view section 2.1.2.)* that consists in in multiple recordings taking videos under a set of scenarios: shopping in a market, in shopping mall and touring in a museum.

Chapter 3 – EgoMon Gaze & Video Dataset

The main contribution of this project is the construction of a new and innovative dataset using as reference other datasets constructed before *(view chapter 2)*. The dataset was created in different environments.

3.1. Acquisition

3.1.1. Data acquisition with Tobii Glasses

The data patterns recorded with an eye gaze tracker is influenced by the profile of the user. There are many studies [3, 4] of how the participant (depending on their age, visual problems, etc.) can change the accuracy of the calibration or the speed of the movement of the eyes. Another study [5] talks of how to improve the accuracy in a post-calibration procedure.

3.1.2. Calibration process of the Tobii Glasses

The calibration process of the glasses requires some small devices, the IR Markers *(figure 4)*, that when connected to an IR Marker Holder, communicate their exact position to the glasses using a visible green light. This IR Markers are used to delineate an Area of Analysis (AOA).

In the calibration procedure requires two people, the *participant* (wearer of the glasses) and the *experimenter* (person who hold the IR Marker). First of all, the recording device asks to the participant to stand at a distance of one meter from a wall. Then, when an auto adjustment of the glasses (look straight to a fix point) is completed, the recording device displays a 3x3 grid of points. The experimenter has to move the IR Marker and match all the points. The participant has to follow the green light with the eyes without moving their head.

For this project was recorded and uploaded on YouTube a video [21] *(figure 7)* explaining the calibration process of the Tobii eye-tracker glasses.

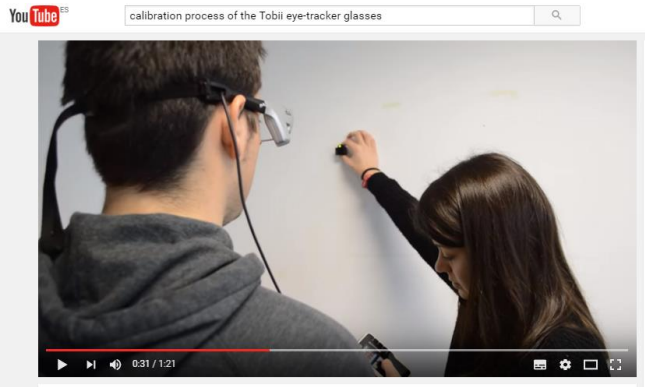

Figure 7: *Keyframe of the video tutorial.*

Once the calibration process is finished, it shows two types of results: accuracy and tracking. Depending on these results the eye-tracker will be more or less accurate. For example, if the calibration is correct, when the participant reads a sentence in a poster, the point of gaze will be in the sentence *(figure 8(a))*; but if the calibration failed, when the participant follows a green light, the point of gaze may be far of the light *(figure 8(b))*.

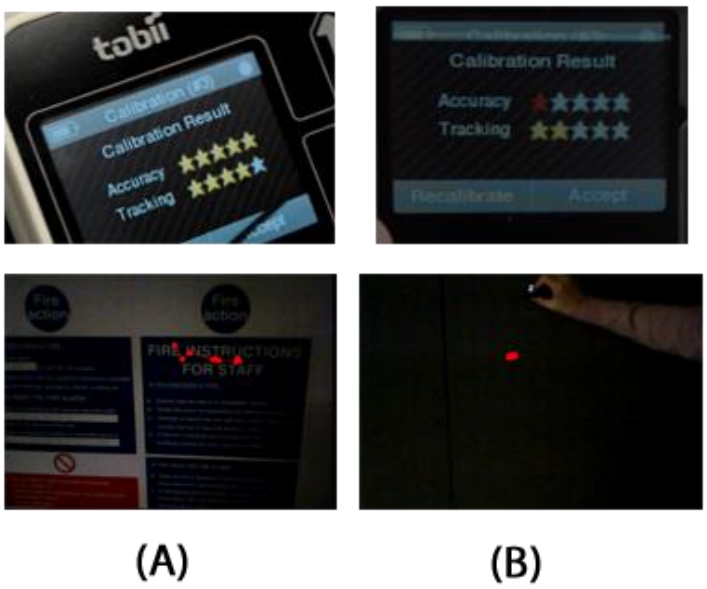

Figure 8. *Results of the calibration process. (a) Results of a good calibration. (b) Results of a bad calibration.*

3.1.3. The Dataset

EgoMon Gaze & Video Dataset is an Egocentric (first person) Dataset that consists of 7 videos of 30 minutes on average *(table 3)*. One of these videos were recorded using two

types of wearables cameras (Tobii glasses and Narrative clip, *view section 1.2.*). This dataset located in our Github [1]Repository and consists in a total of:

7 videos	Recorded with the Tobii eye-tracker Glasses. With gaze information plotted on them
7 videos	Clean (without the gaze information plotted on them)
13428 images	Each image corresponds to each frame per second of all these videos.
7 text files	Gaze data extracted from each video
73 Images corresponding of one video	Taken with the Narrative Clip.

Table 2. *Main features of the EgoMon Gaze & Video Dataset*

The acquisition of the data was carried out by three different wearers of the Tobii glasses *(figure 9)*. One of them is the author of this report and the two others were also mobility students at Insight Dublin City University. The videos were recorded in Dublin (Ireland) during the months of March to May 2016.

Figure 9. *Wearers of the glasses.*

The final has a total of 7 videos (bold), but a total of 9 videos were recorded. Two of them were discarded because corrupted data acquisition or because they were too similar to another one.

Repositori of Egocentric-saliency in GitHub [online] Available: https://github.com/imatge-upc/egocentric-saliency

VIDEO	Status	Outdoors / Indoors	Duration
Oral Presentation	Used	Indoors	45:17
Albert College and DCU Park	Used	Outdoors	22:07
Spanish Omelette	Used	Indoors	29:25
Playing cards	Used	Indoors	30:10
Botanic Gardens	Used	Outdoors	36:13
Bus Ride (option 1)	Not Used	Outdoors	-
Bus Ride (option 2)	Used	Outdoors	26:29
In the Office	Not Used	Outdoors	-
Walking to the office	Used	Outdoors	33:30

Table 3: *EgoMon Gaze & Video Dataset*

The description of each sequence is as follows:

Oral presentation

This video was recorded in an office during two presentations of a progress report of two different thesis. First of all, the person who wore the camera was presenting her project and in the half part of the video another person started the presentation of his project and the person who wore the glasses watched his presentation and the questions round.

Figure 10. *Oral presentation. (a) Fixing the point of view on the judge. (b) Fixing the point of view on the slides. (c) Fixing the point of view on the conversation between the judge and the student.*

In the first part of the video we can see how the participant of the glasses, during her presentation, fixes her attention in the judge *(figure 10(a))* or in the slides that are in her back *(figure 10(b))*. And in the second part of the video we can see how the participant is looking the questions back and forth between the judge and the other student *(figure 10(c))*.

Dublin City University and Albert College Park

Figure 11.*Some examples of the fixation point in the recording.*

This video was recorded outdoors. The participant was walking in a park with two friends *(figure 11)*. In this video we can see some different situations that change the fixation point of view. For example, looking straight and that appears a runner or a dog. Or we can also see the participant looking a map and his friends pointing some parts of the map.

Spanish Omelette

This video *(figure 12)* was recorded cooking the traditional Spanish food "tortilla de patatas". The video is located in a kitchen. We can see in the video how the participant follows with her gaze the different processes to cook this traditional food.

Figure 12. *Some frames of this video.*

Playing cards

This video was recorded during a card game with four players. We can see how the participant fixes their attention in different persons and situations and their cards *(figure 13).*

Figure 13. *Some different situations of this video*

Botanic gardens, Dublin

Figure 14. *Botanic Gardens*

This video *(figure 14)* was recorded outside. The participant was walking in the Botanic Gardens situated in Dublin. This video shows different situations that can change the point of gaze. For example, entering a greenhouse or follow a squirrel, etc. The Dataset also contains images of this video taken with the Narrative Clip *(viev section 1.2.3.)*.

Bus ride (Option 1, Discarded)

This video was discarded because it was a problem with the calibration process and the data extraction.

Bus ride (Option 2)

This video *(figure 15)* was recorded in the front and in the second floor of a Dublin bus. In the video the participant is static but the recorded scene is in motion.

Figure 15. *Bus ride*

In the office (Discarded)

This video was recorded walking around the Insight Centre of Dublin City University. It was discarded because it was the very similar to the oral presentation video.

Walking to the office

This video *(figure 16)* was recorded walking around the streets of the Dublin 9 district from an apartment to the Insight Centre of Dublin City University.

Figure 16. *Walking to the office*

3.1.4. Privacy

Considering the dataset created in this project is published, the videos had to be recorded taking into account the European Directive of Data protection that forbids showing people who can be recognized or car license plates. Special care and review was invested to satisfy these restrictions.

3.1.5. Problems with the Gaze (Losses)

The Tobii eye-tracker glasses required of a good calibration process and a quiet recording to capture all the gaze points. It is possible that during extreme movements of the eyes or the head, the eye-tracker of the Tobii glasses could have losses. Also, if the wearer of the device is looking a vast landscape or a very open scene, the eye-tracker of the glasses could loss the gaze point. The dataset constructed in this project has some frames with this problem, which have been carefully annotated so that the dataset users can choose to test their algorithms with or without these realistic errors.

Each video in the dataset of this project has the next percentage of losses:

Video	Percentage of Losses
Botanic Gardens	12 %
Bus Ride	41 %
DCU and Albert College Park	32 %
Playing Cards	23 %
Oral Presentation	12 %
Spanish Omelette	17 %
Walking to the Office	37 %

Table 4. *Percentage of lost gaze frames for each video.*

These results show that the videos recorded outdoors have more lost frames that the videos recorded inside and in static positions. For example, the video of the Bus Ride was recorded looking a very changing scene in a high velocity (the velocity of the bus). For that reason, this is the video with a higher number of lost frames. On the other side, the video of the Oral Presentation was one with a smaller number of losses because the wearer of the camera was almost all the time static without sudden movements of the head and the video was recorded indoors, in a non-very changing scene.

3.2. Processing

3.2.1. Eye Gaze Data

The analysis of the eye gaze tracking data of the Egomon Gaze & Video Dataset requires extracting these data from the Tobii Studio tool. The resulting data saved in a tab-separated value file (.tsv) or in a Microsoft Excel file (.xls). In this specific case, the data was extracted in a Microsoft Excel file *(figure 17)*. The data export output file can be easily imported into other software such Matlab.

RecordingDuration	MediaWidth	MediaHeight	RecordingTimestamp	GazePointX (MCSpx)	GazePointY (MCSpx)
1327033	640	480	0	246	236
1327033	640	480	33	244	240
1327033	640	480	67	243	238
1327033	640	480	100	243	239
1327033	640	480	133	247	241
1327033	640	480	167	241	237
1327033	640	480	200	241	236
1327033	640	480	233	240	237
1327033	640	480	267	243	235
1327033	640	480	300	243	235
1327033	640	480	333	243	236
1327033	640	480	367	244	238
1327033	640	480	400	243	237
1327033	640	480	433	243	237
1327033	640	480	467	242	236
1327033	640	480	500	242	235
1327033	640	480	533	242	234
1327033	640	480	567	241	234
1327033	640	480	600	242	234
1327033	640	480	633	240	233
1327033	640	480	667	242	234
1327033	640	480	700	240	232
1327033	640	480	733	241	232
1327033	640	480	767	241	233

Figure 17. *Gaze data file extracted from Tobii Studio.*

The Data Export tool in the Tobii Studio allows to choose the Recording (full recording, Segments, or Media), the data types and the export data file properties.

This tool allows choosing between a total of 87 data types to be exported. For that project was extracted a total of 6 data types:

Recording Duration	It shows the duration of the full recording (in milliseconds).
Media Width	The width of the recording (in pixels). The width of the extracted video from the Tobii Glasses is always 640.
MediaHeight	The height of the recording (in pixels). The height of the extracted video from the Tobii Glasses is always 480.
Recording Timestamp	Timestamp counted from the start of the recording (t0 = 0) to the recording duration, that data type shows a timestamp (in milliseconds) of the full recording with a ratio of 33 milliseconds. This timestamp is synchronized with the eye-tracker clock.
Gaze coordintae X	The coordinate (in pixels) in the X axis where the gaze was focused in every recording timestamp.
Gaze coordinate Y	The coordinate (in pixels) in the Y axis where the gaze was focused in every recording timestamp.

Table 5. *Eye Gaze Data extracted for this project*

This gaze data allows creating a ground truth of the fixation point for each frame. Ground truth, in the field of machine learning, is a term to refer to the information provided by direct observation. This information will be compared with a following prediction.

3.2.2. Temporal Subsampling of Video Frames

For this project was chosen a sample rate that would generate an amount of images similar to the one contained in the SALICON dataset (15,000 images) [22] used for visual saliency prediction. This way, both datasets would be comparable in terms of size. Choose a frame rate of 1 fps provided the desired amount of images.

VIDEO SEQUENCE	DURATION	EXTRACTED FRAMES
Oral Presentation	0:45:17	2718
Albert College and DCU Park	0:22:07	1328
Spanish Omelette	0:29:25	1766
Playing cards	0:30:10	1811
Botanic Gardens	0:36:13	2174
Bus Ride	0:26:59	1620
Walking to the Office	0:33:30	2011
TOTAL	**3:43:41**	**13428**
AVERAGE	**0:34:30**	**1918**

Table 6. *Frames extracted from EgoMon Gaze & Video Dataset*

3.2.3. Annotation of the Dataset

The annotation of the dataset was automatic and real from the eye-tracker of the head-mounted recording device. This head-mounted device allows giving the feeling of the video is recorded from someone's eyes. If the wearer of the camera moves their head, the scene moves with the head.

3.3. Results

The EgoMon dataset created in this project has notable differences between the ones described in the *section 2.1*. This dataset contains 7 videos situated in outdoors and indoors environments. The indoors videos contains daily actions as cooking and specific and leisure activities like following a conversation and playing a card game. The outdoors videos contain different situations: the wearer of the glasses walking around the city, seated in a bus looking the moving scene through a window, etc.

The main difference between the others dataset is the others show daily actions, and the EgoMon dataset show both daily and non-daily actions in indoors and outdoors environments.

In the next table is compared the main features of the state of art datasets and the EgoMon Gaze & Video Dataset.

Dataset	Amount of Data	Resolution	Recorder Device	Outdoor/ Indoor	Number of participants
GTEA	17 sequences	640 x 480	Tobii Eye-tracker Glasses	Indoor	14
UT Ego Dataset	4 videos	320 x480	Looxcie wearable (head-mounted) camera	Outdoor and indoor	4
EgoMon Gaze & Video Dataset	7 videos with gaze information plotted on them 7 clean videos 13428 frames 7 files with the gaze data 73 images took with the Narrative Clip	640 x 480	Tobii Eye-tracker Glasses and Narrative Clip	Outdoor and Indoor	3

Table 7. *Comparison of the main features of the state of art datasets.*

Chapter 4 – Visual Saliency Prediction

To have a baseline results over the dataset constructed for this project, it was required to compute saliency models of the frames extracted of this dataset. The tool SalNet was used for this purpose. This chapter introduce this Convolutional Neural Network.

4.1. Saliency Predictor: SalNet

The saliency estimator used for this project is SalNet [23] from Kevin McGuiness. This trained algorithm allows calculating saliency maps from a group of images.

SalNet is a trained 9-layers Convolutional Neural Networks *(figure 18)* and was implemented using Caffe library. The dataset of normal images which this CNN was trained with the SALICON dataset [22], but its first layers were pre-trained with ImageNet dataset [29] on a VGG model [30].

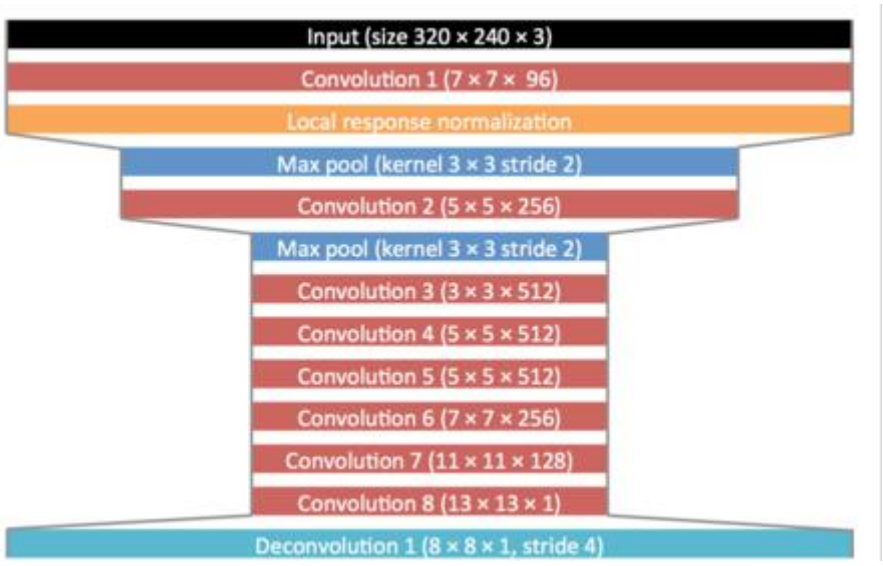

Figure 18. *Architecture of the Deep Convolutional Network.*

The result of running this convnet is a 2D-grey image that shows the saliency predicted model of each corresponding input image.

Figure 19 shows how the adopted state of the art saliency predictor, SalNet, generates a saliency map for non-egocentric images.

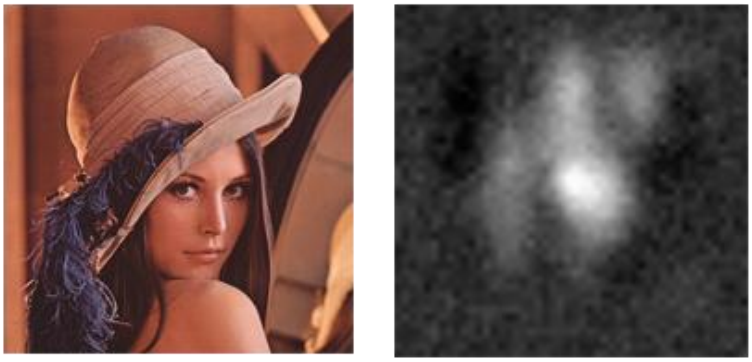

Figure 19. *Result of the saliency estimator for a normal image.*

4.2. Quantitative evaluation: Comparison Metric

The saliency maps generated by SalNet must be compared with the gaze points captured with the eye tracker. The different nature of these two data sources requires a specific metric.

There exist different metrics to evaluate the quality of visual saliency maps [24][25]. These metrics can be separated between two main groups: location-based metrics and distribution-based metrics.

Location-based	**Distribution-based**
AUC-Judd, sAUC, NSS	SIM, CC, EMD, KL

Table 8. *Comparison metrics*

The main difference between these two groups is the types of data being compared. The first category of metrics compares a saliency one or several gaze points from the ground truth. And the second category of metrics directly compares the predicted saliency map with another saliency map generated from ground truth data, typically convolving the gaze points with a Gaussian filter.

In our problem, there is only one eye fixation for each video frame, so we must adopt a location-based metric. The Normalized Scanpath Saliency (NSS) was adopted to run our experiments. This metric was used in this project instead of the other location-based metrics because SalNet got a better result using this metric in the MIT300 Saliency Benchmark [24].

The NSS metric, introduced by Petters and Itti [26], is measured as the mean value of the normalized saliency map at fixation locations.

$$NSS(\rho) = \frac{\mathrm{SM}(\rho) - \mu_{SM}}{\sigma_{SM}}$$

ρ: location of one fixation.
SM: saliency map normalized to have a zero mean and unit standard deviation.

Then, the NSS score is the average of NSS(ρ) for all fixation, in our case, all the fixations in each video (only one ground truth fixation per frame).:

$$NSS = \frac{1}{N} * \sum_{\rho=1}^{N} NSS(\rho)$$

N: total number of eye fixations.

If NSS score is low, it means that the locations with the eye fixations were predicted with a low probability by the predictor, which is undesirable. It shows that the saliency estimator won't be very predictive. Therefore, a good visual predictor is associated to a high value, being 1.0 upper bound.

4.2.1. Implementation

NSS computation

For this project was used the implementation of the NSS metric provided by the MIT300 saliency Benchmark [24].

The input parameters for this code are two:

- Human fixation point binary matrix: Binary matrix of the same dimensions as the input image. All the values zero except the human fixation points (one).

- Saliency map: Predicted saliency map (2-D grey image).

Eye fixations parsing

It was necessary to implement a script to read the gaze data extracted file and adapt it to calculate the NSS script referred in the section before.

This script takes each coordinate for each millisecond that forms a second and makes a binary matrix with all the values 0 except for the gaze coordinate. That matrix will be used to calculate the NSS such is explained in the last section. This script is allocated in the GitHub repository [7] of this project with the name of: reading_gazedata.m.

Results script

As detailed previously *(view section 3.2.4.)* when the videos that form the dataset were recorded, there were a few numbers of losses of the gaze frames. For that reason, it does not make sense to compare the saliency map corresponding to these frames because there are no gaze data for them. These frames are discarded when computing NSS.

This script returns a mean NSS value for each video without having in consideration the losses (NSS = 0 for default of the previous script). Also returns the percentage of frames that have losses. This script is allocated in the GitHub repository [7] of this project with the name of: results.m.

4.3. Results of the Dataset

The dataset constructed in this project contain *(figure 20)*:

- 7 videos recorded with Tobii eye-tracker Glasses.
- 7 files with the data gaze of each frame and.
- 13428 input images for the convnet SalNet.
- 13428 output 2D-grey images of the convnet.

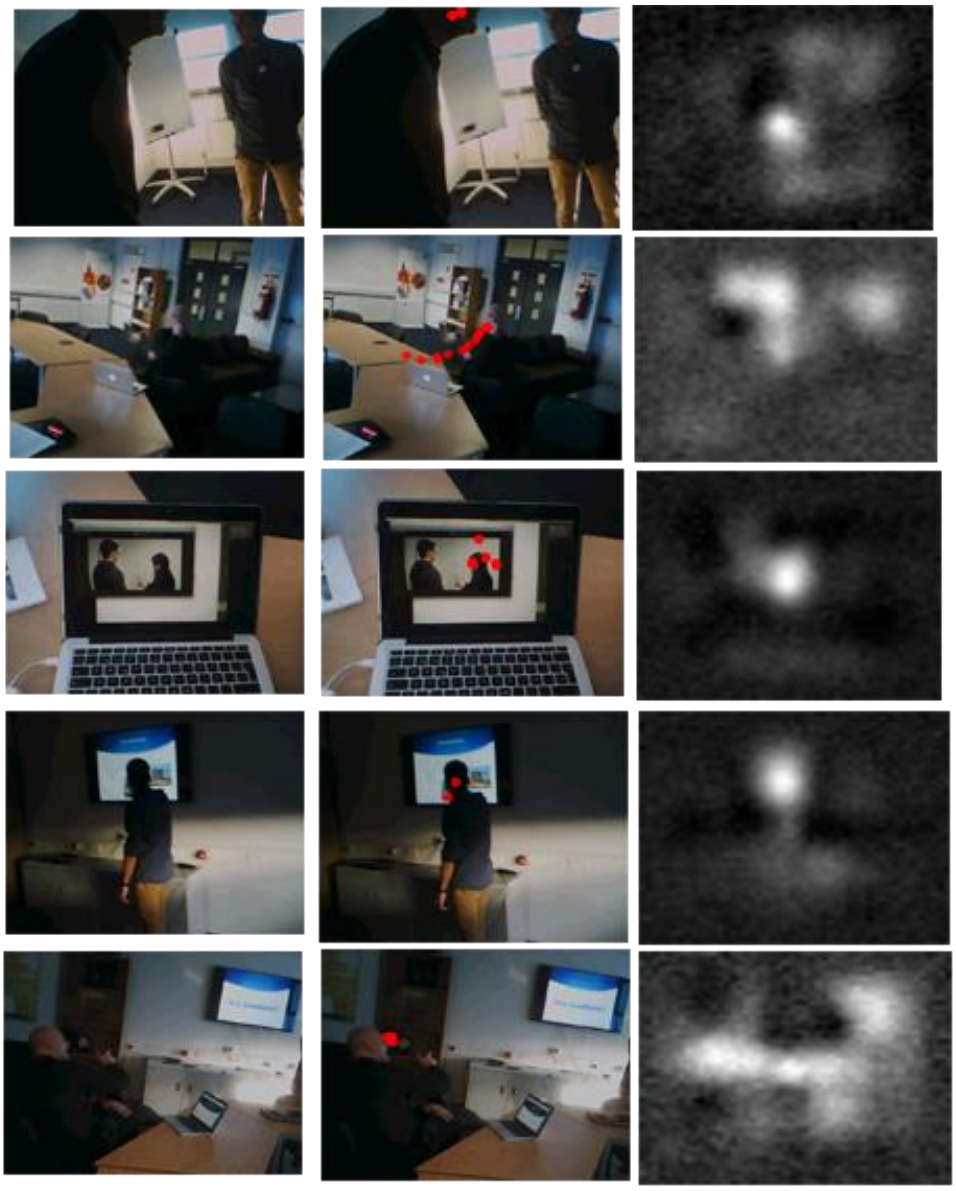

Figure 20. *Example of the results of the oral presentation video. From left to right: extracted frames (input of the convnet), extracted frames with plotted point of gaze and 2-D grey image (output of the convnet).*

4.4. Quantitative Evaluation: Reults

In this section are compared the results returned by the script detailed in a previous section *(view 4.2.1.)*. This script returns a mean NSS value for each video that forms the EgoMon Video & Gaze Dataset created in this project.

VIDEO	NSS
Botanic Gardens	0.661
Bus Ride	0.742
DCU and Albert College Park	0.494
Playing Cards	0.684
Oral Presentation	1.275
Spanish Omelette	1.475
Walking to the Office	0.554
AVERAGED NSS	**0.841**

Table 9. *Quantitative results*

These results can be compared *(figure 21)* with the NSS result that SalNet obtained in the MIT300 Saliency Benchmark [25]:

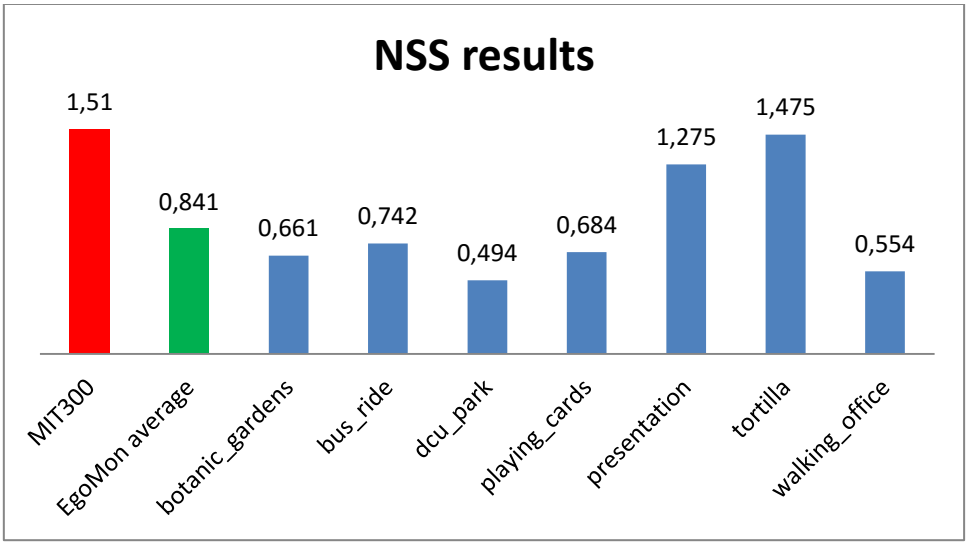

Figure 21. *Normalized Saliency Scanpath results of the Egomon datset and MIT300.*

Taking into account the previous results, the videos included in the dataset created for this project can be divided in two groups:

- The EgoMon videos with a mean NSS value similar to the one obtained for MIT300 (spanish omelette and presentation): these videos share a main feature, they were recorded indoors and doing specific activities (cooking, following a conversation).

- The EgoMon videos with a mean NSS value much lower than the MIT300 NSS value (DCU Park, walking to the office, botanic gardens): these videos were recorded outdoors and walking around the city, depicting vast landscapes, etc.

In average, could be affirmed the results of the NSS for the egocentric videos are worse than the MIT300 result.

4.5. Qualitative Evaluation

This section presents examples of the frames with the best and worst NSS.

Indoor recordings: The best results *(figure 22)* of the Spanish Omelette video are the frames with a very specific activity and view. The worst results are due because the wearer of the glasses looking at a location different from one that contained most of the objects, which tend to attract the attention of SalNet. For example, chopping potatoes and fixing the gaze on the knife but the saliency model focuses into the potatoes *(figure 23).*

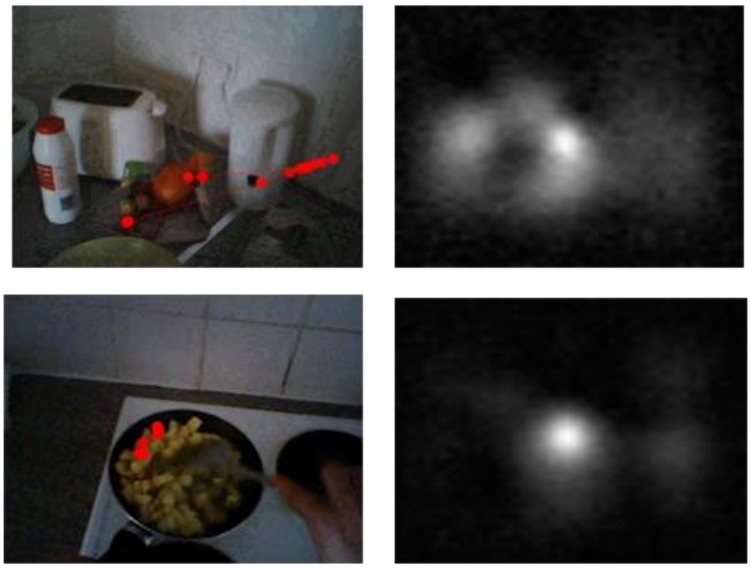

Figure 22. *Frames with best results of the Spanish Omelette video*

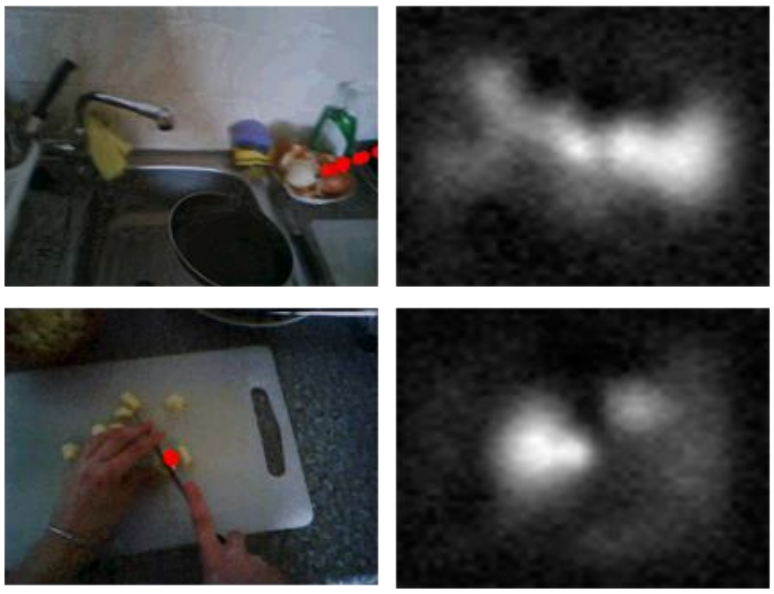

Figure 23. *Frames with the worst results of the Spanish Omelette video*

Figure 24. *Frames with the worst results of the Botanic gardens video.*

Outdoor recordings: Other examples of the worst results *(figure 24)* are some frames of the Botanic gardens video. For example, a scene where a person appears but the wearer of the glasses was looking at a tree, which was not highlighted by the visual saliency predictor.

Chapter 5 –Conclusions and Future works

5.1. Conclusions

This project has presented to contributions. The first one is the construction and publication of a new and innovative dataset called EgoMon Gaze & Video Dataset. This is a large dataset for egocentric visual saliency prediction which has notable differences with others egocentric datasets already created. These notable differences are divided in the next main features: more and different types of recording environments (indoor and outdoor) and a big difference (movements of the wearer of the camera and changes and movements in the environment) between each video that forms the dataset.

The other contribution of this project is the assessment of the state of art saliency predictor for this specific case of egocentric images. This SalNet visual saliency estimator was run and computed the saliency models of the frames extracted from the dataset. The main conclusion is that the saliency estimator SalNet does not perform the same for non-egocentric images than for egocentric ones. But if each case is studied individually, it can conclude that the videos recorded doing a static activity, where all the frames are more similar than a normal image and without a lot of ego-motion *(view section 2.2.)* in them, give better results than the videos with a higher motion

A limiting factor of this work was the numerous errors while acquiring the eye fixations. The videos recorded outdoors with a lot of motion and ego-motion in the scene or the wearer looking a very open scene presented more losses than the videos where the wearer was focused on an activity.

5.2. Future works

This project offers an open window for future works. But two of them are especially promising:

- Calculate the validity of SalNet with egocentric images comparing the results with different metrics.
- Fine-tunning (adaptation) of SalNet (retrain the CNN with egocentric images).

5.2.1. Change the comparison metric [26]

As it was explained in other sections *(view chapter 4)*, there exist different types of Location-based comparison metrics. Apart from the normalized saliency scanpath metric (NSS), there exist two more Location-based metrics (AUC-Judd and sAUC). A possible future work for this project could be comparing the results of that two metrics with the results of SalNet in the MIT300 saliency Benchmark.

5.2.2. Fine-tuning of the saliency estimator

The performance of SalNet over EgoMon could be improved by fine-tuning the model for egocentric images.

A fine-tuning of a Convolutional Neural Network is an adaptation procedure for a specific new case. This procedure is based on the concept of transfer learning [28].

References

[1], Sergi Imedio, "An Investigation of eye gaze tracking data utilities in image logo recognition". Bachelor thesis report. Dublin City University, 2016.

[2], Tobii Glasses User Guide [online] Available: http://www.acuity-ets.com/downloads/Tobii%20Glasses%20User%20Guide.pdf

[3], Morgante, J. D., Zolfaghari, R., & Johnson, S. P. (2012). A critical test of temporal and spatial accuracy of the Tobii T60XL eye tracker. Infancy, 17(1), 9-32.

[4], Nyström, M., Andersson, R., Holmqvist, K., & van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. Behavior research methods, 45(1), 272-288.

[5], Blignaut, P., Holmqvist, K., Nyström, M., & Dewhurst, R. (2014). Improving the accuracy of video-based eye tracking in real time through post-calibration regression. In Current Trends in Eye Tracking Research (pp. 77-100). Springer International Publishing.

[6], Image Processing Group (GPI) [online] Available: https://imatge.upc.edu/web/

[7], Repositori of Egocentric-saliency in GitHub [online] Available: https://github.com/imatge-upc/egocentric-saliency

[8], GTEA (Georgia Tech Egocentric Activities) – Gaze Dataset [online] Available: http://ai.stanford.edu/~alireza/GTEA_Gaze_Website/

[9], Fathi, Alireza, Yin Li, and James M. Rehg. "Learning to recognize daily actions using gaze." In Computer Vision–ECCV 2012, pp. 314-327. Springer Berlin Heidelberg, 2012

[10], UT (University of Texas) Ego Dataset [online] Available: http://vision.cs.utexas.edu/projects/egocentric_data/UT_Egocentric_Dataset.html

[11], Su, Y. C., & Grauman, K. (2016). Detecting Engagement in Egocentric Video. arXiv preprint arXiv:1604.00906.

[12], Looxcie wearable camera manual [online] Available: http://www.bhphotovideo.com/lit_files/45530.pdf

[13], DogCentric Activity Dataset [online] Available: http://robotics.ait.kyushu-u.ac.jp/~yumi/db/first_dog.html

[14], GoPro manual [online] Available: http://cbcdn1.gp-static.com/uploads/product_manual/file/278/UM_HERO_ENG_REVB_WEB.pdf

[15], Yamada, Kentaro, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. "Can saliency map models predict human egocentric visual attention?." In Computer Vision–ACCV 2010 Workshops, pp. 420-429. Springer Berlin Heidelberg, 2010.

[16], Itti, L., Dhavale, N., Pighin, F., et al.: Realistic avatar eye and head animation using a neurobiological model of visual attention. In: SPIE 48th AnnualInternational Symposiumon Optical Science and Technology. Volume 5200. (2003) 64–78.

[17], Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. Advances in Neural Information Processing Systems 19 (2006) 545–552

[18], Matsuo, Kenji, Kentaro Yamada, Satoshi Ueno, and Sei Naito. "An attention-based activity recognition for egocentric video." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 551-556. 2014.

[19], H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[20], M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[21], "Calibration Process of the Tobii eye-tracker glasses" [YouTube Video Online]: https://www.youtube.com/watch?v=8QLuRJAgQIM

[22], M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In IEEE conference on Computer Vision and Pattern Recognition, 2015.

[23], SalNet (Convolutional Neural Network) by Kevin McGuiness [online] Available: https://github.com/imatge-upc/saliency-2016-cvpr/blob/master/README.md

[24], MIT Saliency Benchmark [online] Available: http://saliency.mit.edu/results_mit300.html

[25], Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013). Saliency and human fixations: state-of-the-art and study of comparison metrics. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1153-1160).

[26], Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. Vision research, 45(18), 2397-2416.

[27], Code to compute the saliency comparison metrics [online] Available: https://github.com/cvzoya/saliency/tree/master/code_forMetrics

[28], Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. Unsupervised and Transfer Learning Challenges in Machine Learning, 7, 19.

[29], J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[30], K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In International Conference on Learning Representations, 2014.