UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Alan Carvalho Neves

Gaze-Based Semantic Hyperlapse

Belo Horizonte
2019

Alan Carvalho Neves

Gaze-Based Semantic Hyperlapse

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Erickson Rangel do Nascimento
Co-Advisor: Mario Fernando Montenegro Campos
Co-Advisor: Michel Melo da Silva

Belo Horizonte
2019

Neves, Alan Carvalho.

N518g Gaze-based semantic hyperlapse [recurso eletrônico] / Alan Carvalho Neves – 2019.
1 recurso online (78 f. il, color.) : pdf.

Orientador: Erickson Rangel do Nascimento
Coorientador: Mário Fernando Montenegro Campos
Coorientador: Michel Melo da Silva.

Dissertação (Mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f. 63-69

1. Computação – Teses. 2. Visão por computador– Teses. 3. Vídeos em primeira pessoa – Teses. 4. Web semântica – Teses. I. Nascimento, Erickson Rangel do. II. Campos, Mário Fernando Montenegro. III. Silva, Michel Melo da. IV. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. V. Título.

CDU 519.6*82.10(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

GAZE-BASED SEMANTIC HYPERLAPSE

ALAN CARVALHO NEVES

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ERICKSON RANGEL DO NASCIMENTO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. MARIO FERNANDO MONTENEGRO CAMPOS - Coorientador
Departamento de Ciência da Computação - UFMG

DR. MICHEL MELO DA SILVA - Coorientador
DCC - UFMG

PROF. ALEXEI MANSO CORREA MACHADO
Departamento de Ciência da Computação - PUC-MG

PROF. GUILLERMO CÁMARA CHÁVEZ
Departamento de Computação - UFOP

PROF. HÉLIO PEDRINI
Instituto de Computação - UNICAMP

Belo Horizonte, 28 de Novembro de 2019.

Acknowledgments

First of all, I thank God for the gift of life. I thank my parents and sister for their love and compassion, which helps me to navigate the complex maze of life. To my wife, Jessica, for her sweetness and patience, bringing us together despite our distance.

To the masters by the teachings, allowing us to build a solid foundation of knowledge. Along our journey, they contribute to suppressing the darkness of our ignorance with their light. Special thanks to my advisors, whom I have brought all kinds of possible doubts, from the simplest to the most complex.

Also, I thank the friends I have made on UFMG, great people aiming to build a better world, making the barriers of science move forward. Special thanks to VerLab friends, who have always behaved as a united team, caring and helping each other. I am proud to be part of this team. In particular, I thank the Semantic Hyperlapse team, which was with me along this time discussing problems, ideas, facing many "deadline parties", and when possible, having fun.

I thank the agencies CAPES, CNPq, FAPEMIG, and Petrobras for funding different parts of this work. We also thank NVIDIA Corporation for the donation of a Titan XP GPU used in this research.

"*If I have seen further, it is by standing upon the shoulders of giants.*"
(Sir Isaac Newton)

Resumo

O crescente compartilhamento de dados e a cultura de registro de informações cotidianas têm conduzido a um aumento sem precedentes na quantidade de vídeos de primeira pessoa não editados. Enquanto dispositivos vestíveis reduzem o esforço na aquisição de dados, eles tornam desafiadora a tarefa de recuperar e acessar informações dos dados coletados. Nesta dissertação, buscamos resolver o problema de acessar informação relevante em vídeos de primeira pessoa, através da enfatização dos momentos considerados importantes pelo portador da câmera. Diferente de trabalhos de sumarização, aceleração de vídeos e hyperlapse que tem como semântica um conjunto definido de assuntos/objetos, propomos um modelo de atenção baseado em gaze e análise visual da cena. Rastreando os objetos da cena que interagem com o olhar do usuário, juntamente com a avaliação de suas características temporais e espaciais, nosso modelo pode inferir dinamicamente os interesses do usuário. Além disso, empregando uma estratégia de novidade de cena em nosso modelo de atenção, evitamos assistir excessivamente segmentos de vídeo no vídeo acelerado. O modelo de atenção resultante é usado para calcular a relevância de cada frame do vídeo de entrada. Foram realizadas diversas avaliações experimentais em dois conjuntos de dados de vídeos egocêntricos publicamente disponíveis: o A*STAR Ego-Gaze e Georgia Tech Egocentric Activity. As avaliações mostram que na cobertura de tarefas que necessitam de atenção do usuário, nosso método apresenta um resultado médio superior de $9,6$ pontos percentuais em relação ao melhor competidor. Considerando a carga semântica presente no vídeo acelerado, nosso método capturou 15% mais objetos na vizinhança do gaze que o melhor competidor. Desta forma, nossa metodologia é capaz de acelerar vídeos egocêntricos de maneira automática quando o portador da câmera interage visualmente com os componentes da cena, reforçando o aspecto de diversidade das informações recuperadas.

Palavras-chave: visão computacional; gaze; vídeos de primeira pessoa; vídeos egocêntricos; aceleração de vídeos; informação semântica.

Abstract

The growing data sharing and life-logging cultures are driving an unprecedented increase in the amount of unedited first-person videos. While wearable devices reduce the effort in the data acquisition, they make it challenging to retrieve information and browse through the collected data. In this thesis, we address the problem of accessing relevant information in first-person videos by emphasizing the important moments to the wearer/recorder. Unlike works of summarization, fast-forward, and hyperlapse that have semantics as a set of hard defined subjects, we propose an attention model based on gaze and visual scene analysis. Tracking the objects of the scene that interacts with the user's gaze and evaluating their temporal and spatial characteristics, our model can infer the wearer interests dynamically. Moreover, employing a scene novelty strategy in our attention model, we avoid overly watching video segments in the accelerated video. The resulting attention model is used to compute the relevance of each frame of the input video. Several experimental evaluations were performed on two publicly available first-person video datasets that contain gaze data: the A*STAR Ego-Gaze, and Georgia Tech Egocentric Activity dataset. The evaluation shows that in the coverage of tasks that need user attention, our method shows a better average result of 9.6 percentage points to the best competitor. Also, considering the semantic load present on the accelerated video, our method captured 15% more objects in the gaze surroundings than the best competitor. Therefore, our methodology can automatically fast-forward videos emphasizing moments when the recorder visually interact with scene components while enforcing the diversity aspect of retrieved information.

Keywords: computer vision; gaze; first-person videos; egocentric videos; fast-forwarding videos; semantic information.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

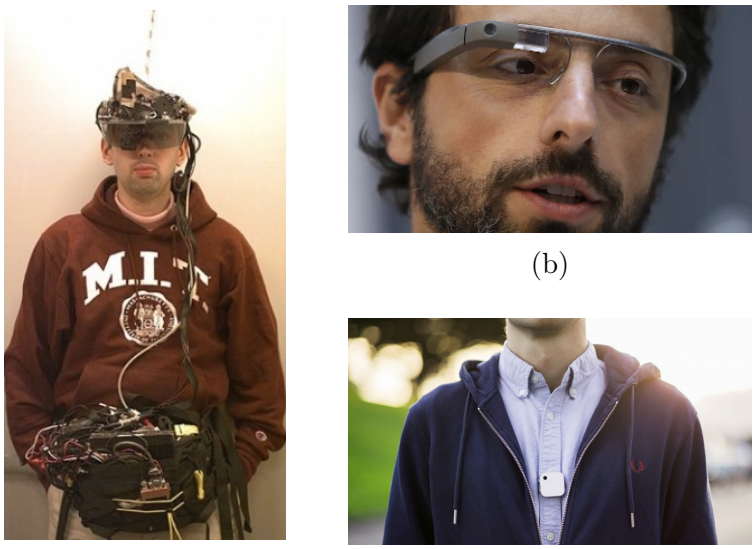
Since old age, humanity has made a registry of everything that happened to their surroundings through the available technology. Our ancestors painted on cave walls to describe situations such as their hunts, social events, and habits in a rudimentary way to store information. In the modern era, it was possible to perpetuate events through the development of the photography camera that enabled us to record aerospatial images, historical events, biological structures, and even ordinary life events, as depicted in Figure 1.1.

Many attempts were made to integrate cameras and other multimedia devices in a portable way to enhance the sense of visual awareness of a person. In the nineties, Mann [1998] developed one of the first modern wearable devices — the WearCam. This device grouped and interfaced devices such as cameras, microphones, and earphones with a computer — primarily to collect data for research, occupying the size of a backpack. With social acceptance and technological advances, wearable cameras increased their capacity and reduced their size over time, as presented in Figure 1.2.

(a) (b) (c)

Figure 1.1: Photography applications. (a) Aerospatial photograph of an HTV-7 vehicle used to resupply the International Space Station. Reprinted from *flickr*, by Alexander Gerst[1]. (b) Photograph of soldiers entrenched in the First World War at the Battle of Fromelles. Reprinted from BBC News[2]. (c) Neutrophils with segmented nuclei surrounded by erythrocytes and platelets. Reprinted from *Wikipedia*, by Dr Graham Beards[3].

[1] https://www.flickr.com/photos/astro_alex/43960102770
[2] https://www.bbc.com/news/uk-35963387
[3] https://commons.wikimedia.org/wiki/File:Neutrophils.jpg
[4] http://wearcam.org/steve5.htm
[5] http://shorturl.at/hFGHW
[6] http://shorturl.at/pvGQ5



(a) (b) (c)

Figure 1.2: Wearables evolution. (a) Steve Mann on MIT by the early 1990s. Extracted from WearCam.org[4]. (b) Sergey Brinn using *Google Glass*™. Obtained at NPR.org[5]. (c) *Narrative Clip 2*™ camera. Reprinted from TechCrunch[6].

(a) (b)

Figure 1.3: FPV configurations. (a) Camera mounted on head. (b) Chest mount.

Nowadays, wearable cameras have become a successful global product thanks to technological advances and the growing culture of instant sharing and life-logging. Life-logging records are usually done in a First-Person Video (FPV) configuration, with the camera mounted on the user's body, *e.g.*, head or chest, allowing to record the user activities for long periods, letting the user hands-free, in a way to approximate the wearer experience, as pointed by Molino et al. [2017] and showed in Figure 1.3.

1.1 Problem

Wearable cameras allow people to spread their achievements, exhibit reached sports milestones, advertise hazards, depict catastrophes in real-time, record celebrations and holidays, or log their daily activities. However, thanks to the easiness of only pressing the shutter button to record a long event and by pressing the sharing button to directly publish it on a social network, uncountable hours of long, monotonous, and unedited First-Person Videos are being generated. Moreover, long First-Person Videos store events in which the recorder is interested in a particular object, person, or advertisement, as well as monotonous segments such as a long idle segment in a stop sign. There is no explicit intention of something on the scene or even a constraint of which kind of things should be recorded. No specific object is focused, as there is no limit to what should be recorded - it is just a continuous sequence of events. Some examples of First-Person Videos are presented in Figure 1.4.

Naturally, not all events in our day-to-day life or leisure time are worthwhile to be recovered in a late evening watching session. The former recorder and now watcher wants fast access to the relevant information in his/her videos, and he/she wants the relevant according to what he/she saw and experienced in the recording time. Recent techniques address the problem of providing quick access to the information through summarization, semantic fast-forward, or visual storylines.

Summarization techniques aim to create a summarized version of the original video by selecting frames using an importance criteria, such as the research works of Lee et al. [2012]; Xu et al. [2015]; Varini et al. [2017] and de Avila et al. [2011]. Semantic fast-forward preserves video content, creating an emphasis effect through frame selection as done by Okamoto and Yanai [2014]; Lai et al. [2017]; Silva et al. [2018b] and Ramos et al. [2016]. Visual storylines methods perform information retrieval of relevant video parts according to user inputs, as shown by Chen et al. [2012] and Xiong et al. [2015]. The drawback

(a)

(b)

Figure 1.4: Examples of First-Person Videos. (a) A man riding a horse. Obtained at YouTube[7]. (b) Motorcycling. Reprinted from NBC News[8].

[7] https://www.youtube.com/watch?v=LUThBDjQocU

[8] https://www.nbcnews.com/news/amp-video/mmvo42688581735

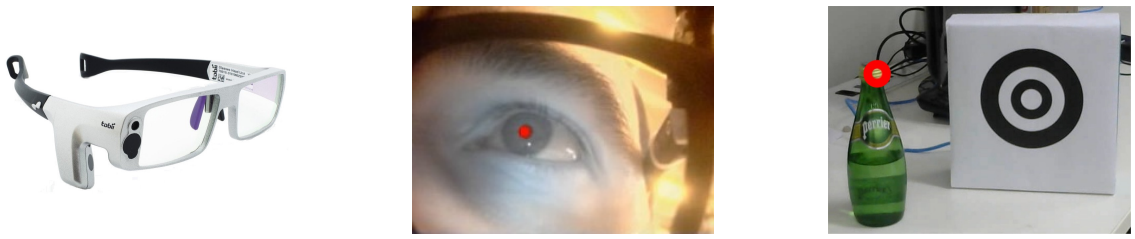
of these techniques is the requirement of a prior definition of relevance. The method must previously know what is relevant to prune out less important frames, limiting these approaches to specific purposes.

In First-Person Videos, relevant elements of the video, *i.e.*, the recorder's interests and intentions, are naturally dynamic as it interacts with the environment. Much effort was made to extract the camera wearer's insights, mainly to characterize the actions performed, using multisensory data such as image, depth, inertial, and gaze. We highlight gaze, the eye focus on a specific scene region, frequently used when describing eye movements. The gaze is intrinsically related to human attention when performing tasks, as demonstrated by Yarbus [1967], providing cues of wearer attention. Gaze data is obtained through an eye-tracker, a device capable of detecting the eye fovea's orientation and projecting the gaze onto a 2D plane, as shown in Figure 1.5.

Using image and gaze data, we built a model able to describe the user attention along a first-person video recording. This model creates a score profile of user attention to be used together with acceleration methods, generating a shortened version of the video with a focus on what was relevant to the camera's wearer. Our model attempts to solve the problem of dynamically describing the recorder's interests along a video, in an implicit approach, using the gaze data captured during the video recording. This work has some applications, such as personalized summarization and object retrieval on a video.

1.2 Research Goal

Hard-defined semantics cannot correctly represent the set of important elements on a video to the camera recorder, given the dynamic nature of a person's intent over time. Our research focuses on capturing the implicit user intents when recording a first-person video, with gaze data, to generate an accelerated version of the video. Our research has the following objectives:



(a) (b) (c)

Figure 1.5: Eye-tracking device and outputs. (a) Tobi eye-tracker. (b) Pupil detection. (c) Gaze in bottle cap.

- Create a user-based evaluation of the frames given gaze data;
- Evaluate the accelerated video given the attention of the user into selected segments.

1.3 Thesis Statement

We argue that when selecting the most relevant frames, the eyes of the beholder, *i.e.*, the recorder, play a crucial role in extracting cues of the significance of each frame to the recorder. Thus, we propose to leverage the semantic extraction from frames by using the recorder's gaze.

1.4 Contributions

The main contributions of this thesis are:

- a gaze-based attention model designed to emphasize the relevant information regarding the behavior of the recorder;
- a new frame scoring methodology combining the attention model with a scene novelty modeling to avoid spending more time than necessary to understand the moment context;
- the modeling and step-by-step instructions on how to build a wearable eye-tracker device using consumer cameras, enabling the capture of gaze-based First-Person Video datasets.

In summary, we propose a gaze-based semantic hyperlapse method for first-person videos. We model the recorder's attention fusing gaze and scene components information, as well as penalizing long unvarying video segments, to accelerate first-person videos, as depicted in Figure 1.6. Our approach was evaluated in two datasets, showing that in the coverage of tasks that need user attention, our method shows a better average result of 9.6 percentage points to the best competitor and captured 15% more objects in the gaze surroundings. Visual results presented our method's ability to capture the user's intentions, as well as the underlying behavior of the subject. This work has many applications in everyday situations, mainly when it involves user interaction, *e.g.*, finding lost objects or

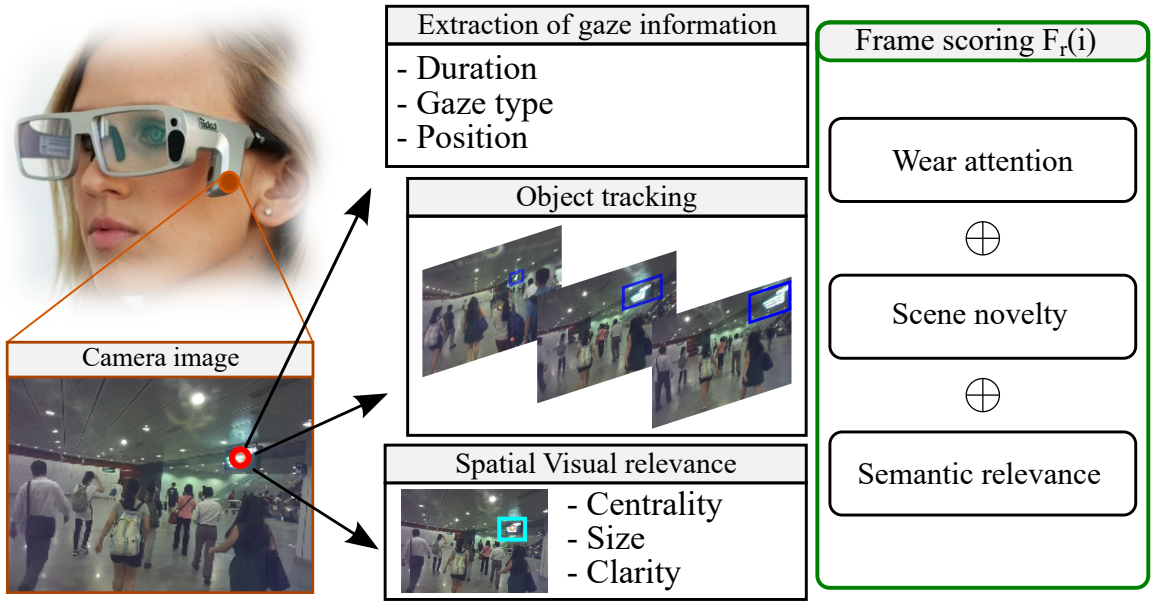


Figure 1.6: Proposed gaze-based frame scoring. In our work, we infer the frame importance based on three factors: attention of the recorder, novelty, and relevance. These factors are obtained by crossing gaze data with scene information, taking into account the interaction of the user's gaze with the scene along the time. A frame is considered relevant if the user's gaze remains consistently in a region of a detected object for a period of time, indicating interest.

reviewing previous situations that caught our attention. Moreover, when capturing only the important parts to the user, a side effect is the original video reduction/compression, a desirable feature for a First-Person Video.

Part of this work was accepted for presentation at the Sixth International Workshop on Egocentric Perception, Interaction and Computing at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (EPIC@CVPR) 2020.

1.5 Organization

This thesis is divided as follow: Chapter 2 discusses background concepts involved on this work, Chapter 3 presents related works, Chapter 4 describes our methodology, Chapter 5 shows the experiments, Chapter 6 discuss the results, Chapter 7 concludes our work and Appendix A presents the steps to build an affordable eye tracker.

Chapter 2

Background

In this chapter, we introduce the concepts underlying the proposed methodology, to give the reader the basis to understand the problem and the choices behind the design of our method.

Given the massive amount of videos available today, it is necessary to provide a suitable way to retrieve useful data to the user, balancing relevance and diversity. Relevance is related to what is considered important to the user, *i.e.*, important events along with the video. Diversity aims at reducing redundancy on the information found. The literature has plenty of techniques capable of indexing and extracting Third-Person Video information. However, First-Person Videos pose an additional challenge to these techniques, as First-Person Videos videos are generally long, continuous, unconstrained, and usually contain blurry and shaking scenes, as pointed by Molino et al. [2017]. We can highlight summarization, fast-forward, and hyperlapse techniques like the ones able to extract relevant information from First-Person Videos.

2.1 Summarization

According to Lu and Grauman [2013], video summarization techniques aim to create a short version of the input video, preserving relevant information. There are two mainly used approaches to generate video summaries, as pointed by de Avila et al. [2011]: Storyboards and Video skims. Storyboards extract highlights in the form of keyframes able to describe the video content. These keyframes may be selected by their position on the video, color, sharpness, or any other representative characteristic. Video skims segment the input video into subshots and selects those containing the most relevant information. Subshots are blocks of contiguous frames separated by previously defined criteria. The mentioned approaches are illustrated in Figure 2.1.

The problem with summarization techniques relies on the understanding of the generated videos. In many situations, it is necessary to preserve the temporal dependency

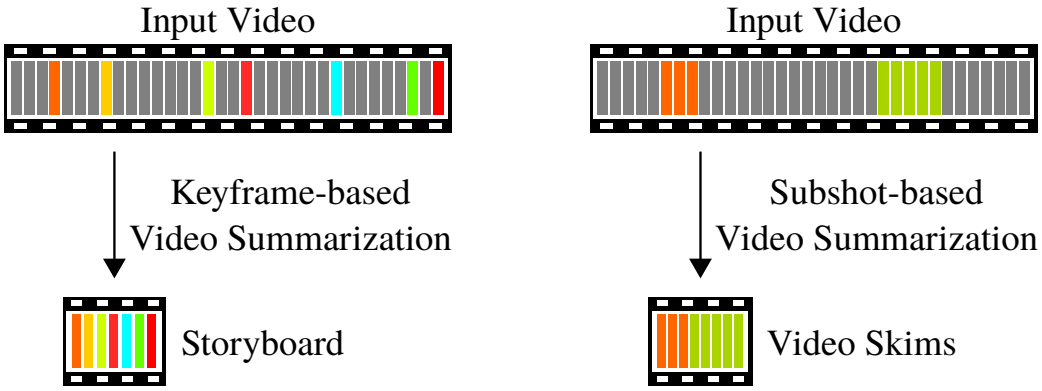


Figure 2.1: Summarization methods. (a) Storyboards select keyframes from the input video according to specific criteria. (b) Video skimming chooses segments from the input video to compose the final summary.

of the video parts to understand their context, such as in adventure and extreme sports videos.

2.2 Fast-forward

Fast-forward methods add the continuity restriction to the summarization problem, sampling frames at a rate, called speed-up, that creates an acceleration effect in the output video – preserving the temporal consistency.

Traditional fast-forward, well known as timelapse or naive fast-forward, is obtained by sampling frames at a fixed rate. It considers all events that occurred on a video as equally relevant. The semantic fast-forward, from its turn, segments the video according to defined criteria of importance and applies variable speed-up rates onto these regions, as realized by Okamoto and Yanai [2014]. Both strategies are presented in Figure 2.2.

Egocentric videos present a natural shakiness related to the body movements of the person on which the camera is attached. This natural shakiness is increased with frame removal employed by fast-forwarding techniques, turning the videos unwatchable.

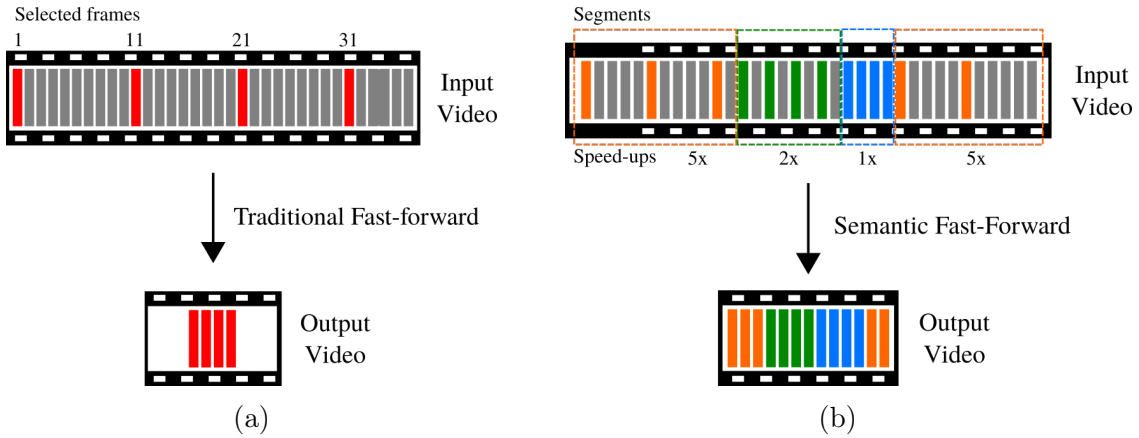


Figure 2.2: Fast-forward methods. (a) Frames are sampled at a fixed rate from the input video, composing the accelerated version. (b) For each segment, different speed-up rates are applied, producing a variable acceleration effect.

2.3 Hyperlapse

Hyperlapse methods attempt to create egocentric fast-forwarded videos concerning visual smoothness, using reconstruction methods or a careful selection of frames on the original video.

Some techniques use reconstruction methods to extract frame features and create a 3D model of the scene. Through this model, they plan an optimal camera path, on which traveling leads to a smooth accelerated video, as in the work of Kopf et al. [2014]. Other methods treat the smoothness problem as a careful selection of frames, selecting the frames in which transition is smoother as possible – as performed by Poleg et al. [2015]. This is achieved by calculating the cost of transition between video frames, through the analysis of visual and/or motion features. The set of frames selected to compose the hyperlapse video is that one that minimizes the transition costs.

The limitation of hyperlapse methods is the absence of a semantic definition, *i.e.*, all video frames are treated as equally relevant concerning semantic information.

2.4 Semantic Hyperlapse

Semantic hyperlapse techniques add relevance constraint to the hyperlapse problem, highlighting important parts of the video without compromising the continuity, smoothness, and desired speed-up of the output video. These techniques could treat

the frame selection problem through a graph optimization approach - modeling video frames as vertices, and the transition cost as edges, they search for the minimum path of the graph. The transition cost takes into account all constraints together, trying to keep constant speed-up, as in the works of Ramos et al. [2016] and Silva et al. [2016], or weighted by their relative size as done by Silva et al. [2018b]. Or even solve the frame sampling as a Minimum Sparse Reconstruction problem, where each video frame is converted into a descriptor that encodes their characteristics, and frames are selected to minimize the reconstruction error of the original video, as performed by Silva et al. [2018a].

These works define semantics as the presence of specific elements in the video such as faces, pedestrians, the coolness aspect of an image, or the presence of the user's object of interest. Therefore, they share the need for a prior definition of relevance – being limited to this defined set.

2.5 Gaze

Instead of using a hard defined semantic set of objects, we propose to use gaze to highlight in a dynamic way, what is relevant to the wearer when interacting with the world. Gaze is the point of the world targeted by fovea, *i.e.*, their focus, as shown in Figure 2.3. The fovea is a small and high-resolution region in the center of the retina, able to capture fine detailed visual information from the world, as explained by Land [2006]. As fovea is small, the eye rotates to puts fovea in a specified direction – getting the desired information.

The first registry from the study of eye movements remains from 1898 by Delabarre [1898], while investigating the relationship between eye movements and optical illusions. From there, many techniques were developed to record eye movements accurately. Furthermore, two works increased the interest in the field, as it revealed physiology relations with the visual system. The first one was the work of Yarbus [1967], who discovered that eye movements pattern change to satisfy the demands of a task. The second work, from Ballard et al. [1992], perceived that the eye always looks to objects involved in a task, as well as it anticipates the related motor action, acquiring the necessary information.

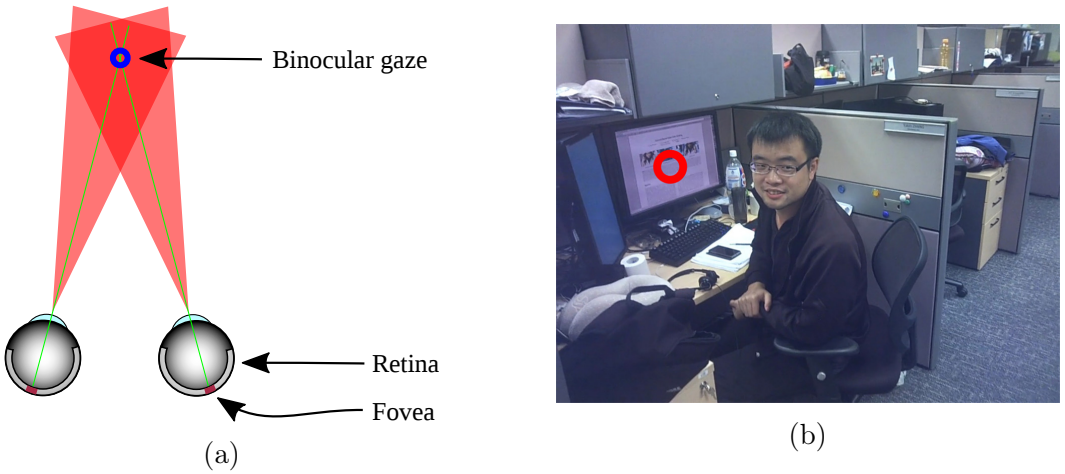
Our eyes perform different kinds of movements, with variable speeds, according to specific strategies. We can highlight three main eye movements: fixation, saccades, and blinks. Fixation is in the highest level of visual focus, allowing the capture of detailed information, characterized by the exposure of the eyes in a small region of the scene. Saccade represents the act of discovering new data through space, with fast eye movements covering a wide range, but with less acquisition of detailed information, as pointed by Land

[2006] and presented on Figure 2.4 . Blink is related to the complete lack of attention, happening when the user is blinking or rambling. These movements differ from the level of information to be acquired, with our vision swapping between fixation and saccades.

However, our eyes do not move freely by the environment. As the foveal vision is a limited resource, covering a small area of the scene, it searches for relevant information guided by visual selection mechanisms. The bottom-up mechanism is performed into the eye, detecting features related to the image such as contours, colors, luminance, and faces, triggering eye movements to these regions without passing through the brain. On the other hand, the top-down mechanism is under the explicit control of the observer's brain, being used to perform tasks. During our experience in the world, these selective mechanisms compete to acquire better information that can prevent imminent danger, as well as allow the execution of complex operations. Another feature of our visual system is their linkage to reinforcement paths on the brain, indicating that many of the eye movements are learned, as mentioned by Hayhoe and Ballard [2005].

2.5.1 Eye trackers

Eye tracker studies came from the early twentieth century when psychologists were researching about human behavior. Initially, eye trackers were invasive, big, and occupied entire rooms. Modern eye trackers are small, light, and can even be carried on the user's head, the so-called wearable eye trackers. Eye trackers are mainly divided into screen-



(a) (b)

Figure 2.3: Binocular gaze. (a) The green line represents the information captured by the fovea. The point of this line that intersects something into the world is the gaze. The intersection of the foveal region on both eyes is the binocular gaze. (b) The red circle is the binocular gaze, captured by an eye-tracker at A*STAR Ego-Gaze (ASTAR) dataset.

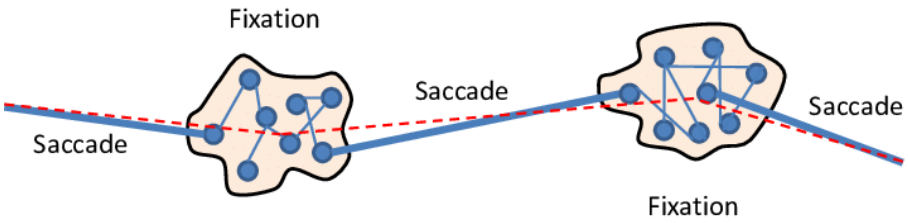


Figure 2.4: Eye movements. The eyes are in constant motion between regions of interest. While staying in these regions, the slight eye movement performed is called *fixation*. The movement performed between these regions is called saccadic movement, also known as *saccade*. Reprinted from Krueger et al. [2016].

(a) (b) (c)

Figure 2.5: Modern screen-based eye trackers. (a) Eyeteck VT3 mini. (b) Gazepoint GP3. (c) Tobii Eye X.

based and wearables. Screen-based remains fixed near a TV monitor and capture gaze data from a person looking to a stimulus coming from the monitor in a third-person view, as presented in Figure 2.5. Wearable devices are usually mounted on a glass frame, recording everything that the user sees in an unconstrained first-person view, as shown in Figure 2.6. These camera-based devices allow us to capture the wearer's gaze by creating a 3D model of the eye, given a sufficient number of observations. Using algorithms that detect the user's pupil and corneal reflection, it projects a location in the space where it predicts that the beholder is looking at – the gaze.

Gaze research was used to investigate neurological and psychological relations of eye movements in attention and memory studies and to aid in diagnosing mind conditions. In marketing analysis, it was used to characterize the importance and appearance of developed products; in simulation to capture insights about user interactions such as their attention and perceived objects. More recently gaming industry is integrating eye-trackers to their games, letting users control their characters through their eyes. In computer vision, gaze was employed on many tasks such as analysis of collaborative

(a) (b) (c)

Figure 2.6: Modern wearable eye trackers. (a) SMI. (b) Ergoneers Dikablis. (c) Pupil Core.

scenarios, action recognition, and user interaction.

Despite the increasing popularity, eye-trackers still remains expensive, in special the wearable eye trackers. One of the cheapest industrial eye trackers, the Pupil Core, costs around €2740 – turning difficult to produce gaze-enabled egocentric datasets. In this thesis, we propose to mount an affordable eye tracker, using consumer cameras and a 3D printer. We present step-by-step instructions at Appendix A. The software used for gaze processing is the open-source solution Pupil Core from Pupil Labs[1].

[1] `https://pupil-labs.com/`

Chapter 3

Related Work

Techniques capable of extracting and compiling information from First-Person Videos are becoming a must-have method due to the exponential growth in the amount of video data. In this chapter, we present a comprehensive view of the area by exploring remarkable works of video summarization, fast-forward, hyperlapse and gaze.

3.1 Video Summarization

The ultimate goal of summarization methods is to produce a compiled story composed of several segments of the input video. After segmenting the video, these methods rank each segment concerning a definition of what is relevant, hereinafter referred to as semantic. As pointed out by Molino et al. [2017], the semantic definition of recent works is grouped into *important to the viewer* or *important to the wearer*. The majority of summarization techniques are in the first group.

3.1.1 Important to the viewer

Techniques present in this group generally rely on features obtained from the input video, such as chromatic, saliency and motion.

In the work of de Avila et al. [2011], they proposed to use k-means clustering algorithm to select the most representative frames of Third-Person Videos. From an equally distributed sample of the original video, color histograms in HSV space are obtained and delivered to k-means - grouping frames likely to be near. The nearest frame of each cluster centroid, the keyframe, is the most representative frame of that cluster. A storyboard of the original video is composed using the keyframes.

Focusing on the most important objects and people which the camera wearer interacts, Lee et al. [2012] proposed to build a storyboard using high-level saliency cues. For each frame, a set of egocentric and object features were calculated in candidate regions to describe their importance. These features were related to hands distance, frame center distance, object-likeness, size, and other geometrical features. A regressor was trained on the obtained features to predict the frame's importance and select events on the original video aiming for a representative storyboard.

Usual video skimming approaches focus on the diversity or representativeness of subshots to build a summary. Instead, Lu and Grauman [2013] aimed to create a coherent chain of video subshots. Given an egocentric video, subshots were extracted based on motion features classification among frames. Objects present on each subshot were scored considering geometrical aspects, as well as their influence and presence among the other subshots, used to calculate each subshot chain relevance, over the set of possible subshot chains. The output video is composed of the optimal chain of subshots, that contains the best event connectivity.

With the advent of Deep Neural Networks (DNNs), many tasks in computer vision were benefited from methods capable of learning directly from data. In video summarization, it is essential to capture the video structure information to understand the high-level semantic present along with the video. Zhang et al. [2016] proposed to use a Long Short-term Memory (LSTM) model able to create summarized videos in the form of storyboards and subshots. Two Bi-LSTM were trained using deep visual features obtained from a GoogLeNet model, being able to predict frame-level importance scores - creating summarizations that ensure diversity maintaining semantic information.

Another approach to create summarized videos involve the direct input of the users about their preferences, generally using text descriptions, as in the work of Choi et al. [2018]. A pairwise ranking model built over a DNN model was trained using image and text data is capable of learning a semantic embedding function that maps frames and sentences to a joint embedding space. Video summaries were built assigning relevant subshots to each sentence on the user input text. A hidden Markov model restricts the subshots to be temporally aligned.

3.1.2 Important to the wearer

Works in this group derive the frame relevance from cues representing wearer intentions, inferred from external sensors information as well as visual-based features.

In the work of Aizawa et al. [2001], brain waves were used to capture the wearer

intents in order to create video subshots. Specifically, α and β waves, as they present specific patterns when a person feels awake, interested, or excited - a wearer cue to describe relevance to a video. These brain waves can be used to directly select frames and construct a storyboard. Another approach consists in filtering a set of subshots previously obtained by segmenting a video, according predefined motion categories: *moving*, *turning*, and *standing*. α waves were used to score the subshots into five levels of importance, and the selected ones are those with a score higher than the third level of importance.

Using an eye-tracker sensor to measure camera wearer attention in an egocentric video, Xu et al. [2015] created subshots that attend to representativity and diversity. The wearer's attention was captured using gaze, the focus of the attention of a person. A region descriptor is computed around the gaze movement position for each frame. Consequent frames similar to each other are grouped. These groups form the initial subshots highlighted by gaze. The final set of subshots, *i.e.*, optimal set, was modeled as a constrained modular maximization problem that takes into account diversity, compactness, and the representativeness of each subshot - measured by the gaze fixation movements quantity on each subshot.

Another sensor commonly used in action recognition tasks is the Inertial Measurement Unit. In the work of Li et al. [2017], three-axis accelerometer and gyroscope data were used by to select keyframes on egocentric videos. Video and sensor features were embedded into a joint space and used as input to sparse representation methodology for each video. The dictionary created minimizes reconstruction error, while keeping the sparseness constraint. Keyframes were selected by minimizing the reconstruction error of original frames using the sparse coefficients - composing the summarized version of the video.

Global Positioning System (GPS) data were used in the work of Varini et al. [2017], in which mixed wearer attention behavior and viewer preferences to build a summarized video. Attention behavior used motion and frame quality features fused with positioning data to calculate the wearer attention behavior score. GPS coordinates were also used to calculate viewer preferences score, searching into the web for related places near that position. The storyboard was composed of the optimal choice of subshots that maximize the attention and viewer preference scores.

3.2 Fast-Forward

Fast-forward techniques create shorter versions of the original video that attends the restriction of continuity. Traditional fast-forward consists of applying a constant

speed-up rate on the original video. Semantic fast-forward segments the video into regions given importance criteria and apply different speed-up rates on them. Usually, the emphasis on semantic information is given by reducing the desired speed-up rate on important segments.

To address the problem of creating a guidance route video, Okamoto and Yanai [2014] proposed a method to dynamically control video playing speed based on the semantic content of video sections. Using egomotion information and pedestrian crosswalk detection, video sections were classified as relevant if they contain crosswalks or present turning movements, moments that the guide needs attention - reducing the playback speed.

Yao et al. [2016] tackled the problem of video highlighting detection with applications on video summarization. Video highlighting is related to the discovery of memorable moments for the user into a video. A highlight score is calculated for each frame of an input video, using a pairwise deep ranking model, composed of a two-stream Convolution Neural Network (CNN) that learns from frames appearance and temporal dynamics. The obtained video scores were used as input data to summarization techniques such as video skimming and timelapse.

Instead of performing the segmentation step, commonly used on semantic fast-forward techniques, Lan et al. [2018] proposed an online method able to learn how to perform fast-forward. Using summarization labels as ground-truth, a reinforcement learning approach is able to choose when video frames should be skipped, in real-time, as it reads the video stream. However, the method fails in static scenes, as the model learns how to skip frames without being aware of their content.

3.2.1 Hyperlapse

Egocentric videos present an inherent shake related to camera wearer motion. Fast-forwarded versions amplify this shakiness, leading to hard to see videos. Hyperlapse techniques arose to solve the stability problem on fast-forwarded egocentric videos while keeping the desired overall speed-up rate.

The seminal work of Kopf et al. [2014] proposes to reconstruct 3D geometry of the scene, in order to build a First-Person timelapse video with a smoothly moving camera - the so-called hyperlapse. Primarily, the 3D scene is reconstructed for every frame on input video with structure-from-motion algorithms. Over the reconstructed scene, the camera path is planned to be as smooth as possible along with the video. The last step of the method uses rendering, stitching, and blending techniques, to generate a timelapse

stabilized version of the original video. The output video presents impressive visual results; however, it contains some reconstruction artifacts and has a high computational cost.

Instead of dealing with complex scene geometries, Poleg et al. [2015] used an adaptive frame sampling to fast-forward an egocentric video addressing the smoothness constraint. The frame sampling is modeled as an energy minimization problem into a directed acyclic graph whose nodes correspond to the frames of the input video. An edge represents the transition cost among frames, taking into account the shakiness, velocity, and visual appearance between them. The shortest path of this graph leads to the hyperlapse version of the input video. The technique fails when there is no translation of the camera, as is built to deal with shakiness induced by moving cameras.

Parallel to the previous work, Joshi et al. [2015] attempted to solve the hyperlapse problem using dynamic programming, achieving real-time processing without sensors, and handling significantly camera motion. Their method uses Dynamic-Time-Warping (DTW) to jointly model the smoothness and time restriction, being able to select frames that balance matching the target speed-up rate and minimizing frame-to-frame motion in the output video. The transition cost between frames was calculated as a weighted sum of frame-matching, velocity, and acceleration costs. The frame-matching cost reflects the reprojection error and lack of overlap among frames. Velocity address the desired speed-up, while acceleration cost prevents visual jumps in the output video. Transition cost is used in the DTW technique, whose output is the set of frames that composes the hyperlapse. A final step consists of 2D smoothing, using frame matching and cropping. The method presents good visual results and is surprisingly fast, running on a mobile phone. However, like all other techniques presented in this subsection, all frames of the video are considered equally relevant, without any semantic consideration.

3.2.2 Semantic Hyperlapse

Semantic Hyperlapse methods attempt to create video summaries that attend to restrictions of continuity, visual smoothness, and relevance. These methods prioritize sections of the video which contains information related to relevant aspects, highlighted by a de-acceleration in the hyperlapsed video.

In one of the first works in the area, Ramos et al. [2016] segmented the egocentric video into semantic and non-semantic regions and generate a semantic hyperlapsed version of the video. The authors considered the presence of faces as semantic content, assigning relevance to a frame based on a linear combination of the face classifier confidence, spatial

centrality, and bounding box size. A semantic thresholded value is used to segment video portions due to their semantic load, applying different speed-ups to emphasize semantic ones. The frame sampling is modeled as an optimization problem through graph-based approach, similar to the work of Poleg et al. [2015], where nodes correspond to the frames of the input video and edges represent the transition cost among frames, taking into account the shakiness, velocity and visual appearance, semantic load and desired speed-up violation between them. The output of the shortest path algorithm onto the graph is the semantic hyperlapse video.

Extending the previous work, Silva et al. [2018b] proposed a parameter-free and fully automatic fast-forward technique, using a Convolution Neural Network to score video frames and a multi-importance approach to segment relevant regions of the input video. The proposed CNN, named Coolnet, was trained on images that compose the most liked videos on the Internet, being used to score video frames due to their *coolness*. Then, the video is segmented using a multi-importance approach, that considers different levels of relevance, assigning different speed-up rates to them. They also proposed to automatically select the set of hyperparameters of the transition cost, one step of the frame sampling problem, using the Particle Swarm Optimization algorithm from Kennedy and Eberhart [1995].

Modeling the adaptive frame selection problem as a Minimum Sparse Reconstruction problem, Silva et al. [2018a] used a Sparse Coding approach to select the set of frames that better accelerates the input video, while preserves semantic content. The sparse set of frames selected from the input video minimizes the reconstruction error based on movement, appearance, content and sequence features. A smoothing frame transition step solves gaps that may appear in high frame transitions. Thus, combining a sparse-based frame sampling technique with a smoothing transition step creates an end-user visually pleasant video. Moreover, using Locality-constrained Linear Coding (LLC) formulation, the problem can be solved analytically, leading to a low computational cost.

Applying personalized-based techniques to solve the problem of defining the frame relevance to the wearer does not work properly, such as one object could catch the wearer attention even if it is not in his/her preference set. For example, in risky situations, where rarely the object that represents the imminent danger will be one of our favorites. Another possible scenario is the contact with previously unknown objects that may provoke curiosity. Moreover, an object that relies on user preferences may not attract its attention in a particular situation, turning complex the task of describing the user behavior by a fixed set of elements. To tackle this lack of certainty about the user focus, we propose to use the gaze in the context of the hyperlapse problem.

3.3 Gaze

One of the first attempts to track and record eye movements were made by Delabarre [1898], trying to explain geometrical optical illusions. In his experiment, a plaster ring was attached to the eye, connected to levers and pulleys, recording eye movements into a kymographic cylinder. Nowadays, modern eye-tracking devices are camera-based, using geometry calculus to track and record eye-gaze - exploring relations between eye movements and motor actions, as pointed by Land [2006].

Many researchers tried to find the connection between gaze and behavioral mechanisms. In the sixty's, Yarbus [1967] analyzed eye movements along with cognitive processes, revealing that seeing is inextricably linked to the observer's goals. Subjects looking at pictures were submitted to answer related questions, evoking different patterns of eye movements on them, clearly related to information required to solve the problem. These eye behaviors suggested a top-down attention mechanism, goal-driven; as opposed to reflexive, stimulus-driven, bottom-up attention mechanisms known until them. Many task-oriented studies, as mentioned by Hayhoe and Ballard [2005], discovered that given the demands of a task, eyes are positioned in elements related spatio-temporally with the task to satisfy its demands. Fixations often precede the motor control, in a *just-in-time* strategy - acquiring specific information for the action to be realized. This could be observed in visuo-motor coordination tasks such as cooking, walking, and driving.

Despite the crescent availability of eye-tracking devices, the options are yet limited as well as expensive. Thus, the need for predicting the gaze emerged. On static scenes, a way to measure human attention is to calculate a saliency map, which describes regions that attract human attention when looking for a picture. This bottom-up attention model is based on the feature integration theory, were distinct visual features such as color, intensity, and contrast on specific areas of the image leads to high saliency regions. Mixing pixel-data, object properties, and semantic attributes, Xu et al. [2014] predicted the human gaze on images through a linear SVM classifier. The training phase involved the mentioned features as well as human fixation maps, an average of eye-tracking information from persons when free viewing images, to predict saliency maps for static images. Using DNN models, Huang et al. [2015] performed a transfer learning approach using fixation maps to achieve state-of-art performance on saliency prediction. As an alternative to eye tracking to describe human attention, Jiang et al. [2015] proposed a psychophysical paradigm using observer interaction to approximate the human gaze in visual exploration. Human eye-fixations on images were also predicted using LSTM-based saliency model, such as in the work of Cornia et al. [2018]. The presented network focuses on the most salient regions of the image, iteratively refining the predicted saliency map, through an attention mechanism.

However, when performing tasks, saliency models alone were inefficiently to describe/predict were in the scene the human attention is. As the bottom-up and top-down attention mechanisms compete for eye dominance along with our interaction of the world, a deeper understanding of their relationship could bring better results. In the work of Li et al. [2013], egocentric cues from object manipulation tasks are used to estimate egocentric gaze on videos. Based on psychophysical experiments that indicate that gaze, head, and hand acts in a coordinated manner in many tasks, cues such as head motion, hands position, and center bias are feed to a random regression forest to predict gaze location. Their results show the applicability of the method in gaze prediction, as well as boosting segmentation and action recognition tasks. Exploring the human attention transition when performing hand-eye coordination tasks, Huang et al. [2018] proposed a hybrid CNN model that integrates bottom-up and top-down attention information to predict gaze on an egocentric video. A two-stream network pre-trained on ImageNet dataset extract temporal and spatial features, that feed bottom-up and top-down attention modules, modeled through saliency and attention transition. A late fusion network merges the result, outputting the predicted gaze. Modeling gaze uncertainties as a distribution, Li et al. [2018] predicted gaze as well as actions on egocentric video. Using a Two-Stream 3D ConvNet (I3D), it jointly learns how to predict gaze and action recognition tasks, with the output of specific convolutional layers of the referred network. The joint training boosted action recognition results - given the relationship between gaze and actions performed. However, gaze prediction is modeled in a bottom-up fashion, leading to saliency-like performance.

Gaze information has attracted attention in many areas, such as product and system design, visual advertisement, customer behavior analysis, shopping layout, and computer-human interaction. In computer vision, many tasks were boosted with the use of gaze information. In collaborative scenarios, Higuch et al. [2016] demonstrated that gaze improves the task completion time by a worker when seeing the collaborator's gaze+gestures in a workspace. They performed three user studies evaluating the performance of participants when realizing assembling, identifying, and arranging objects tasks. Gaze helped workers to capture collaborator's implicit intents as well as immediate instructions, boosting activity completion time, and minimizing the occurrence of mistakes. Focusing on activities that require eye-hand coordination, Fathi et al. [2012] used gaze and visual features to train an SVM able to recognize daily actions on egocentric videos. The spatiotemporal relation among gaze, scene features (objects classification, color histogram, segmentation) and action was modeled onto a probabilistic generative model - able to recognize actions and predict gaze. Gaze information was also used recently to study the relationship between the bottom-up visual attention and task-based visual analysis Polatsek et al. [2018]. Through three low-level analytical tasks, it was analyzed the eye-tracking of 47 participants to determine the influence of bottom-up image features when a person performs exploratory and confirmatory analysis and presentation of data.

To the task of video-guidance, the work of Damen et al. [2016] propose a unsupervised method to discover objects and their usage by multiple users in order to create an assistance system. It extracts useful data from egocentric videos such as objects appearance and position, triggered by the presence of user gaze fixations. When assisting novice users, it presents the most common objects usage snippets, as well as the most likely object to be used next, automatically. On health-related studies, Gu et al. [2017] created a vision-aided system based on visual analysis of gaze tracking to illustrate reading patterns and help specialists on detection and verification of mind wandering cases. It created an eye-tracking graph (ETGraph), where nodes represent clustered fixation patterns, and edges are the saccades between them. ETGraph-based system aid users in understanding reading patterns of persons and identify anomalous behaviors that can be characterized as mind wandering.

This thesis takes a step towards emphasizing the relevant moments to the wearer, enabling quick access to the information therein. We propose an attention model based on the gaze that fast-forwards First-Person Videos while keeping the visual smoothness and the required speed-up.

Chapter 4

Methodology

Semantic hyperlapse methods highlight video segments given a previously defined semantic, constraining these techniques to be used for specific purposes. As the wearer intents are dynamic along with his/her interaction with the environment, methods must be able to deal with a set of preferences that may change along the time, to extract useful information. In this chapter, we present a methodology able to highlight the fluid interests of the wearer using gaze information.

Our model is based on four main factors: i) the visual interaction of the wearer with scene components; ii) temporal and iii) spatial-visual relevance of the scene component; iv) avoidance of over watching video portions.

Modern wearable eye-tracking devices provide video and gaze data, which are used to assign relevance to each frame according to the recorder gaze. In the first step of our method, the visual interaction step, the object in focus is identified among all objects detected, using gaze and image information. Its trajectory is tracked in the temporal-visual relevance step using a tracking algorithm. Visual and geometrical features of the object in focus are obtained in the spatial-visual relevance step, also contributing to the frame scoring. In the last step of our method, a novelty term penalizes objects that have been focused for a long time. After scoring all frames of the video and feeding the semantic hyperlapse technique, a new video is created, emphasizing the relevant parts to the wearer. Figure 4.1 illustrates the overview of the methodology.

4.1 Visual interaction

In First-Person Videos, the capture device is usually attached to the recorder's body or head and follows its movements. Additionally, when attached to the recorder's head through an eye-tracker device, gaze data provides an important clue to infer the wearer preference at that moment, since gaze is deeply related to the user's intentions and interactions when performing tasks, as described by Yarbus [1967]. The recorder's

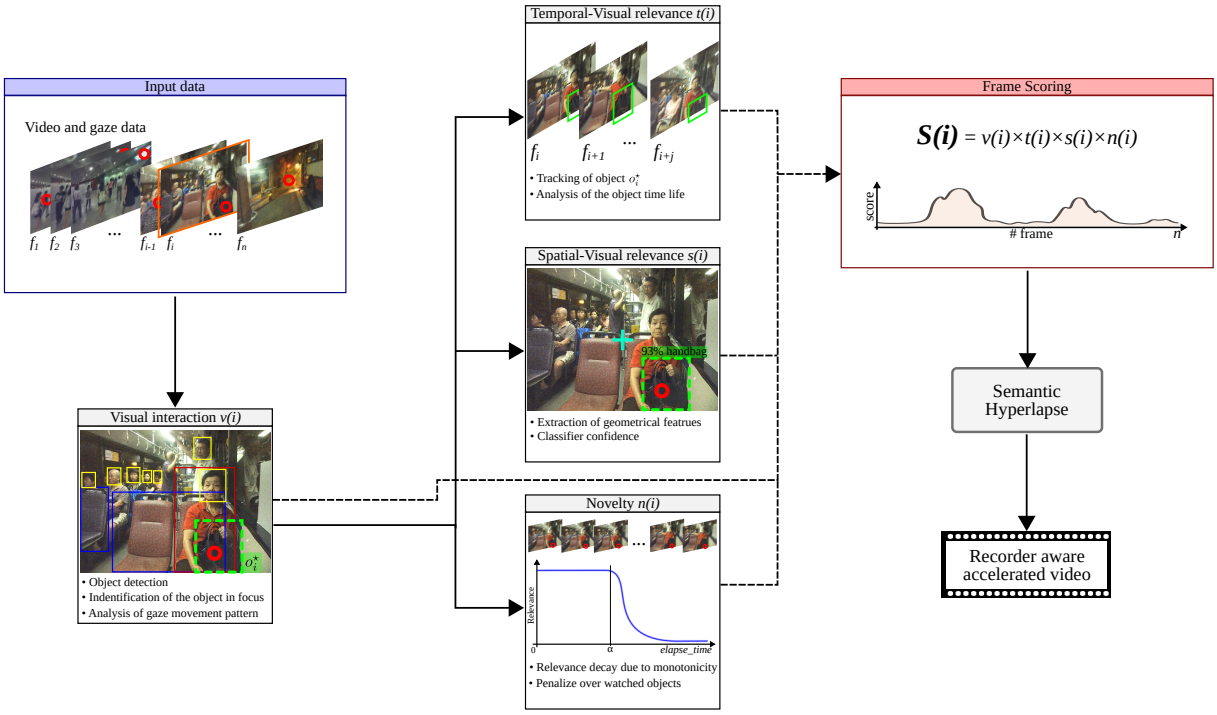


Figure 4.1: Proposed gaze-based semantic hyperlapse methodology. Using gaze and video data provided by wearable eye-tracking devices, the object in focus $o_i^\star$ is identified (green dashed box) and its trajectory is tracked while it gots user attention. Analysis of temporal, visual and geometrical features of $o_i^\star$ composes the frame score, which feeds a semantic hyperlapse technique, resulting in a new video that emphasizes the relevant parts to the wearer.

visual interaction is modeled through the analysis of the scene components and the gaze position. The scene components are defined using CNN object detectors.

Let f_i be the i-th frame of the video, $\mathbf{g}_i = [x, y]^T$ be the gaze position regarding frame i coordinate system and $\mathcal{O}_i$ the set of bounding boxes of the detected objects in the frame i. To infer which object is observed by the wearer, it is verified whether the gaze position $\mathbf{g}_i$ is lying inside of the bounding boxes of objects in the set $\mathcal{O}_i$. The result is the set of objects $\mathcal{O}_k \subset \mathcal{O}_i$.

The non-occluded object with the smallest area is selected, assumed as the foreground object, as follows

$$o_i^\star = \underset{o_j \in \mathcal{O}_k}{\operatorname{argmin}} A(o_j^{bb}), \tag{4.1}$$

where the function $A(x^{bb})$ returns the area of the bounding box x^{bb} around the object x. After inferring the scene component that the recorder is looking at, the visual interaction component $v(i)$ of the i-th frame is obtained regarding the eye-movement pattern $mp(\mathbf{g}_i)$. The $mp(\mathbf{g}_i)$ is provided by the eye-tracker through calculation of raw tracking data, considering the eye position along the time. It is a classification of the eye movement at the

frame i. The $v(i)$ term is defined as follows:

$$v(i) = \begin{cases} 1.0, & \text{if } mp(\mathbf{g}_i) = \text{Fixation;} \\ 0.5, & \text{if } mp(\mathbf{g}_i) = \text{Saccade;} \\ 0.0, & \text{if } mp(\mathbf{g}_i) = \text{Blink or } \mathcal{O}_k = \emptyset. \end{cases} \tag{4.2}$$

The presented values for the visual interaction term $v(i)$ were chosen to express the associated levels of attention, with the higher value assigned to fixation movement, a medium value to saccade, and a null value to blink patterns.

At the end of this step, the term related to the visual interaction $v(i)$ and the object $o_i^\star$ which the user interacts in the i-th frame of the video are defined.

4.2 Temporal-Visual relevance

It is expected that objects of interest remain longer in the recorder's field of view. This behavior is modeled as the temporal-visual relevance of the semantic information, penalizing scene components that have a short time life.

The life time of the object $o_i^\star$ is computed by tracking the object along with the video. While tracking, many information are kept, such as the object identification, its bounding box, the classifier's confidence regarding the object class, tracking start timestamp, tracking duration, and position. The tracking is performed from the start of the visual interaction until the attention got dispersed, or the object disappear. Tracking data is used to compute the temporal relevance of the object $o_i^\star$, at the frame i as

$$t(i) = \frac{tr(o_i^\star)}{T_{max}}, \tag{4.3}$$

where $tr(o_i^\star)$ indicates the tracking duration in frames of the object $o_i^\star$, and T_{max} is the longest tracking duration considering all objects of a video V composed of n frames,

$$T_{max} = \max\left(tr(o_i^\star)\right) \ \forall \ o_i^\star \in V \mid i = 1, \ldots, n. \tag{4.4}$$

$t(i)$ ranges from 0 to 1 and has the highest value when evaluating the longest tracking.

The tracking also brings robustness to our methodology since temporary oclusion or sensor failures during the visual interaction will not mask the real wearer intention. For example, if the recorder is focused looking to television and someone passes in front of the television, the tracking process still holds the object position for a few frames, waiting for interaction with the user's gaze.

4.3 Spatial-Visual relevance

In many computer vision tasks, such as saliency and semantic fast-forward, visual analysis of the semantic elements such as their centrality, size, complexity, and color-related characteristics are used to describe image importance. We highlight the works of Judd et al. [2009], Xu et al. [2014], and Ramos et al. [2016]. In a similar fashion, geometrical and visual features were embedded in our model as follows:

Relative area ($a_{o_i^\star}$). The first geometrical feature is the ratio of the focused object $o_i^\star$ box bounding area w.r.t. the whole image, as presented on Figure 4.2:

$$a_{o_i^\star} = A(o_i^{\star bb})/A(f_i), \tag{4.5}$$

where x^{bb} indicates de box around object x, and $A(.)$ is a function that returns the bounding box area. $a_{o_i^\star}$ ranges from 0 to 1, having the higher value for objects closer to the camera.

Centrality ($c_{o_i^\star}$). Visual perception experiments performed by Bindemann [2010] suggest that observers tend to fixate the center of the observable area initially. This strong center bias persists even when visual features are present off-center location on scenes. This way, being central in an image captured by a wearable device also represents a fundamental cue about the scene component relevance for the wearer. The centrality is given by

$$c_{o_i^\star} = \frac{1}{1 + \left\| C(o_i^{\star bb}) - C(f_i) \right\|_2}, \tag{4.6}$$

where $C(.)$ is a function that returns the central point of a rectangle, and $\|.\|_2$ is the euclidean norm. The maximum value of 1 happens when the center points are overlapping each other. Ilustrated at Figure 4.3.

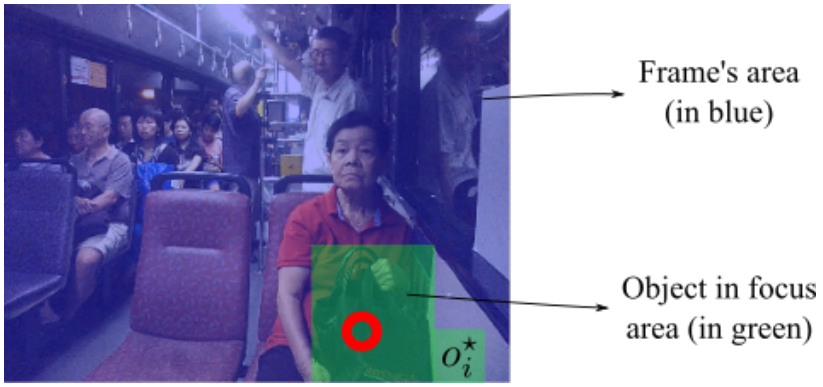


Figure 4.2: Relative area metric. Ratio of focused object area by image area, representing relative proximity to the recorder.

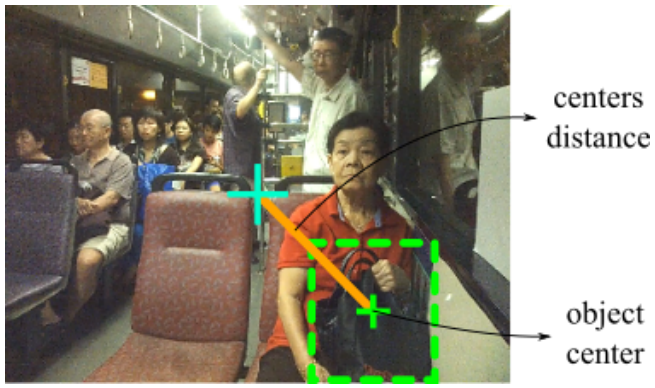


Figure 4.3: Centrality metric. Euclidian distance between image center and object center.

Focus ($m_{o_i^\star}$). The third geometric feature used to compose the visual relevance is the focus on the object $o_i^\star$. This feature is derived from centrality, assuming that the recorder is fully concentrated in an object when he is looking at any point of the vertical centerline of the bounding box, as showed on Figure 4.4. As far as the gaze is horizontally from the of the central section of the object, lower the focus is

$$m_{o_i^\star} = \frac{1}{1 + \left|\mathbf{g}_{i_x} - C(o_i^{\star bb})_x\right|}, \tag{4.7}$$

where $|.|$ is the absolute diference. Values range from 0 to 1 with maximum value when the user is looking directly to the middle of the object.

The geometrical features obtained are based on information acquired by an object detector, being directly related to the classifier confidence ($d_{o_i^\star}$). We consider the $d_{o_i^\star}$ as a weighting factor, once high confidence generally implies in high definition, non-occluded and facing forward objects. The final spatial-visual relevance score s_i of the i-th frame is given by the sum of the three previous defined geometric features weighted by the visual feature:

$$s(i) = (a_{o_i^\star} + c_{o_i^\star} + m_{o_i^\star}) \times d_{o_i^\star}. \tag{4.8}$$

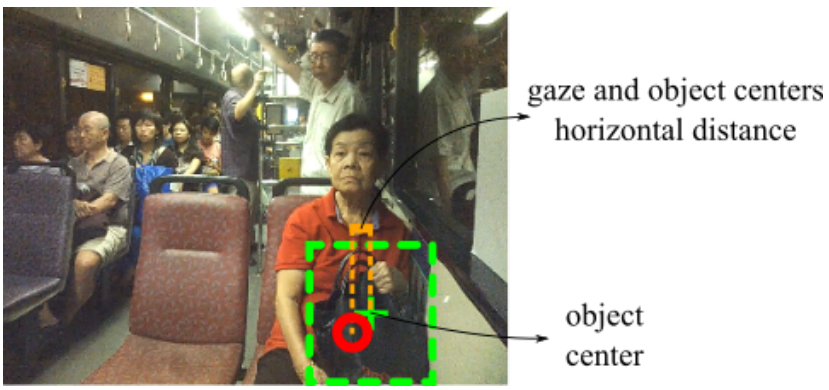


Figure 4.4: Focus metric. Horizontal distance between image center and gaze.

4.4 Novelty model

Since First-Person Videos rely on an unconstrained video domain, in some cases, the recorder could perform a long task that requires visual focus, *e.g.*, a video of a software developer looking to the computer screen, a gamer playing, or an attendant talking with customers. Given the design of the visual interaction term, when infering frame relevance, all frames in a video segment could receive high scores, even when the task itself is not relevant in its entirety. To avoid overly watching these segments, we propose to use a novelty weighting factor $n(i)$ — based on the tracking of object $o_i^\star$ and a decay function. The decay function is applied to video segments, penalizing frames after a defined period α. A conditional exponential decay function was chosen to decrease frame importance after the α period rapidly.

The novelty weighting factor $n(i)$ is defined as:

$$n(i) = \begin{cases} 1 & \text{, if } et_i < \alpha; \\ e^{-\frac{et_i \ - \ \alpha}{2}} & \text{, otherwise.} \end{cases} \tag{4.9}$$

The term et_i is given by

$$et_i = i - tr(o_i^\star, \text{start time stamp}), \tag{4.10}$$

where $tr(o_i^\star, \text{start time stamp})$ returns the initial frame when the user first focused in object $o_i^\star$. The condition α is the number of frames considered as fresh information, after that the frame relevance will be penalized by the lack of novelty. Values range from 0 to 1 with lower values related to tedious moments of the video.

4.5 Frame scoring

Given each term defined before, for the i-th frame, we assign the score

$$S_i = v(i) \times t(i) \times s(i) \times n(i). \tag{4.11}$$

In the visual interaction term $v(i)$, the attention level of the wearer is evaluated, preferring moments where the beholder's eyes remain stationary, *i.e.*, fixation movements. Temporal-visual relevance term $t(i)$ explores the continuity of visual contact of the recorder with scene components, as an object that attracted the recorder's attention

for a significant amount of time should be important – highly scoring frames in a segment with long-time tracking objects. Scene components far from the wearer, in the image border, or visually noisy, will penalize the frame score, as they do not present substantial interactions or interest to the wearer - performed by spatial-visual relevance term $s(i)$. Although long-time tracking objects represent relevant video elements, they can become the most important elements in the video. To address diversity in our method, these long appearing objects are penalized, *i.e.*, frames with newly focused objects will score higher than the frames in which the same object has appeared enough. All the terms from the Equation 4.11 can be obtained separately, however, the most important ones should occur jointly, *e.g.*, a long tracked object should be more important if it is near and in a central position of the image. In that way, terms dependency is modeled through a multiply operation among them. By processing all frames in the video, a relevance profile of the video is created.

4.6 Semantic Hyperlapse

To achieve our goal of creating a video emphasizing the relevant moments for the recorder, semantic hyperlapse algorithms are used, feeding it with our gaze-based video relevance profile as semantic information. These techniques create the semantic hyperlapse video by segmenting it into relevant and non-relevant segments according to the semantic score level assigned to each frame of the video. Using an optimization function based on the length of each segment, different speed-up rates are defined for each type of segment. This process creates the emphasis effect by assigning lower rates to relevant segments.

A frame sampling approach is applied to select an optimal set of frames regarding visual smoothness, the amount of semantic information, and length of the final video. The result is a visually pleasant, hyperlapse video with the relevant segments highlighted. Since our method can infer the user's attention during the recording, and this model fed the technique, the output is a video emphasizing the moments that caught the user's attention.

In this thesis, we feed the state-of-the-art techniques regarding semantic hyperlapse, Multi-Importance Fast-Forward (MIFF), and Sparse Adaptive Sampling (SAS) along with our gaze-based video semantic profile S. MIFF is a graph-based optimization method, presented in the work of Silva et al. [2018b], that splits the original video into semantic segments, by levels of semantic information; while the SAS method of Silva et al. [2018a] accelerates a video by applying an adaptive frame sampling scheme based on sparse coding formulation and minimum reconstruction problem.

Chapter 5

Experiments

In this Chapter, we describe the datasets used in the experimental evaluation, the implementation details, and the baselines techniques.

5.1 Datasets

Were used two datasets composed of videos recorded using eye-tracking wearable devices: Georgia Tech Egocentric Activity (GTEA Gaze+)[1] and ASTAR[2].

The GTEA Gaze+ dataset was presented in the work of Fathi et al. [2012] and its data collected in Georgia Tech's AwareHome - an instrumented house with a kitchen containing all standard appliances and furnishings. The dataset is composed of videos of seven meal preparation activities performed by six different subjects, as shown in Figure 5.1. Participants must prepare meal following to food recipes: American Breakfast, Turkey Sandwich, Cheese Burger, Greek Salad, Pizza, Pasta Salad, and Afternoon Snack. Each activity takes around 10-15 minutes to be completed on average. The dataset was built for action recognition and gaze prediction tasks, taking advantage of top-down attention mechanisms when executing coordinated hand-eye tasks - such as cooking. In coordinated hand-eye tasks, the eye anticipates hand movements aiding their execution to attend the goal, being participative along with task execution. The dataset is composed of 37 videos in HD resolution (1280×960) at 24 frames per second (fps) with gaze tracking data, action annotations, and hand masks. Gaze location was obtained using SMI eye-tracking glasses at a sampling rate of 30 Hz. Actions were annotated with ELAN, a linguistic annotator, using predefined verbs and nouns. An action is defined as a short temporal segment on the video, such as putting sauce on pizza crust or washing the mushrooms. GTEA Gaze+ dataset is used in action recognition, summarization, gaze prediction, and other tasks.

[1] Publicity available at `http://www.cbi.gatech.edu/fpv`.
[2] Available under request.

Figure 5.1: Recipes preparation in Georgia Tech Egocentric Activity dataset. Gaze is presented as a red circle.

The ASTAR dataset is the second used in our experiments, published by Ma et al. [2012]. It is an unconstrained egocentric dataset with gaze annotation. This dataset was recorded in an opened scenario, where participants perform free daily activities, such as socializing, going to work, manipulating objects, and cleaning the house, as presented in Figure 5.2. Unlike GTEA Gaze+, there were no instructions about what activities should be performed during the recording. In that way, ASTAR is very challenging, as the wearer performs various types of activities, triggering top-down and bottom-up attention mechanisms, *i.e.*, being attracted by visual characteristics of the scene as well things related to an internal state. The gaze is the only information available to present user intent. ASTAR contains 12 videos in HD resolution (1280×960) at 24 fps with gaze tracking data and annotated actions, recorded by six subjects. The gaze information was captured using SMI eye-tracking glasses, sampled at 30 Hz. Annotated actions were made by three volunteers, describing overall information of the scene as the time of day, place, a summary, and an activity such as social, walk, object, transit, and observe.

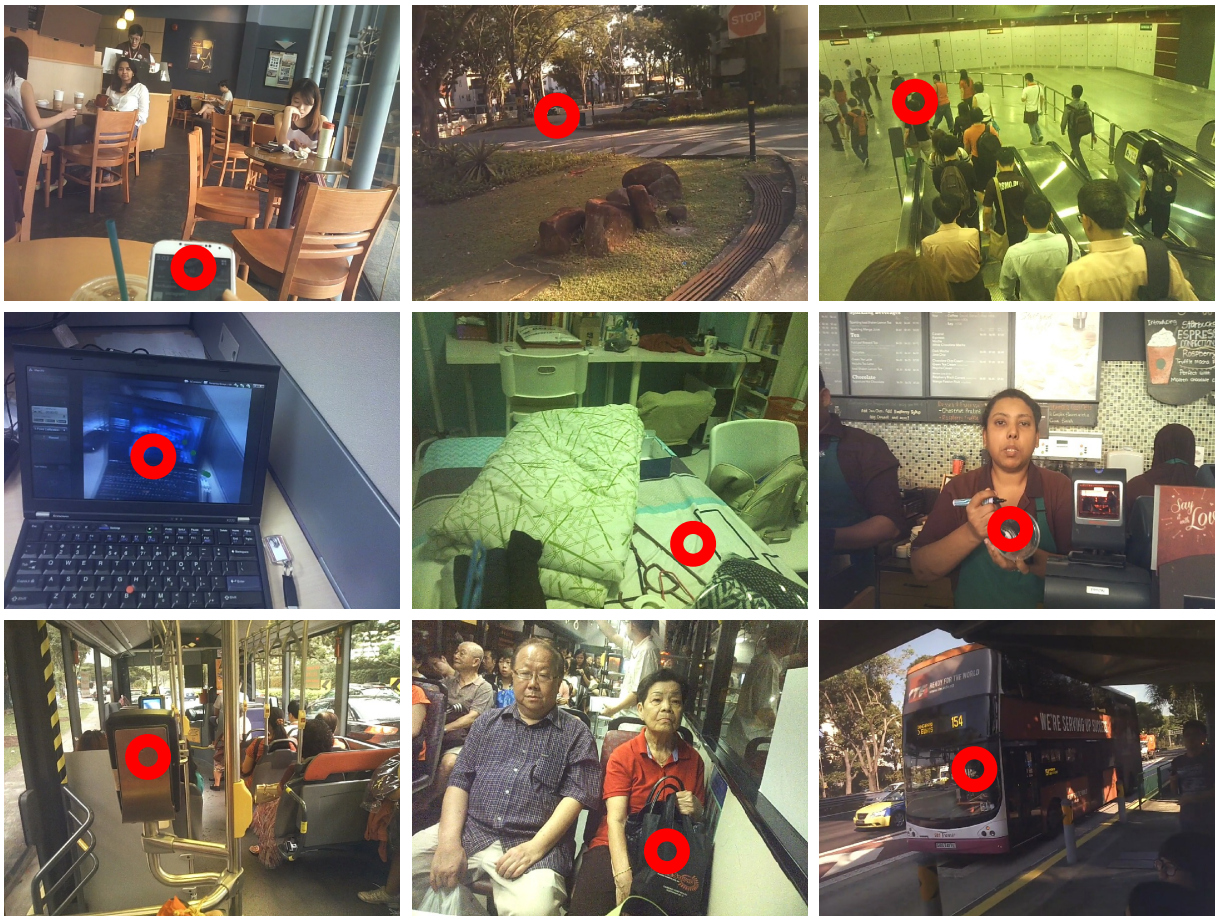


Figure 5.2: Unconstrained recording of daily living activities in A*STAR Ego-Gaze dataset. Gaze is presented as a red circle.

5.2 Implementation details

The gaze data present on both datasets were sampled at 30 Hz, while the video stream was sampled at 24 fps. To register both data, we apply a linear interpolation on gaze data and generate new gaze location coordinates sampled at 24 Hz. For each interpolated gaze point, if it belongs to an existent interval of fixation or saccade, the respective eye movement type is assigned. Otherwise, the gaze point is considered as being a blink. In the end, we performed a cleaning step, by removing fixations with negative coordinates as well transforming fixations with less than 500 ms (12 frames) in saccades, because of sensor failures.

As discussed in Section 4.1, the scene component analysis was performed through an object detector. For each frame, the selected object detector returns a set of bounding boxes with the candidate objects with their respective confidence, excluding those below a predefined threshold. We used the default confidence threshold of 0.5 of the selected

method, the You Only Look Once (YOLO)[3]. YOLO is a CNN designed to the task of object detection proposed by Redmon and Farhadi [2017]. The premise of YOLO is the accurate and fast detection of objects, using a custom CNN able to perform image classification together with bounding box offset predicition in one pass, given a previous known dataset distribution. The custom network, so-called *Darknet-19* is mainly composed of 19 convolutional layers and 5 maxpooling layers, as depicted in Table 5.1. We use the YOLO pre-trained version on Microsoft Common Objects in Context (MSCOCO)[4] dataset, being able to detect up to 80 classes, as defined on the research work of Lin et al. [2014].

Given the set of detected objects and the gaze information, we look for the object in focus and track it until the user stops the visual interaction, or the object disappear. The tracking is performed by using the tracking-by-detection algorithm Simple Online and Realtime Tracking (SORT)[5], proposed by Bewley et al. [2016]. SORT is as Multiple Object Tracking algorithm that predicts the object location using Kalman filters and Hungarian algorithm, to be used in realtime applications. We set the parameters *min_hits* and *max_hits* as 3 and 7, respectively. With this setting, the tracking process starts when there are at least 3 intersections between the gaze and the detected object. Also, it keeps the tracking of the object for at least 7 frames, as soon as the visual interaction is interrupted, bringing robustness to abrupt changes. During our experiments, these values

Table 5.1: Darknet-19 architecture

Layer Type	Filters	Size/Stride	Output
Convolutional	32	3×3	224×224
Maxpool	-	$2 \times 2/2$	112×112
Convolutional	64	3×3	112×112
Maxpool	-	$2 \times 2/2$	56×56
Convolutional	128	3×3	56×56
Convolutional	64	1×1	56×56
Convolutional	128	3×3	56×56
Maxpool	-	$2 \times 2/2$	28×28
Convolutional	256	3×3	28×28
Convolutional	128	1×1	28×28
Convolutional	256	3×3	28×28
Maxpool	-	$2 \times 2/2$	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Maxpool	-	$2 \times 2/2$	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7

[3]Publicity available at `https://github.com/pjreddie/darknet`.
[4]Publicity available at `http://cocodataset.org`.
[5]Publicity available at `https://github.com/abewley/sort`.

balanced the quantity and quality of tracked objects, respectively.

Regarding video content and frame-rate, in our experiments, we use $\alpha = 96$ in the novelty model (Section 4.4), as the videos were recorded 24 fps, treating the first 4 seconds as fresh information.

To choose a good value to saccade eye movement in visual interaction $v(i)$ score term, we performed a grid-search method for values from 0.0 to 1.0 with a step of 0.1 on two videos of each dataset. On ASTAR, the videos were *S007_C02* and *S002_C03*. On GTEA Gaze+, the chosen were *Shaghayegh_American* and *Alireza_Snack*. After executing the methodology with the varying parameter, we evaluate the semantic quantity present on videos generated, achieving better results with saccade equal to 0.5 - being the chosen value.

After computing the semantic video profile (a curve with semantic scores of each frame), we used a Gaussian function with $\sigma = 30$ to smooth abrupt changes in the score caused by fast camera viewpoint changes. As the last step, a Min-Max normalization is applied, fitting values between 0 and 1.

5.3 Competitors

We compared our method against two semantic information extractors, *i.e.*, CoolNet and YOLO-based, from two state-of-the-art semantic hyperlapse techniques: the MIFF and SAS method.

MIFF is presented in the work of Silva et al. [2018b]. It is a fully automatic fast-forward method for egocentric videos, that provides a balance between smoothness and emphasis on relevant parts. By using a multi-importance approach, their methodology splits the original video into semantic segments (composed of relevant frames), by levels of semantic information. The adaptive frame sampling step involves creating graphs for each segment. On a graph, a node represents a frame and each edge represents the transition costs among them. It takes into account terms related to instability, appearance, velocity and semantics. The frame sampling step is processed by computing the shortest path, where the selected nodes represent the frames chosen for the fast-forwarded video. To classify the semantic contents of a video, the authors proposed a CNN trained on web video statistics of *coolness* — the CoolNet. This network architecture is based on the VGG16 model, trained on MIT Places205 dataset and fine-tuned on a defined domain. The CoolNet receives as input a frame f_i and returns a value that indicates the probability of this frame be considered *cool* by general users — being out first challenger.

The Sparse Adaptive Sampling method of Silva et al. [2018a] accelerates a video

by applying an adaptive frame sampling scheme based on sparse coding formulation and minimum reconstruction problem. An advantage of the SAS method is its capability of leveraging the smoothness restrictions related to abrupt camera movements, which are common in First-Person Videos. The semantics presented by the authors was defined as the presence of faces captured by the Normalized Pixel Difference (NPD) face detector algorithm. However, it is not fair to use face semantics in datasets like GTEA Gaze+ Gaze+, as there are no faces present in videos. Instead, we use YOLO-based semantic in our experiments, obtained through the weighted sum of spatial features, such as relative size and centrality, from detected objects at an image feed to the YOLO network.

In our experiments, we evaluated the power of extracting semantic information based on the gaze and compare the results using YOLO and Coolnet semantic approaches. Thus, we fed the two methods of adaptive frame sampling, *i.e.*, sparse coding (SAS) and sampling based on graph optimization (GS), and we analyzed the quality of the fast-forwarded videos. The parameter setting of both techniques was performed as recommended by the authors in the original works.

Chapter 6

Results

In this chapter, we compared the accelerated videos generated extracting semantic information using our gaze-based method, the CoolNet and YOLO-based method. We performed several quantitative and qualitative evaluations by analyzing the emphasis on high attention tasks, diversity of objects being focused, and fast-forward metrics.

6.1 Results on visual attention tasks

We borrow the idea of this first quantitative analysis from the work of Xu et al. [2015]. In their work, the authors evaluated the quality of the subshot extraction by overlapping the subshots with ground truth task segmentation. The motivation of this analysis relies on the premise that if a moment is labeled, then it has a meaning to the watcher. In this metric, we analyzed the overlap of the emphasized video segments and ground-truth tasks, which caught the user's attention.

6.1.1 Evaluation metric

Each person proceeds in a different way to perform a task, delivering more or less attention to elements on the scene, *e.g.*, a person who does not cook has difficulties when cutting legumes, differently from a chef who can cut it without looking at the legumes entirely. Instead of just evaluating the intersection of selected segments with ground truth tasks, we perform a filtering step on ground truth tasks, selecting those on which the eye-tracking monitor logged at least 50% of the task duration as gaze fixation - defined as high attention tasks. Semantic fast-forward techniques define the speed-up rate of each segment regarding its relevance. So as high is the score, the lower is the speed-up rate. In

our experiments, we found that some fail in identification or even low scores could lead to an emphasis moment. However, the speed-up applied to these misclassified emphasis segments is a value close to the speed-up of the non-emphasized segments. Therefore, we only consider a true emphasis when the acceleration rate applied to the segment is smaller than half of the required speed-up. The *emphasized actions metric* is calculated by counting overlaps between segments of true emphasis on the accelerated video and high attention tasks, given by

$$Ea = |E \cap \mathcal{H}|, \tag{6.1}$$

where $\mathcal{H}$ is the set of high attention tasks of the video, E is the set of true emphasized segments on accelerated video and $|.|$ is the cardinality of a set of objects.

For more detailed performance assessment of the semantic extractor influence in the creation of the final video, we fed the graph frame selection algorithm from the work of Silva et al. [2018b], *i.e.*, MIFF, with three semantic score extractors: scores given by the CoolNet, the YOLO-based method, and our gaze-based scoring methodology. The experiments were performed on the GTEA Gaze+ dataset due to annotations' detail. For the ASTAR dataset, we could not use the annotations since they are related to general and continuous actions - such as walking, looking around, opening the door, and so on.

6.1.2 Discussion

Table 6.1 shows the number of high attention tasks that were truly emphasized by the sampling techniques using different semantic scores. The values were presented as a percentage between truly emphasized moments and high attention tasks for each video.

Accelerating videos with CoolNet presented the worst result, being unable to capture and emphasize user attention when performing tasks. The result is related to defined semantic - what is considered "cool" by the network on presented videos. YOLO-based method detects many kitchen elements on the scene; however, it does not emphasize any specific object at all - creating a fast-forwarded video with high acceleration rates (speed-up). Our method models the temporal interaction of the wearer with the surrounding environment, better-selecting segments containing high attention tasks, being more relevant to the wearer, in 28 out of 33 videos of the dataset.

Table 6.1: Percentage of truly emphasized tasks on accelerated videos, among different semantic score extractors, on GTEA Gaze+ dataset. For each video, we evaluate the number of high attention tasks that were truly emphasized by our proposed method and the competitors. Best in bold.

Videos	CoolNet	YOLO-based	Ours
Ahmad_American	1%	5%	**14**%
Ahmad_Burger	**18**%	1%	6%
Ahmad_Greek	1%	13%	**26**%
Ahmad_Pasta	1%	10%	**16**%
Ahmad_Pizza	4%	2%	**12**%
Ahmad_Snack	4%	**8**%	4%
Ahmad_Turkey	**12**%	4%	7%
Alireza_American	8%	2%	**14**%
Alireza_Burger	10%	10%	**12**%
Alireza_Greek	**24**%	6%	16%
Alireza_Pasta	**23**%	7%	17%
Alireza_Pizza	6%	6%	**16**%
Alireza_Turkey	**17**%	5%	6%
Carlos_American	3%	3%	**20**%
Carlos_Burger	5%	4%	**18**%
Carlos_Greek	2%	0%	**27**%
Carlos_Pasta	7%	7%	**8**%
Carlos_Pizza	6%	29%	**42**%
Carlos_Snack	3%	9%	**27**%
Carlos_Turkey	6%	6%	**14**%
Rahul_American	0%	4%	**20**%
Rahul_Burger	12%	0%	**16**%
Rahul_Greek	3%	12%	**20**%
Rahul_Pasta	1%	8%	**19**%
Rahul_Pizza	5%	3%	**19**%
Rahul_Snack	8%	**16**%	**16**%
Rahul_Turkey	0%	6%	**16**%
Shaghayegh_Pizza	2%	17%	**20**%
Shaghayegh_Snack	13%	12%	**15**%
Yin_American	14%	6%	**19**%
Yin_Burger	11%	10%	**17**%
Yin_Greek	**26**%	7%	18%
Yin_Pizza	0%	0%	**12**%
Yin_Snack	8%	15%	**26**%
Yin_Turkey	3%	0%	**26**%
Mean	7.6%	7.2%	**17.2**%

6.2 Diversity of objects in focus

In this evaluation, we analyzed the capacity of the methods providing a rich diversity set of objects. We quantified the detected objects in a Region of Interest (ROI) around the user's attention focus, through the gaze.

6.2.1 Evaluation Metric

We are always gazing at something into the scene, to perform an action, or even explore, visually interacting with objects in the environment. On this metric, our objective is to quantify the amount of semantic information captured by the accelerated video, that is useful to the camera wearer, using the desired semantic extractor. Given an accelerated video, we extracted the gaze position on each frame and defined a ROI based on gaze movement pattern, *i.e.*, fixation, saccade, or blink. A ROI is defined around the gaze coordinates to deal with gaze measurements imprecision, along the eye-tracking process. For fixations, we used a ROI of 100×100 pixels sized centered on the gaze location. Saccades have a smaller area the fixation, with 50×50 pixels, because of the dynamic behavior of the movement, while the 0-area ROI is assigned to blinks. The next step is to detect all objects in a frame using YOLO and obtaining the objects that intersect with the defined ROI. Objects that intersect with ROI are considered of the potential interest of the user. We also discard objects in the class person since when performing daily activities in crowded environments, the user attention focus could not correspond to their real intention or desire. The diversity metric is defined by

$$Do = \sum_{i=1}^{N} \left| R(g_i) \cap \mathcal{O}_i \right|, \tag{6.2}$$

where N is the number of frames on the accelerated video, $\mathcal{O}_i$ is the set of detected objects on the i-th frame, g_i is the gaze position on the i-th frame, $R(.)$ is a function that returns the ROI around the gaze position and $|.|$ is the cardinality of the resulting set of objects.

The metric outputs the number of objects in the surroundings of the wearers' gaze, along with the video. We used the videos from the ASTAR dataset and the graph frame selection algorithm (GS) to fast-forward the videos based on the three mentioned semantic scores: CoolNet, Yolo, and our model.

6.2.2 Discussion

Table 6.2 presents the results of the diversity of objects in the focus metric. Low interaction of objects and user gaze was detected when using CoolNet and YOLO-based semantic data, showing that selected frames were unable to highlight user intents. Our model, instead, better-selected frames that contain objects in the vicinity of gaze position, detecting more objects in 7 out of 10 videos. This result shows that the proposed method

capture scene elements that attract the interest of the wearer, different from the other techniques that only detect their presence.

6.3 Fast-forward analysis

The evaluation of fast-forwarding techniques relies on two quality factors of the produced final video: the length and the frame transition shakiness. The length factor is related to the adaptive frame sampling process. An ideal fast-forward technique should sample the number of frames determined by the required speed-up while keeping the transition between frames as smooth as possible. Shakiness is the more challenging issue for fast-forwarding techniques since it turns the final video unwatchable and even nauseating.

6.3.1 Evaluation metric

Instability is an index to measure how visually pleasing is the video considering the shakiness of frame transitions. It calculates the cumulative sum over the standard deviation of pixels in a sliding window over the video, defined as

$$In = M\left(\frac{1}{N}\cdot\sum_{i=1}^{N}\sqrt{\frac{\sum_{j\in B_i}(f_j-\overline{f}_i)^2}{(N_B-1)}}\right), \tag{6.3}$$

where N is the number of frames in the video, B_i is the i-th buffer composed by N_B

Table 6.2: Object detection measurement among techniques on ASTAR Dataset. We present the object count for each video and technique for a defined ROI. Best in bold.

Videos	CoolNet	YOLO-based	Ours
S001_C01	112	**123**	96
S001_C02	115	124	**128**
S001_C03	129	135	**157**
S002_C01	53	41	**88**
S002_C02	330	369	**606**
S003_C02	345	**517**	418
S003_C03	44	58	**136**
S004_C02	72	55	**154**
S004_C03	140	107	**176**
S007_C01	**350**	342	197
Mean	169.0	187.1	**215.6**

temporal neighborhood frames, f_j is the j-th frame of the video, $\overline{f}_i$ is the average frame of the buffer B_i , $M(\cdot)$ is a function that returns the mean value for the pixels of a given image and In indicates the instability index of the video. The less the In value, smoother the resulting video.

The Speed-up metric measure how well the hyperlapse techniques respect the required speed-up. It is given by the absolute difference of the achieved speed-up rate from the required speed-up rate. The achieved speed-up rate is the ratio between the number of frames in the original video and its fast-forward version, as showed on Equation 6.4 In this thesis, we used 10 as the required speed-up. We performed frame sampling using SAS and MIFF. Both methods were fed with our proposed frame scoring in GTEA Gaze+ and ASTAR datasets.

$$Sp = \left| \frac{N}{N_a} - r_s \right|, \tag{6.4}$$

where N is the number of frames in the video, N_a is the number of frames in the accelerated video, r_s is the required speed-up rate, $|.|$ is absolute value of a scalar and Sp is the deviating speed-up value from the required speed-up. Both metrics were defined by Silva et al. [2018b].

6.3.2 Discussion

Table 6.3 presents the speed-up and instability values for the output videos created by all three techniques. These results demonstrate that our gaze-based frame scoring is feasible to be combined with the semantic fast-forward techniques since it did not negatively affect the frame sampling step. It is worth noting that the combination of the

Table 6.3: Comparison of the methods regarding the achieved speed-up and video instability. We combined our scoring with the frame sampling step of SAS and MIFF. Best in bold.

	Dataset	**SAS**			**MIFF**		
		CoolNet	YOLO-based	Ours	CoolNet	YOLO-based	Ours
Instability	ASTAR	30.54	32.50	**30.26**	33.67	35.34	32.46
	GTEA Gaze+	22.32	**20.74**	23.15	27.18	27.36	28.66
Speed-up	ASTAR	**0.14**	**0.14**	0.16	0.27	1.62	0.38
	GTEA Gaze+	0.48	0.80	0.26	0.17	**0.02**	0.10

proposed scoring methodology with the SAS adaptive sampling achieved the best average value for Instability in the ASTAR dataset and the second-best for GTEA Gaze+. For the Speed-up metric, excluding a few outliers, all values represent decimal differences.

6.4 Qualitative analysis

In this section, we discuss the visual results when accelerating the input video using the different frame scoring (Ours gaze-based, YOLO-based, and CoolNet) combined with MIFF frame sampling on ASTAR and GTEA Gaze+ datasets.

To highlight the frame scoring capability of our technique, we present the top-10 scored frames according to CoolNet, gaze-based, and YOLO-based methods, for one video from each dataset. We select the top-10 frames performing a local maxima search, with a restriction of at least 240 frames (10 seconds) separating each peak.

Figure 6.1 presents the top-10 frames according to the semantic scores obtained using three methods: Coolnet, YOLO-based, and gaze-based - for the video *S007_C01*. This video records the path traveled by a person using an eye-tracker, from a workplace to a coffee shop. When using the CoolNet, the presence of radical sports and beautiful landscapes related elements could trigger a high score on evaluated frames. However, in this video, it highlights frames related to brighter regions and with wooden appearance. These frames should contain useful objects; indeed, it is not guaranteed that they lie on the user's attention focus. For example, in the *6th* frame, a person is in the central region of the scene, while the camera wearer looks to the column - an object that usually should not catch attention. Other frames such as 2, 5 and 10 present image-like landscapes, with far from view images, on which the wearer is blinking or looking outside the visual recorder camera area.

The YOLO-based semantic score is obtained through the sum of detected objects size and position on the frame. In this way, frames with higher scores contain many objects, preferably those near the camera wearer. It should be noted that the two most important frames have little or no interaction with the user gaze, being irrelevant to the user in the video context, as well as in frames 4, 9 and 10. Moreover, some of them do not bring any useful information about a user interest, such frames 1 and 8.

The top-10 frames obtained by the gaze-based scoring method highlights scenes were the user gaze interacts with objects detected on the scene. As our technique takes into account the object size, relative position, temporal relevance, and attention level - it is expected that emphasizes the implicit interaction of the user with scene related elements. This could be seen in interaction with people face to face, on frames 1, 2, 6 and

8; or when using hands to handle an object as in frames 3, 5 and 10.

CoolNet and YOLO-based semantics were unable to model the importance of the objects through time. These methods rely on pre-defined semantics that cannot be modified during their execution, *i.e.*, their context is static. The most important video segment selected by those techniques is defined by the presence of near elements on a single frame. Since they use only image characteristics, it may highlight video regions without relevant information to the camera wearer. Instead, our approach can dynamically select a subset of objects which are relevant to the camera wearer along with the video, without explicit intervention, just employing user attention — the eyes of the camera wearer act as a selector of what is important.

We show the top-10 frames according to the semantic scores obtained using three methods mentioned before, on *Yin_American* video, as exhibited in Figure 6.2. The semantic score obtained using CoolNet presents a noisy behavior, triggering high scores on scenes in which a white wall and a kitchen cabinet table appear. Four of these frames are a little blurred (1, 3, 5 and 8) or do not contain detectable objects in the central region of the image (4 and 10).

YOLO-based semantic score presents frames that contain important visual information; however, 6 out of 10 frames highlight the same object that occupies a large part of the scene - an oven. These frames also contain many objects that contribute to their high score. Our methodology gives high scores to frames that contain camera wearer interactions with objects, as well as not being affected just by the presence of the scene objects.

Gaze-based scoring methodology was capable of modeling the user interest dynamics when interacting with objects in the environment, even though we are always in contact with visual information. The tracking of the objects of interest, as well as the novelty model, are able to filter the visual information flood that we are exposed - as visualized into the graphs presented with the figures. While Coolnet and YOLO-based semantic information are noisy, gaze-based is smooth; being able to highlight different parts of the video relevant to the camera wearer.

In Figure 6.3, we present the semantic profile (blue line) and the speed-up rate for two specific frames from video *S003_C02* of the ASTAR dataset when using: CoolNet, YOLO-based, and our gaze-based attention model. The frame on the left received a high value in our model because of the interaction between the bus and the wearers' gaze. A crescent value is assigned when using YOLO-based information as a semantic score due to bus detection and its crescent proximity throughout the time. CoolNet output high semantic values due to the presence of trees in the background. In the frame on the right, our method assigned low values of relevance since there is no interaction with the user's attention (red circle) and a scene component. The YOLO-based semantics returned a higher score due to the presence of several people on the scene. For CoolNet, the low

scores obtained should be related to the indoor-like scenes, as the recorder is inside a bus, with little portion of *coolness* in the video such the presence of trees. YOLO detected

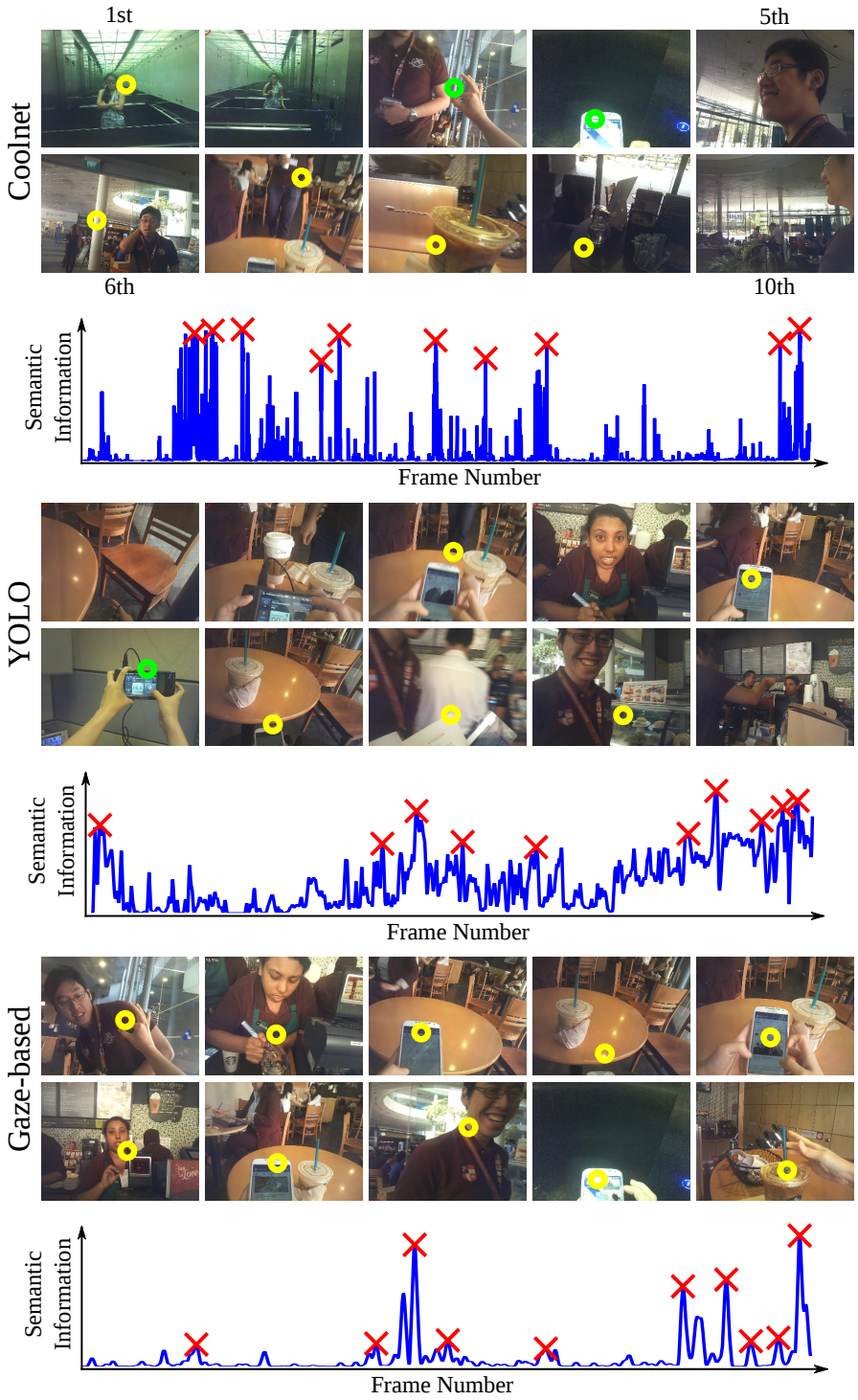


Figure 6.1: Top-10 frames, according to the semantic score computed by three different semantic information extraction approaches on *S007_C01* from ASTAR dataset. Frame importance is ordered from left to right, top to down. Wearers' gaze is presented as a circle, with fixations and saccades showed on green and yellow colors, respectively. Frames without circle mean blink. The score graph with the red crosses highlights the top-10 frames.

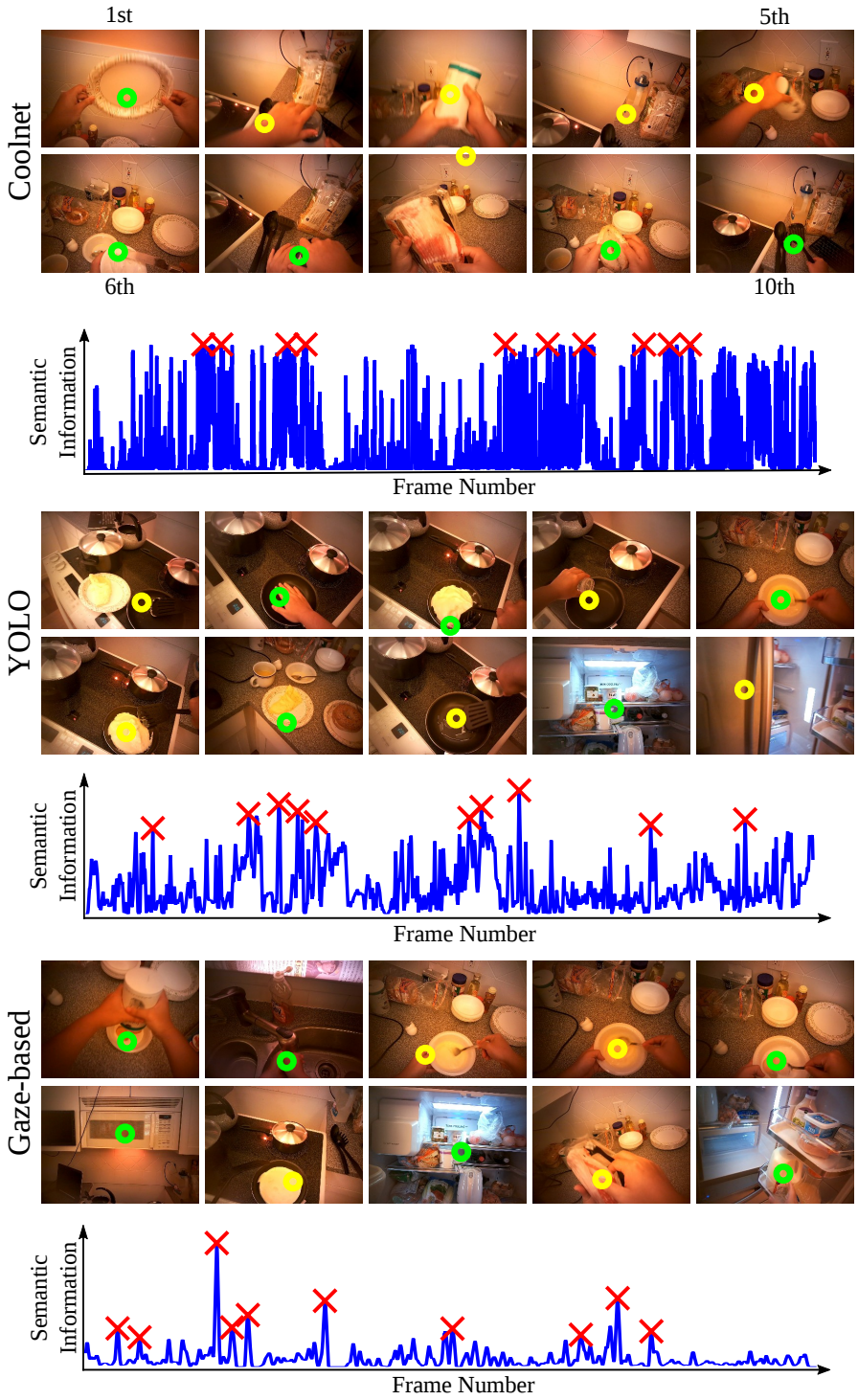


Figure 6.2: Top-10 frames, according to the semantic score computed by the three different semantic information extraction approaches on *Yin_American* from GTEA Gaze+ dataset. Frame importance is ordered from left to right, top to down. Wearers' gaze is presented as a circle, with fixations and saccades showed on green and yellow colors, respectively. Frames without circle mean blink. The score graph with the red crosses highlights the top-10 frames.

many objects in both images, leading to high speed-up rates on interesting segments while highlighting a moment in the bus where the gaze is lost. CoolNet gives high scores to

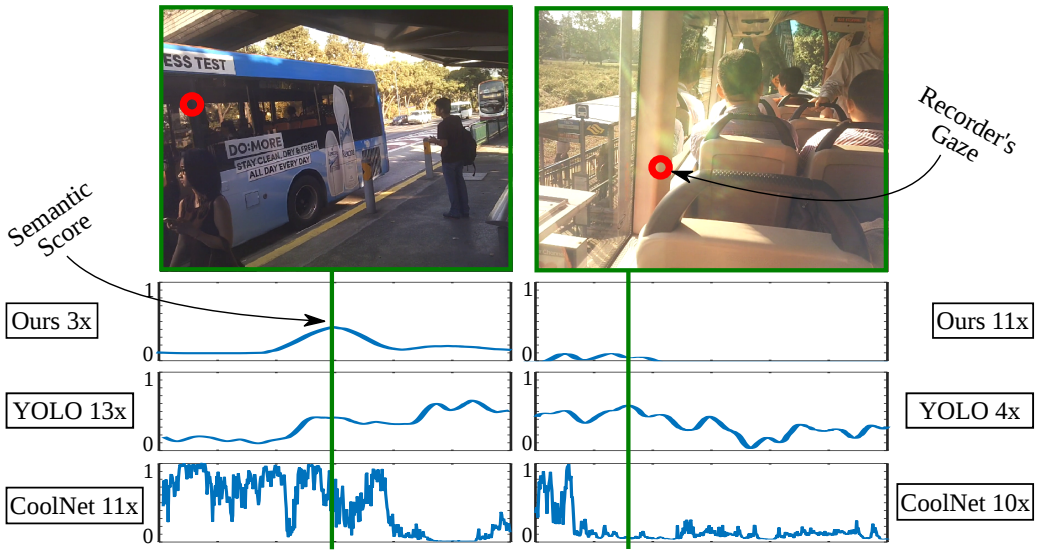


Figure 6.3: Example of two scenes and the semantic score computed by three different semantic information extraction approaches.

many regions of the video, being unable to really emphasize specific moments of them. Both images contain distinguishable scene elements, but our method remains unaffected by their presence unless it arouses the interest of the wearer.

Figure 6.4 shows parts of two videos recorded by different subjects preparing a snack receipt. The snack should be prepared rigorously following the recipe. We used the videos *Carlos_Snack* and *Shaghayegh_Snack* from GTEA Gaze+ Dataset. In this example, we present the semantic scores of our method for two different moments of the receipts execution. For the first subject, a high semantic score is assigned at the moment the person interacts with the microwave oven. The high value is related to the attention focus on the object along the time. When preparing the peanut butter, a small value is obtained, due to the absence of focus along with the action. When considering the video of the second subject, we observe the opposite. The microwave does not attract the attention of the user, receiving a low semantic score, differently of the bowl with the peanut butter, which received a high score for P2. Even following a recipe, each of the recorders acts in their way to complete the tasks with the same available resources, based on their experience and feelings. Our gaze-based methodology was capable of capturing this behavior, as highlighted by the wearer.

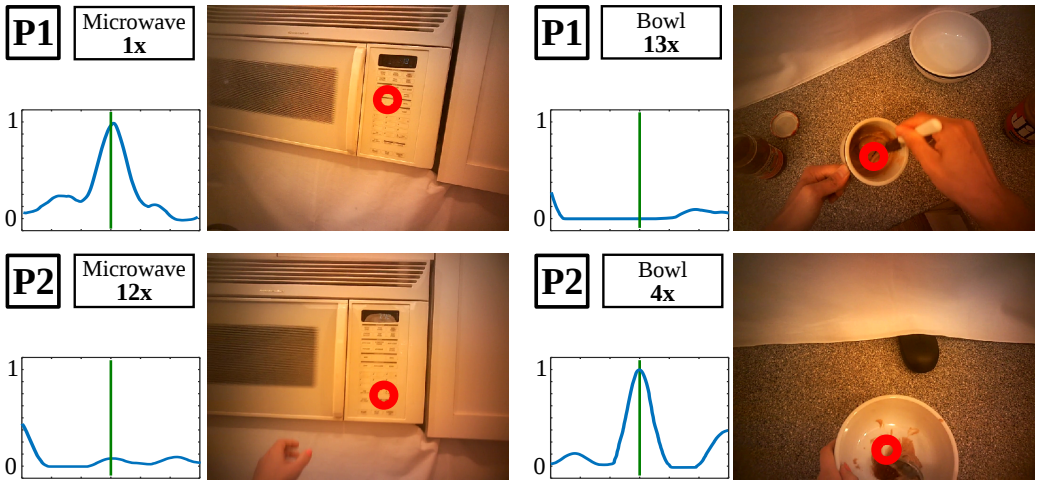


Figure 6.4: Two different persons (**P1** and **P2**) cooking a snack receipt. The blue curve shows the semantic score for each scene computed by our method. Red circle represents the gaze.

Chapter 7

Conclusion

In this thesis, we presented a new attention model based on the fusion of gaze and visual information, and a scene novelty modeling that is used for emphasizing relevant moments to the camera wearer on FPV. The attention model is based on visual and geometrical information of the image. Different from previous works, where the semantics are defined *a priori*, there is no requirement of predefining the semantic information in our model. The semantics are extracted by the wearer's behavior observed in the gaze. We evaluated our approach in two datasets, against two state-of-the-art methods on fast-forwarding first-person videos. They show that in the coverage of tasks that need user attention, our method shows a better average result of 9.6 percentage points to the best competitor. Futhermore, considering the semantic load present on the accelerated video, our method captured 15% more objects in the gaze surroundings than the best competitor. Still, our method does not impair the frame sampling step from the used hyperlapse methods. Visual results presented our method ability to capture the user's intentions, as well as the underlying behavior of the subject. Thus, our methodology outperformed the state-of-the-art techniques, generating accelerated videos that emphasize the user's behavior, generalizing well in many contexts.

7.1 Limitations

One major limitation of our method is the requirement of egocentric gaze data, as it gives clues about camera wearer intentions. This requirement constrains our methodology to be applied to a small set of egocentric videos - that one made using wearable eye trackers.

Moreover, wearable eye trackers are expensive, leading to a limited number of egocentric datasets containing gaze data, even more with medium to long videos, suitable to generate accelerated videos. These factors constrain our experiments to just a few datasets.

Another limitation is that the method performance is related to the object detection technique, as , *i.e.*, the number of classes that the detector is able to classify and its accuracy. Recent object detectors are generally trained on well-known datasets such as PASCAL-VOC and MS-COCO datasets, with 20 and 80 object classes, respectively. Moreover, as our methodology considers a visual interaction when the user gaze intersects an object perimeter, some objects can be over considered by us - their real perimeter is smaller than the detected bounding box.

Although our model penalizes overexposed segments of a tracked object on the video, it cannot handle the uniqueness of these objects. For diversity reasons, an already detected object should not appear anymore in the selected segments. For example, if the recorder interacts with an object in different instants of the video, being this object uninformative for the understanding of the video – such as a coffee machine.

At last, our scoring framework, summarized in Equation 4.11, does not weigh the frame scoring associated terms, leading all of them with the same importance. Some of them should have more impact when expressing the user's attention. Moreover, the multiplication strategy can put noise on the final score value just by the presence of small values in one of the terms of the equation.

7.2 Future work

One approach to solving the gaze dependency problem is to use gaze prediction algorithms, providing egocentric gaze data to feed our algorithm. Although, it should be noted that many of these methods predict gaze from image features, based on bottom-up attention mechanisms, not caring about the temporal dependency of gaze. Recent methods, as the work of Huang et al. [2018], mix bottom-up and top-down cues of a specific-domain task to predict the gaze. The obtained results are promising, indeed they can be applied only to videos of the same domain. Another approach is to predict the frame importance based on our methodology. An initial attempt was made using a 3D CNN, based on the work of Tran et al. [2015]. We modified the C3D architecture to perform gaze-based frame score regression, using the values obtained by our method. Other modifications include the size reduction of some layers to adjust the needed for data. Some tests were performed, and despite the initial results, some adjustments need to be made for a proper result in many scenarios. Using these approaches, may it should be possible to overcome the barrier of gaze annotated videos.

For network training and methodology evaluation, many more datasets providing gaze data were needed. Following previous attempts, an affordable eye tracker can be built

using consumer cameras, in order to create egocentric gaze-enabled datasets. Following the works of Kassner et al. [2014]; Schneider et al. [2011], we built an eye-tracker prototype, described in Appendix A.

To mitigate the problem of limited object detection in current techniques, an object detection model can be trained in other recently released datasets that provide up to 1200 categories, such as LVIS[1] dataset. Furthermore, to solve the problem related to our visual interaction detection, an object instance segmentation networks could be used instead of object detectors, such as Mask R-CNN from He et al. [2017], better selecting visually interacted objects.

To handle the object's uniqueness along with the accelerated video, dealing with object diversity, we can extract tracked object features, as well as scene features, in order to build a video index and allow objects to appear once in the video.

Dealing with the weighting problem of our frame scoring model, we can study the impact of each term and assign parameters that regularize their influence on the frame relevance. Multi-task learning approach can be employed to explore the parameters space, achieving an optimal set of parameters capable of best expressing the user's desire. To guide such approach, as well as to better understand the user's intentions, a qualitative study on the results obtained should be conducted.

[1] `https://www.lvisdataset.org/`

References

K. Aizawa, K. Ishijima, and M. Shiina. Summarizing wearable video. In *IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 398–401, Thessaloniki, GR, October 2001. doi: 10.1109/ICIP.2001.958135.

Jason S. Babcock and Jeff B. Pelz. Building a lightweight eyetracking headgear. In *Symposium on Eye Tracking Research & Applications*, ETRA '04, pages 109–114, San Antonio, Texas, 2004. ISBN 1-58113-825-3. doi: 10.1145/968363.968386.

Dana H. Ballard, Mary M. Hayhoe, Feng Li, and Steven D. Whitehead. Hand-eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 337(1281):331–339, 1992. doi: 10.1098/rstb.1992.0111.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, Phoenix, USA, 2016. doi: 10.1109/ICIP.2016.7533003.

Markus Bindemann. Scene and screen center bias early eye movements in scene viewing. In *Vision Research*, pages 2577–2587, 2010. doi: https://doi.org/10.1016/j.visres.2010.08.016. Vision Research Reviews.

Tao Chen, Aidong Lu, and Shi-Min Hu. Visual storylines: Semantic visualization of movie sequence. *Computers & Graphics*, 36(4):241–249, 2012. ISSN 0097-8493. doi: 10.1016/j.cag.2012.02.010. Applications of Geometry Processing.

J. Choi, T. Oh, and I. S. Kweon. Contextually customized video summaries via natural language. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1718–1726, Lake Tahoe, NV, USA, March 2018. doi: 10.1109/WACV.2018.00191.

Jaeger Claus and Siedersbeck Alfons. *Eye Safety of IREDs used in lamp applications*. OSRAM - Opto Semiconductors, 10 2018.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing (ITIP)*, 27(10):5142–5154, October 2018. ISSN 1941-0042. doi: 10.1109/TIP.2018.2851672.

Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding (CVIU)*, 149:98–112, 2016. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2016.02.016.

Sandra Eliza Fontes de Avila, Ana Paula Brandão Lopes, Antonio da Luz, and Arnaldo de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011. ISSN 0167-8655. doi: 10.1016/j.patrec.2010.08.004. Image Processing, Computer Vision and Pattern Recognition in Latin America.

E. B. Delabarre. *A method of recording eye-movements*, pages 572–574. University of Illinois Press, 1898. ISBN 978-1-4899-5379-7. doi: 10.2307/1412191.

Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision (ECCV)*, pages 314–327, Firenze, IT, 2012. doi: 10.1007/978-3-642-33718-5_23.

Yi Gu, Chaoli Wang, Robert Bixler, and Sidney D'Mello. Etgraph: A graph-based approach for visual analytics of eye-tracking data. *Computers & Graphics*, 62:1–14, 2017. ISSN 0097-8493. doi: 10.1016/j.cag.2016.11.001.

Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, April 2005. doi: 10.1016/j.tics.2005.02.009.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Veneza, ITA, 2017. IEEE. doi: 10.1109/ICCV.2017.322.

Keita Higuch, Ryo Yonetani, and Yoichi Sato. Can eye help you?: Effects of visualizing eye fixations on remote collaboration scenarios for physical tasks. In *Conference on Human Factors in Computing Systems CHI*, pages 5180–5190, San Jose, USA, 2016. doi: 10.1145/2858036.2858438.

X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 262–270, December 2015. doi: 10.1109/ICCV.2015.38.

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *European Conference on Computer Vision (ECCV)*, pages 789–804, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01225-0.

ICNIRP. Icnirp statement on far infrared radiation exposure. *Health physics*, 91(6): 630–645, December 2006. doi: 10.1097/01.hp.0000240533.50224.65.

M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, June 2015.

Neel Joshi, Wolf Kienzle, Mike Toelle, Matt Uyttendaele, and Michael F. Cohen. Real-time hyperlapse creation via optimal frame selection. *ACM Transactions on Graphics (TOG)*, 34(4):63:1–63:9, July 2015. ISSN 0730-0301. doi: 10.1145/2766954.

T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2106–2113, September 2009. doi: 10.1109/ICCV.2009.5459462.

Moritz Kassner, William Patera, and Andreas Bulling. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 1151–1160, Seattle, Washington, 2014. ISBN 978-1-4503-3047-3. doi: 10.1145/2638728.2641695.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4, 1995. doi: 10.1109/ICNN.1995.488968.

Johannes Kopf, Michael F. Cohen, and Richard Szeliski. First-person hyper-lapse videos. *ACM Transactions on Graphics (TOG)*, 33(4):78:1–78:10, July 2014. ISSN 0730-0301. doi: 10.1145/2601097.2601195.

Nikolaos Kourkoumelis and Margaret Tzaphlidou. Eye safety related to near infrared radiation exposure to biometric devices. *The Scientific World Journal*, 11:520–8, 03 2011. doi: 10.1100/tsw.2011.52.

Robert Krueger, Steffen Koch, and Thomas Ertl. Saccadelenses: Interactive exploratory fltering of eye tracking trajectories. In *IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)*, Baltimore, USA, October 2016. doi: 10.1109/ETVIS.2016.7851162.

W. S. Lai, Y. Huang, N. Joshi, C. Buehler, M. H. Yang, and S. B. Kang. Semantic-driven generation of hyperlapse from 360° video. *IEEE Transactions on Visualization and Computer Graphics*, PP(99), 2017. ISSN 1077-2626. doi: 10.1109/TVCG.2017.2750671.

S. Lan, R. Panda, Q. Zhu, and A. Roy-Chowdhury. Ffnet: Video fast-forwarding via reinforcement learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, June 2018.

Michael F. Land. Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25(3):296–324, 2006. ISSN 1350-9462. doi: https://doi.org/10.1016/j.preteyeres.2006.01.002.

Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346–1353, June 2012. doi: 10.1109/CVPR.2012.6247820.

Dongheng Li, Jason Babcock, and Derrick J. Parkhurst. openeyes: A low-cost head-mounted eye-tracking solution. In *Symposium on Eye Tracking Research & Applications*, ETRA '06, pages 95–100, San Diego, California, 2006. ISBN 1-59593-305-0. doi: 10.1145/1117309.1117350.

Y. Li, A. Kanemura, H. Asoh, T. Miyanishi, and M. Kawanabe. Extracting key frames from first-person videos in the common space of multiple sensors. In *IEEE International Conference on Image Processing (ICIP)*, pages 3993–3997, Beijing, China, September 2017. doi: 10.1109/ICIP.2017.8297032.

Yin Li, Alireza Fathi, and James M. Rehg. Learning to predict gaze in egocentric video. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3216–3223, Sydney, AU, 2013. ISBN 978-1-4799-2840-8. doi: 10.1109/ICCV.2013.399.

Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision (ECCV)*, pages 639–655, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01228-1.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, DEU, 2014.

Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, Columbus, USA, June 2013. doi: 10.1109/CVPR.2013.350.

Keng-Teck Ma, Rosary Lim, Peilun Dai, Liyuan Li, and Joo-Hwee Lim. Unconstrained ego-centric videos with eye-tracking data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop on Scene Understanding (SUNw)*, Providence, USA, 2012.

S. Mann. 'WearCam' (the wearable camera): personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic

memory prosthesis. In *IEEE International Symposium on Wearable Computing*, pages 124–131, Pittsburgh, USA, October 1998. doi: 10.1109/ISWC.1998.729538.

Ana Garcia del Molino, Cheston Tan, J. H. Lim, and A. H. Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, February 2017. doi: 10.1109/THMS.2016.2623480.

Fiona Mulvey, Arantxa Villanueva, David Sliney, Robert Lange, Sarah Cotmore, and Mick Donegan. Exploration of safety issues in eyetracking. Technical report, COGAIN EU Network of Excellence, 2008.

Masaya Okamoto and Keiji Yanai. Summarization of egocentric moving videos for generating walking route guidance. In *Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, pages 431–442, Guanajuato, MX, October 2014. doi: 10.1007/978-3-642-53842-1_37.

Patrik Polatsek, Manuela Waldner, Ivan Viola, Peter Kapec, and Wanda Benesova. Exploring visual attention and saliency modeling for task-based visual analysis. *Computers & Graphics*, 72:26–38, 2018. ISSN 0097-8493. doi: 10.1016/j.cag.2018.01.010.

Yair Poleg, Tavi Halperin, Chetan Arora, and Shmuel Peleg. Egosampling: Fast-forward and stereo for egocentric videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4768–4776, Boston, MA, USA, June 2015. doi: 10.1109/CVPR.2015.7299109.

Washington Luis Souza Ramos, Michel Melo Silva, Mario Fernando Montenegro Campos, and Erickson Rangel Nascimento. Fast-forward video based on semantic extraction. In *IEEE International Conference on Image Processing (ICIP)*, pages 3334–3338, Phoenix, USA, September 2016. doi: 10.1109/ICIP.2016.7532977.

Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, Honolulu, USA, July 2017. doi: 10.1109/CVPR.2017.690.

Nicolas Schneider, Peter Bex, Erhardt Barth, and Michael Dorr. An open-source low-cost eye-tracking system for portable real-time and offline tracking. In *Conference on Novel Gaze-Controlled Applications*, NGCA '11, pages 8:1–8:4, Karlskrona, Sweden, 2011. ISBN 978-1-4503-0680-5. doi: 10.1145/1983302.1983310.

M. M. Silva, W. L. S. Ramos, J. P. K. Ferreira, F. C. Chamone, M. F. M. Campos, and E. R. Nascimento. A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, June 2018a.

Michel M. Silva, Washington L. S. Ramos, Felipe C. Chamone, João P. K. Ferreira, Mario F. M. Campos, and Erickson R. Nascimento. Making a long story short: A multi-importance fast-forwarding egocentric videos with the emphasis on relevant objects. *Journal of Visual Communication and Image Representation (JVCI)*, 53:55–64, 2018b. ISSN 1047-3203. doi: 10.1016/j.jvcir.2018.02.013.

Michel Melo Silva, Washington Luis Souza Ramos, Joao Pedro Klock Ferreira, Mario Fernando Montenegro Campos, and Erickson Rangel Nascimento. Towards semantic fast-forward and stabilized egocentric videos. In *European Conference on Computer Vision Workshop (ECCVW)*, pages 557–571, Amsterdam, NL, October 2016. ISBN 978-3-319-46604-0. doi: 10.1007/978-3-319-46604-0_40.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, December 2015. doi: 10.1109/ICCV.2015.510.

P. Varini, G. Serra, and R. Cucchiara. Personalized egocentric video summarization of cultural tour on user preferences input. *IEEE Transactions on Multimedia*, 19(12): 2832–2845, December 2017. ISSN 1520-9210. doi: 10.1109/TMM.2017.2705915.

Bo Xiong, Gunhee Kim, and Leonid Sigal. Storyline representation of egocentric videos with an applications to story-based search. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4525–4533, December 2015. doi: 10.1109/ICCV.2015.514.

J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2235–2244, Boston, USA, June 2015. doi: 10.1109/CVPR.2015.7298836.

Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28, 2014. doi: 10.1167/14.1.28.

Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 982–990, Las Vegas, USA, June 2016. doi: 10.1109/CVPR.2016.112.

Alfred L. Yarbus. *Eye Movements During Perception of Complex Objects*, pages 171–211. Springer US, Boston, MA, 1967. ISBN 978-1-4899-5379-7. doi: 10.1007/978-1-4899-5379-7_8.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European Conference on Computer Vision (ECCV)*, pages 766–

782, Amsterdam, NL, 2016. ISBN 978-3-319-46478-7. doi: 10.1007/978-3-319-46478-7_47.

Appendix A

Wearable eye-tracking glass construction

In this appendix we provide instructions of how to build a wearable eye-tracking glass using consumer cameras and a free and open-source software provided by Pupil Labs[1].

Wearable eye-trackers glasses allow to capture egocentric gaze, describing the eye movements. However, they are costly and with limited options. At most, there are few datasets with gaze data publicity available - even more with medium to long sized videos.

To overcome this, we propose to build an eye-tracker device, following the ideas of Babcock and Pelz [2004]; Li et al. [2006]; Schneider et al. [2011] and Kassner et al. [2014]. The afforementioned software Pupil Apps[2] would be responsible to detect eye fovea, create the eye tridimensional model and project the gaze into 2D image.

To build the eye-tracker, two types of cameras are needed: a world and an eye camera. A World camera replicates the vision of the wearer of the camera, capturing the environment images with same viewpoint of the wearer - with the highest possible field-of-view. The eye camera is used to record eye movements, providing data for the eye's 3D model and pupil's detection algorithm. To enhance the results obtained by the pupil's detection algorithm, reducing ilumination influence, sensible IR cameras are generally used. It is possible to track one or both eyes, in a monocular or binocular configuration, respectively. In this Appendix, we will build a monocular eye-tracker.

A.1 Hardware requirements

The following items are necessary:

1. Near IR LED

[1] `https://pupil-labs.com/`

[2] `https://github.com/pupil-labs/pupil/releases/latest`

2. Visible light filter
3. Plastic clamps
4. Eyeglasses frame + camera frames
5. World camera
6. Eye camera
7. M2x12mm flat screw and nuts

Near IR LED Digital camera containing CCD and CMOS sensors are sensitive to light with wavelengths between $400 - 1100$nm, as showed on Figure A.1. Many times, IR light on environment is insuficient to highlight the pupil, being necessary to use external ilummination such as LEDs. However, as any radiation source, IR light could be hazardous to the user of the eye-tracker device. Some research already defined safe limiars to near-ir light exposition into the eye, such as the works of Mulvey et al. [2008]; Kourkoumelis and Tzaphlidou [2011]; ICNIRP [2006] and Claus and Alfons [2018]. An irradiance level less than $10mW/cm^2$ is considered safe for chronic IR exposure (more than 10000 seconds) in the $720 - 1400$nm wavelength range.

If not provided, it should be necessary to caracterize the irradiance and wavelength of the used IR LED - using a radiation meter and a spectrometer. We used commom 3mm Near IR leds, with a resistor of 200Ω. LED outputs at $\lambda = 940$nm with a radiance flux of 0.291mW at 4cm, as presented on Figure A.2. As our sensor diameter is $\varnothing = 9.5$mm, the sensor area is 0.94cm^2, leading to a calculated irradiance of $0.31mW/\text{cm}^2$ - far bellow the safe limits.

Visible light filter As we intent to just capture Near IR wavelenght light, an visible light filter should be placed between lens and the image sensor. We used a home made filter, with used photography camera film (Figure A.3). Two or three layers could be sufficient to block visible light.

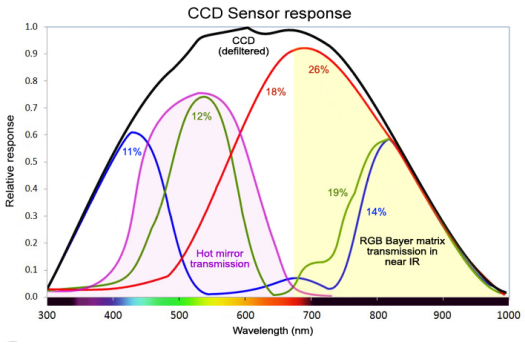


Figure A.1: CCD Spectral response. Image obtained from `http://www.astrosurf.com/luxorion/photo-ir-uv3.htm`.

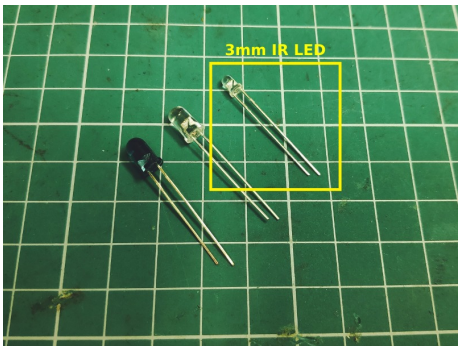


Figure A.2: (a) 3mm IR LED. (b) Radiation meter measuring IR LED power at 4cm distance.

Plastic clamps Plastic clamps were used to hold camera and light cables. Ilustration on Figure A.4.

Eyeglasses frame + camera frames We use an eyeglass frame as suport to cameras and LED. Original camera cases were removed, and printed ones which better adapt to the eyeglass are used. We provide these models, and it should printed using ABS plastic with the following 3D printer configurations:

- Layer Height: 0.1 mm
- Infill Density: 50%
- Infill Pattern Type: Grid
- Bottom Solid Fill Layers: 4
- Top Solid Fill Layers: 8
- Gap filling: Fill Gaps in Shells

The eyeglass frame[3] model is publicity available on Thingiverse comunity site. It is under CC 3.0 License, and no modifications were performed on the model. After printed,

Figure A.3: Photographic Film.

[3] `https://www.thingiverse.com/thing:2386143/files`

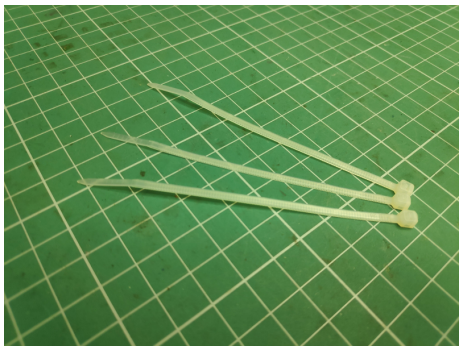


Figure A.4: Plastic clamps.

it should appear as the Figure A.5. To fix the world and eye camera supports into the eyeglass frame, two holes must be drilled on the right temple of the eyeglass.

Logitech C270 camera case was built on top of *Logitech C270 Webcam Cover Replacement (with GoPro Mount)* model[4], with some modifications. This model is also under CC 3.0 License and it is modified version can be found in our github repository[5], as well as Microsoft Vx-7000 camera case and supports.

World camera Any camera could be used as world camera, however it must be UVC compliant (a webcam specification) to be detected on Pupil software. It's prefered to use cameras with a wide field-of-view (FOV) and higher resolution. We used a Logitech C270, an HD resolution (1280×960) camera with 30 fps and 60° diagonal FOV.

Eye camera Choose an eye camera UVC compliant and with at least 30 fps. The recomended was a camera with $120 \sim 200$ fps without IR Filter - the camera could have low resolution (such as 320×240). If the desired camera contains a IR filter, be sure that it can be removed. Dark Pupil detection employed by the software needs IR ilumination together with enhanced IR capture - thus needing this removal. We use a Microsoft VX-7000, configured in 320×240 resolution and 30 fps.

Figure A.5: Eyeglass frame.

[4] `https://www.thingiverse.com/thing:3748407`
[5] `https://github.com/verlab/EyeTracking_Glasses`

M2x12mm flat screw and nuts M2 screws were used to attach frame parts. We need at least 10 screws and nuts.

A.2 Cameras dissasembly

A.2.1 World camera dissasembly

We describe the dissasembly of Logitech C270 camera to put into a printed camera case, that includes a support for the world camera, being attached to the eyeglass frame. To realize it, just perform the steps bellow, as visualized on Figure A.6.

1. Remove the outer plastic from frontal case.
2. Without the cap, remove three screws on top of the frontal case.
3. Remove the two screws that holds circuit into the back case, as well as the nut that holds the circuit wire.
4. Use M2x10mm (2x) screws to attach printed camera case to the world support.
5. Put the camera circuit into printed camera case and screw it using original circuit screws.
6. Put the front part on the camera. The result is a world camera mounted into printed case with a support.

A.2.2 Eye camera dissasembly

Dark Pupil algorithms performs better when using IR images. When using consumer cameras, proper ilumination as well as adequation of necessary camera filters should be done. RBG cameras generally contains IR-blocking filters, that must be replaced by IR-pass ones. The Figure A.7 depicts the Microsoft VX-7000 dissasembly steps to turn it a suitable camera for eye-tracking.

1. Remove screw from the camera back plastic.
2. Put the camera front using a tool.
3. Remove the screw holding the usb wire.
4. Remove the two usb connections, plug and soldering.
5. Remove screws present on metal plate containing the CMOS sensor and the circuit. Detach the circuit using a tool.

Figure A.6: C270 dissasembly steps. (a) Frontal outer plastic removal. (b) Frontal case screws removal. (c) Circuit screws removal. (d) Nut holding circuit wire. (e) Original camera parts. (f) Camera circuit and printed case (g) World camera support and camera case (h) Final parts. (i) Camera circuit assembled on printed camera case. (j-k) Final result.

6. The lens holder is on the top of the sensor area. It contains lens and a Near IR filter. A small quantidity of glue marks the factory focus adjustment. The glue must be removed to properly remove lens holder.

7. Remove the IR filter (the red glass on bottom of lens holder), with extreme care. The set of lens is near of IR filter; damaging it would turn camera invalid.

8. Cut photographic film in round circles to replace IR filter, blocking the visible light and being an IR-pass filter.

9. Mount the camera in printed eye-camera frame.

10. Test the camera output.

A.3 Eye tracker results

Following the described steps, the eye-tracking glass should appear as the one on Figure A.8. Before executing the calibration steps on Pupil software, it should be possible to see eye gaze as well the pupil detection model.

Figure A.7: VX-7000 dissasembly steps. (a) Camera back screw removal. (b-c) Camera front removal using a tool. (d) USB wire removal. (e) Disconnection of USB cable. (f) Sensor area screws and detaching from front camera case. (g) Camera circuit and front camera part. (h) Camera circuit. (i) Lens holder on top of CMOS sensor. (j) Lens holder and camera circuit. (k) Top view from CMOS Sensor, IR Filter and Lens set. (l) Camera parts and cutted photographic film. (m) Image obtained from assembled camera, with IR-pass filter.

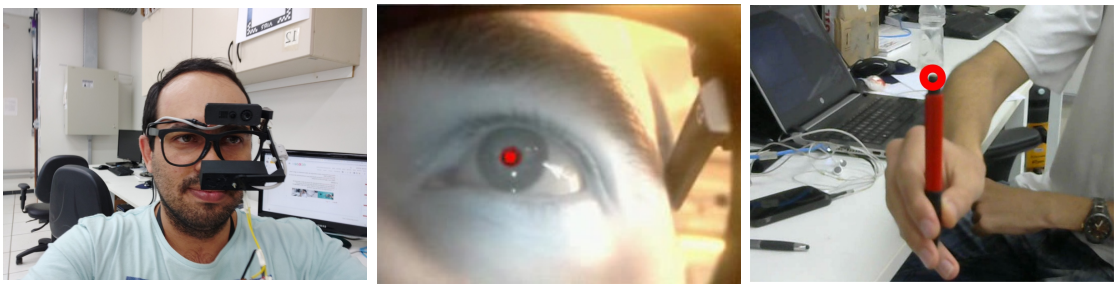


Figure A.8: The wearable eye tracker. (a) A person using the eye tracker. (b) Pupil detection. (c) Gaze into viewed image.