



UNIVERSITAT DE
BARCELONA

Lifestyle Understanding through the Analysis of Egocentric Photo-streams

Estefanía Talavera Martínez

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

*Lifestyle Understanding through the
Analysis of Egocentric Photo-streams*

Estefanía Talavera Martínez

This research has been conducted at the Intelligent Systems group of Johann Bernoulli Institute for Mathematics and Computer Science (Onderzoeksintituut JBI) of the University of Groningen and at the Department of Mathematics and Computer Science of the University of Barcelona.

This work was partially founded by projects TIN2015-66951-C2, RTI2018-095232-B-C2, SGR 1742, CERCA, Nestore Horizon2020 SC1-PM-15-2017 (num. 769643), Validithi EIT Health Program, and ICREA Academia 2014. The founders had no role in the study design, data collection, analysis, and preparation of the manuscript. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of several Titan Xp GPU used for this research.

Lifestyle understanding through the analysis of egocentric photo-streams

Estefanía Talavera Martínez

ISBN: 978-94-034-2313-5 (printed version)

ISBN: 978-94-034-2312-8 (electronic version)



rijksuniversiteit
 groningen



UNIVERSITAT DE
 BARCELONA

Lifestyle Understanding through the Analysis of Egocentric Photo-streams

PhD thesis

to obtain the degree of PhD of the
 University of Groningen
 on the authority of the
 Rector Magnificus Prof. C. Wijmenga
 and in accordance with
 the decision by the College of Deans

and

to obtain the degree of PhD of the
 Universitat de Barcelona
 on the authority of the
 Rector Dr. Joan Elias i Garcia,
 and in accordance with
 the decision by the College of Deans

Double PhD degree

This thesis will be defended in public on

Friday 14 February 2020 at 11.00 hours

by

Estefanía Talavera Martínez

born on 21 September 1990
 in Úbeda, Spain

Supervisors

Prof. N. Petkov

Prof. P. Radeva

Assessment Committee

Prof. M. Biehl

Prof. C. N. Schizas

Prof. J. Vitrià

Prof. G. M. Farinella

To my beloved parents and sister / A mis queridos padres y hermana

Contents

| | |
|--|-----------|
| List of Figures | iv |
| List of Tables | v |
| 1 Introduction | 1 |
| 1.1 Scope | 1 |
| 1.1.1 Societal impact | 3 |
| 1.1.2 Privacy issues | 4 |
| 1.2 Background | 4 |
| 1.2.1 Temporal Segmentation | 6 |
| 1.2.2 Routine Discovery | 7 |
| 1.2.3 Food Related scene classification | 9 |
| 1.2.4 Inferring associated sentiment to images | 10 |
| 1.2.5 Social pattern analysis | 11 |
| 1.3 Objectives | 12 |
| 1.4 Research Contributions | 12 |
| 1.5 Thesis Organization | 15 |
| 2 Egocentric Photo-streams temporal segmentation | 17 |
| 2.1 Introduction | 18 |
| 2.2 Related works | 19 |
| 2.3 Approach | 20 |
| 2.3.1 Features | 21 |
| 2.3.2 Temporal Segmentation | 26 |
| 2.4 Experiments and Validation | 29 |
| 2.4.1 Data | 30 |
| 2.4.2 Experimental setup | 32 |
| 2.4.3 Experimental results | 37 |
| 2.4.4 Discussion | 40 |
| 2.5 Conclusions and future work | 41 |

| | | |
|----------|---|------------|
| 3 | Routine Discovery from Egocentric Images | 43 |
| 3.1 | Introduction | 44 |
| 3.2 | Related works | 47 |
| 3.2.1 | Routine from manual annotation | 47 |
| 3.2.2 | Routine from sensors | 47 |
| 3.2.3 | Routine from conventional images | 48 |
| 3.2.4 | Routine from egocentric images | 49 |
| 3.3 | Unsupervised routine discovery following an outlier detection approach | 50 |
| 3.3.1 | Experiments | 52 |
| 3.4 | Unsupervised routine discovery relying on topic models | 58 |
| 3.4.1 | Experimental Framework and Results | 62 |
| 3.5 | Discussions | 73 |
| 3.6 | Conclusions | 74 |
| 4 | Hierarchical approach to classify food scenes in egocentric photo-streams | 75 |
| 4.1 | Introduction | 76 |
| 4.1.1 | Our aim | 76 |
| 4.1.2 | Personalized Food-Related Environment Recognition | 78 |
| 4.2 | Related works | 79 |
| 4.2.1 | Scene classification | 79 |
| 4.2.2 | Classification of egocentric scenes | 80 |
| 4.2.3 | Food-related scene recognition in egocentric photo-streams | 81 |
| 4.3 | Hierarchical approach for food-related scenes recognition in egocentric photo-streams | 82 |
| 4.4 | Experiments and Results | 85 |
| 4.4.1 | Dataset | 85 |
| 4.4.2 | Experimental setup | 89 |
| 4.4.3 | Dataset Split | 90 |
| 4.4.4 | Evaluation | 91 |
| 4.4.5 | Results | 92 |
| 4.5 | Discussions | 95 |
| 4.6 | Conclusions | 98 |
| 5 | Recognition of Induced Sentiment when Reviewing Personal Egocentric Photos | 101 |
| 5.1 | Introduction | 102 |
| 5.2 | Related works | 102 |
| 5.3 | Sentiment detection by global features analysis | 105 |

CONTENTS

| | | |
|----------|--|------------|
| 5.3.1 | Experimental Setup | 107 |
| 5.4 | Sentiment detection by semantic concepts analysis | 109 |
| 5.4.1 | Sentiment Model | 111 |
| 5.4.2 | Experimental Setup | 112 |
| 5.5 | Discussion and conclusions | 114 |
| 6 | Towards Egocentric Person Re-identification and Social Pattern Analysis | 117 |
| 6.1 | Introduction | 118 |
| 6.2 | Related works | 119 |
| 6.3 | Social Patterns Characterization | 120 |
| 6.3.1 | Person Re-Identification | 120 |
| 6.3.2 | Social Profiles Comparison | 122 |
| 6.4 | Experiments | 122 |
| 6.4.1 | Dataset | 122 |
| 6.4.2 | Experimental setup | 123 |
| 6.4.3 | Results | 124 |
| 6.5 | Conclusions | 124 |
| 7 | Summary and Outlook | 127 |
| 7.1 | Work Summary | 127 |
| 7.2 | Outlook | 129 |
| | Bibliography | 133 |
| | Summary | 147 |
| | Samenvatting | 149 |
| | Resumen | 151 |
| | Acknowledgements | 153 |
| | Research Activities | 155 |
| | About the Author | 159 |

List of Figures

| | | |
|------|---|----|
| 1.1 | Illustration of collected photo-streams | 2 |
| 1.2 | Wearable camera - Narrative Clip. | 5 |
| 1.3 | Examples of wearable cameras | 6 |
| 1.4 | Illustration of the temporal segmentation of a collected photo-stream | 7 |
| 1.5 | Illustration of behaviours that describe the routine of a person | 8 |
| 1.6 | Illustration of food-related daily habits | 9 |
| 1.7 | Illustration of a camera user reviewing his or her collected events, being affected by their associated sentiment. | 10 |
| 1.8 | Pipeline for the analysis of social patterns | 11 |
| 2.1 | Example of temporal segmentation of an egocentric sequence | 18 |
| 2.2 | General scheme of the SR-Clustering method | 21 |
| 2.3 | Graph obtained after calculating similarities of the concepts of a day's lifelog and clustering them | 23 |
| 2.4 | Example of the final semantic feature matrix obtained for an egocentric sequence | 24 |
| 2.5 | Example of extracted tags on different segments | 25 |
| 2.6 | General scheme of the semantic feature extraction methodology. | 26 |
| 2.7 | Change detection by the different algorithms implemented | 28 |
| 2.8 | Different segmentation results obtained by different subjects | 33 |
| 2.9 | LCE and GCE of the manual segmentations | 34 |
| 2.10 | Correlation of the LCE and GCE among sets | 35 |
| 2.11 | LCE and GCE of the manual segmentations - excluding the camera wearer segmentation | 36 |
| 2.12 | Correlation of the LCE and GCE among sets - excluding the camera wearer segmentation | 37 |
| 2.13 | Examples of different segments and the top 8 found concepts | 38 |
| 3.1 | Example of images recorded by one of the camera wearers. | 44 |
| 3.2 | Pipeline of the proposed model. | 50 |
| 3.3 | Average number of images per recorded egocentric photo-stream. We give the number of collected days per user between parenthesis. | 53 |

| | | |
|------|--|-----|
| 3.4 | Histograms showing the occurrence of activities throughout the days | 56 |
| 3.5 | Visualization of the obtained classification results | 57 |
| 3.6 | Illustration of the proposed Topics-based model | 58 |
| 3.7 | Illustration of a photo-stream/document described by proportion of topics | 60 |
| 3.8 | Average number and variance of egocentric images per recorded photo-stream for the 7 users | 63 |
| 3.9 | Example of selected images throughout some of the recorded photo-streams of User1. | 63 |
| 3.10 | Number of <i>Routine</i> and <i>Non-Routine</i> days for each user (U) in the <i>EgoRoutine</i> dataset. | 64 |
| 3.11 | Example of given photo-streams, sample images at several time-slots, their representative topics, and the concepts that compose them. | 71 |
| 3.12 | Affinity matrix (DTW) and the later discrimination as Routine or Non-Routine related days (SpClust) of collected days by users 3 and 7 | 72 |
| 4.1 | Examples of images of each of the proposed food-related categories present in the introduced EgoFoodPlaces dataset. | 77 |
| 4.2 | The proposed semantic tree for food-related scenes categorization. | 84 |
| 4.3 | Total number of images per food-related scene class. | 86 |
| 4.4 | Illustration of the variability of the size of the events for the different food-related scene classes. | 87 |
| 4.5 | Visualization of the distribution of the classes using the t-SNE algorithm. | 88 |
| 4.6 | Mean Silhouette Score for the samples within the studied food-related classes | 88 |
| 4.7 | Confusion matrix with the classification performance of the proposed hierarchical classification model. | 94 |
| 4.8 | Examples of top 5 classes for the images in the test set | 95 |
| 4.9 | Illustration of detected food-related events in egocentric photo-streams | 97 |
| 5.1 | Examples of Positive, Negative and Neutral images. | 106 |
| 5.2 | Architecture of the proposed method combining global and semantic features | 107 |
| 5.3 | Examples of the automatic event sentiment classification | 109 |
| 5.4 | Sketch of the proposed method for semantic concepts analysis | 110 |
| 6.1 | Architecture of the proposed model | 118 |
| 6.2 | Samples of the clusters obtained from recorded days | 121 |
| 6.3 | Obtained social profiles as a result of applying our method | 125 |
| 7.1 | Future directions of research | 130 |

List of Tables

| | | |
|-----|---|-----|
| 1.1 | Comparison of some popular wearable cameras. | 6 |
| 2.1 | Table summarizing the main characteristics of the datasets used in this work: | 30 |
| 2.2 | Average FM results of the state-of-the-art works on the egocentric datasets | 39 |
| 2.3 | Average FM score on each of the tested methods using our proposal of semantic features on the dataset presented in (Poleg et al., 2014). | 40 |
| 3.1 | Description of the collected Egoroutine dataset by 5 users. | 52 |
| 3.2 | Summary of the labelling results for the Egoroutine dataset. | 53 |
| 3.3 | Performance of the different methods implemented for the discovery of routine and non-routine days. | 55 |
| 3.4 | Total number of recorded days and collected images per user. | 62 |
| 3.5 | Summary of the agreement among the 6 individuals that labelled the collected photo-streams into Routine or Non-Routine related days. | 64 |
| 3.6 | Results of the proposed pipeline and baseline models | 67 |
| 3.7 | Results of the proposed pipeline for the best setting of the parameters | 68 |
| 3.8 | Example of detected concepts in a given recorded day by User 1 | 68 |
| 3.9 | Comparison between our previous work and the model here proposed | 72 |
| 4.1 | Food-related scene classification performance. | 93 |
| 4.2 | Classification performance at different levels of the proposed semantic tree for food-related scenes categorization. | 93 |
| 5.1 | Different image sentiment ontologies. | 103 |
| 5.2 | Description of the UBRUG-EgoSenti dataset. | 108 |
| 5.3 | Performance results achieved at image and event level. | 108 |
| 5.4 | Examples of clustered concepts based on their semantic similarity, initially grouped following the distance computed by the WordNet tool. | 111 |
| 5.5 | Parameter-selection results | 113 |
| 5.6 | Test set results | 114 |

| | | |
|-----|--|-----|
| 6.1 | Average Precision, Recall, and F-Measure result for each of the tested methods on the extended test-set composed by egocentric images. . . | 123 |
| 6.2 | This table shows the social behavioural traits obtained from the detected social interactions for the different camera wearer. | 124 |

Chapter 1

Introduction

How can we improve and contribute to the people's quality of life? The personal development process is described as the assessment of people's qualities and behaviours. By tracking people's daily behaviours we can help them draw a picture of their lifestyle. The obtained information can be used to later improve their personal development (Ryff, 1995). However, the self-awareness and personal development process are not trivial. They include the enhancement of the following activities: self-knowledge, health, strengths, aspirations, social relations, enhancing lifestyle, quality of life and time-management, among others (Ryff, 1995). For instance, the quantification of their daily activities helps to define goals for future changes and/or advances in their personal needs and ambitions.

This thesis addresses the development of automatic computer vision tools for the study of people's behaviours. To this end, we rely on the analysis of egocentric photo-streams recorded by a wearable camera. These pictures show an egocentric view of the camera wearers' experiences, allowing an objective description of their days (Bolaños et al., 2017). They describe the users' daily activities, including people they meet, time spent working on their computers, outdoor activities, sports, eating, or shopping. The first-view perspective shown by the images describes how the lives of the camera wearers look like. We believe that this data is a powerful source of information since it is a raw description of the behaviours of people in society. Our goal is to demonstrate that egocentric images can help us draw a picture of the days of the camera wearer, that can be used to improve the healthy living of individuals.

1.1 Scope

This thesis aims to develop and introduce automatic computer vision tools that allow the study and characterization of the lifestyle of people. To do so, we rely on egocentric images recorded by wearable cameras, see Fig. 1.2. An egocentric photo-stream or egocentric photo-sequence is defined as a collection of temporal consecutive images. Fig 1.1 illustrates a collection of photo-streams recorded by a camera

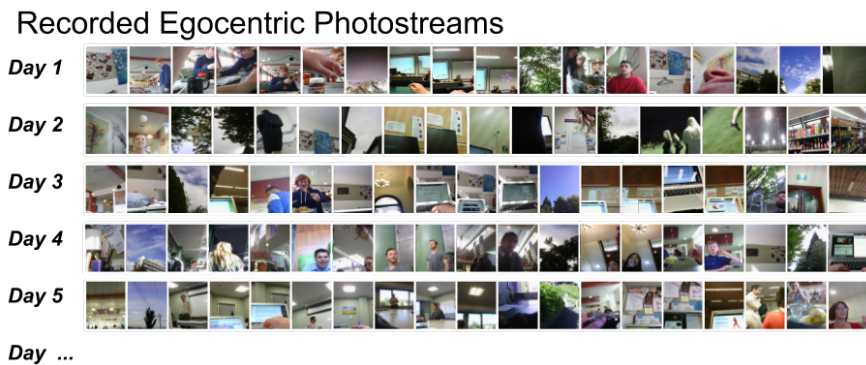


Figure 1.1: Illustration of recorded days in the form of egocentric photo-streams. These images were collected by the Narrative Clip wearable camera and describe the life of the camera wearer.

wearer. The information that we can obtain from the recorded photo-streams is broad because of the wide range of applications that can be addressed. More specifically, in this work, we focus on the analysis of the following behavioural traits:

- **Temporal segmentation:** Days are composed of moments when the camera wearer spends time at certain environments. To find such moments, we look for sequences of similar images. Given an egocentric photo-sequence, our model decides the temporal boundaries that divided the photo-stream into moments based on the global and semantic features of the images.
- **Routine discovery:** Implement an automatic tool for the discovery of Routine-related days among days recorded by different users. To this end, we evaluate the role of semantics extracted from the egocentric photo-streams.
- **Recognition of food-related scenes:** Identify food-related environments where the user spends time to describe food-related activity routines.
- **Sentiment retrieval:** Given images describing recorded scenes by the user the aim is to determine their sentiment associated based on the extraction of either visual features or semantic concepts with sentiment associated, or their combination.
- **Social pattern characterization:** Provide an automated description of patterns of the experienced social interactions, according to the detection of people and the occurrence of their appearance throughout the recorded photo-streams.

Egocentric images describe from an egocentric point of view the wearer's life. The extracted information allows us to get insight into the lifestyle of the camera users, for the later improvement of their health. Moreover, wearable cameras are lightweight, financially affordable and with potential for other applications to assist or improve the quality of life of people.

1.1.1 Societal impact

Nowadays, describing people's lives has become a hot topic in several disciplines. In psychology, this topic is addressed aiming to help ordinary people, and especially people with some kind of need (Martin et al., 1986; de Haan et al., 1997; Yesavage, 1983), where an automatic evaluation of lifestyle would be of much help for the practitioners.

Healthy ageing is of relevance due to the ever-increasing number of elderly people in the population. These collections of digital data can serve as cues to trigger autobiographical memory about past events and can be used as an important tool for prevention or hindrance of cognitive and functional decline in elderly people (Doherty et al., 2013), and memory enhancement (Lee and Dey, 2008). In (Sellen et al., 2007), it was discussed that if memory cues are provided to people suffering from Mild Cognitive Impairment (MCI), they would be helped to mentally 're-live' specific past life experiences. Studies have shown how different cues, such as: time, place, people, and events, trigger autobiographical memories, suggesting that place, events, and people are the stronger ones. A collaboration with neuropsychologist from the Hospital of Terrassa, Spain, shows the good acceptability of older adults of wearable devices, where the potential benefits for memory outweigh concerns related to privacy (Gelonch et al., 2019). Our novel proposed system will contribute to healthy ageing by improving peace of mind of elderly people. The developed models that we applied in such situations have shown promising outcomes.

In the last few years, there has been an exponential increase in the use of self-monitoring devices (Trickler, 2013) by ordinary people who want to get to know themselves better. These devices offer information about daily habits, by logging daily data of the user, such as: how many steps the user walks (Cadmus-Bertram et al., 2015), how and how long smartphones and apps are used (Wei et al., 2011), heart-rate with the use of smart bracelets or watches (Reeder and David, 2016), to name some. People want to increase their self-knowledge automatically, expecting that it will lead to psychological well-being and the improvement of their lifestyle (Ryff, 1995). Self-knowledge is a psychology term that describes a person's answers to the question "How am I like?" (Neisser, 1988). To answer this, there is often need of external information mainly because of two causes. On one side, it is a difficult

task to describe our behavioural patterns. On the other side, we tend to alter and not be accurate when describing what it is like (Silvia and Gendolla, 2001).

From another point of view, big companies started looking for information about their employees and clients with the aim of improving productivity and customer acquisition (Chin et al., 2011; Sanlier and Seren Karakus, 2010; Spiliopoulou et al., 1999). Furthermore, behavioural psychologists from the University of Otago, New Zealand, already shown their interest in this tool since they are working on the characterization of the lifestyle of students. The identification of which whom students tend to interact and the duration of such interactions is of high importance when aiming to understand their daily habits, and ultimately improve them.

1.1.2 Privacy issues

Personal data relates to any information that can be obtained from the living of an individual. The use of wearable devices to track our lifestyle can be seen as intrusive, but can help to promote life-enhancing. Following the General Data Protection Regulation (EU) 2016/679 (GDPR), we consider data protection and ensure personal data privacy from different perspectives:

- *Researchers*: People working on the analysis of the collected data were asked to sign a consent form confirming that they will use the data for research purposes, respecting the privacy of the participants.
- *Participants*: Camera wearers were asked to give their written consent for the later use of their collected data. The collected data is then linked to an identifier that ensures the anonymization of the camera user. In the case of models where detected faces are needed for the analysis, we do not blur the identity of the persons with whom the participant interacts, but we do ask for their consent to be part of the dataset. The participants have the right to revoke their consent at any time.

1.2 Background

In this section, we describe the main concepts that we refer to throughout this thesis, such as lifelogging, wearable cameras, egocentric vision, and egocentric photo-streams. Moreover, we briefly introduce the framework of the different applications that we later describe and address in the following chapters of this thesis.

Before the emergence of static and wearable sensors, people's daily habits were manually recorded. For instance, *Activities of Daily Living (ADL)* were manually

annotated by either individual users and/or specialists, as in (Andersen et al., 2004; Wood et al., 2002). In (Andersen et al., 2004), manually recorded information about the ability of someone's ADL performance was examined to classify the patients' dependence, as either dependent or independent.

Lifelogging Nowadays, the development of new wearable technologies allows to automatically record data from our daily living. *Lifelogging* appeared in the 1960s as the process of recording and tracking personal activity data generated by the daily behaviour of a person. Through the analysis of recorded visual data, information about the lifestyle of the camera wearer can be obtained and retrieved. By recording people's own view of the world, lifelogging opens new questions and goes a step forward to the desired and personalized analysis of the lifestyle of individuals. The objective perspective offered by the recorded data of what happened during different moments of the day, represents a robust tool for the analysis of the lifestyle of people.



Figure 1.2: Wearable camera - Narrative Clip.

Cameras Among the advances in wearable technology during the last few years, wearable cameras specifically have gained more popularity (Bolaños et al., 2017). In Fig. 1.3 we present some examples of wearable cameras that are available on the market. These cameras are used for different purposes and have different specifications (see Table 1.1). All the mentioned devices allow capturing high-quality images in a hands-free fashion from the first-person point of view.

Wearable video cameras, such as GoPro and Looxcie, which have a relatively high frame rate, ranging from 25 to 60 fps, are mostly used for recording the user activities for a few hours. Instead, wearable photo cameras, such as the Narrative

Clip and SenseCam, capture only 2 or 3 fpm and are therefore mostly used for image acquisition during longer periods of time (e.g. a whole day). By using wearable cameras with a low temporal resolution, the camera wearer captures each day up to 1000 egocentric photo-streams.

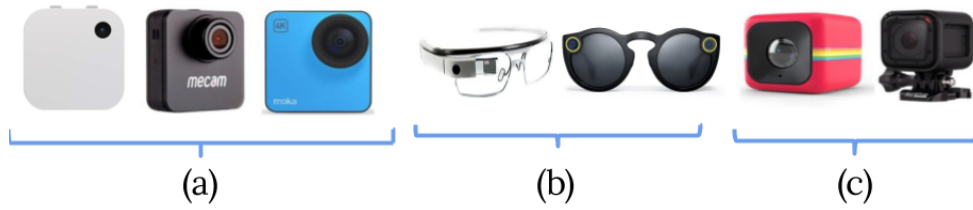


Figure 1.3: Some of the available wearable cameras that can be found on the market. While the a) torso mounted cameras are commonly used for visual diary creation and security, the (b) glass mounted wearable cameras are often used for augmented reality [Google Glasses and Spectacles]. Finally, the (c) head mounted cameras are used for recording sports and leisure activities [GoPro and Polaroid Cube].

Table 1.1: Comparison of some popular wearable cameras.

| Camera | Main use | Temporal Resolution (FPS/FPM) | Worn on | Size (mm) | Weight (gr.) |
|------------------|-------------------|-------------------------------|----------------|------------------|--------------|
| GoPro Hero5 | Entertainment | High (60fps) | Head and Torso | 38x38 | 73 |
| Google Glasses | Augmented Reality | High (60fps) | Head | up to 133.35x203 | 36 |
| Spectacles | Social Networks | High (60 fps) | Head | 53x145 | 48 |
| Axon Body 2 | Security | High (30 fps) | Torso | 70x87 | 141 |
| Narrative Clip 2 | Lifelogging | Low (2-3fpm) | Torso | 36x36 | 19 |
| SenseCam | Lifelogging | Low (2 fpm) | Torso | 74x50 | 90 |
| Autographer | Lifelogging | Low (2-3fpm) | Torso | 90x36 | 58 |

Egocentric Photo-streams The recorded photo-streams offer a first-person view of the world (see Fig. 1.1). The big advantage of image-based lifelogging is that it gives rich information able to generate explanations and visualize the circumstances of the person's activities, scenes, state, environment and social context that influence his/her way of life, as it defines the contextual information. Through the analysis of images collected by continuously recording the user's life, information about daily routines, eating habits, or positive memories can be obtained and retrieved.

1.2.1 Temporal Segmentation

Egocentric photo-streams generally appear in the form of long unstructured sequences of images, often with a high degree of redundancy and abrupt appearance changes even in temporally adjacent frames, that harden the extraction of semantically meaningful content. Temporal segmentation, the process of organizing un-

structured data into homogeneous chapters, provides a large potential for extracting semantic information. Video segmentation aims to temporally divide the video into different groups of consecutive images called *events* or *scenes* that describe the performance of an activity or a specific environment where the user is spending time (see Figure 1.4). Many segmentation techniques have been proposed in the literature in an attempt to deal with this problem, such as video summarization based on clustering methods or object detection. The work described in (Goldman et al., 2006) was a first approach where the user selected the frames considered important as key-frames (considered as the frame that best represents the scene), generating the storyboard that reported object’s trajectory. Other studies incorporate audio or linguistic information (Nam and Tewfik, 1999; Smith, 1997) to the segmentation approach looking for the semantic meaning of the video.



Figure 1.4: Example of temporal segmentation of an egocentric sequence based on what the camera wearer sees. In addition to the segmentation, our method provides a set of semantic attributes that characterize each segment.

We believe that the division of the photo-stream into a set of homogeneous and manageable segments is important for the better characterization of the collection of images. Each segment can be represented by a small number of key-frames and indexed by semantic features. This division provides a basis for understanding the semantic structure of the event. Hence, in this work, we aim to study and discuss the following related research lines: *Can we obtain a good enough division of the recorded photo-streams into events? Which are the features that help us achieve the best temporal segmentation? Is the manually temporal segmentation process robust on its own?* These questions will be developed in Chapter 2 of this thesis.

1.2.2 Routine Discovery

Human behaviour analysis is of high interest in our society and a recent research area in computer vision. Routine-related days have common patterns that describe situations of the daily life of a person. More specifically, routine was described as

regularity in the activity in (Sevtsuk and Ratti, 2010). Fig 1.5 is an illustration of what can be considered as the routine of a person. Social psychologists exposed in (Society for Personality and Social Psychology, 2014) that each day 40% of people’s daily activities are performed in similar situations. However, Routine has no concrete definition, since it varies depending on the lifestyle of the individual under study. Therefore, supervised approaches are not useful due to the need for prior information in the form of annotated data or predefined categories. For the discovery of routine-related days, unsupervised methods are necessary to enable an analysis of the dataset with minimal prior knowledge. Moreover, we need to apply automatic methods that can extract and group the days of an individual using correlated daily elements. We address the discovery of routine-related days following two different approaches:

- On one side, we evaluate outlier detection methods for the discovery of clusters corresponding to routine-related days, i.e. outliers to non-routine related days. In this approach, days are described as the aggregation of the images’ global features.
- On the other side, we propose a novel automatic unsupervised pipeline for the identification and characterization of routine-related days from egocentric photo-streams. We perform an ablation study at different levels of the proposed architecture for the characterization and comparison of days.

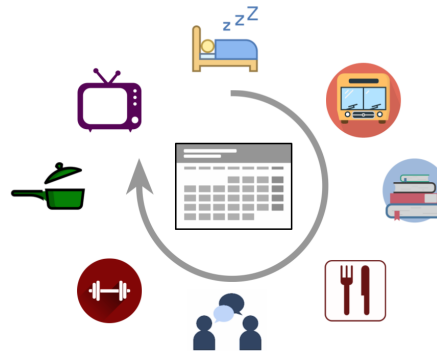


Figure 1.5: The routine of the camera wearer is described by his or her performed activities throughout the days. We aim to discover the daily habits of people to get a better understanding of their behaviour.

Together with the proposed models, we introduce EgoRoutine, a new egocentric dataset composed of a total of 100,000 images, from 104 days. Further description of the proposed methodology and experiments can be found in Chapter 3.

1.2.3 Food Related scene classification

From another perspective, nutritional habits are of importance for the understanding of the lifestyle of a person. Recent studies in nutrition argue that it is not only important *what people eat* but also *how/where people eat* (Laska et al., 2015). We propose the analysis of collected egocentric photo-streams for the automatic characterization and monitoring of the health habits of the camera wearer. To this end, we focus on the classification of 15 different food-related scenes. Scenes recorded by an egocentric perspective and related to food consumption, acquisition or preparation share visual information, which makes difficult to distinguish them. Therefore, we propose a hierarchical classification model that organizes the classes based on their semantic relation. We illustrate the three main food-related activities and some of the scenes in Fig. 1.6. The intermediate probabilities help to improve the final classification by re-enforcing the predictions of the classifiers. There are no previous works on this field, and therefore, the proposed model represents the baseline for food-related scenes classification.

Moreover, we propose and make publicly available EgoFoodPlaces, a dataset composed of more than 33,000 images representing food-related scenes. We describe EgoFoodPlaces, the proposed model and the performed experiments in Chapter 4 of this thesis.

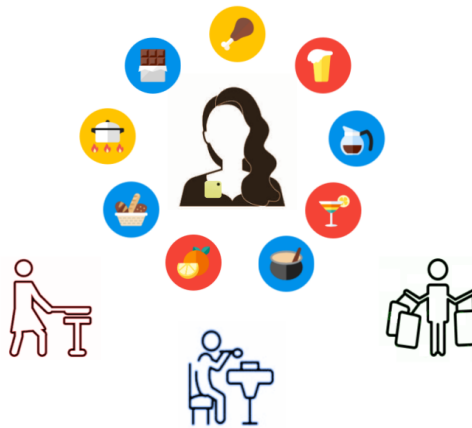


Figure 1.6: Daily health habits related to food consumption, acquisition or preparation can be studied by the examination of recorded egocentric photo-streams. The analysis of food-related scenes and activities can help us understand the lifestyle of the camera wearer for the improvement of his or her nutritional behaviour.

1.2.4 Inferring associated sentiment to images

Understanding emotions plays an important role in personal growth and development, and gives insight into how human intelligence works. Moreover, selected memories can be used as a tool for mental imagery, which is described as the process in which the feeling of an experience is imagined by a person in the absence of external stimuli. The process of reliving previous experiences is illustrated in Fig. 1.7. Therapists assumed it is directly related to emotions (Holmes and et al., 2006), opening some questions when images describing past moments of our lives are available: *Can an image facilitate the process of mental imagery?* or *Can specific images help us to retrieve or imply feelings and moods?* Semantic concepts extracted from the collection of egocentric images help us describing the emotions related to memories that the photo-streams capture.

Part of the recorded egocentric images are redundant, non-informative or routine and thus without special value for the wearer to be preserved. Usually, users are interested in keeping special moments, images with sentiments that will allow them in the future to re-live the personal moments captured by the camera. An automatic tool for sentiment analysis of egocentric images is of high interest to make possible the processing of the big collection of lifelogging data and keeping out just the images of interest i.e. of high charge of positive sentiments. To the best of our knowledge, no previous works had addressed this topic from egocentric photo-streams in the literature. In Chapter 5, we study how egocentric images can be analyzed to discover events that would invoke positive, neutral or negative feelings to the user.

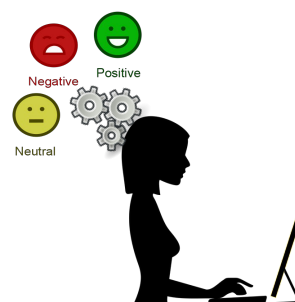


Figure 1.7: Illustration of a camera user reviewing his or her collected events, being affected by their associated sentiment.

1.2.5 Social pattern analysis

Human social behaviour involves how people influence and interact with others, and how they are affected by others. This behaviour varies depending on the person and is influenced by ethics, attitudes, or culture (Allport, 1985). Understanding the behaviour of an individual is of high interest in social psychology. In (House et al., 1988), the authors addressed the problem of how social relationships affect health and demonstrated that social isolation leads to major risk factors for mortality. Moreover, in (Yang et al., 2016), the authors observed that lack of social connections is associated with health risks in specific life stages, such as the risk of inflammation in adolescence, or hypertension in old age. Also, as in (Kawachi and Berkman, 2001) it was highlighted that social ties have a beneficial effect when maintaining psychological well-being.

Considering the importance of the matter, automatic discovery and understanding of the social interactions are of high importance to the scientists, as they remove the need for manual labour. On the other hand, egocentric cameras are useful tools as they offer the opportunity to obtain images of the daily activities of users from their own perspective. Therefore, providing a tool for automatic detection and characterization of social interactions through these recorded visual data can lead to personalized social pattern discoveries, see Fig. 1.8. We discuss the proposed model and findings in Chapter 6.

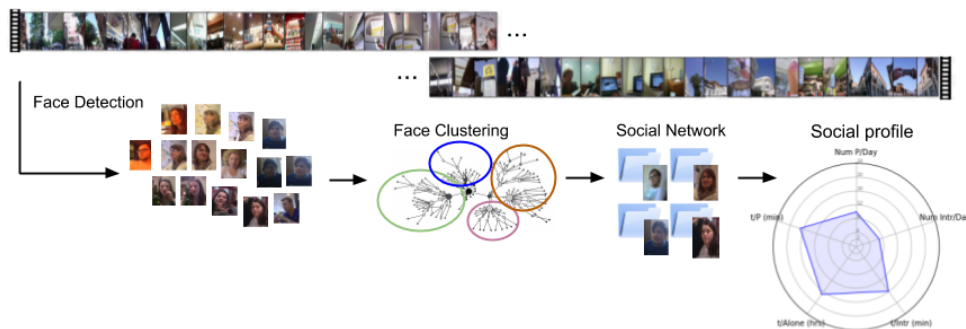


Figure 1.8: Example of social profile given a set of collected photo-streams associated with one person. First, we detect appearing faces in the photo-streams. Later, we apply the OpenFace tool to convert the faces into feature vectors. We propose to define the re-identification problem as a clustering problem with a later analysis of the grouped faces occurrence.

1.3 Objectives

The main goal of this dissertation is to give appropriate tools for the analysis and interpretation of egocentric photo-streams for the understanding of the behavioural patterns of the camera wearer. Given the previous general lines that represent the ground of this thesis, we defined the following particular objectives:

- To temporally segment egocentric photo-streams into moments within the day for their later analysis according to global and semantic features extracted from the images.
- To provide an automatic tool for routine discovery through the recognition of days with similar patterns within the egocentric photo-streams collection.
- To automatically classify egocentric photo-streams into food-related scenes to get an understanding of the user's eating habits.
- To define a simple social pattern analysis framework to compare different user's social behavioural patterns.
- To identify the sentiment that a retrieved moment would provoke the users when reviewing it.

1.4 Research Contributions

This thesis argues that behavioural patterns can be analysed in the domain of egocentric photo-streams since they represent a first-person perspective of the life experiences of the camera user. The analysis of egocentric photo-streams allows us to extract information which gives us insight into the lifestyle of the camera wearer. Our contributions aim to improve a person's lifestyle. The presented models can be easily adapted for personalized behavioural patterns analysis from images recorded from a first-person view.

Specifically, the contributions of this thesis can be summarized as follows:

1. Due to the free movement of the camera and its low frame rate, abrupt changes are visible even among temporally adjacent images (see Fig. 2.1 and Fig. 2.8). Under these conditions motion and low-level features such as colour or image layout are prone to fail for event representation, hence urges the need to incorporate higher-level semantic information. Instead of representing each image by its contextual global features, which capture the basic environment appearance, we detect segments as a set of temporally adjacent images with the same

contextual representation in terms of semantic visual concepts. Nonetheless, not all the semantic concepts in an image are equally discriminant for environment classification: objects like trees and buildings can be more discriminant than objects like dogs or mobile phones, since the former characterizes a specific environment such as forest or street, whereas the latter can be found in many different environments. In this paper, we propose a method called Semantic Regularized Clustering (SR-Clustering), which takes into account semantic concepts in the image together with the global image context for event representation. These are the contributions within this line of research:

- Methodology for the description of egocentric photo-streams based on semantic information.
- Set of evaluation metrics applied to ground truth consistency estimation.
- Evaluation on an extensive number of datasets, including our own, which was published with this work.
- Exhaustive evaluation on a broader number of methods to compare with.

The proposed model for temporal segmentation was published as (Talavera et al., 2015) and (Dimiccoli et al., 2017).

2. We address for the first time the discovery of routine-related days from egocentric photo-streams. With this aim, we propose two different approaches. On one hand, we propose an unsupervised and automatic model for the discovery of routine days following a novelty detection approach. This model is based on the analysis of the aggregation of descriptors of the images within the photo-stream. We tested the proposed model over a home-made collected egocentric dataset. This dataset describes the daily life of the camera wearers. It is composed of a total of 73,000 images, from 72 recorded days by 5 different users. We name this dataset *EgoRoutine*. This work was published in a conference as (Talavera et al., 2019). On the other hand, we introduce a novel automatic unsupervised pipeline for the identification and characterization of Routine-related days from egocentric photo-streams. In our proposed model we first extract semantic features from the egocentric images in terms of detected concepts. Later, we translate them to documents following the temporal distribution of the labels. To do so, the detected words in images that were recorded during pre-defined time-slots define a document. Then, we apply the topic modelling technique to the created documents to find abstract topics related to the person's behaviour and his/her daily habits. We prove that topic modelling is a powerful tool for pattern discovery when addressing Bag-of-Words representation of photo-streams. Later, Dynamic Time Warping

(DTW) and Spectral Clustering are applied for the unsupervised routine discovery. We prove that using DTW and Distance-based clustering is a robust technique to detect the cluster of routine days being tolerant to small temporal differences in the daily events. The proposed pipeline is evaluated over an extension of the previous *EgoRoutine* dataset, which is composed of more than 100,000 images, from 104 days collected by 7 different users. This work was submitted and is currently under review.

3. A novel model for food-related scenes classification is introduced in Chapter 4. Food-related scenes that commonly appear in the collected egocentric photo-streams tend to be semantically related. There exists a high intra-class variance in addition to not a high inter-class similarity, leading to a challenging classification task. To face this classification problem the contributions of the chapter are three-fold. On one side, we define a taxonomy with the relation of the studied classes, where food-related environments are organized in a fine-grained way that take into account the main food-related activities (eating, cooking, buying, etc). On the other side, we propose a hierarchical model composed of different layers of deep neural networks. The model is adapted to the defined taxonomy for food-related scenes classification in egocentric photo-streams. Our hierarchical model can classify at the different levels of the taxonomy. Finally, we introduce a new egocentric dataset of more than 33,000 images describing 15 food-related environments. We call it FoodEgoPlaces and along with its ground-truth are publicly available in <http://www.ub.edu/cvub/dataset/>. This work is published as (Talavera et al., 2014).
4. We present innovative models for emotion classification in egocentric photo-streams setting, see Chapter 5. In this chapter, we present two models: one is based on the analysis of semantic concepts extracted from images that belong to the same event, while the other analyses the combination of semantic concepts and general visual features of such images. In our proposed analysis, we evaluate the role of considered semantic concepts in terms of Adjective-Noun-Pairs (ANPs), given that they have sentiment values associated (Borth et al., 2013), and their combination with general visual features extracted with a CNN (Krizhevsky, Sulskever and Hinton, 2012). With this work, we prove the importance of such a combination in the invoked sentiment detection. Moreover, we test our method on a new egocentric dataset of 12,088 pictures with ternary sentiment values acquired from 3 users in a total of 20 days. Our contribution is an analytic tool for positive emotion retrieval seeking events that best represent a pleasant moment to be invoked within the whole set of a

day photo-stream. We focus on the event’s sentiment description from an objective point of view of the moment under analysis. The results given in this chapter are published in two conferences (Talavera, Radeva and Petkov, 2017; Talavera, Strisciuglio, Petkov and Radeva, 2017).

5. We propose a method that enables us to automatically analyse and answer to questions such as *Do I socialize throughout my days?* or *With how many people do I interact daily?*. To do so, we rely on the analysis of egocentric photo-streams. Given sets of captured days by camera wearers, our proposed model employs a person re-identification model to achieve social pattern descriptions. First, a Haar-like feature-based cascade classifier is applied (Viola et al., 2001) to detect the appearing faces in the photo-streams. Detected faces in this step are converted into feature descriptors by applying the OpenFace tool (Amos et al., 2016). Finally, we propose to define the person re-identification problem as a clustering problem. The clustering is applied over the pile of photo-streams recorded by the users along the days to find the recurrent faces within photo-streams. Shaping an idea about the social behaviour of the users becomes possible through referring to the time and day when the recurrences were appearing. The proposed work was presented in a conference as (Talavera et al., 2018).

1.5 Thesis Organization

The remaining chapters of this thesis are organised as follows: Chapter 2 describes our proposed temporal segmentation method, which divides egocentric sequences into sequential similar images, that we call *events*. In Chapter 3, we present an automatic model for the discovery of routine-related days from the photo-stream collection of a user. Following, in Chapter 4, we introduce a hierarchical network for the classification of images into food-related scenes. Later, in Chapter 5, we address the recognition of what an image would invoke to the camera wearer. In Chapter 6, we focus on the analysis of social interactions of the user to then infer a social pattern that describes his or her social daily behaviour. Finally, Chapter 7 provides a summary of the thesis and gives an outlook of how the proposed techniques can be developed further and applied in different computer vision applications.

Most of this chapter was published as:

M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, G. Stavri, P. Radeva, "SR-Clustering: Semantic Regularized Clustering for Egocentric Photo-Streams Segmentation", International Journal Computer Vision and Image Understanding (CVIU), Pages 55-69, Vol. 155, 2016.

Section 2.2.2 is taken from:

E. Talavera, M. Dimiccoli, M. Bolaños, M. Aghaei, P. Radeva, "R-Clustering for Egocentric Video Segmentation," 7th Iberian Conference on Pattern Recognition and Image Analysis (IBPRIA), pp. 327-336, Pattern Recognition and Image Analysis, Chapter Springer Verlag, 2015.

Chapter 2

Egocentric Photo-streams temporal segmentation

Abstract

While wearable cameras are becoming increasingly popular, locating relevant information in large unstructured collections of egocentric images is still a tedious and time-consuming process. This paper addresses the problem of organizing egocentric photo streams acquired by a wearable camera into semantically meaningful segments, hence making an important step towards the goal of automatically annotating these photos for browsing and retrieval. In the proposed method, first, contextual and semantic information is extracted for each image by employing a Convolutional Neural Networks approach. Later, a vocabulary of concepts is defined in a semantic space by relying on linguistic information. Finally, by exploiting the temporal coherence of concepts in photo streams, images which share contextual and semantic attributes are grouped together. The resulting temporal segmentation is particularly suited for further analysis, ranging from event recognition to semantic indexing and summarization. Experimental results over an egocentric set of nearly 31,000 images, show the prominence of the proposed approach over state-of-the-art methods.

2.1 Introduction

Among the advances in wearable technology during the last few years, wearable cameras specifically have gained more popularity (Bolaños et al., 2017). These small light-weight devices allow capturing high-quality images in a hands-free fashion from the first-person point of view. Wearable video cameras such as GoPro and Looxcie, by having a relatively high frame rate ranging from 25 to 60 fps, are mostly used for recording the user activities for a few hours. Instead, wearable photo cameras, such as the Narrative Clip and SenseCam, capture only 2 or 3 fpm and are therefore mostly used for image acquisition during longer periods of time (e.g. a whole day).

The images collected by continuously recording the user’s life can be used for understanding the user’s lifestyle and hence they are potentially beneficial for prevention of non-communicative diseases associated with unhealthy trends and risky profiles (such as obesity, depression, etc.). Besides, these images can be used as an important tool for prevention or hindrance of cognitive and functional decline in elderly people (Doherty et al., 2013). However, egocentric photo streams generally appear in the form of long unstructured sequences of images, often with high degree of redundancy and abrupt appearance changes even in temporally adjacent frames, that harden the extraction of semantically meaningful content. Temporal segmentation, the process of organizing unstructured data into homogeneous chapters, provides a large potential for extracting semantic information. Indeed, once the photo-stream has been divided into a set of homogeneous and manageable segments, each segment can be represented by a small number of key-frames and indexed by semantic features, providing a basis for understanding the semantic structure of the event.



Figure 2.1: Example of temporal segmentation of an egocentric sequence based on what the camera wearer sees. In addition to the segmentation, our method provides a set of semantic attributes that characterize each segment.

2.2 Related works

State-of-the-art methods for temporal segmentation can be broadly classified into works with focus on what-the-camera-wearer-sees (Castro et al., 2015; Doherty and Smeaton, 2008; Talavera et al., 2015) and on what-the-camera-wearer-does (Poleg et al., 2014, 2016). As an example, from the what-camera-wearer-does perspective, the camera wearer spending time in a bar while sit, will be considered as a unique event (sitting). From the what-the-camera-wearer-sees perspective, the same situation will be considered as several separate events (waiting for the food, eating, and drinking beer with a friend who joins later). The distinction between the aforementioned points of view is crucial as it leads to different definitions of an event. In this respect, our proposed method fits in the what-the-camera-wearer-sees category. Early works on egocentric temporal segmentation (Doherty and Smeaton, 2008; Lin and Hauptmann, 2006) focused on what the *camera wearer sees* (e.g. people, objects, foods, etc.). For this purpose, the authors used as image representation, low-level features to capture the basic characteristics of the environment around the user, such as color, texture or information acquired through different camera sensors. More recently, the works in (Bolaños et al., 2015) and (Talavera et al., 2015) have used Convolutional Neural Network (CNN) features extracted by using the AlexNet model (Krizhevsky, Sutskever and Hinton, 2012) trained on ImageNet as a fixed feature extractor for image representation. Some other recent methods infer from the images what the *camera wearer does* (e.g. sitting, walking, running, etc.). Castro et al. (Castro et al., 2015) used CNN features together with metadata and color histogram.

Most of these methods use as image representation ego-motion (Lu and Grauman, 2013; Bolaños et al., 2014; Poleg et al., 2014, 2016), which is closely related to the user motion-based activity but cannot be reliably estimated in photo streams. The authors combined a CNN trained on egocentric data with a posterior Random Decision Forest in a late-fusion ensemble, obtaining promising results for a single user. However, this approach lack of generalization, since it requires to re-train the model for any new user, implying to manually annotate large amount of images. To the best of our knowledge, except the work of Castro et al. (Castro et al., 2015), Doherty et al. (Doherty and Smeaton, 2008) and Tavalera et al. (Talavera et al., 2015), all other state-of-the-art methods have been designed for and tested on videos.

We proposed an unsupervised method, called *R-Clustering* in (Talavera et al., 2015). Our aim was to segment photo streams from the what-the-camera-wearer-see perspective. The proposed methods rely on the combination of Agglomerative Clustering (AC), that usually has a high recall, but leads to temporal over-segmentation, with a statistically founded change detector, called ADWIN (Bifet and Gavaldà, 2007), which despite its high precision, usually leads to temporal

under-segmentation. Both approaches are integrated into a *Graph-Cut (GC)* (Boykov et al., 2001) framework to obtain a trade-off between AC and ADWIN, which have complementary properties. The graph-cut relies on CNN-based features extracted using AlexNet, trained on ImageNet, as a fixed feature extractor to detect the segment boundaries.

Later, we extend our previous work by adding a semantic level to the image representation. Due to the free motion of the camera and its low frame rate, abrupt changes are visible even among temporally adjacent images (see Fig. 2.1 and Fig. 2.8). Under these conditions motion and low-level features such as color or image layout are prone to fail for event representation, hence urges the need to incorporate higher-level semantic information. Instead of representing images simply by their contextual CNN features, which capture the basic environment appearance, we detect segments as a set of temporally adjacent images with the same contextual representation in terms of semantic visual concepts. Nonetheless, not all the semantic concepts in an image are equally discriminant for environment classification: objects like trees and buildings can be more discriminant than objects like dogs or mobile phones, since the former characterizes a specific environment such as forest or street, whereas the latter can be found in many different environments. In this paper, we propose a method called Semantic Regularized Clustering (SR-Clustering), which takes into account semantic concepts in the image together with the global image context for event representation.

This manuscript is organized as follows: Section 2.3 provides a description of the proposed photo stream segmentation approach discussing the semantic and contextual features, the clustering and the graph-cut model. Section 2.4 presents experimental results and, finally, Section 2.5 summarizes the important outcomes of the proposed method providing some concluding remarks.

2.3 Approach

A visual overview of the proposed method is given in Fig. 2.2. The input is a day-long photo-stream from which contextual and semantic features are extracted. An initial clustering is performed by AC and ADWIN. Later, GC is applied to look for a trade-off between the AC (represented by the bottom colored circles) and ADWIN (represented by the top colored circles) approaches. The binary term of the GC imposes smoothness and similarity of consecutive frames in terms of the CNN image features. The output of the proposed method is the segmented photo-stream. In this section, we introduce the semantic and contextual features of SR-clustering and provide a detailed description of the segmentation approach.

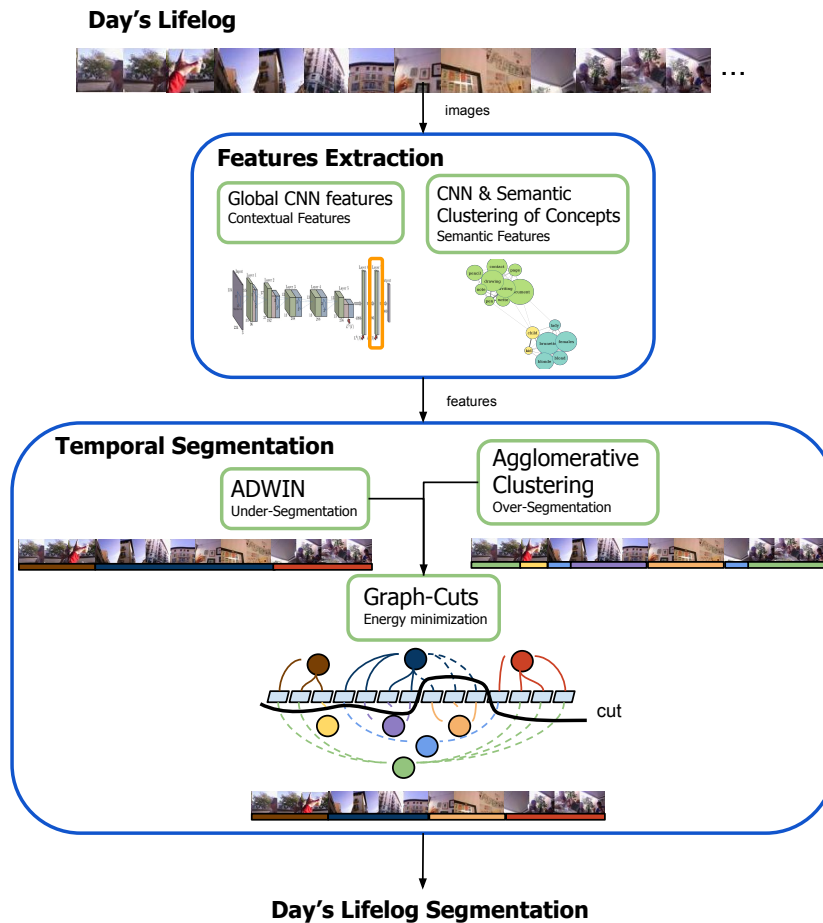


Figure 2.2: General scheme of the Semantic Regularized Clustering (SR-Clustering) method.

2.3.1 Features

We assume that two consecutive images belong to the same segment if they can be described by similar image features. When we refer to the features of an image, we usually consider low-level image features (e.g. color, texture, etc.) or a global representation of the environment (e.g. CNN features). However, the objects or concepts that semantically represent an event are also of high importance for the photo stream segmentation. Below, we detail the features that semantically describe the egocentric images.

Semantic Features

Given an image I , let us consider a tagging algorithm that returns a set of objects/tags/concepts detected in the images with their associated confidence value. The confidence values of each concept form a semantic feature vector to be used for the photo streams segmentation. Usually, the number of concepts detected for each sequence of images is large (often, some dozens). Additionally, redundancies in the detected concepts are quite often due to the presence of synonyms or semantically related words. To manage the semantic redundancy, we will rely on WordNet (Miller, 1995), which is a lexical database that groups English words into sets of synonyms, providing additionally short definitions and word relations.

Given a day's lifelog, let us cluster the concepts by relying on their synset ID in WordNet to compute their similarity in meaning, and following, apply clustering (e.g. Spectral clustering) to obtain 100 clusters. As a result, we can semantically describe each image in terms of 100 concepts and their associated confidence scores. Formally, we first construct a semantic similarity graph $\mathcal{G} = \{V, E, W\}$, where each vertex or node $v_i \in V$ is a concept, each edge $e_{ij} \in E$ represents a semantic relationship between two concepts, v_i and v_j and each weight $w_{ij} \in W$ represents the strength of the semantic relationship, e_{ij} . We compute each w_{ij} by relying on the meanings and the associated similarity given by WordNet, between each appearing pair. To do so, we use the max-similarity between all the possible meanings m_i^k and m_j^r in M_i and M_j of the given pair of concepts v_i and v_j :

$$w_{ij} = \max_{m_i^k \in M_i, m_j^r \in M_j} \text{sim}(m_i^k, m_j^r).$$

To compute the Semantic Clustering, we use their similarity relationships in the spectral clustering algorithm to obtain 100 semantic concepts, $|C| = 100$. In Fig. 2.3, a simplified example of the result obtained after the clustering procedure is shown. For instance, in the purple cluster, similar concepts like 'writing', 'document', 'drawing', 'write', etc. are grouped in the same cluster, and 'writing' is chosen as the most representative term. For each cluster, we choose as its representative concept, the one with the highest sum of similarities with the rest of the elements in the cluster.

The semantic feature vector $f^s \in \mathbb{R}^{|C|}$ for image I is a 100-dimensional array, such that each component $f^s(I)_j$ of the vector represents the confidence with which the j -th concept is detected in the image. The confidence value for the concept j , representing the cluster C_j , is obtained as the sum of the confidences r_I of all the concepts included in C_j that have also been detected on image I :

$$f^s(I)_j = \sum_{c_k \in \{C_j\}} r_I(c_k)$$

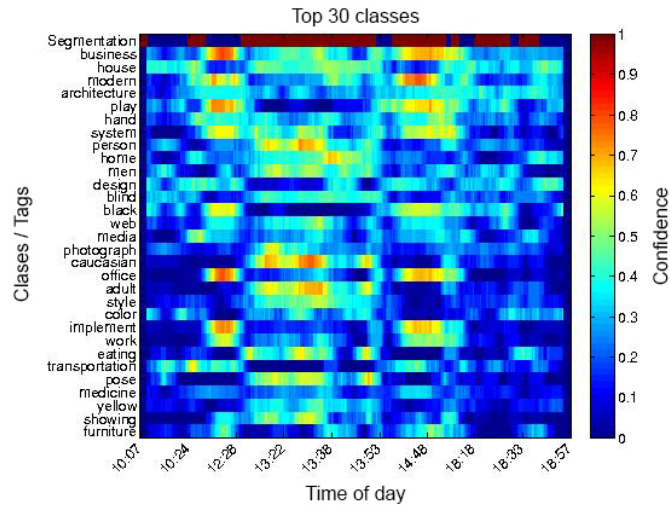


Figure 2.4: Example of the final semantic feature matrix obtained for an egocentric sequence. The top 30 concepts (rows) are shown for all the images in the sequence (columns). Additionally, the top row of the matrix shows the ground truth (GT) segmentation of the dataset.

Taking into account that the camera wearer can be continuously moving, even if in a single environment, the objects that can be appearing in temporally adjacent images may be different. To this end, we apply a Parzen Window Density Estimation method (Parzen, 1962) to the matrix obtained by concatenating the semantic feature vectors along the sequence to obtain a smoothed and temporally coherent set of confidence values. Additionally, we discard the concepts with low variability of confidence values along the sequence which correspond to non-discriminative concepts that can appear on any environment. The low variability of the confidence value of a concept may correspond to constantly having high or low confidence value in most environments.

In Fig. 2.4, the matrix of concepts (semantic features) associated with an egocentric sequence is shown, displaying only the top 30 classes. Each column of the matrix corresponds to a frame and each row indicates the confidence with which the concept is detected in each frame. In the first row, the ground truth of the temporal segmentation is shown for comparison purposes. With this representation, repeated patterns along a set of continuous images correspond to the set of concepts that characterizes an event. For instance, the first frames of the sequence represent an indoor scene, characterized by the presence of people (see examples Fig. 2.5). The whole process is summarized in Fig. 2.6.



Figure 2.5: Example of extracted tags on different segments. The first one corresponds to the period from 13.22 - 13.38 where the user is having lunch with colleagues, and the second, from 14.48 - 18.18, where he/she is working in the office with the laptop.

In order to consider the semantics of temporal segments, we used a concept detector based on the auto-tagging service developed by Imagga Technologies Ltd. Imagga’s auto-tagging technology¹ uses a combination of image recognition based on deep learning and CNNs using very large collections of human-annotated photos. The advantage of Imagga’s Auto Tagging API is that it can directly recognize over 2,700 different objects and in addition return more than 20,000 abstract concepts related to the analyzed images.

Contextual Features

In addition to the semantic features, we represent images with a feature vector extracted from a pre-trained CNN. The CNN model that we use for computing the images’ representation is the AlexNet, which is detailed in (Krizhevsky, Sutskever and Hinton, 2012). The features are computed by removing the last layer corresponding to the classifier from the network. We used the deep learning framework Caffe (Jia, 2013) in order to run the CNN. Due to the fact that the weights have been trained on the ImageNet database (Deng et al., 2009), which is made of images containing single objects, we expect that the features extracted from images containing multiple objects will be representative of the environment. It is worth to remark that we did not use the weights obtained using a pre-trained CNN on the scenes from Places 205 database (Zhou et al., 2014), since the Narrative camera’s field of view is narrow, which means that mostly its field-of-view is very restricted to characterize the whole scene. Instead, we usually only see objects on the foreground. As detailed in (Talavera et al., 2015), to reduce the large variation distribution of the CNN features, which results in problems when computing distances between vectors, we used a signed root normalization to produce more uniformly distributed data (Zheng et al., 2014).

¹<http://www.imagga.com/solutions/auto-tagging.html>

2.3.2 Temporal Segmentation

Due to the low-temporal resolution of egocentric videos, as well as to the camera wearer's motion, temporally adjacent egocentric images may be very dissimilar between them. Hence, we need robust techniques to group them and extract meaningful video segments. In the following, we detail each step of our approach that relies on an AC regularized by a robust change detector within a GC framework.

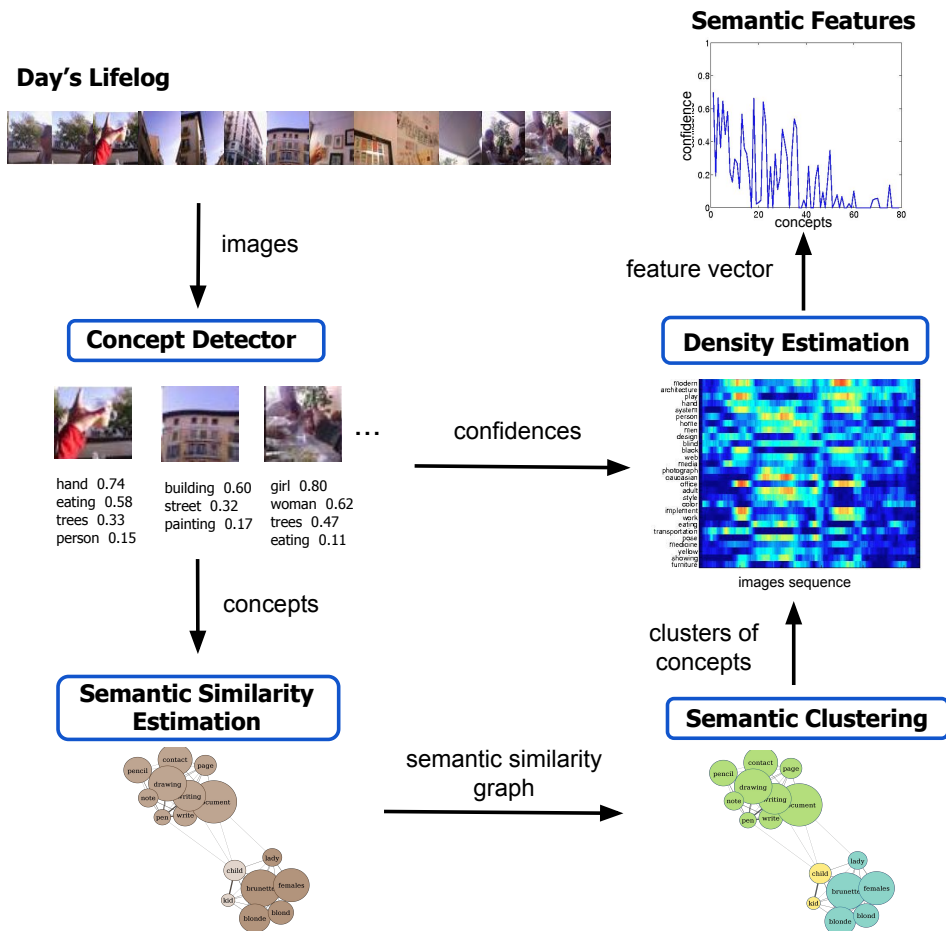


Figure 2.6: General scheme of the semantic feature extraction methodology.

Clustering methods:

The AC method follows a general bottom-up clustering procedure, where the criterion for choosing the pair of clusters to be merged in each step is based on the distances among the image features. The inconsistency between clusters is defined through the *cut* parameter. In each iteration, the most similar pair of clusters are merged and the similarity matrix is updated until no more consistent clustering are possible. We chose the Cosine Similarity to measure the distance between frames features, since it is a widely used measure of cohesion within clusters, especially in high-dimensional positive spaces (Tan et al., 2005). However, due to the lack of incidence for determining the clustering parameters, the final result is usually over-segmented.

Statistical bound for the clustering:

To bound the over-segmentation produced by AC, we propose to model the video as a multivariate data stream and detect changes in the mean distribution through an online learning method called Adaptive Windowing (**ADWIN**) (Bifet and Gavaldà, 2007). ADWIN works by analyzing the content of a sliding window, whose size is adaptively recomputed according to its rate of change: when the data is stationary the window increases, whereas when the data is statistically changing, the window shrinks. According to ADWIN, whenever two large enough temporally adjacent (sub)windows of the data, say W_1 and W_2 , exhibit distinct enough means, the algorithm concludes that the expected values within those windows are different, and the older (sub)window is dropped. *Large enough* and *distinct enough* are defined by the Hoeffding's inequality (Hoeffding, 1963), testing if the difference between the averages on W_1 and W_2 is larger than a threshold, which only depends on a pre-determined confidence parameter δ . The Hoeffding's inequality guarantees rigorously the performance of the algorithm in terms of false-positive rate.

This method has been recently generalized in (Drozdzal et al., 2014) to handle k -dimensional data streams by using the mean of the norms. In this case, the bound has been shown to be:

$$\epsilon_{cut} = k^{1/p} \sqrt{\frac{1}{2m} \ln \frac{4}{k\delta'}}$$

where p indicates the p -norm, $|W| = |W_0| + |W_1|$ is the length of $W = W_1 \cup W_2$, $\delta' = \frac{\delta}{|W|}$, and m is the harmonic mean of $|W_0|$ and $|W_1|$. Given a confidence value δ , the higher the dimension k is, the more samples $|W|$ the bound needs to reach assuming the same value of ϵ_{cut} . The higher the norm is used, the less important is the dimensionality k . Since we model the video as a high dimensional multivariate data stream, ADWIN is unable to predict changes involving a small number of samples,

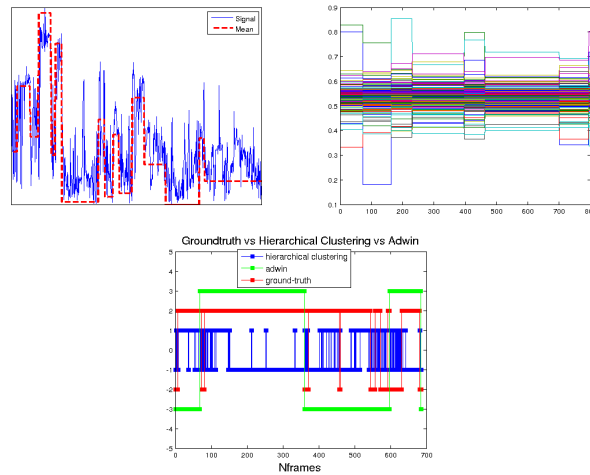


Figure 2.7: Left: change detection by ADWIN on a $1 - D$ data stream, where the red line represents the estimated mean of the signal by ADWIN; Center: change detection by ADWIN on a $500 - D$ data stream, where, in each stationary interval, the mean is depicted with a different color in each dimension; Right: results of the temporal segmentation by ADWIN (green) vs AC over-segmentation (blue) vs ground-truth shots (red) along the temporal axis (the abscissa).

which often characterizes life-logging data, leading to under-segmentation. Moreover, since it considers only the mean change, it is enabled to detect changes due to other statistics such as the variance. The ADWIN under-segmentation represents a statistical bound for the AC (see Fig.2.7 (right)). We use GC as a framework to integrate both approaches and to regularize the over-segmentation of AC by the statistical bound provided by ADWIN.

Graph-Cut regularization of egocentric videos:

GC is an energy-minimization technique that minimizes the energy resulting from a weighted sum of two terms: the *unary energy* $U(-)$, that describes the relationship of the variables to a possible class and the *binary energy* $V(-, -)$, that describes the relationship between two neighbouring samples (temporally close video frames) according to their feature similarity. GC has the goal to smooth boundaries between similar frames, while attempts to keep the cluster membership of each video frame according to its likelihood. We define the unary energy as a sum of 2 parts ($U_{ac}(f_i)$ and $U_{adw}(f_i)$) according to the likelihood of a frame to belong to segments coming from the AC and ADWIN. The GC energy to minimize is as follows:

$$E(f) = \sum_i ((1 - \omega_1)U_{ac}(f_i) + \omega_1 U_{adw}(f_i)) + \omega_2 \sum_{i,n \in N_i} \frac{1}{N_i} V_{i,n}(f_i, f_n)$$

where $f_i, i = \{1, \dots, m\}$ are the set of image features, N_i are the temporal frame neighbours of image i , ω_1 and ω_2 ($\omega_1, \omega_2 \in [0, 1]$) are the unary and the binary weighting terms respectively. Defining how much weight do we give to the likelihood of each unary term (AC and Adwin, always combining the events split of both methods), and balancing the trade-off between the unary and the pairwise energies, respectively. The minimization is achieved through the max-cut algorithm, leading to a temporal video segmentation with similar frames having as large likelihood as possible to belong to the same event, while maintaining video segment boundaries in neighbouring frames with high feature dissimilarity.

Features:

As image representation for both segmentation techniques, we used the CNN features (Jia, 2013). The CNN features trained on ImageNet (Krizhevsky, Sutskever and Hinton, 2012) have demonstrated to be successfully transferred to other visual recognition tasks such as scene classification and retrieval. In this work, we extracted the 4096-D CNN vectors by using the Caffe (Jia, 2013) implementation trained on ImageNet. Since each CNN feature has a large variation distribution in its value, and this could be problematic when computing distances between vectors, we used a signed root normalization to produce more uniformly distributed data (Zheng et al., 2014). First, we apply the function $f(x) = \text{sign}(x)|x|^\alpha$ on each dimension and then we l_2 -normalize the feature vector. In all the experiments, we take $\alpha = 0.5$. Following we apply a PCA dimensionality reduction keeping 95% of the data variance. Only in the GC pair-wise term we use a different feature pre-processing, where we simply apply a 0-1 data normalization.

$$\epsilon_{cut} = k^{1/p} \sqrt{\frac{1}{2m} \ln \frac{4}{k\delta'}}$$

2.4 Experiments and Validation

In this section, we discuss the datasets and the statistical evaluation measurements used to validate the proposed model and to compare it with the state-of-the-art methods. To sum up, we apply the following methodology for validation:

1. Three different datasets acquired by 3 different wearable cameras are used for validation.

2. The F-Measure is used as a statistical measure to compare the performance of different methods.
3. Two consistency measures to compare different manual segmentations is applied.
4. Comparison results of SR-Clustering with 3 state-of-the-art techniques is provided.
5. Robustness of the final proposal is proven by validating the different components of SR-Clustering.

2.4.1 Data

To evaluate the performance of our method, we used 3 public datasets (EDUB-Seg, AIHS and Huji EgoSeg’s sub dataset) acquired by three different wearable cameras (see Table 2.1).

| Dataset | Camera | FR | SR | #Us | #Days | #Img |
|-------------|--------------|--------|-----------|-----|-------|--------|
| EDUB | Narrative | 2 fpm | 2592x1944 | 7 | 20 | 18,735 |
| AIHS-subset | SenseCam | 3 fpm | 640x480 | 1 | 5 | 11,887 |
| Huji EgoSeg | GoPro Hero3+ | 30fps* | 1280x720 | 2 | 2 | 700 |

Table 2.1: Table summarizing the main characteristics of the datasets used in this work: frame rate (FR), spatial resolution (SR), number of users (#Us), number of days (#Days), number of images (#Img). The Huji EgoSeg dataset has been subsampled to 2 fpm as detailed in the main text.

EDUB-Seg: is a dataset acquired by people from our lab with the Narrative Clip, which takes a picture every 30 seconds. Our Narrative dataset, named EDUB-Seg (Egocentric Dataset of the University of Barcelona - Segmentation), contains a total of 18,735 images captured by 7 different users during overall 20 days. To ensure diversity, all users were wearing the camera in different contexts: while attending a conference, on holiday, during the weekend, and during the week. The EDUB-Seg dataset is an extension of the dataset used in our previous work (Talavera et al., 2015), that we call EDUB-Seg (Set1) to distinguish it from the newly added in this paper EDUB-Seg (Set2). The camera wearers, as well as all the researchers involved in this work, were required to sign an informed written consent containing a set of moral principles (Wiles et al., 2008; Kelly et al., 2013). Moreover, all researchers of the team have signed to do not publish any image identifying a person in a photo stream without his/her explicit permission, except for unknown third parties.

AIHS subset: is a subset of the daily images from the database called *All I Have Seen* (AIHS) (Jojic et al., 2010), recorded by the SenseCam camera that takes a picture every 20 seconds. The original AIHS dataset ² has no timestamp metadata. We manually divided the dataset in five days guided by the pictures the authors show in the website of their project and based on the daylight changes observed in the photo streams. The five days sum up a total of 11,887 images. Comparing both cameras (Narrative and SenseCam), we can remark their difference with respect to the cameras' lens (fish eye vs normal), and the quality of the images they record. Moreover, SenseCam acquires images with a larger field of view and significant deformation and blurring. We manually defined the GT for this dataset following the same criteria we used for the EDUB-Seg photo streams.

Huji EgoSeg: due to the lack of other publicly available LTR datasets for event segmentation, we also test our temporal segmentation method to the ones provided in the dataset Huji EgoSeg (Poleg et al., 2014). This dataset was acquired by the Go-Pro camera, which captures videos with a temporal resolution of 30fps. Considering the very significant difference in frame-rate of this camera compared to Narrative (2 fpm) and SenseCam (3 fpm), we applied a sub-sampling of the data by just keeping 2 images per minute, to make it comparable to the other datasets. In this dataset, several short videos recorded by two different users are provided. Consequently, after sub-sampling all the videos, we merged the resulting images from all the short videos to construct a dataset per each user, which consists of a total number of 700 images. The images were merged following the numbering order that was provided by the authors to their videos. We also manually defined the GT for this dataset following the same used criteria for the EDUB-Seg dataset.

In summary, we evaluate the algorithms on 27 days with a total of 31,322 images recorded by 10 different users. All datasets contain a mixture of highly variable indoor and outdoor scenes with a large variety of objects. We make public the EDUB-Seg dataset³, together with our GT segmentations of the datasets Huji EgoSeg and AIHS subset. Additionally, we release the SR-Clustering ready-to-use complete code⁴.

²<http://research.microsoft.com/en-us/um/people/jojic/aihs/>

³<http://www.ub.edu/cvub/dataset/>

⁴<https://github.com/MarcBS/SR-Clustering>

2.4.2 Experimental setup

Following (Li et al., 2013), we measured the performances of our method by using the F-Measure (FM) defined as follows:

$$FM = 2 \frac{RP}{R + P},$$

where P is the precision defined as ($P = \frac{TP}{TP+FP}$) and R is the recall, defined as ($R = \frac{TP}{TP+FN}$). TP , FP and FN are the number of true positives, false positives and false negatives of the detected segment boundaries of the photo stream. We define the FM, where we consider TPs the images that the model detects as boundaries of an event and that were close to the boundary image defined in the GT by the annotator (given a tolerance of 5 images in both sides). The FPs are the images detected as events delimiters, but that were not defined in the GT, and the FNs the lost boundaries by the model that are indicated in the GT. Lower FM values represent a wrong boundary detection while higher values indicate a good segmentation. Having the ideal maximum value of 1, where the segmentation correlates completely with the one defined by the user.

The annotation of temporal segmentations of photo streams is a very subjective task. The fact that different users usually do not perform the same when annotating, may lead to bias in the evaluation performance. The problem of the subjectivity when defining the ground truth was previously addressed in the context of image segmentation (Martin et al., 2001). In (Martin et al., 2001), the authors proposed two measures to compare different segmentations of the same image. These measures are used to validate if the performed segmentations by different users are consistent and thus, can be served as an objective benchmark for the evaluation of the segmentation performances. In Fig. 2.8, we report a visual example that illustrates the urge of employing this measure for temporal segmentation of egocentric photo streams. For instance, the first segment in Fig. 2.8 is split into different segments when analyzed by different subjects although there is a degree of consistency among all segments.

Inspired by this work, we re-define the local refinement error, between two temporal segments, as follows:

$$E(S_A, S_B, I_i) = \frac{|R(S_A, I_i) \setminus R(S_B, I_i)|}{|R(S_A, I_i)|},$$

where \setminus denotes the set difference and, S_A and S_B are the two segmentations to be compared. $R(S_X, I_i)$ is the set of images corresponding to the segment that contains the image I_i , when obtaining the segmentation boundaries S_X .

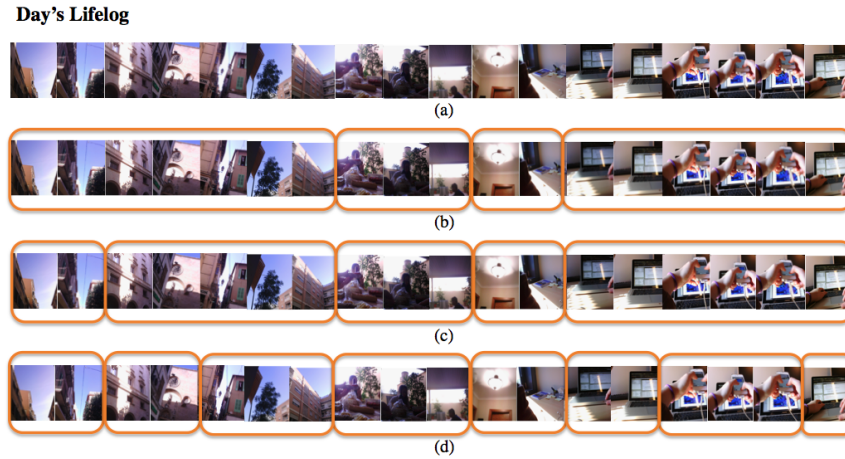


Figure 2.8: Different segmentation results obtained by different subjects. (a) shows a part of a day. (b), (c) and (d) are examples of the segmentation performed by three different persons. (c) and (d) are refinements of the segmentation performed by (b). All three results can be considered as being correct, due to the subjective intrinsic of the task. As a consequence, a segmentation consistency metric should not penalize different, yet consistent results of the segmentation.

If one temporal segment is a proper subset of the other, then the images lie in one interval of refinement, which results in the local error of zero. However, if there is no subset relationship, the two regions overlap in an inconsistent manner that results in a non-zero local error. Based on the definition of local refinement we provided above, two error measures are defined by combining the values of the local refinement error for the entire sequence. The first error measure is called Global Consistency Error (GCE) that forces all local refinements to be in the same direction (segments of segmentation A can be only local refinements of segments of segmentation B). The second error measure is the Local Consistency Error (LCE), which allows refinements in different directions in different parts of the sequence (some segments of segmentation A can be of local refinements of segments of segmentation B and vice versa). The two measures are defined as follows:

$$GCE(S_A, S_B) = \frac{1}{n} \min \left\{ \sum_i^n E(S_A, S_B, I_i), \sum_i^n E(S_B, S_A, I_i) \right\}$$

$$LCE(S_A, S_B) = \frac{1}{n} \sum_i^n \min \{ E(S_A, S_B, I_i), E(S_B, S_A, I_i) \}$$

where n is the number of images of the sequence, S_A and S_B are the two different temporal segmentations and I_i indicates the i -th image of the sequence. The GCE and the LCE measures produce output values in the range $[0, 1]$ where 0 means no error.

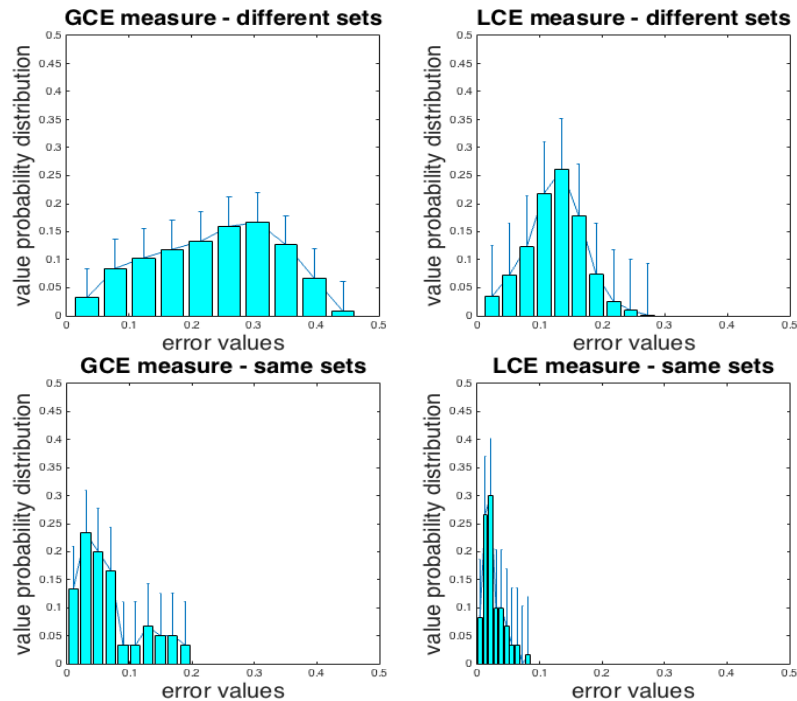


Figure 2.9: GCE (left) and LCE (right) normalized histograms with the error values distributions, showing their mean and variance. The first row graphs represent the distribution of errors comparing segmentations of different sequences while the second row graphs show the distribution of error when comparing segmentations of the same set, including the segmentation of the camera wearer.

To verify that there is consistency among different people for the task of temporal segmentation, we asked three different subjects to segment each of the 20 sets of the EDUB-Seg dataset into events. The subjects were instructed to consider an *event* as a semantically perceptual unit that can be inferred by visual features, without any prior knowledge of what the camera wearer is actually doing. No instructions were given to the subjects about the number of segments they should annotate. This process gave rise to 60 different segmentations. The number of all possible pairs of segmentations is 1800, 60 of which are pairs of segmentations of the same set. For each pair of segmentations, we computed GCE and LCE. First, we considered only

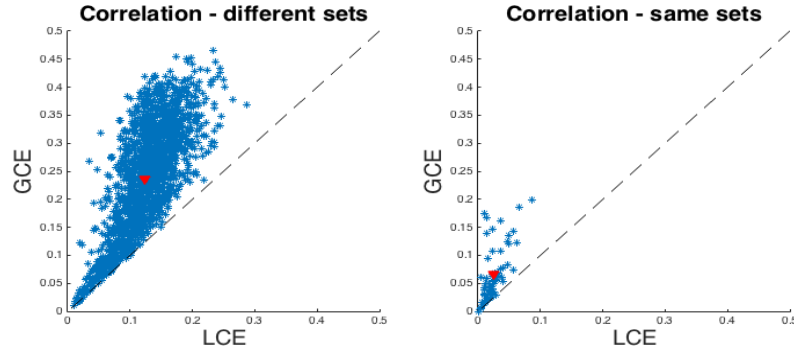


Figure 2.10: LCE vs GCE for pairs of segmentations of different sequences (left) and for pairs of segmentations of the same sequence (right). The differences w.r.t. the dashed line $x=y$ show how GCE is a stricter measure than LCE. The red dot represents the mean of all the cloud of values, including the segmentation of the camera wearer.

pairs of segmentations of the same sequence and then, considered the rest of possible pairs of segmentations in the dataset. The first two graphics in Fig. 2.9 (first row) show the GCE (left) and LCE (right) when comparing each set segmentations with the segmentations applied on the rest of the sets. The two graphics in the second row show the distribution of the GCE (left) and LCE (right) error when analyzing different segments describing the same video. As expected, the distributions that compare the segmentations over the same photo-stream have the center of mass to the left of the graph, which means that the mean error between the segmentations belonging to the same set is lower than the mean error between segmentations describing different sets. In Fig. 2.10 we compare, for each pair of segmentations, the measures produced by different datasets segmentations (left) and the measures produced by segmentations of the same dataset (right). In both cases, we plot LCE vs. GCE. As expected, the average error between segmentations of the same photo-stream (right) is lower than the average error between segmentations of different photo-streams (left). Moreover, as indicated by the shape of the distributions on the second row of Fig.2.10 (right), the peak of the LCE is very close to zero. Therefore, we conclude that given the task of segmenting an egocentric photo-stream into events, different people tend to produce consistent and valid segmentation. Fig. 2.11 and 2.12 show segmentation comparisons of three different persons (not being the camera wearer) that were asked to temporally segment a photo-stream and confirm our statement that different people tend to produce consistent segmentations.

Since our interpretation of events is biased by our personal experience, the segmentation done by the camera wearer could be very different by the segmentations done by third persons. To quantify this difference, in Fig. 2.9 and Fig. 2.10 we

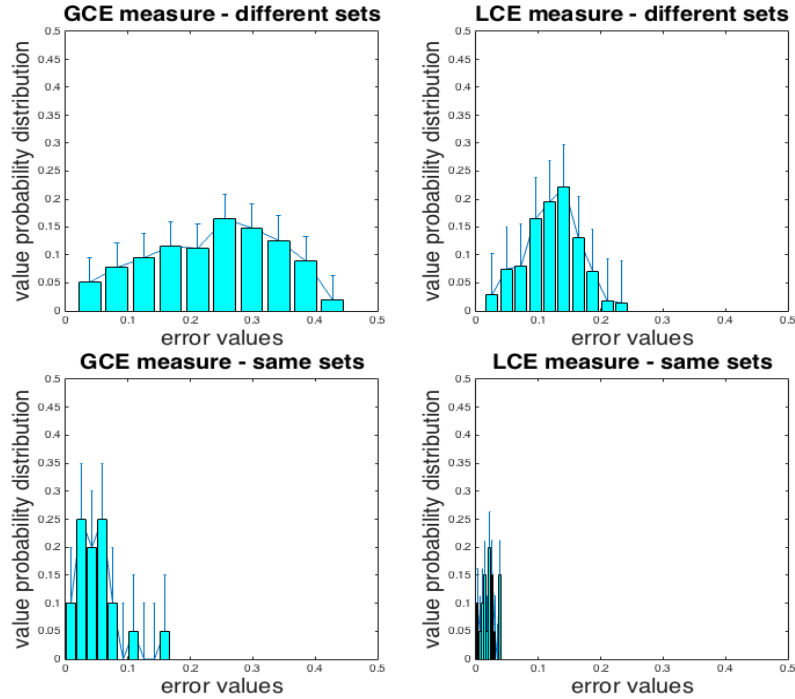


Figure 2.11: GCE (left) and LCE (right) normalized histograms with the error values distributions, showing their mean and variance. The first row graphs represent the distribution of the errors comparing segmentations of different sequences while the second row graphs show the distribution of the errors when comparing segmentations of the same set, excluding the segmentation of the camera wearer.

evaluated the LCE and the GCE including also the segmentation performed by the camera wearer. From this comparison, we can observe that the error mean does not vary but that the degree of local and global consistency is higher when the set of annotators does not include the camera wearer as it can be appreciated by the fact that the distributions are slightly shifted to the left and thinner. However, since this variation is of the order of 0.05%, we can conclude that event segmentation of egocentric photo-streams can be objectively evaluated.

When comparing the different segmentation methods w.r.t. the obtained FM (see section 2.4.3), we applied a grid-search for choosing the best combination of hyper-parameters. The set of hyper-parameters tested are the following:

- AC linkage methods $\in \{ward, centroid, complete, weighted, single, median, average, \}$

- AC cutoff $\in \{0.2, 0.4, \dots, 1.2\}$,
- GraphCut unary weight ω_1 and binary weight $\omega_2 \in \{0, 0.1, 0.2, \dots, 1\}$,
- AC-Color $t \in \{10, 25, 40, 50, 60, 80, 90, 100\}$.

2.4.3 Experimental results

In Table 2.2, we show the FM results obtained by different segmentation methods over different datasets. The first two columns correspond to the datasets used in (Talavera et al., 2015): AIHS-subset and EDUB-Seg (Set1). The third column corresponds to the EDUB-Seg (Set2) introduced in this paper. Finally, the fourth column corresponds to the results on the whole EDUB-Seg. The first part of the table (first three rows) presents comparisons to state-of-the-art methods. The second part of the table (next 4 rows), shows comparisons to different components of our proposed clustering method with and without semantic features. Finally, the third part of the table shows the results obtained using different variations of our method.

In the first part of Table 2.2, we compare to state-of-the-art methods. The first method is the Motion-Based segmentation algorithm proposed by Bolaños et al. (Bolaños et al., 2014). As can be seen, the average results obtained are far below SR-Clustering. This can be explained by the type of features used by the method, which are more suited for applying a motion-based segmentation. This kind of segmentation is more oriented to recognize activities and thus, is not always fully aligned with the event segmentation labeling we consider (i.e. in an event where the user

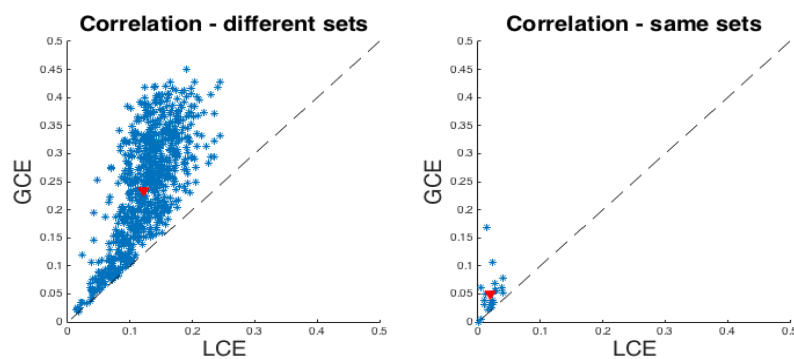


Figure 2.12: LCE vs GCE for pairs of segmentations of different sequences (left) and for pairs of segmentations of the same sequence (right). The differences w.r.t. the dashed line $x=y$ show how GCE is a stricter measure than LCE. The red dot represents the mean of all the cloud of values, excluding the segmentation of the camera wearer.

goes outside of a building, and then enters to the underground tunnels can be considered "in transit" by the Motion-Based segmentation, but be considered as three different events in our event segmentation). Furthermore, the obtained FM score on the Narrative datasets is lower than the SenseCam's for several reasons: Narrative has lower frame rate compared to Sensecam (AIHS dataset), which is a handicap when computing motion information, and a narrower field of view, which decreases the semantic information present in the image. We also evaluated the proposal of Lee and Grauman (Lee and Grauman, 2015) (best with $t = 25$), where they apply an Agglomerative Clustering segmentation using LAB color histograms. In this case, we see that the algorithm is even far below the obtained results by AC, where the Agglomerative Clustering algorithm is used over contextual CNN features instead of colour histograms. The main reason for this performance difference comes from



Figure 2.13: Illustration of our SR-Clustering segmentation results from a subset of pictures from a Narrative set. Each line represents a different segment. Below each segment we show the top 8 found concepts (from left to right). Only a few pictures from each segment are shown.

| | AIHS (Jojic et al., 2010) | EDUB-Seg Set1 | EDUB-Seg Set2 | EDUB-Seg |
|--------------------------------------|---------------------------|---------------|---------------|-------------|
| Motion (Bolaños et al., 2014) | 0.66 | 0.34 | | |
| AC-Color (Lee and Grauman, 2015) | 0.60 | 0.37 | 0.54 | 0.50 |
| R-Clustering (Talavera et al., 2015) | 0.79 | 0.55 | | |
| ADW | 0.31 | 0.32 | | |
| ADW-ImaggaD | 0.35 | 0.55 | 0.29 | 0.36 |
| AC | 0.68 | 0.45 | | |
| AC-ImaggaD | 0.72 | 0.53 | 0.64 | 0.61 |
| SR-Clustering-LSDA | 0.78 | 0.60 | 0.64 | 0.61 |
| SR-Clustering-NoD | 0.77 | 0.66 | 0.63 | 0.60 |
| SR-Clustering | 0.78 | 0.69 | 0.69 | 0.66 |

Table 2.2: Average FM results of the state-of-the-art works on the egocentric datasets (first part of the table); for each of the components of our method (second part); and for each of the variations of our method (third part). The last line shows the results of our complete method. AC stands for Agglomerative Clustering, ADW for ADWIN and ImaggaD is our proposal for semantic features, where D stands for Density Estimation.

the high difference in features expressiveness, that supports the necessity of using a rich set of features for correctly segmenting highly variable egocentric data. The last row of the first section of the table shows the results obtained by our previously published method (Talavera et al., 2015), where we were able to outperform the state-of-the-art of egocentric segmentation using contextual CNN features both on AIHS-subset and on EDUB-Seg Set1. Another possible method to compare with would be the one from Castro et al. (Castro et al., 2015), although the authors do not provide their trained model for applying this comparison.

In the second part of Table 2.2, we compare the results obtained using only ADWIN or only AC with (ADW-ImaggaD, AC-ImaggaD) and without (ADW, AC) semantic features. One can see that the proposed semantic features, leads to an improved performance, indicating that these features are rich enough to provide improvements on egocentric photo-stream segmentation.

Finally, on the third part of Table 2.2, we compared our segmentation methodology using different definitions for the semantic features. In the SR-Clustering-LSDA case, we used a simpler semantic features description, formed by using the weakly supervised concept extraction method proposed in (Hoffman et al., 2014), namely LSDA. In the last two lines, we tested the model using our proposed semantic methodology (Imagga’s tags) either without Density Estimation, SR-Clustering-NoD or with the final Density Estimation (SR-Clustering), respectively.

Comparing the results of SR-Clustering and R-Clustering on the first two datasets (AIHS-subset and EDUB-Seg Set1), we can see that our new method is able to outperform the results adding 14 points of improvement to the FM score, while keeping

nearly the same FM value on the SenseCam dataset. The improvement achieved using semantic information can be also corroborated, when comparing the FM scores obtained on the second half of EDUB-Seg dataset (Set2 on the 3rd column) and on the complete version of this data (see the last column of the Table).

| Huji EgoSeg (Poleg et al., 2014) LTR | |
|--------------------------------------|-------------|
| ADW-ImaggaD | 0.59 |
| AC-ImaggaD | 0.88 |
| SR-Clustering | 0.88 |

Table 2.3: Average FM score on each of the tested methods using our proposal of semantic features on the dataset presented in (Poleg et al., 2014).

In Table 2.3 we report the FM score obtained by applying our proposed method on the sub-sampled Huji EgoSeg dataset to be comparable to LTR cameras. Our proposed method achieves high performance, being 0.88 of FM for both AC and SR-Clustering when using the proposed semantic features. The improvement of the results when using the GoPro camera with respect to Narrative or SenseCam can be explained by two key factors: 1) the difference in the field of view captured by GoPro (up to 170°) compared to SenseCam (135°) and Narrative (70°), 2) the better image quality achieved by the head-mounted camera.

In addition to the FM score, we could not consider the GCE and LCE measures to compare the consistency of the automatic segmentations to the ground truth, since both methods lead to a number of segments much larger than the number of segments in the ground truth and therefore these measures would not descriptive enough. This is due to the fact that any segmentation is a refinement of one segment for the entire sequence, and one image per segment is a refinement of any segmentation. Consequently, these two trivial segmentations, one segment for the entire sequence and one image per segment, achieve error zero for LCE and GCE. However, we observed that on average, the number of segments obtained by the method of Lee and Grauman (Lee and Grauman, 2015) is about 4 times bigger than the number of segments we obtained for the SenseCam dataset and about 2 times bigger than for the Narrative datasets. Indeed, we achieve a higher FM score with respect to the method of Lee and Grauman (Lee and Grauman, 2015), since it produces a considerable over-segmentation.

2.4.4 Discussion

The experimental results detailed in section 2.4.3 have shown the advantages of using semantic features for the temporal segmentation of egocentric photo-streams. Despite the common agreement about the inability of low-level features in provid-

ing understanding of the semantic structure present in complex events (Habibian and Snoek, 2014), and the need of semantic indexing and browsing systems, the use of high level features in the context of egocentric temporal segmentation and summarization has been very limited. This is mainly due to the difficulty of dealing with the huge variability of object appearance and illumination conditions in egocentric images. In the works of Doherty et al. (Doherty and Smeaton, 2008) and Lee and Grauman (Lee and Grauman, 2015), temporal segmentation is still based on low level features. In addition to the difficulty of reliably recognizing objects, the temporal segmentation of egocentric photo-streams has to cope with the lack of temporal coherence, which in practice means that motion features cannot reliably be estimated. The work of Castro et al. (Castro et al., 2015) relies on the visual appearance of single images to predict the activity class of an image and on meta-data such as the day of the week and hour of the day to regularize over time. However, due to the huge variability in appearance and timing of daily activities, this approach cannot be easily generalized to different users, implying that for each new user re-training of the model and thus, labelling of thousand of images is required.

The method proposed in this paper offers the advantage of being needless of a cumbersome learning stage and offers a better generalization. The employed concept detector, has been proven to offer a rich vocabulary to describe the environment surrounding the user. This rich characterization is not only useful for better segmentation of sequences into meaningful and distinguishable events, but also serves as a basis for event classification or activity recognition among others. For example, Aghaei et al. (Aghaei et al., 2016a, 2015, 2016b) employed the temporal segmentation method in (Talavera et al., 2015) to extract and select segments with trackable people to be processed. However, incorporating the semantic temporal segmentation proposed in this paper, would allow, for example, to classify events into social or non-social events. Moreover, using additional existing semantic features in a scene may be used to differentiate between different types of a social event ranging from an official meeting (including semantics such as laptop, paper, pen, etc.) to a friendly coffee break (coffee cup, cookies, etc.). Moreover, the semantic temporal segmentation proposed in this paper is useful for indexing and browsing.

2.5 Conclusions and future work

This paper proposed an unsupervised approach for the temporal segmentation of egocentric photo-streams that is able to partition a day's lifelog in segments sharing semantic attributes, hence providing a basis for semantic indexing and event recognition. The proposed approach first detects concepts for each image separately by

employing a CNN approach and later, clusters the detected concepts in a semantic space, hence defining the vocabulary of concepts of a day. Semantic features are combined with global image features capturing more generic contextual information to increase their discriminative power. By relying on these semantic features, a GC technique is used to integrate a statistical bound produced by the concept drift method, ADWIN and the AC, two methods with complementary properties for temporal segmentation. We evaluated the performance of the proposed approach on different segmentation techniques and on 17 day sets acquired by three different wearable devices, and we showed the improvement of the proposed method with respect to the state-of-the-art. Additionally, we introduced two consistency measures to validate the consistency of the ground truth. Furthermore, we made publicly available our dataset EDUB-Seg, together with the ground truth annotation and the code. We demonstrated that the use of semantic information on egocentric data is crucial for the development of a high-performance method.

Further research will be devoted to exploiting the semantic information that characterizes the segments for event recognition, where social events are of special interest. Additionally, we are interested in using semantic attributes to describe the camera wearer context. Hence, opening new opportunities for the development of systems that can take benefit from contextual awareness, including systems for stress monitoring and daily routine analysis.

M. Dimiccoli, M. Bolaños, **E. Talavera**, M. Aghaei, G. Stavri, P. Radeva, "SR-Clustering: Semantic Regularized Clustering for Egocentric Photo-Streams Segmentation", **Author Contributions:** Conceptualisation, P.R. and M.D. and M.B. and E.T. ; implementation, M.B. and E.T. ; writing - original draft preparation, E.T. and M.B. and M.A. and M.D. ; writing - review and editing, E.T. and M.B. and G.S. and M.A. and M.D. and P.R. ; supervision, P.R.

Most of this chapter is from:

E. Talavera, C. Wuerich, N. Petkov, P. Radeva, "Topic Modelling for Routine Discovery from Egocentric Photo-streams," (Submitted), 2019.

Section 3.3 is taken from:

E. Talavera, N. Petkov, P. Radeva, "Unsupervised routine discovery in egocentric photo-streams", 18th International conference on Computer Analysis of Images and Patterns (CAIP), published in the proceedings of the conference, Springer LNCS series, 2019.

Chapter 3

Routine Discovery from Egocentric Images

Abstract

Developing tools to understand and visualize lifestyle is of high interest when addressing the improvement of habits and well-being of people. Routine, defined as the usual things that a person does daily, helps describe the individuals' lifestyle. With these works, we are the first ones to address the development of novel tools for automatic discovery of routine days of an individual from his/her egocentric images. In the proposed model, sequences of images are firstly characterized by semantic labels detected by pre-trained CNNs. Then, these features are organized in temporal-semantic documents to later be embedded into a topic models space. Finally, Dynamic-Time-Warping and Spectral-Clustering methods are used for final day routine/non-routine discrimination. Moreover, we introduce a new EgoRoutine-dataset, a collection of 104 egocentric days with more than 100.000 images recorded by 7 users. Results show that routine can be discovered and behavioural patterns can be observed.

3.1 Introduction

The characterization of people’s life has become an active area of research with the increasing availability of wearable sensors (Doherty et al., 2013). Lifelogging is the process of collecting data about the life of people; this data can describe their activities, emotions and interactions along the day. It offers a rich source of information that allows understanding of the lifestyle of a person. More specifically, by using wearable cameras, images can be automatically collected from a first-person, a.k.a. egocentric point of view of the camera wearer’s. Egocentric images are a valuable source of information in many domains due to the similarity to human perception and memory. However, egocentric collections use to be large (of order of thousands of pictures per day), which makes difficult its analysis. In this work, we rely on long temporal resolution (2fpm) egocentric images for the discovery and study of Routine-related days of people since they allow to monitor and visualize most of their day. The discovery of *Routine* and *Non-Routine* days from these photostreams is an important step for several applications, such as: self-awareness, how does my daily life look like?; monitoring patients or health-care and assistance of elderly people, it is essential to know the person’s common behaviour and *Routine*; or, for memory enhancement and rehabilitation, which benefits from structuring the photo-stream into *Routine* and *Non-Routine* to easily find important events used in memory reminiscence therapy and interventions (Oliveira-Barra et al., 2017).



Figure 3.1: Example of images recorded by one of the camera wearers.

Routine-related days have common patterns that describe situations of the daily life of the person. However, Routine has no concrete definition, since it varies depending on the lifestyle of the individual under study. Therefore, supervised approaches are not useful due to the need for prior information in the form of annotated data or predefined categories. For the discovery of routine-related days, un-

supervised methods are necessary to enable an analysis of the dataset with minimal prior knowledge. Moreover, we need to apply automatic methods that can extract and group the days of an individual using correlated daily elements. We address the discovery of routine-related days following two different approaches:

- In Section 3.3, we propose a personalized and automatic tool for the discovery of routine related days within recorded photo-streams by a camera wearer. We hypothesize that discovering routine related days can be addressed as a clustering problem where methods such as k -means with, for instance, $k = 2$ could potentially classify the days in terms of the behaviour they represent. However, some days present abnormal behaviour. These days correspond to non-routine related days. Most of the time they are not related to each other, which can be interpreted as outliers within the user’s recorded photo-streams. Experience has shown that it is difficult to describe what non-routine related days are for a given photo-stream collection. In the context of outlier detection, samples considered as outliers do not form the cluster with higher density when representing the days in a feature space. We propose an unsupervised classification method that assumes that outliers are situated in low-density areas. Outlier detection methods are commonly used in data mining to indicate variability in measurements, errors or novel samples (Ding and Fei, 2013; Hodge and Austin, 2004). Among their applications are fraud detection (Ghosh and Reilly, 1994) and satellite image analysis (Alvera-Azcárate et al., 2012). However, up to our knowledge for first-time routine detection is defined through an outlier detection approach. Within the available outlier detection algorithms, we propose Isolation Forest algorithm (Liu et al., 2008). This method has shown a good performance when detection outliers in multi-dimensional space, not seeking normal data points but identifying anomalies. Our model is unsupervised because routine differs per person and our aim is to propose a generic model able to discover routine of unknown users. However, since we have the labels of the recorded photo-streams that compose our dataset, we use them to validate if we are able to discover their routine related days.
- In Section 3.4, we apply Topic Modelling (TM) technique to help us detect correlated elements of the individual’s day (e.g. objects that use to appear together in the environment of the wearer). We use TM as an unsupervised approach for the analysis of behavioural habits with the final goal of detecting *Routine* from egocentric images and thus, to describe and understand the daily patterns of conduct of the camera wearer. The analysis of the appearing topics throughout recorded days allows the understanding of the different situations

where the user spends time: working, shopping, walking outside, etc. These elements define the context of the person's lifestyle. Our goal is to address the routine discovery by analyzing the appearance of these patterns in the life of a person. Our goal is to address the routine discovery by analyzing the appearance of these patterns in the life of a person. This pattern give us the opportunity to compare and evaluate days. They also allow us to describe what Routine represents for a person given a collection of his or her days.

In this work, we propose to apply TM to our problem by translating collected egocentric photo-streams into documents. We select this technique because it has demonstrated to be a powerful tool for the discovery of abstract topics appearing in collections of documents. The input images are translated to a Bag-of-Word (BoW) representation, where an image is described by the objects around the wearer, activities of the wearer and the scene the image depicts. Next, the BoW is converted to a new representation of the day in terms of a set of discovered probabilistic topics. Then, the following step is to discover similar days. Routine can present daily small variations thus, the similarity measure use to compare performed activities during the day by the camera wearer should be tolerant to small differences. For instance, having breakfast at 6am and going to work from 7am to 5pm exhibits the same *Routine* as having breakfast at 7am and working from 9am to 7pm. We argue that this allows flexibility in the occurrence of performed activities during the day while temporal order among day elements is maintained. Therefore, in our model, we define similarities among days by evaluating distances between time-slots of a certain duration. To discover similar days we use Dynamic Time Warping for the computation of similarities/distances among the collected photo-streams, allowing that daily habits are tolerant to small differences in starting time and duration.

The contributions of this chapter are the following:

- We address for the first time the problem of routine extraction from egocentric data.
- We propose an unsupervised and automatic model for the analysis of routine days following an anomaly detection approach. This model is based on the aggregation of the descriptors of the images within the photo-stream.
- We introduce an automatic unsupervised pipeline for the identification and characterization of Routine-related days from egocentric photo-streams. This pipeline can be adapted to different characterizations of days. Our model

is based on the topics that describe the day-by-day from egocentric photo-streams for their classification into *Routine* and *Non-Routine* days.

- We present a new egocentric dataset describing the daily life of the camera wearers. It is composed of a total of 100.000 images, from 104 days recorded by 7 different users. We call it *EgoRoutine* and together with its ground-truth are publicly available in <http://www.ub.edu/cvub/dataset/>.

This chapter is organized as follows: in Section 3.2, we highlight relevant work related to the routine discovery. In Section 3.3 and 3.4, we describe the approaches proposed for *Routine*. Within the approach section we also described the proposed dataset, outline the experiments performed and the results obtained, and discuss the achieved results. Finally, in Sections 3.5 and 3.6, we globally discuss our findings and present our conclusions, respectively.

3.2 Related works

In this section, we describe how the routine behaviour of people was studied before the raise of wearable devices and what has been studied since then.

3.2.1 Routine from manual annotation

The manual annotation of daily habits tend to be common practise for its later analysis by either the own person (Andersen et al., 2004) or physicians (Wood et al., 2002). In (Andersen et al., 2004), manually recorded information about the ability of someone performing ADL was examined to classify the patients' dependence, as either dependent or independent. Also, in (Wood et al., 2002) the authors studied diaries from 70 undergraduate students, who rated the assiduity of activity during the previous month through a questionnaire.

3.2.2 Routine from sensors

With the increasing availability of wearable sensors, the aim for automatic data collecting and understanding the behaviour of people have become active areas of research. These sensors allow the automatic collection of big amount of data describing the life of the person who uses them. One of the first works on analyzing regularities in human behaviour from a large scale dataset in an unsupervised manner was presented in (Eagle and Pentland, 2006). The model relied on information

from mobile phones, such as locations, Bluetooth device proximity, application usage, and phone status. Other works relied on data collected by sensors placed in smart homes, such as the one in (Li et al., 2015).

One of the seminal works on routine discovery was presented in (Seiter et al., 2015) that applied a Latent Dirichlet Allocation (LDA) model for detecting activities and a subsequent assessment of the similarity of a person's days. There, topic modelling was employed to discover daily life activities related to rehabilitation patients from wearable sensors. Specific activity groups were applied to define the user's routine. The main 6 categories are eating/leisure (social interactions, eating, playing games), cognitive training (using pc, puzzles), medical fitness, kitchen work (household activities), motor training, and rest. In (Farrahi and Gatica, 2011), the authors focused on *Routine* discovery by analyzing the localization patterns in a phone location dataset collected by 97 people over one year. Their proposed model is based on LDA and word analyses that are built based on location sequences. Sequences of words are defined by translating the pre-defined locations 'home', 'work', 'others' and 'no reception' to H, W, O, and N, respectively. Combining a fine-grain (30 minutes) and coarse-grain (several hours) consideration, they construct a bag representation of location sequences. Every location sequence consists of three consecutive location labels for the fine-grain intervals, followed by a number indicating the coarse-grain time-slot. This approach identifies *Routines* which dominate the entire group's behaviour such as 'going to work late' or 'working non-stop'. Furthermore, they characterize or classify individuals by those *Routines*. From another perspective, in (Biagioni and Krumm, 2013), the behaviour information comes from phone GPS location and is used to assess the similarity of a person's day. The authors applied a modified version of Dynamic Time Warping (DTW) (Keogh and Pazzani, 2001) method to sequences of GPS points sampled at an interval of 10 seconds. Thereafter, a spectral clustering algorithm is employed to cluster similar days and find anomalous behaviours. The authors in (Yürüten et al., 2014) proposed a model for the discovery of clusters of daily activity routines based on accelerometer data, which describes the expenditure data and steps. The model applies a low rank and sparse decomposition of the data signal to later isolate routine and deviations as two different sets of clusters. DTW and hierarchical clustering are used for the computation of pairwise distances and final classification, respectively.

3.2.3 Routine from conventional images

In (Xu and Damen, 2018), the authors addressed the problem of recognition of routine changes from short-term video sequences. Short-term refers to shortly defined time-slots while long-term tends to define the continuous collection throughout the

day. The dataset was recorded by a static camera at the entrance of a kitchen and for periods of time in 6 consecutive days, in 3 different years. In their approach, they first proposed to define a model per year. This model represents the structure of the sequential activities performed by the individual during that week and makes use of Dynamic Bayesian Network to estimate the similarity among sliding windows of the collected video sequences against the evaluated model. By evaluating the differences between each time frame and the model, their algorithm detects the changes between years in the performed activities when the person is in the kitchen. Although the excellent results of this work, this method is applied on strongly controlled environments under the field of view of the static camera and so are not applicable to detect routine days of individuals.

3.2.4 Routine from egocentric images

The availability of wearable cameras allows to collect large amount of egocentric photo-streams, showing a first-view perspective of the performed activities by the camera wearer. Since the egocentric vision field emerged, several works have addressed the analysis of such collections of data from different perspectives: activity recognition (Furnari et al., 2017, 2015, 2016), social interactions characterization (Aghaei et al., 2017; Alletto et al., 2015; Talavera et al., 2018), food-scenes classification (Sarker et al., 2018; Talavera et al., 2014), photo-stream segmentation (Dimiccoli et al., 2017) and summarization (Bolaños et al., 2015), and sentiment analysis (Talavera, Strisciuglio, Petkov and Radeva, 2017). Especially difficult is the problem of analysis of long-term egocentric photo-streams (e.g. activity recognition), as they are recorded with a lower frame rate (2 fpm) and therefore provide sparser contextual information. Other related works mainly focus on the analysis of ADL. For instance, the works presented in (Ermes et al., 2008) and (Furnari et al., 2016) analyze egocentric images, focusing on recognizing the activities the camera wearer was performing. These studies do not go deeper into the analysis of how regularly the recognized activities or environment appear in the recorded photo-streams. Such pattern of appearance is what we believe will allow us to discover *Routine-related* days.

Whereas most of the long-term *Routine* analysis approaches rely on mobile phone locations or sensor data, our approach models patterns of behaviour based on visual data from egocentric images. This source of data allows us to understand the surrounding world and to give a visual explanation to our findings. In contrast with the mentioned above, this chapter goes some steps further by automatically discovering routines as well as visualizing and describing behavioural patterns of the camera wearer from his or her collected photo-streams.

3.3 Unsupervised routine discovery following an outlier detection approach

In this section, we propose an innovative and unsupervised routine discovery method. Its application scheme is given in Fig. 3.2.

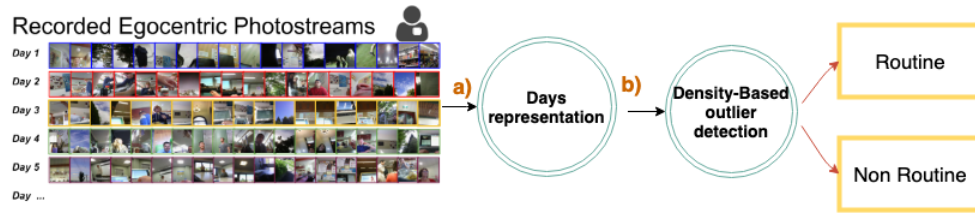


Figure 3.2: The pipeline of the proposed model. Given a set of recorded days, a) they are translated to a set of global or semantic features. Later, b) days are considered as routine or non-routine based on their resemblance.

Our proposed method is based on an outlier detection algorithm. For outlier detection models, an outlier sample is known as a sample outside the ‘boundary’ of the known classes. In our case, these samples relate to non-routine related days. Hence, we assume that routine related days define a class, of which the samples are close to each other within the feature space. The proposed model indicates routine of the person by detecting the sample days that can be clustered together. In the following subsections, we describe the steps in the proposed pipeline as shown in Fig. 3.2.

a) From days to feature vectors

As mentioned above, a day is described by a collection of images and takes the form of photo-stream. We address the day classification by translating the recorded photo-streams into feature vectors for their later analysis and comparison.

Based on the high accuracy recently achieved for the classification of daily activities in egocentric images in (Cartas et al., 2017), we use their proposed network for the characterization of the recorded days. Given an image, this network classifies it into 21 Activities of Daily Living. A day of the user is represented by $Day = \frac{\sum_i^N image_i}{N}$, where N is the number of images within a day, and $image$ represents the feature vector of the recorded images.

We consider the following descriptors obtained from the collected photo-streams:

1. Activity occurrence within the day: We consider the occurrence of activities throughout the day for the characterization of routine, i.e. bag-of-activities.

This feature vector gives an overview of the activities the user performs in a day. However, it does not include temporal information.

2. Global descriptors: We use the ResNet CNN model (He et al., 2016) to extract global descriptors from the images. We use the activation over the entire image given by the last fully connected layer. Given an image, we obtain a 2048 features vector.
3. We concatenate the mentioned features in 1) and 2).

b) Routine related days recognition

More specifically, we rely on the unsupervised outlier detection *Isolation Forest* (Liu et al., 2008) algorithm, and use its available implementation in Scikit-learn (Pedregosa et al., 2011). It is a tree ensemble method that analyses the density of the space to ‘isolate’ outliers. The algorithm works as follows:

First, it randomly selects a feature. Then, for the selected feature, it randomly selects a split value between its maximum and minimum value. By recursive partitioning, it can be represented by a tree structure. As the number of trees increases, the algorithm reaches the convergence. The length of the path from the root to the end node can be considered as the number of splittings needed to isolate a sample. By randomly partitioning the data, the paths for anomalies become shorter. Therefore, samples with shorter path lengths are likely to be anomalies. Later, the anomaly score is calculated per sample based on the averaged and normalized distance of the path. Finally, samples considered as outliers have an anomaly score of 1, while samples with values close to 0 are considered as regular.

The *Isolation Forest* algorithm, given a set of n samples and an observation x , computes the anomaly score $s(x)$ as follows:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}, \quad (3.1)$$

where $h(x)$ is the path length of a point (x) measured by the number of edges that the point traverses from the root node until the last external node. $E(h(x))$ corresponds to the average of $h(x)$ from a collection of isolation trees. $c(n)$ is the average path length, and it is defined as follows:

$$c(n) = 2H(n-1) - (2(n-1)/n), \quad (3.2)$$

where $H(i)$ is the harmonic number and it can be estimated by $\ln(i) + 0.5772156649$ (Euler’s constant).

| User ID | #1 | #2 | #3 | #4 | #5 | Total |
|----------------|-----|----|-----|-----|-----|-------|
| Num Days | 14 | 10 | 16 | 19 | 13 | 72 |
| Images per day | 20k | 8k | 21k | 13k | 11k | 73k |

Table 3.1: Description of the collected Egoroutine dataset by 5 users.

To *summarize*, given a collection of photo-streams recorded by a camera wearer, our proposed personalized and automatic tool will detect the non-routine related days by computing the density within the feature space. The proposed *Isolation Forest* algorithm considers as routine related days if their samples are in a dense region of samples. In contrast, samples that represent non-routine related days correspond to points in a low-density area. This will have as an output the distinction among days, giving insight into the daily habits and lifestyle of the person.

3.3.1 Experiments

In this section, we describe the experimental setup, the metrics used to evaluate the analysis, and the obtained results.

Dataset

We collected data from 5 different subjects who were asked to record their daily life during at least a week. To this end, the users worn the *Narrative Clip* camera¹ fixed to their chest, with a resolution of 2 fpm. The introduced dataset consists of 100k images, from a total of 72 recorded days, see Table 3.1. They captured information about their daily routine, such as the people with whom they interacted, the activities they performed or how often they walked outside. Since there is no training involved in this approach, the whole dataset is analysed by our proposed model. Moreover, in order to show the variance among collected days, Fig. 3.3 shows the average number of images per day. We can observe how the amount of images differs per day and user.

Process of creating the Ground-truth

The annotators got the following definition of “*Life routine*; a sequence of actions which are followed regularly, or at specific intervals of time, daily or weekly”. Next, they were shown mosaics of images representing days of the user. They were asked to first have a look at all the mosaics to get an impression of how routine looks like for that specific user. Later, they gave a binary label: routine or non-routine related.

In Table 3.2, we present the summary of the labels given by the different annotators. From the labelling results, we can deduce that defining what is routine and

¹<http://getnarrative.com/>

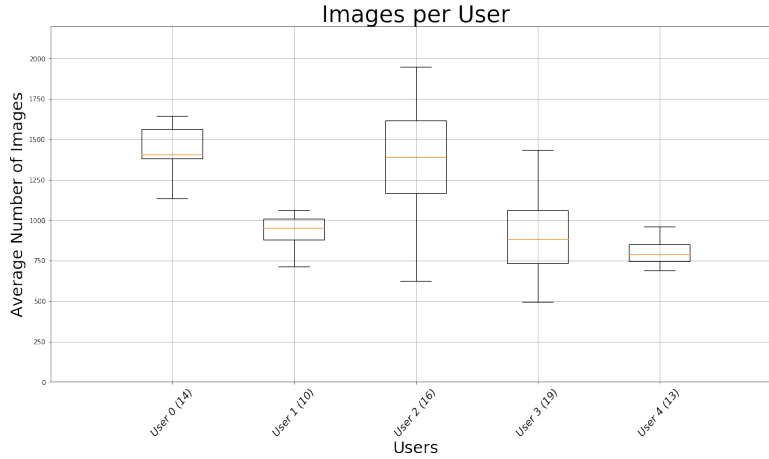


Figure 3.3: Average number of images per recorded egocentric photo-stream. We give the number of collected days per user between parenthesis.

| Class | Six Agree | Five Agree | At Least Four Agree | At Least Three Agree | Total |
|-------------|-----------|------------|---------------------|----------------------|-------|
| All | 34 | 21 | 11 | 6 | 72 |
| routine | 28 | 16 | 7 | 0 | 51 |
| non-routine | 6 | 5 | 4 | 6 | 21 |

Table 3.2: Summary of the labelling results for the Egoroutine dataset.

non-routine is not an easy task. Routine can be easily described in general terms, but it becomes challenging when sequences of images describing a long time period are classified. We can observe how in the majority of cases, the annotators agreed when it comes to label days as routine. However, the non-routine related days are more difficult to perceive leading to disagreement among the annotators. Finally, we have considered as routine related days when >4 of the labels agreed. In case of a draw, the day is labelled as non-routine related. Therefore, from a total of 72 recorded days, 51 days are routine related, and 21 are non-routine related. If we extrapolate to a common life scenario, 72 days correspond to almost 15 recorded weeks. If the users followed what could be considered common routine (a week has 5 working days and 2 weekend days or holiday), in 10 weeks we have 20 weekend days and 50 working days.

Validation

We evaluate the performance of the proposed model and compare it with the baseline models by computing the *Accuracy* (Acc), *Recall* (R), *Precision* (P), and *F-Score* metrics, where: $F - Score = 2 \cdot \left(\frac{P \cdot R}{P + R} \right)$. *Precision* computes the ratio between True Positive (TP) samples and False Positive (FP) samples following: $TP / (TP + FP)$. *Recall* evaluates the ratio of TP and False Negative (FN), showing the ability of the model to find the positive samples, the formula is $TP / (TP + FN)$. Due to the unbalanced dataset we calculate and compare their ‘macro’ and ‘weighted’ mean. The ‘weighted’ mean evaluates the true classification per label, while ‘macro’ calculates the unweighted mean per label. The weighted measures provide the strength of the classifier when applied to unbalanced data.

Experimental setup

To the best of our knowledge, no previous works have addressed the recognition of routine discovery from egocentric photo-streams. Therefore, we evaluate the performance of the proposed model and compare it with what we introduced as baseline methods. We select several outlier detection algorithms namely: Robust Covariance, and One-class SVM. Moreover, we propose to apply unsupervised clustering techniques that allow the identification of outliers or isolation of samples outside the high-density space. These methods allow the recognition of non-similar samples or with non-convex boundaries within the sample collection. Specifically, we evaluate the performance of DBSCAN and Spectral clustering.

Here we give a brief explanation of how these baseline methods work:

- Robust Covariance (Rousseeuw and Driessen, 1999), also called elliptic envelope, assumes that the data follow Gaussian distribution and learns an ellipse. Its drawback is that it degrades when the data is not uni-modal.
- One-class SVM (Platt et al., 1999) is an unsupervised algorithm that estimates the support of the dimensional distribution.
- DBSCAN (Ester et al., 1996), short for Density-Based Spatial Clustering of Applications with Noise, finds samples with high density and defines them as the centre of a cluster. From the center, it expands the cluster. Its *eps* parameter determines the maximum distance between samples to be considered as in the same cluster. Outliers are samples that lie alone in low-density regions.
- Spectral Clustering (Yu and Shi, 2003) works on the similarity graph between samples. It computes the first k eigenvectors of its Laplacian matrix and defines a feature vector per sample. Later, k -Means is applied to these feature

| Methods | Feature Vector | All Users | | | | | | |
|--------------------|---------------------------|-------------|-------------|------|------|-------------|------|------|
| | | Acc | Weighted | | | Macro | | |
| | | | F-Score | P | R | F-Score | P | R |
| Robust covariance | Activity Occurrence (Act) | 0.61 | 0.49 | 0.50 | 0.50 | 0.59 | 0.59 | 0.61 |
| | Global Features (Glo) | 0.71 | 0.60 | 0.63 | 0.60 | 0.69 | 0.70 | 0.71 |
| | Act - Glo | 0.54 | 0.39 | 0.39 | 0.41 | 0.52 | 0.51 | 0.54 |
| One-Class SVM | Activity Occurrence (Act) | 0.72 | 0.65 | 0.69 | 0.65 | 0.70 | 0.70 | 0.72 |
| | Global Features (Glo) | 0.67 | 0.56 | 0.60 | 0.57 | 0.64 | 0.67 | 0.67 |
| | Act - Glo | 0.65 | 0.58 | 0.59 | 0.58 | 0.64 | 0.64 | 0.65 |
| DBSCAN | Activity Occurrence (Act) | 0.61 | 0.51 | 0.55 | 0.55 | 0.57 | 0.60 | 0.61 |
| | Global Features (Glo) | 0.69 | 0.41 | 0.34 | 0.50 | 0.56 | 0.48 | 0.69 |
| | Act - Glo | 0.63 | 0.56 | 0.57 | 0.60 | 0.60 | 0.62 | 0.63 |
| SpectralClustering | Activity Occurrence (Act) | 0.66 | 0.48 | 0.50 | 0.51 | 0.61 | 0.61 | 0.66 |
| | Global Features (Glo) | 0.66 | 0.55 | 0.64 | 0.62 | 0.63 | 0.72 | 0.66 |
| | Act - Glo | 0.62 | 0.46 | 0.50 | 0.50 | 0.57 | 0.61 | 0.62 |
| Isolation Forest | Activity Occurrence (Act) | 0.69 | 0.61 | 0.62 | 0.62 | 0.68 | 0.67 | 0.69 |
| | Global Features (Glo) | 0.76 | 0.68 | 0.71 | 0.68 | 0.74 | 0.75 | 0.76 |
| | Act - Glo | 0.76 | 0.68 | 0.71 | 0.68 | 0.74 | 0.75 | 0.76 |

Table 3.3: Performance of the different methods implemented for the discovery of routine and non-routine days.

vectors to separate them into k classes. In our case, we set $k = 2$, so we evaluate its performance when addressing routine vs non-routine classification.

For the last two proposed unsupervised model, DBSCAN and Spectral Clustering, the closeness among the recorded days is computed based on their shared similarities, following an all-vs-all strategy. To do so, we use the well-known Euclidean metric. The computed similarity matrix is fed to the unsupervised classifier algorithm for the detection of outliers within the set samples. The outlier detection methods are fed with the feature matrix describing the samples.

Results

We present the obtained classification accuracy at day level for the performed experiments in Table 3.3. The proposed model, based on the Isolation Forest algorithm and with global features as descriptors of the recorded days, achieved the best performance with respect to the rest of the tested baseline methods. Our model achieves an average of 76% Accuracy and 68% Weighted F-Score for all the users, outperforming the rest of the tested methods. The highest performance is when analysing global features, which cover most of the possible present activities.

Moreover, in Fig. 3.5 we visualize the days as points in the feature space drawn by the first two principal components of the dataset. We can see the Ground-truth indicated with the boundaries of the circles and the prediction of the model, for both cases red corresponds to routine related days and blue to non-routine related. As it can be observed, our model is the one that obtains the best results.

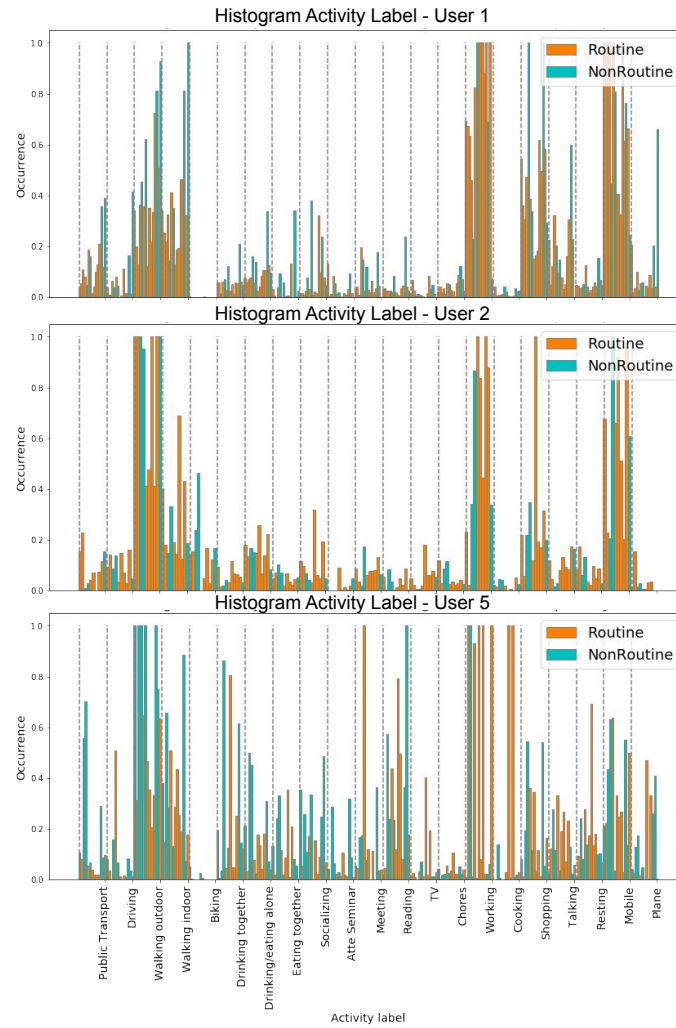


Figure 3.4: Histograms showing the occurrence of activities throughout the days of 3 of the 5 users that worn the camera. As we can appreciate, some activities are more related to non-routine related days, while ‘working’ and ‘walking indoor’ characterizes routine related days.

In Fig. 3.4 we can observe the occurrence of activities per day in the form of a histogram. This representation allows us to better infer and understand how routine (orange) and non-routine (blue) related days vary for the different camera wearers. From this representation we can confirm our initial assumptions: i) the set of ac-

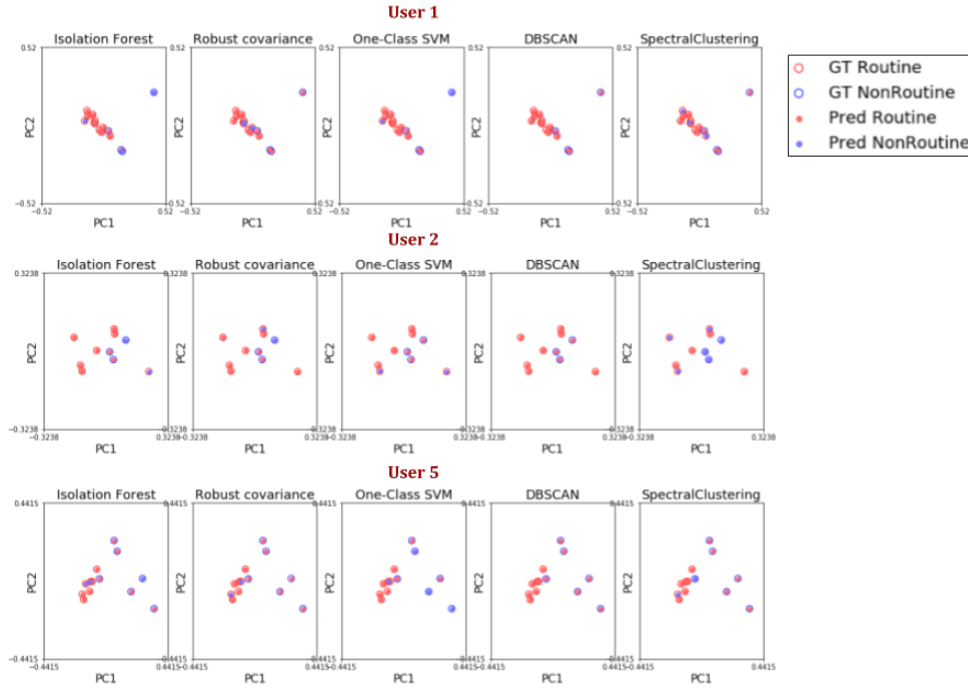


Figure 3.5: Visualization of the obtained classification results based on the analysis of the histogram of activities occurring throughout the day for User1, User2 and User5. We show the classification per user and per studied method. Each dot in the graph corresponds to one day recorded by the user. Each of the 4 subplots shows the classification into routine or non-routine by the baseline methods. The colour of the boundaries of the dots represents the given Ground-truth and the filling the classification label; Red routine and Blue non-routine.

tivities performed as routine and non-routine related days differs per person, ii) a subset of activities is commonly shared when it comes to routine, such as ‘working’, which is mostly described by a laptop/pc as central object in the scene, or ‘using mobile’. In contrast, some activities are specific per user: The routine of *User 5* is characterized by ‘cooking’, ‘reading’, and ‘meeting’. In contrast, for *User 2* ‘walking outdoor’, ‘shopping’, and ‘mobile’ are the more representative activities.

Limitations: The presented analysis can be improved in several directions as by augmenting the number of subjects and the amount of collected data. We believe this is a good starting point for this new field of unsupervised routine analysis from a first-person perspective. Moreover, and even though in this work we consider that there exists one routine per person, future lines will address the discovery of several routines. However, for that, it is needed a bigger amount of data.

3.4 Unsupervised routine discovery relying on topic models

In this section, we describe our proposed model for the characterization of egocentric photo-streams for their later classification into *Routine* and *Non-Routine* related days. Fig. 3.6 illustrates the main steps that our model follows given a set of collected long-term temporal resolution photo-streams. Below, we describe in detail how they are implemented.

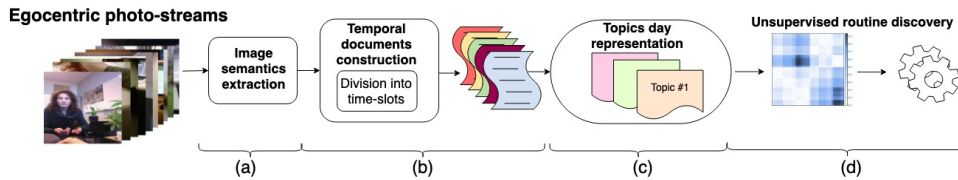


Figure 3.6: Illustration of the proposed pipeline for the discovery of routine from sets of egocentric photo-streams collected by a user. The model proceeds as follows: (a) image semantics extraction, (b) temporal documents construction, (c) topics day representation, and finally, (d) unsupervised routine discovery.

a) *Image semantics extraction*

Describing sequences of photo-streams is not a trivial task due to the unknown visual content. In this work, we propose to describe our daily recorded images through detected concepts by an already pre-trained CNN. For a broad analysis of the scene depicted on a given image, we make use of CNNs pre-trained for the recognition of objects (Chollet, 2017; Redmon, 2018), places (Zhou et al., 2017), and activities (Cartas et al., 2017).

Let us consider that for each image I the CNNs return, L_r , labels related to a total of R concepts found in the images; objects, scene, and activities of the wearer. Thus, each image is represented by a Bag-of-Words composed of these detected semantic concepts (CNN labels).

b) *Temporal documents construction*

To model the patterns of behaviour of the camera wearer, we embed the detected semantic labels extracted from the egocentric images into a temporal document. The detected concepts by the CNNs represent the words that describe the day i.e. that form the document.

In order to maintain the temporal information about the appearance of the extracted semantics, we define J time intervals within the day (e.g. from 7-9h,

9-11h, etc.). For each time-interval we estimate the frequency of appearing of each concept ($L_r, r = 1 \dots R$). For the time-intervals in which no images are taken, we create a dummy variable. Hence, each day is represented by a vector of $J \times R$ dimension.

Given a set I_u of egocentric photo-streams (days) for user u , a matrix $M_{i,j}$ is constructed where each of its elements (i,j) corresponds to day $i = 1, \dots, |I_u|$, and $j = 1, \dots, J \times R$. This temporal document is composed of the concepts detected in the images recorded at a specific range of time. Thus, the proposed model translates a recorded day that is composed of a sequence of egocentric images, to a temporal document represented by the matrix $M_{i,j}$ defined in terms of the frequency of the detected concepts (words) in the photo-stream.

c) *Topics day representation*

Topic modelling allows the transformation of the dataset by factorisation of a set D of documents. A document is composed of a vector of words frequencies, and at the same time, it is assumed that it defines a certain number, K , of topics. In this work, we rely on Latent Dirichlet Allocation (LDA)(Blei et al., 2003), a topic modelling approach that is a generative probabilistic model applied to explain multinomial observations using unsupervised learning. The LDA method follows a generative process described as follows (Blei et al., 2003):

- (a) Choose $\theta_i \sim \text{Dirichlet}(\alpha)$, where $i \in \{1, \dots, D\}$.
- (b) For each of the N_i words w_{ij} in document i :
 - i. choose a topic $z_{ij} \sim \text{Multinomial}(\theta_i)$
 - ii. choose a word w_{ij} from $P(w_{ij}|z_{ij}, \beta) \sim \text{Multinomial}$ probability on the topic z_{ij} .

where the parameters of the multinomials for topics in a document θ_i and words in a topic z_{ij} have Dirichlet priors, $\text{Dir}(\alpha)$ and $\text{Dir}(\beta)$ respectively. The probability of a corpus with D documents is defined as follows:

$$P(D|\alpha, \beta) = \prod_{i=1}^{|D|} \int P(\theta_i|\alpha) \left(\prod_{j=1}^{N_i} \sum_{z_{ij}} P(w_{ij}|z_{ij}, \beta) P(z_{ij}|\theta_i) \right) d\theta_i$$

where the parameters α and β are sampled only once in the process of generating the corpus, while the variables θ_i are sampled once per document. Lastly, the variables z_{ij} and w_{ij} are word-level variables which are sampled once per word j in each document i .

As a result, given a corpus (set) of D documents and K topics to be discovered, LDA gives (Blei et al., 2003):

- the structure or combination of words that best fits the number of topics, by giving a *topic-word matrix* $P(w_{ij}|z_{ij}, \beta)$ where each element of it defines the probability of assigning word w_{ij} to topic z_{ij} .
- a *document-topic matrix* $P(z_{ij}|\theta_i)$ so that each element of it defines the probability of a topic z_{ij} for given a document θ_i .

In our case, we apply the LDA to decompose the elements $M_{i,j}$ of the temporal documents M corresponding to day i and time-slot j . LDA returns a document-topic matrix $P(z_{ij}|M_{i,j})$ with the probabilities of all K topics associated with each element $M_{i,j}$ and the topic-words matrix $P(w_{ij}|z_{ij})$ that defines the relations between topics and words. This is illustrated in Fig. 3.7 showing a day represented by the most important topics (with the highest probability) and the relations between topics and words.

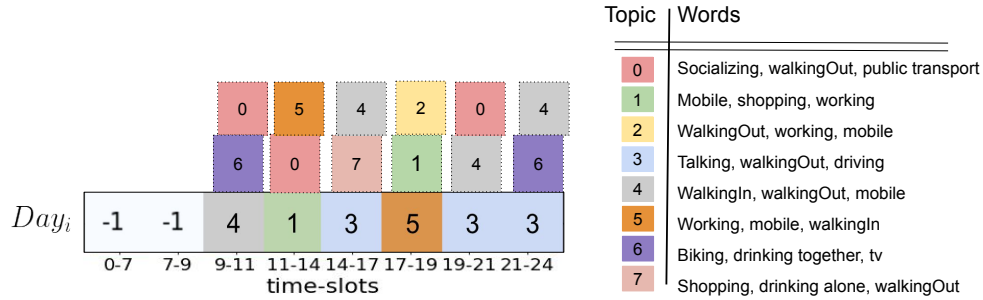


Figure 3.7: Illustration of how a photo-stream/document (Day_i) is described by different proportions of topics throughout the day. We present the winning topic for each time-slot, together with the following $N=2$ topics with the higher representation.

d) Unsupervised routine discovery

Once we have the representation of each day in terms of the most relevant topics with their probabilities, we need to find similarities among days for their later classification as *Routine* or *Non-Routine* days. For example, we expect that days that used to repeat (e.g. defined by topics related to *breakfast*, *metro*, *work*, *lunch*, *work*, *metro*, and *dinner*), appear frequently and thus correspond to a user's routine days.

At this point, a day is represented as a J -dimensional vector, where each element is a K -dimensional vector composed of the probabilities of the detected topics describing it (see Fig. 3.7). In order to find similar days, we need a

metric to compare topics representation. However, it should be tolerant to small temporal differences, since events during the days can begin and last differently. To this purpose, we propose to apply DTW (Keogh and Pazzani, 2001) for computing the similarity of topics representation among days. DTW is an algorithm that computes the optimal alignment between two sequences, where one of them might be stretched or shrunken non-linearly along the time axis. Given two sequences (or vectors) corresponding to two day representations, a warp path (w_1, w_2, \dots, w_Q) is constructed, where Q is the length of the path and every element w_q is a pair $(w_q[1], w_q[2])$ that indicates the mapping of element $w_q[1]$ in the first sequence s' to element $w_q[2]$ in the second one s'' . Further, $w_q[1]$ and $w_q[2]$ have to monotonically increase. The optimal warp path defines the best correspondence of elements of both sequences represented by the path with minimal distance and is computed as follows:

$$dist_{DTW}(s', s'') = \sum_{r=1}^Q dist(s'_{w_q[1]}, s''_{w_q[2]}).$$

In our proposed model, we employ the fastDTW algorithm (Salvador and Chan, 2007), which is an accurate approximation of the DTW method, but has a linear time and space complexity. In contrast to the standard DTW, the fastDTW algorithm shrinks a time series into smaller ones with fewer data points trying to preserve as much information about the original curve as possible. Given two sequences describing two days, the fastDTW algorithm computes the distance among them and gives as output the cost of aligning two days, i.e. their similarity. To compare the topics representation of each time-slot, we apply Euclidean distance.

DTW only gives the distance between pairs of days. Next, we need to discover clusters of similar days. For that purpose, we cannot rely on the days topics representation but on the computed distances among pairs. We apply the *Spectral clustering* algorithm (Yu and Shi, 2003) over the computed affinity matrix of the distances between the days. This method does not make assumptions about the global structure of the data, but bases its decision on local evidence of how likely two elements (days) might belong to the same cluster. From the affinity matrix, the algorithm constructs a weighted graph $G = (Vn, E, We)$, being Vn the set of nodes, E the set of edges and We the weights of the edges. The global optimum is then computed by eigen-decomposition. This clustering method relies on k -Means for the final classification and thus, needs a number kc of clusters to be defined, which without loss of generality, we set to 2 for the discovery of *Routine* and *Non-Routine* related days.

3.4.1 Experimental Framework and Results

In this section, we detail a newly introduced EgoRoutine dataset. Then, we describe the metrics used for the evaluation of the performed experiments. Next, we depict the experimental setup with the proposed baseline approaches. Finally, we analyze the obtained results at different stages of the proposed pipeline.

EgoRoutine - An egocentric dataset for behaviour analysis

In this work, we propose and make publicly available the *EgoRoutine* dataset². This dataset is composed of recorded days by 7 individuals who wore the Narrative Clip camera³ fixed to their chest and were asked to record their daily life. *EgoRoutine* consists of 115.430 images, from a total of 104 recorded days. In Table 3.4 and Fig. 3.8, we indicate the number of days and images collected per user. The camera wearers captured information about their daily *Routine*, taking pictures of the activities they performed and their occurrence as well as the people with whom they interacted.

| User ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total |
|----------------|-------|------|-------|-------|-------|-------|-------|--------|
| Num Days | 14 | 10 | 16 | 20 | 13 | 18 | 13 | 104 |
| Images per day | 20521 | 9583 | 21606 | 19152 | 17046 | 16592 | 10957 | 115430 |

Table 3.4: Total number of recorded days and collected images per user.

GT evaluation: The collected dataset was labelled by 6 annotators who were asked to classify days into *Routine* or *Non-Routine* related. The annotators got the following definition “Life *Routine* is a sequence of actions which are followed regularly, or at specific intervals of time, daily or weekly”. Days were shown to them in the form of a mosaics.

In Fig. 3.9, we present a representation of some of the collected photo-streams of User 1 with their final routine (R) or Non-Routine (NR) labels given on the right. In Table 3.5, we present the summary of the labels given by the different annotators. From the labelling results we can deduce that defining what is *Routine* and *Non-Routine* is not an easy task. *Routine* can be easily verbally described, but it becomes challenging when we want to classify sequences of images describing a long period of time. We observed that in the majority of cases, the annotators agreed when labelling days related to *Routine*. However, the *Non-Routine* related days were more difficult to perceive leading to disagreement among the annotators. For the final distinction, we have considered as *Routine* related days when more than 4 annotators agreed on the label. In case of a draw, the day is labelled as *Non-Routine*

²<http://www.ub.edu/cvub/dataset/>

³<http://getnarrative.com/>

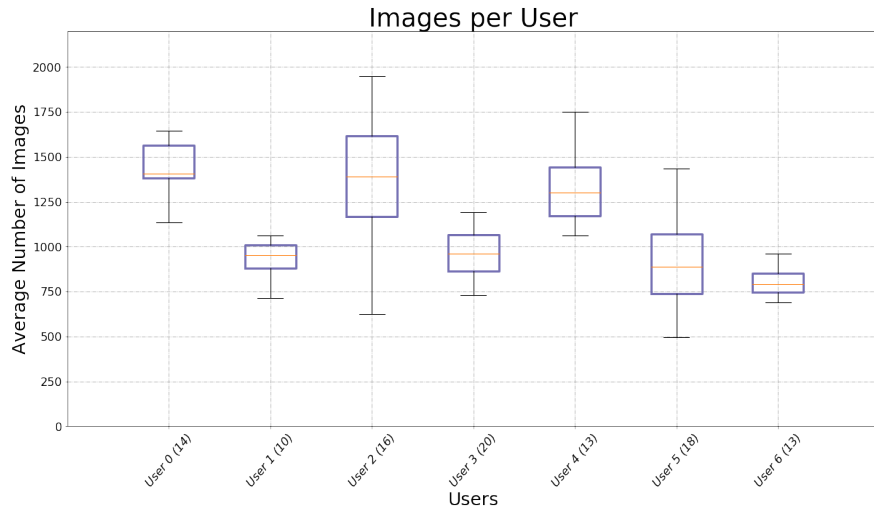


Figure 3.8: Average number and variance of egocentric images per recorded photo-stream for the 7 users. Between parenthesis, we show the number of recorded days per user.



Figure 3.9: Example of selected images throughout some of the recorded photo-streams of User1. On the right, we can see the given ground-truth (R for routine and NR for non-routine) and the predicted binary label by the best combination of parameters (1 for Non-routine and 0 for Routine days).

related. Therefore, from a total of 104 recorded days, 65 days are *Routine* related, and 39 are *Non-Routine* related. In Fig. 3.10 we present the number of labelled days per user into *Routine* and *Non-Routine*. If we extrapolate to a common life scenario, then 104 days correspond to almost 15 recorded weeks. If the users followed what could be considered as common *Routine*, where a week has 5 working days and 2 weekend days, in 15 weeks we have 30 weekend days and 75 working days. This could be an explanation of the resulted labels since it is proportional to the working days reported by the camera wearers.

| Class | Six Agree | Five Agree | At Least Four Agree | At Least Three Agree | Total |
|--------------------|-----------|------------|---------------------|----------------------|-------|
| All | 47 | 29 | 18 | 10 | 104 |
| <i>Routine</i> | 35 | 22 | 8 | 0 | 65 |
| <i>Non-Routine</i> | 13 | 7 | 9 | 10 | 39 |

Table 3.5: Summary of the agreement among the 6 individuals that labelled the collected photo-streams into *Routine* or *Non-Routine* related days.

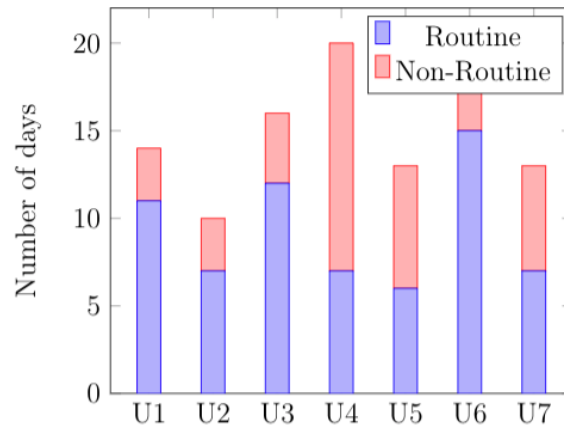


Figure 3.10: Number of *Routine* and *Non-Routine* days for each user (U) in the *EgoRoutine* dataset.

Evaluation

In this section, we describe the metrics that we use to evaluate our proposed model for the discovery of *Routine* and *Non-Routine* related days.

The discovery of routine behaviour is an unsupervised problem with non-trivial evaluation. We evaluate the results in terms of *Accuracy (A)*, *Precision (P)* and *Recall (R)* and F_1 score in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), when classifying days into *Routine* or *Non-Routine*, defined as follows:

$$F_1 = \frac{2P \cdot R}{P + R}, P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.3)$$

Moreover, since the proposed pipeline for the discovery of routine behavioural patterns is composed of several steps, we also present qualitative results of the intermediate steps of our proposal.

Implementation setting

Regarding the concepts detected in the egocentric images, we perform an ablation study using the following different CNNs:

1. *Objects detection*: Detected objects by Yolo (Redmon, 2018) and Xception (Chollet, 2017). These models were trained on the COCO (Lin et al., 2014) and ImageNet dataset (Deng et al., 2009), respectively.
2. *Scene recognition*: We represent an image by the top-1 probability scene label obtained by the VGG16, a pre-trained network previously trained on the Places365 dataset (Zhou et al., 2017).
3. *Activities recognition*: We use the activity labels given by the CNN proposed in (Cartas et al., 2017), which was trained for the recognition of 21 different daily activities. We select the activity label with the highest probability per image.

Concerning DTW, we use the *Euclidean* metric to compute the distance among samples. Finally, with respect to the Spectral clustering, we set k equal to 2 to discover Routine and Non-Routine related days.

Experimental setup

We evaluate the performance of the different steps of our approach:

- **Image semantics extraction** in terms of the detected concepts in the egocentric images by the pre-trained CNNs as descriptors of the egocentric photo-streams.
- **Temporal documents construction** by the conversion of photo-streams concepts to documents. To evaluate the effect of this, we test the following:
 1. *Long duration time-slots*: We define J number of time-slots following the ones proposed in (Farrahi and Gatica, 2011): 0am-7am, 7am-9am, 9am-11am, 11am-2pm, 2pm-5pm, 5pm-7pm, 7pm-9pm, 9pm-12pm.
 2. *Short duration time-slots*: Of one hour each, 00:00-01:00, 01:00-02:00, 02:00-03:00, etc, with a result of 24 time-slots.
- **Topics day representation**, we evaluate the importance and the robustness of the proposal on the number of topics. Moreover, we study the need of individual vs. generic topic models in order to explore if the information about the routine of other users improve the final classification. Given multiple camera users, the LDA model can be computed either using the images of all users (generic) or considering the set of documents collected by each person separately (personalized).
- **Unsupervised routine discovery** of photo-streams. We assess the goodness of the proposed clustering method for the discovery of routine-related days, comparing it to the one achieved when using the *Agglomerative Hierarchical Clustering* (Rokach and Maimon, 2005) for the discrimination among days.

Results and discussions

Next, we present quantitative and qualitative results of the performance on the different stages of our approach for routine discovery validated on our *EgoRoutine* dataset.

- **Image semantics extraction performance**: in terms of the detected concepts: objects, activities and scenes. Within an ablation study we evaluate the performance of the different concept descriptors when they are considered separately or as a combination. In Table 3.6, we depict the performance of the experiments obtained. As it can be observed, the combination of labels of detected objects, activity and places better describes the data leading to the best

| | TimeSlot | Clustering | #Topics | Xception(Chollet, 2017) | | | | Yolo(Redmon, 2018) | | | | Activities(Cartas et al., 2017) | | | | Places(Zhou et al., 2017) | | | | Combination | | | | |
|-------------|----------------------------------|------------|---------|-------------------------|----------------|------|------|--------------------|----------------|------|------|---------------------------------|----------------|------|------|---------------------------|----------------|------|-------------|-------------|----------------|-------------|------|------|
| | | | | Acc | F ₁ | P | R | Acc | F ₁ | P | R | Acc | F ₁ | P | R | Acc | F ₁ | P | R | Acc | F ₁ | P | R | |
| | | | | | | | | | | | | | | | | | | | | | | | | |
| Personalize | Per Hour | SpClus | 2 | 0.72 | 0.68 | 0.70 | 0.71 | 0.71 | 0.68 | 0.73 | 0.75 | 0.72 | 0.70 | 0.72 | 0.73 | 0.68 | 0.65 | 0.69 | 0.70 | 0.72 | 0.69 | 0.70 | 0.72 | 0.72 |
| | | | 4 | 0.75 | 0.73 | 0.74 | 0.77 | 0.72 | 0.71 | 0.74 | 0.77 | 0.72 | 0.69 | 0.70 | 0.71 | 0.78 | 0.76 | 0.77 | 0.81 | 0.75 | 0.72 | 0.74 | 0.75 | |
| | | | 6 | 0.72 | 0.70 | 0.73 | 0.76 | 0.76 | 0.73 | 0.74 | 0.76 | 0.76 | 0.73 | 0.75 | 0.77 | 0.74 | 0.72 | 0.75 | 0.78 | 0.76 | 0.72 | 0.74 | 0.76 | |
| | | | 8 | 0.78 | 0.75 | 0.76 | 0.79 | 0.76 | 0.73 | 0.75 | 0.78 | 0.77 | 0.75 | 0.78 | 0.81 | 0.71 | 0.70 | 0.75 | 0.76 | 0.77 | 0.73 | 0.76 | 0.80 | |
| | | | 10 | 0.73 | 0.72 | 0.75 | 0.78 | 0.73 | 0.70 | 0.72 | 0.74 | 0.69 | 0.66 | 0.69 | 0.71 | 0.72 | 0.69 | 0.72 | 0.74 | 0.74 | 0.71 | 0.74 | 0.75 | |
| | | HierClus | 2 | 0.68 | 0.64 | 0.71 | 0.71 | 0.66 | 0.64 | 0.73 | 0.74 | 0.71 | 0.69 | 0.74 | 0.76 | 0.71 | 0.69 | 0.73 | 0.74 | 0.71 | 0.68 | 0.76 | 0.74 | |
| | 4 | 0.75 | 0.72 | 0.77 | 0.77 | 0.76 | 0.74 | 0.76 | 0.78 | 0.71 | 0.67 | 0.72 | 0.72 | 0.75 | 0.72 | 0.76 | 0.77 | 0.73 | 0.69 | 0.72 | 0.74 | | | |
| | 6 | 0.66 | 0.60 | 0.66 | 0.67 | 0.76 | 0.73 | 0.77 | 0.79 | 0.71 | 0.65 | 0.71 | 0.69 | 0.75 | 0.71 | 0.78 | 0.75 | 0.70 | 0.68 | 0.71 | 0.74 | | | |
| | 8 | 0.79 | 0.75 | 0.83 | 0.79 | 0.72 | 0.68 | 0.71 | 0.71 | 0.72 | 0.66 | 0.73 | 0.72 | 0.77 | 0.75 | 0.81 | 0.82 | 0.75 | 0.72 | 0.78 | 0.77 | | | |
| | 10 | 0.72 | 0.64 | 0.69 | 0.68 | 0.71 | 0.63 | 0.67 | 0.71 | 0.67 | 0.61 | 0.67 | 0.69 | 0.76 | 0.71 | 0.71 | 0.75 | 0.73 | 0.66 | 0.74 | 0.73 | | | |
| | As in (Farrahi and Gatica, 2011) | SpClus | 2 | 0.69 | 0.66 | 0.69 | 0.71 | 0.66 | 0.63 | 0.67 | 0.68 | 0.68 | 0.66 | 0.71 | 0.72 | 0.68 | 0.67 | 0.70 | 0.72 | 0.69 | 0.68 | 0.71 | 0.72 | |
| | | | 4 | 0.72 | 0.71 | 0.74 | 0.77 | 0.75 | 0.72 | 0.75 | 0.77 | 0.74 | 0.72 | 0.74 | 0.77 | 0.75 | 0.73 | 0.77 | 0.79 | 0.77 | 0.75 | 0.77 | 0.80 | |
| 6 | | | 0.77 | 0.75 | 0.77 | 0.80 | 0.71 | 0.68 | 0.72 | 0.74 | 0.72 | 0.68 | 0.70 | 0.72 | 0.74 | 0.71 | 0.74 | 0.76 | 0.80 | 0.77 | 0.79 | 0.82 | | |
| 8 | | | 0.70 | 0.67 | 0.70 | 0.72 | 0.66 | 0.63 | 0.70 | 0.70 | 0.76 | 0.72 | 0.73 | 0.74 | 0.76 | 0.73 | 0.74 | 0.77 | 0.72 | 0.69 | 0.72 | 0.74 | | |
| 10 | | | 0.76 | 0.73 | 0.74 | 0.76 | 0.70 | 0.66 | 0.72 | 0.72 | 0.75 | 0.73 | 0.74 | 0.76 | 0.77 | 0.75 | 0.77 | 0.80 | 0.77 | 0.75 | 0.76 | 0.79 | | |
| HierClus | | 2 | 0.73 | 0.70 | 0.72 | 0.73 | 0.69 | 0.67 | 0.72 | 0.72 | 0.69 | 0.63 | 0.65 | 0.67 | 0.64 | 0.60 | 0.67 | 0.66 | 0.72 | 0.63 | 0.64 | 0.68 | | |
| 4 | 0.70 | 0.68 | 0.72 | 0.74 | 0.70 | 0.68 | 0.71 | 0.74 | 0.69 | 0.68 | 0.72 | 0.74 | 0.68 | 0.65 | 0.69 | 0.71 | 0.74 | 0.73 | 0.75 | 0.77 | | | | |
| 6 | 0.73 | 0.72 | 0.76 | 0.79 | 0.63 | 0.57 | 0.64 | 0.65 | 0.65 | 0.56 | 0.60 | 0.63 | 0.71 | 0.69 | 0.72 | 0.74 | 0.75 | 0.72 | 0.75 | 0.75 | | | | |
| 8 | 0.66 | 0.62 | 0.70 | 0.69 | 0.67 | 0.62 | 0.68 | 0.69 | 0.71 | 0.66 | 0.69 | 0.70 | 0.71 | 0.66 | 0.70 | 0.71 | 0.75 | 0.70 | 0.71 | 0.73 | | | | |
| 10 | 0.67 | 0.59 | 0.61 | 0.66 | 0.72 | 0.64 | 0.69 | 0.69 | 0.67 | 0.60 | 0.68 | 0.68 | 0.71 | 0.69 | 0.72 | 0.75 | 0.73 | 0.66 | 0.71 | 0.71 | | | | |
| Generic | Per Hour | SpClus | 2 | 0.74 | 0.69 | 0.70 | 0.71 | 0.76 | 0.74 | 0.76 | 0.79 | 0.79 | 0.75 | 0.75 | 0.77 | 0.72 | 0.69 | 0.70 | 0.72 | 0.76 | 0.72 | 0.73 | 0.75 | |
| | | | 4 | 0.74 | 0.70 | 0.73 | 0.75 | 0.78 | 0.74 | 0.75 | 0.78 | 0.77 | 0.75 | 0.78 | 0.80 | 0.74 | 0.72 | 0.75 | 0.78 | 0.77 | 0.74 | 0.76 | 0.77 | |
| | | | 6 | 0.76 | 0.72 | 0.74 | 0.76 | 0.75 | 0.71 | 0.73 | 0.76 | 0.74 | 0.73 | 0.76 | 0.79 | 0.76 | 0.74 | 0.75 | 0.78 | 0.75 | 0.71 | 0.73 | 0.75 | |
| | | | 8 | 0.72 | 0.69 | 0.72 | 0.74 | 0.74 | 0.71 | 0.73 | 0.75 | 0.73 | 0.71 | 0.74 | 0.76 | 0.76 | 0.74 | 0.76 | 0.78 | 0.76 | 0.72 | 0.74 | 0.76 | |
| | | | 10 | 0.76 | 0.72 | 0.74 | 0.76 | 0.75 | 0.72 | 0.74 | 0.76 | 0.73 | 0.71 | 0.72 | 0.75 | 0.75 | 0.73 | 0.76 | 0.79 | 0.74 | 0.71 | 0.74 | 0.75 | |
| | | HierClus | 2 | 0.69 | 0.65 | 0.69 | 0.71 | 0.67 | 0.59 | 0.65 | 0.65 | 0.68 | 0.65 | 0.71 | 0.72 | 0.68 | 0.65 | 0.72 | 0.72 | 0.67 | 0.63 | 0.70 | 0.70 | |
| | 4 | 0.75 | 0.71 | 0.78 | 0.76 | 0.74 | 0.68 | 0.70 | 0.73 | 0.75 | 0.72 | 0.77 | 0.76 | 0.67 | 0.63 | 0.70 | 0.69 | 0.74 | 0.70 | 0.72 | 0.74 | | | |
| | 6 | 0.72 | 0.66 | 0.67 | 0.71 | 0.67 | 0.63 | 0.71 | 0.71 | 0.73 | 0.68 | 0.72 | 0.75 | 0.79 | 0.75 | 0.81 | 0.76 | 0.73 | 0.70 | 0.75 | 0.77 | | | |
| | 8 | 0.67 | 0.63 | 0.77 | 0.72 | 0.69 | 0.65 | 0.75 | 0.73 | 0.73 | 0.64 | 0.65 | 0.70 | 0.75 | 0.70 | 0.76 | 0.74 | 0.76 | 0.73 | 0.75 | 0.78 | | | |
| | 10 | 0.68 | 0.66 | 0.73 | 0.75 | 0.74 | 0.67 | 0.70 | 0.70 | 0.70 | 0.63 | 0.71 | 0.70 | 0.73 | 0.69 | 0.76 | 0.73 | 0.76 | 0.70 | 0.77 | 0.74 | | | |
| | As in (Farrahi and Gatica, 2011) | SpClus | 2 | 0.70 | 0.68 | 0.71 | 0.73 | 0.71 | 0.69 | 0.73 | 0.74 | 0.67 | 0.66 | 0.68 | 0.71 | 0.69 | 0.66 | 0.70 | 0.71 | 0.69 | 0.67 | 0.72 | 0.73 | |
| | | | 4 | 0.69 | 0.66 | 0.70 | 0.72 | 0.71 | 0.68 | 0.73 | 0.74 | 0.70 | 0.67 | 0.68 | 0.70 | 0.73 | 0.71 | 0.75 | 0.77 | 0.78 | 0.76 | 0.78 | 0.81 | |
| 6 | | | 0.75 | 0.72 | 0.74 | 0.77 | 0.73 | 0.71 | 0.73 | 0.76 | 0.69 | 0.65 | 0.67 | 0.68 | 0.74 | 0.70 | 0.72 | 0.73 | 0.78 | 0.76 | 0.77 | 0.80 | | |
| 8 | | | 0.74 | 0.71 | 0.72 | 0.75 | 0.69 | 0.64 | 0.67 | 0.68 | 0.72 | 0.68 | 0.70 | 0.73 | 0.72 | 0.70 | 0.73 | 0.75 | 0.73 | 0.72 | 0.74 | 0.76 | | |
| 10 | | | 0.72 | 0.69 | 0.71 | 0.74 | 0.73 | 0.70 | 0.74 | 0.76 | 0.73 | 0.70 | 0.72 | 0.74 | 0.76 | 0.74 | 0.76 | 0.79 | 0.76 | 0.74 | 0.76 | 0.78 | | |
| HierClus | | 2 | 0.73 | 0.68 | 0.71 | 0.73 | 0.67 | 0.65 | 0.70 | 0.71 | 0.73 | 0.70 | 0.71 | 0.73 | 0.70 | 0.64 | 0.69 | 0.70 | 0.65 | 0.63 | 0.70 | 0.70 | | |
| 4 | 0.68 | 0.65 | 0.68 | 0.70 | 0.66 | 0.64 | 0.71 | 0.71 | 0.64 | 0.58 | 0.62 | 0.63 | 0.60 | 0.54 | 0.64 | 0.63 | 0.64 | 0.59 | 0.65 | 0.67 | | | | |
| 6 | 0.74 | 0.67 | 0.68 | 0.72 | 0.69 | 0.64 | 0.69 | 0.70 | 0.70 | 0.65 | 0.73 | 0.70 | 0.69 | 0.63 | 0.75 | 0.69 | 0.72 | 0.67 | 0.68 | 0.73 | | | | |
| 8 | 0.69 | 0.64 | 0.69 | 0.70 | 0.67 | 0.61 | 0.64 | 0.64 | 0.74 | 0.70 | 0.74 | 0.75 | 0.69 | 0.61 | 0.67 | 0.65 | 0.70 | 0.68 | 0.75 | 0.75 | | | | |
| 10 | 0.75 | 0.68 | 0.73 | 0.73 | 0.72 | 0.66 | 0.70 | 0.72 | 0.71 | 0.67 | 0.70 | 0.70 | 0.75 | 0.71 | 0.77 | 0.75 | 0.67 | 0.61 | 0.67 | 0.69 | | | | |

Table 3.6: Results of the proposed pipeline and baseline models. We report results when evaluating different lengths of the time-slots in which we divide the photo-streams: per hour or the ones introduced in (Farrahi and Gatica, 2011). We also quantify the performance when evaluating 2, 4, 6, 8 and 10 topics. Moreover, we present the obtained results when applying Hierarchical (HierClus) and Spectral Clustering (SpClus). Finally, we show the output of the model when evaluating collected days by the user (Personalized) or by the whole set of user (Generic topics).

results when addressing routine discovery, with $Acc = 80\%$ and $F_1 = 77\%$. This makes sense since a richer description of the image helps to better draw the description of the behaviour of people. Depending on the final goal and application, it could be that independently studying information about activities, objects and/or places helps describe better the routine of people.

In Table 3.8, we show the concepts that are detected by the different evaluated CNNs in a given photo-stream. Overall, the detected places given by the network get close enough to reality and therefore are evaluated. In the case of *activity* recognition, and since the network was trained with egocentric images, the results are more consistent. For the detection of objects, *YOLO* seems more consistent when detecting objects of the daily living. We understand that this

| | User 1 | User 2 | User 3 | User 4 | User 5 | User 6 | User 7 | Avg |
|-------|--------|--------|--------|--------|--------|--------|--------|------|
| Acc | 0.79 | 0.74 | 0.75 | 0.90 | 0.92 | 0.56 | 0.92 | 0.80 |
| F_1 | 0.75 | 0.70 | 0.71 | 0.89 | 0.92 | 0.50 | 0.92 | 0.77 |
| P | 0.75 | 0.75 | 0.70 | 0.89 | 0.93 | 0.56 | 0.94 | 0.79 |
| R | 0.86 | 0.79 | 0.75 | 0.89 | 0.93 | 0.60 | 0.92 | 0.82 |

Table 3.7: Results of the proposed pipeline for the best setting of the parameters: analysing the set of collected photo-streams of User1, seeking for 6 topics to describe the data, with time-slots of long duration, and with spectral clustering as the final classifier.

is due to the fact that the CNN was trained with 80 different categories corresponding to Common Objects in Context (COCO (Lin et al., 2014)). In contrast, *Xception* might be able to recognize uncommon objects since it was trained over a bigger dataset composed of 1000 different categories (the ImageNet (Deng et al., 2009)). We can observe some inconsistencies in the classes given by the network trained over *Places365*, such as finding the ‘airplane cabin’ label early in the morning. We explain it by the fact that the network used was not trained with egocentric pictures. The change of perspective modifies how scenes are understood, and lights in the ceiling of an office or corridor can be miss-interpreted as the lights in the cabin of an airplane.

| | Time-slot (h) | | | | | | | | | | | |
|-----------------------------------|-----------------|-----|-----------------|-----|-----------------|-----|----------------|-----|------------------|-----|----------------|-----|
| | 9-11 | | 11-14 | | 14-17 | | 17-19 | | 19-21 | | 21-24 | |
| Xception (Chollet, 2017) | screen | 29 | desktop pc | 266 | desktop pc | 85 | desktop pc | 83 | radio | 16 | photocopier | 80 |
| | menu | 23 | screen | 265 | desktop pc | 75 | screen | 80 | CD player | 16 | desk | 59 |
| | monitor | 19 | monitor | 254 | screen | 51 | monitor | 74 | slot | 16 | projector | 42 |
| Places (Zhou et al., 2017) | airplane cabin | 90 | airplane cabin | 167 | conference room | 49 | office | 41 | airplane cabin | 31 | reception | 28 |
| | atrium/public | 8 | office | 113 | office | 43 | airplane cabin | 26 | bowling alley | 14 | airplane cabin | 26 |
| | office cubicles | 8 | office cubicles | 42 | reception | 37 | computer room | 23 | airport terminal | 10 | hotel room | 14 |
| Activity (Cartas et al., 2017) | WalkingIn | 50 | Mobile | 227 | Mobile | 60 | Working | 78 | Mobile | 30 | Talking | 50 |
| | Shopping | 40 | Shopping | 94 | Talking | 46 | Mobile | 39 | Driving | 25 | WalkingOut | 37 |
| | WalkingOut | 36 | Working | 75 | meeting | 46 | WalkingOut | 32 | WalkingOut | 16 | Mobile | 27 |
| Yolo (Redmon, 2018) | person | 146 | tvmonitor | 383 | person | 202 | person | 132 | person | 107 | person | 198 |
| | laptop | 38 | cup | 354 | laptop | 112 | tvmonitor | 122 | chair | 32 | chair | 155 |
| | chair | 38 | laptop | 334 | chair | 108 | keyboard | 73 | cell phone | 23 | diningtable | 53 |

Table 3.8: Example of detected concepts in a given recorded day by User 1. This table aims to give an idea of how documents develop throughout the day of the person. Each column represents a time slot of a specific duration. Rows present the top-3 concepts detected by the pre-trained networks referenced in the left of the table. The presented numbers describe the numbers of times a concept was present in that time-slot.

- **Evaluation of the Temporal documents construction:** We study the effect on the discovered topics for the final classification when analyzing time-slots of different duration. Time-slots of longer duration might affect the result by smoothing activities happening during a short time. In contrast, fine-grained time-slots might lead to noise in the final classification. From the results shown in Table 3.6, we can observe that the model better performs when the day is described by analyzing the time division proposed in (Farrahi and Gatica,

2011). We deduce that time-slots with a longer duration smooth the activities performed during short periods of time when comparing days. A fine-grained time-slots with an hour duration might include noise to the description of a day.

- **Evaluation of the topics day representation performance:** Topic models discover abstract topics within given documents. A natural question that may arise is the data used for the discovery of topics: should they be discovered from the set involving all users or they should be extracted for each user individually?. A hypothesis is that if more documents are given (joining all data), more robust topics will be discovered, and thus, better they will be able to describe the behavioural patterns of the camera wearer. Thus, when learning the topic-word distribution following the generic approach, we could take advantage of a bigger dataset. A negative aspect of seeking generalization is that user-specific activities can be missed, since they would become not relevant to be detected. In contrast, we assume that individually learned topics might find more personalized representations of every specific activities of the user, since the places of their daily life, e.g. the office desk or living room of different people, might be described differently. Therefore, we evaluate the performance of the model when obtaining the topics just based on the collected photo-streams by the user under study (personalized approach), or when analyzing all the collected photo-streams that compose the EgoRoutine dataset (generic approach). From the results and for the goal of routine discovery, the personalized approach allows the model to better distinguish Routine-related days with a 80% accuracy and 77% F_1 (see Table 3.6).

The goodness of the model when varying the number of topics is also tested. We present results when discovering 2, 4, 6, 8 and 10 topics. As it can be observed, the performance of the classifier is highest when discovering 6 and addressing the time-division proposed in (Farrahi and Gatica, 2011). However, it could be that for a more detailed analysis of what is happening at a specific time, a higher number of fine-grained time-slots might describe in more detail, in terms of objects, activities and places.

- **Evaluation of the Unsupervised routine discovery performance:** We compare the performance of the proposed Spectral Clustering algorithm with the results obtained by the *Agglomerative Hierarchical Clustering* (Rokach and Maimon, 2005) (HC) when classifying into *Routine* or *Non-Routine* related days. HC method follows a bottom-up approach where each data point starts as a single cluster, and pairs of samples are recursively merged following the path that minimally increases the given linkage distance. The process continues

as samples are clustered moving up in the similarity hierarchy. We select the HC since we need to compare against methods that are able to analyse pre-computed distance matrices.

We can observe in Table 3.6 that the Spectral Clustering classifier leads to more accurate discovery of the Routine-related days, outperforming the classification by the HC. We believe this is due to the ability of the Spectral clustering to adapt to complex shapes of the data in the data space.

For a more detailed understanding of the performance at user level, in Table 3.7 we show results of the best performing model. We can observe that for some of the users the classification into *Routine* and *Non-Routine* related days is rather clear, such as for User 5 or User 7, while for User 6 the classification is close to random. This is due to the difference between the lifestyle of the users. Some of them have a clear distribution of routine (e.g. work) and non-routine (e.g. non-work) related activities, while others recorded days for periods when their activities were not following an established routine pattern.

In Fig. 3.9, we present some collected days of User 1 and the predicted label by the best combination of parameters (personalize analysis of documents, combination of labels as images descriptors, 6 topics, and Spectral clustering). Days predicted as Non-Routine related are assigned label '1' and Routine-related days - label '0'. Day 1 is miss-classified as Non-Routine related. From observing the data, we can guess that this user tends to start working at noon until late in the evening. In contrast, on Day 1, User 1 spent much less hours at work and left the office much earlier. This could be a cause of miss-classification by the model. Non-Routine related days contain events where the user works for short periods and spends longer time interacting with colleagues or friends. Day 7 is an example where User 1 is going for dinner to a restaurant right after working for a short time.

- **Final routine characterization and visualization for behaviour modelling:** The characterization of days based on detected concepts and the later inferred topics have demonstrated to be a rich tool for behaviour visualization. In Fig. 3.11 we present how the found topics could be analysed by the wearer or an expert. As an example for visualization, results are shown following a personalized analysis of the data collected by User 1 described with activity labels, and discovering 8 topics. As we can observe, Non-routine related days differ from the Routine-related days as the first one presents Topic 0 and Topic 7, which are composed of activity labels describing social interaction in food-related environments. Routine-related days are mainly described by Topic 1, 3, 4, and 5, which describe working environments. We understand that activ-

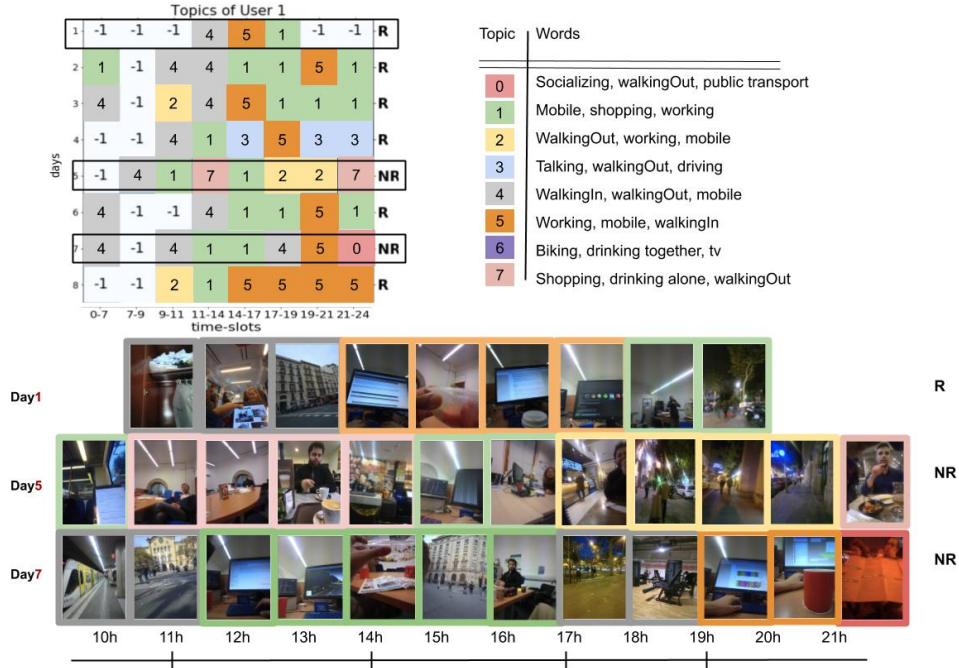


Figure 3.11: Example of given photo-streams, sample images at several time-slots, their representative topics, and the concepts that compose them. We present results with the following combination of the parameters of our model: activity labels, time-slots as in (Farrahi and Gatica, 2011), 8 topics and personalized approach.

ity labels such as *mobile*, *talking*, and *walking Indoor/Outdoor* can be understood as screen, meeting, and commuting, respectively.

To get insight at the classification level, we present in Fig. 3.12 the affinity matrix that the Spectral Clustering uses for the discrimination among the collected days by User 3 and User 7. The given labels for the collected days are indicated in the figure on the right of the matrix, where 'R' correspond to Routine-related and 'NR' to Non-Routine related. In the presented affinity matrix, we highlight the two final clusters with orange and blue. We can observe how in the case of these users clear R-related clusters are defined, while NR-related clusters are scattered. The accuracy for User 3 and User 7 is of 75% and 92%, respectively, which agree with the visual association in Fig. 3.12 between similar days and given labels.

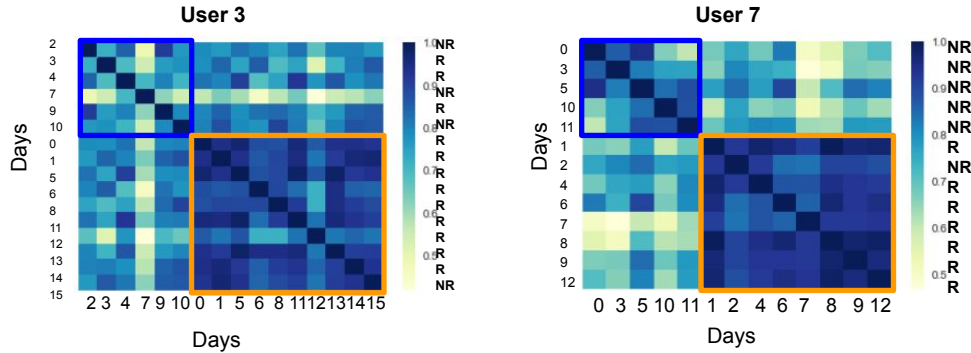


Figure 3.12: Affinity matrix obtained from the distances computed by DTW for the later discrimination as Routine or Non-Routine related days by Spectral Clustering of collected days by users 3 and 7. Days are divided with orange and blue boxes as the two final clusters. On the right, we indicate the ground-truth labels per day.

Finally, in Table 3.9 we compare the obtained results for routine discovery to the routine discovery in (Talavera et al., 2019). As one can see the method in (Talavera et al., 2019) run on 5 users achieved 0.76 of accuracy and 0.69 of F_1 score while the method proposed here achieved 0.81 of accuracy and 0.80 of F_1 score. A possible explanation is that the work proposed in (Talavera et al., 2019) relied on the aggregation of global features of all the images composing a day for its description. In contrast, the model proposed here relies on semantic concepts combined with topic modeling, DTW and spectral clustering, which results also allow understanding of what is happening in the life of the camera user. We also present the results of our method for the subset of five users that were analyzed in (Talavera et al., 2019), with a performance of $Acc = 0.82$ and $F_1 = 0.79$. As we can observe, the results are quite similar: moreover, higher classification performance is achieved when topics modeling DTW and spectral clustering are applied on the collection of documents composed of detected semantic concepts.

| Method | Number of Users | Acc | F_1 |
|---|-----------------|-------------|-------------|
| Routine discovery (Talavera et al., 2019) | 5 | 0.76 | 0.69 |
| Routine discovery propose here | | 0.82 | 0.79 |
| Routine discovery propose here | 7 | 0.81 | 0.80 |

Table 3.9: Comparison between our previous work introduced in (Talavera et al., 2019) and the model here proposed for routine discovery from egocentric photo-streams.

3.5 Discussions

In this work, we presented a new method for the analysis of routine behavioural patterns from collected egocentric visual data. We demonstrated that these images are a rich source of information and that detected concepts from the images can help us draw a picture of the lifestyle of the camera wearer.

One of the important advantages of this work is the unsupervised discovery of routine and non-routine related days. Given a new user, we can discriminate routine days and characterize their collected photo-streams. In particular, given a collection of photo-streams, our model can discover routine-related days by relying on the found topics when considering detected concepts as image descriptors. The input is a Bag-of-Word representation of the images, where an image is described by the objects and the scene it depicts. This is treated as a document for the discovery of abstract topics describing the themes of the lifestyle of the individual under study. Documents are fed to an LDA model that organizes semantic labels into topics computing a topic-word distribution and a document-topic distribution, thus, obtaining topics distribution for each given document. Moreover, we show that using temporal documents based on time-slots into which days are divided, allows flexibility when comparing the behaviour at different times of the day. The distances between the days can be computed using DTW to finally cluster days and assign them into *Routine* and *Non-Routine* ones by applying Spectral clustering.

Moreover, we introduced a new *EgoRoutine* dataset, on which we tested and validated our proposed model. The dataset is composed of a total of 104 days, recorded by 7 users, and we make it publicly available⁴ for the future development of this line of research. The analysis of the model could be improved by the augmentation of the dataset. For further steps in this direction, we need richer data. However, this is not a trivial task and we are working on it. Moreover, more accurate detected concepts would be of help when describing the collected days. For this, we would need trained networks on egocentric images.

We hypothesize that Routine-related days will share similar traits and thus, will represent a cluster. Commonly, Non-routine related days, tend to be the ones non-work related. These days share their own routine-patterns, i.e. there can be more than one routine in the life of people; cleaning, cooking, or going out with friends could describe one of them. A limitation of our work is that Non-Routine related days might not define a cluster. In future works, we plan to evaluate if the combination of outlier detection with topic modelling allows a better understanding of the lifestyle of the camera wearer.

⁴<http://www.ub.edu/cvub/dataset/>

We hope that our proposed dataset and the shown results will be a call for other researchers who aim to study people's behaviour for its understanding and providing tools for lifestyle improvement.

3.6 Conclusions

In this work, we conclude that behavioural analysis from visual data is possible. Moreover, topic models proved to be a powerful tool for the discovery of patterns when addressing Bag-of-Words representation of photo-streams. From the obtained results, we observed that discovered topic models following a personalized approach improve the classification of days. This provides a more detailed explanation of wearer daily behaviour. However, a generic or personalized approach can be applied depending on if the goal is to detect general information or peculiarities of the life of a person. One of the important advantages of this work is the unsupervised discovery of routine and non-routine related days. Given a new user, we can discriminate routine days and characterize their collected photo-streams.

Further works will explore the inclusion of outlier detection techniques and the discovery of specific behaviours, such as: social interactions and nutritional behaviour by studying the appearance of people in certain situations and food-related scenes, respectively. Furthermore, we are interested in studying how topic modelling and CNNs can be interconnected.

Published as:

E. Talavera, M. Leyva-Vallina, Md. M. K. Sarker, D. Puig, N. Petkov and P. Radeva "Hierarchical approach to classify food scenes in egocentric photo-streams," IEEE Journal of Biomedical and Health Informatics, 2019.

Chapter 4

Hierarchical approach to classify food scenes in egocentric photo-streams

Abstract

Recent studies have shown that the environment where people eat can affect their nutritional behaviour (Laska et al., 2015). In this work, we provide automatic tools for a personalised analysis of a person's health habits by the examination of daily recorded egocentric photo-streams. Specifically, we propose a new automatic approach for the classification of food-related environments, that is able to classify up to 15 such scenes. In this way, people can monitor the context around their food intake in order to get an objective insight into their daily eating routine. We propose a model that classifies food-related scenes organized in a semantic hierarchy. Additionally, we present and make available a new egocentric dataset composed of more than 33000 images recorded by a wearable camera, over which our proposed model has been tested. Our approach obtains an accuracy and F-score of 56% and 65%, respectively, clearly outperforming the baseline methods.

4.1 Introduction

Nutrition is one of the main pillars of a healthy lifestyle. It is directly related to most chronic diseases like obesity, diabetes, cardiovascular diseases, and also cancer and mental diseases (Stalonas and Kirschenbaum, 1985; Hopkinson et al., 2006; Donini et al., 2003). Recent studies show that it is not only important *what people eat*, but also *how/where people eat* (Laska et al., 2015). For instance, it is common knowledge that it is advised a person who is on a weight-reduction plan should not go to the supermarket while being hungry (Tal and Wansink, 2013). Social environment also matters; we eat more in certain situations, such as parties than at home (Higgs and Thomas, 2016). If we are exposed to the food we feel the need or temptation to eat, the same feeling of temptation will be experienced at the supermarket (Kemps et al., 2014). Not only the sight plays its role, but also smell: everyone has walked in front of a bakery shop and felt tempted or hungry immediately (de Wijk et al., 2012). The conclusion is that *where we are* can have a direct impact on *what or how we eat* and, by extension, on our health (Larson et al., 2009). However, there is a clear lack of automatic tools to monitor objectively the context of our food intake along time.

4.1.1 Our aim

Our aim is to propose an automatic tool based on robust deep learning techniques able to classify food-related scenes where a person spends time during the day. Our hypothesis is that if we can help people get insight into their daily eating routine, they can improve their habits and adopt a healthier lifestyle. By *eating routine*, we refer to activities related to the acquisition, preparing and intake of food, that are commonly followed by a person. For instance, ‘after work, I go shopping and later I cook dinner and eat’. Or, ‘I go after work directly to a restaurant to have dinner’. These two eating routines would affect us differently, having a direct impact on our health. The automatic classification of food-related scenes can represent a valuable tool for nutritionists and psychologists as well to monitor and understand better the behaviour of their patients or clients. This tool would allow them to infer how the detected eating routines affect the life of people and to develop personalized strategies for behaviour change related to food intake.

The closest approaches in computer vision to our aim focus either on scene classification, with a wide range of generic categories, or on food recognition from food-specific images, where the food typically occupies a significant part of the image. However, food recognition from these pictures does not capture the context of food intake and thus does not represent a full picture of the routine of the person. It mainly exposes what the person is eating, at a certain moment, but not *where, in*

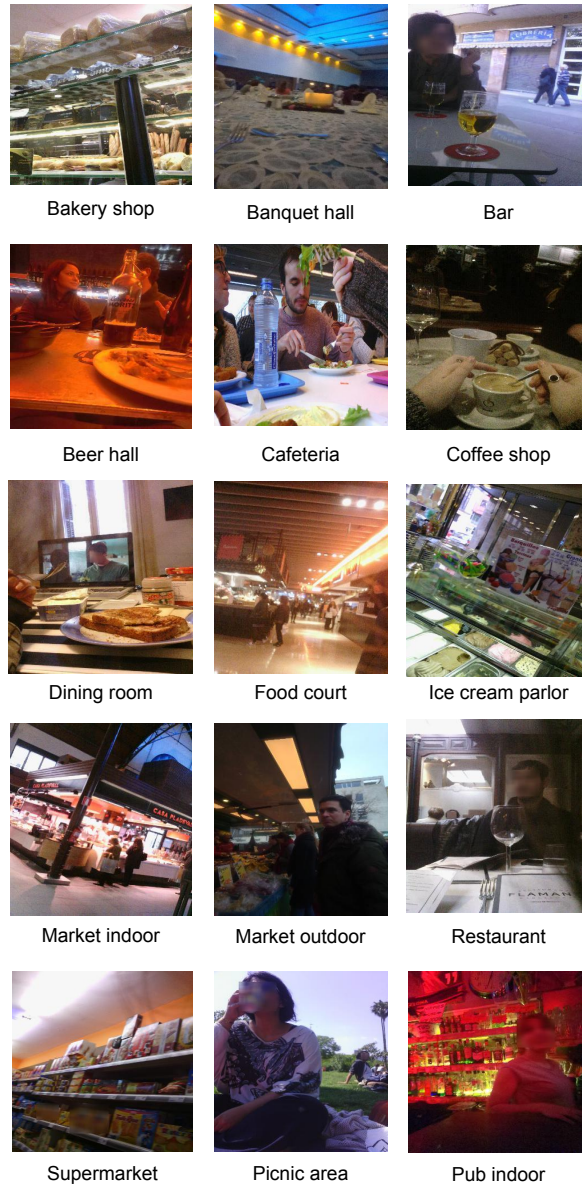


Figure 4.1: Examples of images of each of the proposed food-related categories present in the introduced EgoFoodPlaces dataset.

which environment. These environmental aspects are important to analyze in order to keep track of the people behaviour.

4.1.2 Personalized Food-Related Environment Recognition

In this work, we propose a new tool for the automatic analysis of food-related environments of a person. In order to be able to capture these environments along time, we propose to use recorded egocentric photo-streams. These images provide visual information from a first-person perspective of the daily life of the camera wearer by taking pictures frequently: visual data about activities, events attended, environments visited, and social interactions of the user are stored. Additionally, we present a new labelled dataset that is composed of more than 33000 images, which were recorded in 15 different food-related environments.

The differentiation of food-related scenes that commonly appear in recorded egocentric photo-streams is a challenging task due to the need to recognize places that are semantically related. In particular, images from two different categories can look very similar, although being semantically different. Thus, there exists a high inter-class similarity, in addition to a low intra-class variance (i.e. semantically similar categories, like *restaurant* and *pizzeria*, might look visually similar). In order to face this problem, we consider a taxonomy taking into account the relation of the studied classes. The proposed model for food-related scene classification is a hierarchical classifier that embeds convolutional neural networks emulating the defined taxonomy.

The contributions of this work are three-fold:

- A deep hierarchical network for classification of food-related scenes from egocentric images. The advantage of the proposed network is that it adapts to a given taxonomy. This allows the classification of a given image into several classes describing different levels of abstraction.
- A taxonomy of food-related environments organized in a fine-grained way that takes into account the main food-related activities (eating, cooking, buying, etc.). Our classifier is able to classify the different categories and subcategories of the taxonomy within the same model.
- An egocentric dataset of 33000 images and 15 food-related environments. We call it FoodEgoPlaces and, together with its ground-truth, is publicly available in <http://www.ub.edu/cvub/dataset/>.

The paper is organized as follows: in Section 4.2, we highlight some relevant works related to our topic, in Section 4.3 we describe the approach proposed for

food scene recognition. In Section 4.4, we introduce our FoodEgoPlaces dataset and outline the experiments performed and obtained results. In Section 4.5, we discuss the results achieved. Finally, in Section 4.6, we present our conclusions.

4.2 Related works

Scene recognition has been extensively explored in different fields, namely: robotics in (Falomir, 2012), surveillance in (Makris and Ellis, 2005), environmental monitoring in (Higuchi and Yokota, 2011), or egocentric videos in (Cartas et al., 2017). In this section, we describe previous works addressing this topic.

The recognition and monitoring of food-intake have been previously addressed in the literature, as in (Fontana et al., 2014; Ravì et al., 2015; Liu et al., 2012). For instance, in (Fontana et al., 2014), the authors proposed the use of a microphone and a camera worn on the ear to get insight into the subject’s food intake. On one side, the sound allows the classification of chewing activities, and on the other side, the selection of keyframes create an overview of the food intake that otherwise would be difficult to quantify. A food-intake log supported by visual information allows inferring the food-related environment where a person spends time. However, no work has focused on this challenge so far.

4.2.1 Scene classification

The problem of scene classification was originally addressed in the literature by applying traditional techniques ((Lazebnik et al., 2006; Quattoni and Torralba, 2009), just to mention a few), over handcrafted features. Nowadays, deep learning is the state-of-the-art (Zhou et al., 2017).

As for the former case, one of the latest works on scene recognition using traditional techniques is (Lazebnik et al., 2006), whose aim was to recognize 15 different scenes categories of outdoor and indoor scenes. The proposed model was based on the analysis of image sub-region geometric correspondences by computing histograms of local features. In (Quattoni and Torralba, 2009), the proposed approach focused on indoor scenes recognition, extending the number of recognized scenes to 67, where 10 of them are food-related. Having the hypothesis that similar scenes contain specific objects, their approach combines local and global image features for the definition of prototypes for the studied scenes. Very soon scene recognition was outperformed using deep learning.

Convolutional Neural Networks (CNNs) are a type of feed-forward artificial neural network with specific connectivity patterns. Since Yann LeCun’s LeNet (LeCun et al., 1998) was introduced, many other deep architectures have been devel-

oped and applied to different computer vision known problems, achieving better results than the state-of-art techniques: MNIST (LeCun et al., 1998) (images), Reuters (Lewis, n.d.)(documents) and TIMIT (Garofolo and et al., 1993) (recordings in English), ImageNET (Deng et al., 2009) (Data Sets classification), etc. Within the wide range of recently proposed architectures, some of the most popular are: GoogleNet (Szegedy et al., 2015), AlexNet (Krizhevsky, Sutskever and Hinton, 2012), ResNet (He et al., 2016), or VGGNet (Simonyan and Zisserman, 2015). The use of CNNs for learning high-level features has shown huge progress in scene recognition outperforming traditional techniques like (Quattoni and Torralba, 2009). This is mostly due to the availability of large datasets, those presented in (Quattoni and Torralba, 2009; Yu et al., n.d.) or the ones derived from the MIT Indoor dataset ((Zhou et al., 2014, 2017)). However, the performance at *scene recognition* level has not reached the same level of success as *object recognition*. Probably, this is a result of the difficulty presented when generalizing the classification problem, due to the huge range of different environments surrounding us (e.g. 400 in the Places2 dataset (Zhou et al., 2014)). In (Koskela and Laaksonen, n.d.), CNN activation features were extracted and concatenated following a spatial pyramid structure and used to train one-vs-all linear classifiers for each scene category. In contrast, in (Zhou et al., 2014) the authors evaluate the performance of the responses from the trained Places-CNN as generic features, over several scene and object benchmarks. Also, a probabilistic deep embedding framework, which analyses regional and global features extracted by a neural network, is proposed in (Zheng et al., 2014). In (Wang et al., 2015), two different networks called Object-Scene CNNs, are combined by late fusion; the ‘object net’ aggregates information for event recognition from the perspective of objects, and the ‘scene net’ performs the recognition with help from the scene context. The nets are pre-trained on the ImageNet dataset (Deng et al., 2009) and Places dataset (Zhou et al., 2014) respectively. Recently, in (Herranz et al., 2016) the authors combine object-centric and scene-centric architectures. They propose a parallel model where the network operates over different scale patches extracted from the input image. None of these methods has been tested on egocentric images, which by themselves represent a challenge for image analysis. In this kind of data, the camera follows the user’s movements. This results in big variability on illumination, blurriness, occlusions, drastic visual changes due to the low frame rate of the camera, narrow field of view, among other difficulties.

4.2.2 Classification of egocentric scenes

In order to obtain personalized scene classification, we need to analyze egocentric images acquired by a wearable camera. Egocentric image analysis is a rela-

tively recent field within computer vision concerning the design and development of Computer Vision algorithms to analyze and understand photo-streams captured by a wearable camera. In (Furnari et al., 2016), several classifiers were proposed to recognize 8 different scenes (not all of them food-related). First, they discriminate between food/no-food and later, they train One-vs-all classifiers to discriminate among classes. Later, in (Furnari et al., 2017) a multi-class classifier was proposed, with a negative-rejection method applied. In (Furnari et al., 2016, 2017) they only consider 8 scene categories, just 2 of them are food-related (*kitchen* and *coffee machine*) and without visual or semantic relation.

4.2.3 Food-related scene recognition in egocentric photo-streams

In our preliminary work presented in (Sarker et al., 2018), we proposed a MACNet neural architecture for the classification of food-related scenes. This network input image is scaled into five different resolutions (the original image, with a scale value of 0.5). The five scaled images are fed to five blocks of atrous convolutional networks (Chen et al., 2018) with three different rates (1, 2, and 3) to extract the key features of the input image in multi-scale. In addition, four blocks of pre-trained ResNet are used to extract 256, 512, 1024 and 2048 feature maps, respectively. Each feature maps extracted by an atrous convolutional block is concatenated with the corresponding ResNet block to feed the subsequent block. Finally, the features obtained from the fourth ResNet layer is the final features are used to classify the food places images using two fully connected (FC) layers.

However, the challenge still remains due to the high variance that environments take in real-world places and the wide range of possibilities of how a scene can be captured. In this work, we propose an organization of the different studied classes into semantic groups following the logic that relates them. We define a taxonomy, i.e. a semantic hierarchy relating the food-related classes. Hierarchical classification is an iterative process that groups features or concepts based on their similarity into clusters, until merging them all together. There are two strategies for hierarchical classification: agglomerative (bottom-up) and divisive (top-down). We aim to classify food-related images following a top-down strategy, i.e. from a less to a more specific description of the scene. The proposed hierarchical model supports its final classification on the dependence among classes at the different levels of the classification tree. This allows us to study different levels of semantic abstraction. The different semantic levels (L), Level 1 (L1), Level 2 (L2) and Level 3 (L3), are introduced in Fig. 4.2. In this document, we refer to meta-class as the class whose instances are semantic and visual correlated classes.

Therefore, we organize environments according to the actions related to them: cooking, eating, acquiring food products. We demonstrate that by creating different levels of classification and classifying scenes by the person action, it can serve as a natural prior for more specific environments and thus can further improve the performance of the model. The proposed classification model, implemented following this taxonomy, allows analyzing at different semantic levels of where the camera wearer spends time.

To the best of our knowledge, no previous work has focused on the problem of food-related scenes recognition at different semantic levels, either from conventional or egocentric images. Our work aims to classify food-related scenes from egocentric images recorded by a wearable camera. We believe that these images highly describe our daily routine and can contribute to the improvement of healthy habits of people.

4.3 Hierarchical approach for food-related scenes recognition in egocentric photo-streams

We propose a new model to address the classification of food-related scenes in egocentric images. It follows a hierarchical semantic structure, which adapts to the taxonomy that describes the relationships among classes. The classes are hierarchically implemented from more abstract to more specific ones. Therefore, the model is scalable and can be adapted depending on the classification problem, i.e. if the taxonomy changes.

For the purposes of food-related scene classification, we define a semantic tree which is depicted in Fig. 4.2. We redefine the problem inspired by how humans hierarchically organize concepts into semantic groups. The Level 1 directly related to the problem of physical activities recognition (Cartas et al., 2017): *eating*, *preparing*, and *acquiring food (shopping)*. Note that the recognition of physical activities itself is a well-known and still open research problem in egocentric vision (Cartas et al., 2017). On the other hand, recognition of these three activities has multiple applications like for patients with Mild Cognitive Impairment (MCI) in the Cambridge cognition test (Schmand et al., 2000). There, the decrease of older people's cognitive functions with time is one of the factors to estimate their cognitive capacities by measuring their capacity to prepare food or go for shopping (Petersen et al., 1999). Later it splits eating into eating outdoor or indoor. Some of the subcategories group several classes, such as the subcategory *eating indoor* that encapsulates seven food-related scenes classes: *bar*, *beer hall*, *cafeteria*, *coffee shop*, *dining room*, *restaurant*, and *pub indoor*. In contrast, *preparing* and *eating outdoor* are represented uniquely by *kitchen*

and *picnic area*, respectively. The semantic hierarchy was defined following the collected food-related classes and their intrinsic relation. Thus, the automatic analysis of the frequency and duration of such food-related activities is of high importance when analyzing their behaviour. The environment is differentiated in Level 2. As commented in the manuscript, in (Laska et al., 2015) the authors stated that ‘where you are, affects your eating habits’. Thus, the food routine or habits of camera wearers can be inferred by recognizing the food-related environment where they spend time (e.g. outdoor, indoor, etc.). The classification of scenes is already a scientific challenge, see the dataset Places (Zhou et al., 2017). For us, the novelty is to address the classification of scenes with similar characteristics (food-related) that makes the problem additionally more difficult.

We proposed this taxonomy because we think it represents a powerful tool to address the behaviour of people. Moreover, it could be of interest in order to estimate the cognitive state of MCI patients. We reached this conclusion after previous collaborations with psychologists working on the MCI disorder, and analysing egocentric photo-streams addressing several problems.

The differentiation among classes at the different levels of the hierarchy needs to be performed by a classifier. In this work, we propose to use CNNs for the different levels of classification of our food-related scenes hierarchy. The aggregation of CNNs layers mimics the structure of the food-related scenes presented in Fig. 4.2. Due to the good quality of the scene classification results over the Places2 dataset (Zhou et al., 2016), we made use of the pre-trained VGG16 introduced in (Simonyan and Zisserman, 2015), on which we built our hierarchical model. In this work, we will refer to it as VGG365 network. Note that this approach resembles the DECOC classifier (Pujol et al., 2006) that proves the efficiency of decomposing a multi-class classification problem in several binary classification problems organized in a hierarchical way. The difference with the food-related scene classification is that in the latter case the classes are organized semantically in meta-classes corresponding to nutrition-related activities instead of constructing meta-classes without explicit meaning, but according to the entropy of training data (Pujol et al., 2006).

Given an image, the final classification label is based on the aggregation of estimated intermediate probabilities obtained for the different levels of the hierarchical model, since a direct dependency exists between levels of the classification tree. The model aggregates the chain of probabilities by following the statistical inference method. The probability of an event is based on its prior estimated probabilities.

Let us consider classes C^i and C^{i-1} so that superscript shows the level of the class in the hierarchy and C^{i-1} is the parent of C^i in the hierarchical organization of the tree. Thus, we can write:

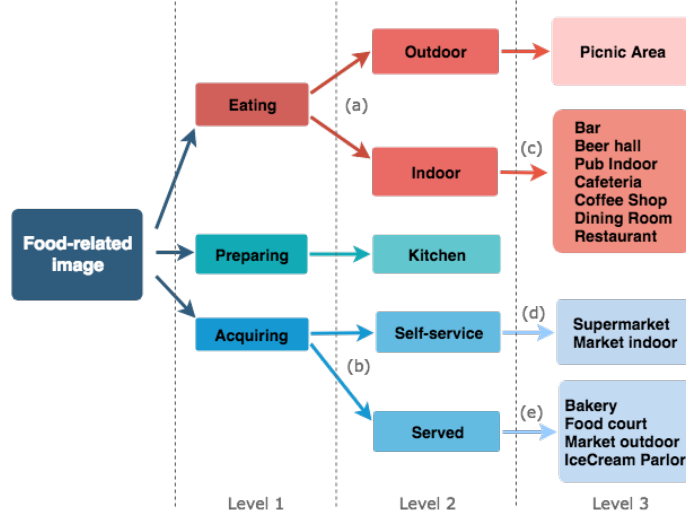


Figure 4.2: The proposed semantic tree for food-related scenes categorization. For their later reference, we mark with dashed lines the different depth levels, and with letters the sub-classification groups.

$$P(C^i, x) = P(C^i, x|C^{i-1}, x) * P(C^{i-1}|x) \quad (4.1)$$

where $P()$ relates to probabilities. $P(C^{i-1}, x|C^i, x)$ represents the likelihood of C^{i-1} , given image x , occurring given that C^i , given image x , is happening, while $P(C^i, x)$ and $P(C^{i-1}, x)$ are marginal probabilities given image x , i.e. the probabilities of independently observing C^i and C^{i-1} , respectively.

Note that we can estimate $P(C^i, x|C^{i-1}, x)$ from the classifier of the network trained to classify the classes children of class C^i , $P(C^{i-1}, x|C^i, x)$ is 1 since C^i is a subclass of C^{i-1} .

$P(C^{i-1}, x)$ can be recursively estimated by considering the estimated probability on C^{i-1} and its class parent. Hence, we obtain that for each node C^i in the hierarchy (in particular, for the leaves), we get:

$$P(C^i, x) = \prod_{j=1}^i P(C^j, x|C^{j-1}, x) * P(C^{j-1}, x) \quad (4.2)$$

Without loss of generality, we consider that the probability of the class in the root is the probability to have a food-related image, ($P(C^0)$), obtained by a binary classifier.

Let us illustrate the process with an example. Following the semantic tree in Fig.4.2, our goal is to classify an egocentric image belonging to the class *dining room*.

We observe that as *dining room* is a subclass of *indoor* and *indoor* is of *eating*, etc. Thus, the probability of *dining room* occurring given image x is computed as:

$$P(\text{diningroom}, x) = P(\text{diningroom}, x | \text{indoor}, x) \dot{P}(\text{indoor}, x | \text{eating}, x) \dot{P}(\text{eating}, x | \text{foodrelated}, x) \dot{P}(\text{foodrelated}, x) \quad (4.3)$$

To summarize, given an image, our proposed model computes the final classification as a product of the estimated intermediate probabilities at the different levels of the defined semantic tree.

4.4 Experiments and Results

In this section, we describe a new home-made dataset that we make public, the experimental setup, the metrics used to evaluate the analysis, and the obtained results.

4.4.1 Dataset

In this work, we present *EgoFoodPlaces*, a dataset composed of more than 33000 egocentric images from 11 users organized in 15 food-related scene classes. The images were recorded by a Narrative Clip camera¹. This device is able to generate a huge number of images due to its continuous image collection. It has a configurable frame rate of 2-3 images per minute. Thus, users regularly record an amount of approximately 1500 images per day. The camera movements and the wide range of different situations that the user experiences during his/her day, lead to new challenges such as background scene variation, changes in lighting conditions, and handled objects appearing and disappearing throughout the photo sequence.

Food-related scene images tend to have an intrinsic high inter-class similarity, see Fig. 4.1. To determine the food-related categories, we selected a subset of the ones proposed for the Places365 challenge (Zhou et al., 2017). We focus on the categories with a higher number of samples in our collected egocentric dataset, disregarding very unlikely food-related scenes, such as *beer garden* and *ice-cream parlor*. Furthermore, we found that discriminating scenes like *pizzeria* and *fast-food restaurant* is very subjective if the scene is recorded from a first-person view, and hence, we merged them into a *restaurant* class.

EgoFoodPlaces was collected during the daily activities of the users. To build the dataset, we select the subset of images from the EDUB-Seg dataset that described food-related scenes, introduced in (Talavera et al., 2015; Dimiccoli et al., 2017), and

¹<http://getnarrative.com/>

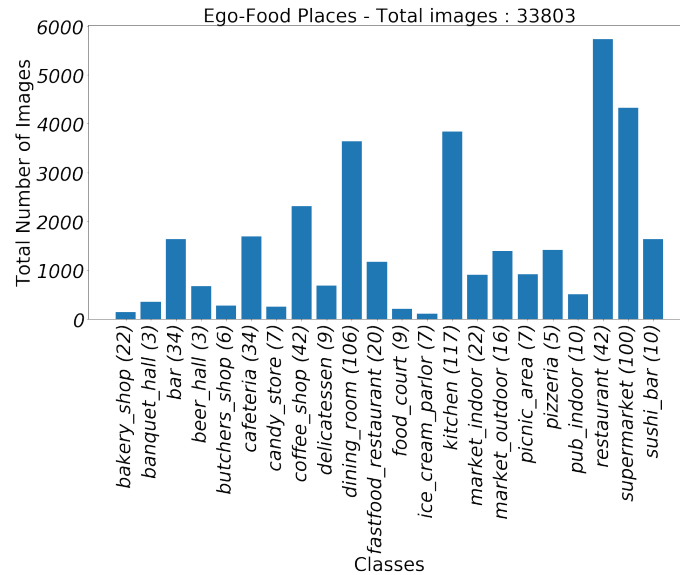


Figure 4.3: Total number of images per food-related scene class. We give the number of collected events per class between parenthesis.

later extended it with new collected frames. The dataset was gathered by 11 different subjects, during a total of 107 days, while spending time in scenes related to the *acquisition*, *preparing* or *consumption* of food. The dataset has been manually labelled into a total of 15 different food-related scenes classes: *bakery*, *bar*, *beer hall*, *cafeteria*, *coffee shop*, *dining room*, *food court*, *ice cream parlour*, *kitchen*, *market indoor*, *market outdoor*, *picnic area*, *pub indoor*, *restaurant*, and *supermarket*. In Fig. 4.3, we show the number of images per different classes. This figure shows the unbalanced nature of the classes in our dataset, reflecting the different prolongation of time that a person spends on different food-related scenes.

Since the images were collected by a wearable camera when performing any of the above-mentioned activities, the dataset is composed of groups of images close in time. This leads to two possible situations. On one hand, images recorded ‘sitting in front of a table while having dinner’ will most likely be similar. On the contrary, in scenes such as ‘walking at the supermarket’ the images vary since they follow the walking movement of the user in a very varying environment.

In Fig. 4.4, we present the dataset by classes and events. This graph shows how the average, maximum and minimum spent time for the given classes differ. Note that this time can be studied since it is directly related to the number of recorded images in the different food-related scenes. As we previously assumed, classes with

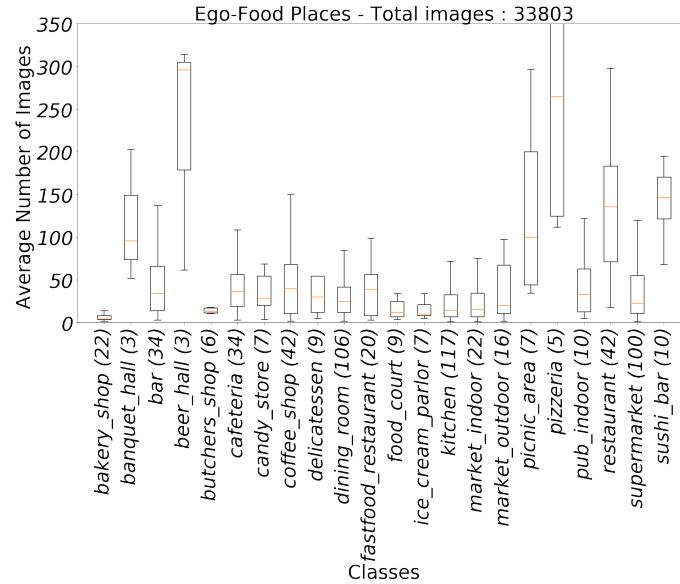


Figure 4.4: Illustration of the variability of the size of the events for the different food-related scene classes. The data is presented by making the width of the box proportional to the size of the group. We give the number of collected events per class between parenthesis. The range of the data of a class is shown by the whiskers extend from its data box.

a small number of images correspond to unusual environments or environments where people do not spend a lot of time in (e.g. *bakery*). In contrast, the most populated classes refer to everyday environments (e.g. *kitchen*, *supermarket*), or to environments where more time is usually spent (e.g. *restaurant*).

Class-variability of the EgoFoodPlaces dataset

To quantify the degree of semantic similarity among the classes in our proposed dataset, we compute the intra- and inter-class correlation. We use the classification probabilities output of the proposed baseline VGG365 network in order to find suitable descriptors for our images for this comparison. This network was trained for the classification of the proposed 15 food-related scenes. These descriptors encapsulate the semantic similarities of the studied classes.

To study the intra-class variability, we compute the mean silhouette coefficient for all samples, that is defined as,

$$\text{Silhouette_score} = (b - a) / \max(a, b) \quad (4.4)$$

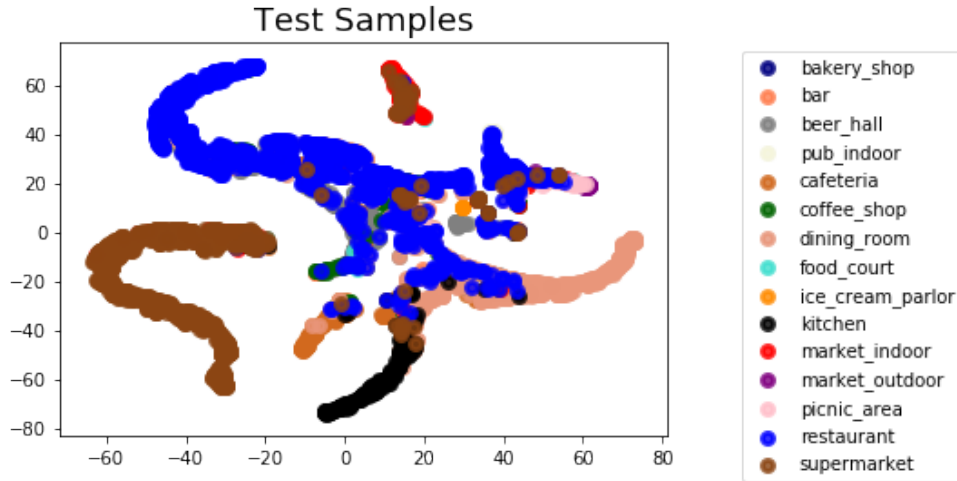


Figure 4.5: Visualization of the distribution of the classes using the t-SNE algorithm.

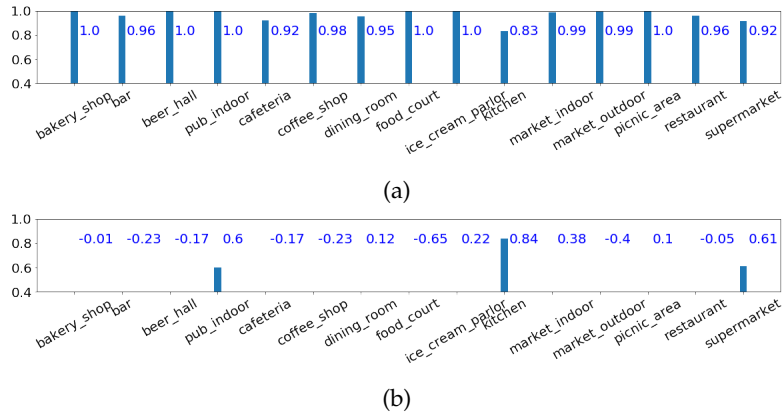


Figure 4.6: Mean Silhouette Score for the samples within the studied food-related classes. The train and test sets are evaluated separately in (a) and (b), respectively. The score is shown with bars and in blue text on top of them.

where (a) corresponds to the intra-class distance per sample, and (b) corresponds to the distance between a sample and the closest class to which the sample is part of. Note that the silhouette takes values from 1 to -1; the highest value represents high density and separated clusters. The value 0 represents overlapping of clusters. Negative values indicate that there are samples with more similar clusters than the

one they have been assigned to. The mean Silhouette score is 0.94 and 0.15 for the train and test samples, respectively. The score is depicted for the different analyzed classes in Fig. 4.6. The high score obtained for the train set is due to the fact that the analyzed descriptors are extracted fine-tuning the network with those specific samples. Thus, their descriptors are of high quality for their differentiation. In contrast, the test set is an unseen set of images. The low value of the test set indicates that the classes are challenging to classify.

Furthermore, we visually illustrate the inter-class variability of the classes by embedding the 15-dimensional descriptor vector to 2 dimensions using the t-SNE algorithm (Maaten and Hinton, 2008). The results are shown in Fig. 4.5. This visualization allows us to better explore the variability among the samples in the test set. For instance, classes such as *restaurant* and *supermarket* are clearly distinguishable as a cluster. In contrast, we can recognize the classes with lower recognition rate, like the ones overlapping with *supermarket* and *restaurant*. For instance, *market indoor* is merged in its majority with *supermarket*. At the same time, the class *restaurant* clearly overlaps with *coffee shop* and *picnic area*.

4.4.2 Experimental setup

In this work, we propose to build the model on top of the VGG365 network (Zhou et al., 2017) since it outperformed state-of-the-art CNNs when classifying conventional images into scenes. We selected this network because it was already pre-trained with images describing scenes, and after evaluating and comparing its performance to the state-of-the-art CNNs. The classification accuracy obtained by the VGG16 (Simonyan and Zisserman, 2015), InceptionV3 (Szegedy et al., 2016), and ResNet50 (He et al., 2016), were 55.07%, 51.22%, and 60.43%, respectively, lower than the 64.02% accuracy achieved by the VGG365 network.

We build our hierarchical classification model by aggregating VGG365 nets over different subgroups of images/classes, emulating the proposed taxonomy for food-related scenes recognition in Fig. 4.2. The final probability of a class is computed by the model, as described in Section 4.3.

The model adapts to an explicit semantic hierarchy that aims to classify a given sample of food-related scenes. Moreover, it aims to further understanding of the relation among the different given classes. Therefore, we compare the performance of the proposed model against existent methodologies that can be adapted to obtain similar classification information.

We compare the performance of the proposed model with the following baseline experiments:

1. FV: Fine-tuning of the VGG365 network with *EgoFoodPlaces*.

2. FV-RF: We use this categorical distribution obtained by the fine-tuned VGG365 in (1) as image descriptors. Later, we train the Random Forest classifier with 200 trees (Ho, 1995).
3. FV-SVM: Fine-tuned VGG365 to obtain image descriptors and Support Vector Machines (Cortes and Vapnik, 1995).
4. FV-KNN: Fine-tuned VGG365 to obtain image descriptors and k-Nearest Neighbors (Altman, 1992) (n=3).
5. SVM-tree: We use the categorical distribution obtained by the fine-tuned VGG365 as images descriptors of the subsets of images that represents the nodes of the tree. Later, we train SVM as nodes of the proposed taxonomy.
6. MACNet (Sarker et al., 2018): We fine-tuned the MACNet network introduced in (Sarker et al., 2018) to fit our proposed dataset.
7. FV-Ensemble: We evaluate the performance of a stack of FV networks that are trained with a different random initialization of the final fully connected weights for classification. The final prediction is the average of the predictions of the networks. We ensemble the same number of CNNs as the number of CNNs included in the proposed hierarchical model, i.e. 6 CNNs.

We perform a 3-Fold cross-validation of the proposed model to verify its ability to generalize and report the average value. The baseline methodologies are also evaluated following a 3-Fold cross-validation strategy.

We make use of the Scikit-learn machine learning library available for Python for the training of the traditional classifiers (SVM, RF, and KNN). For all the experiments, the images are re-sized at size 256x256. For the CNNs, we fine-tuned the baseline CNNs for 10 epochs, with a training batch size of 8, and run the validation set each 1000 iterations. The training of the CNNs was implemented using Caffe (Jia et al., 2014) and its Python interface. The code for the implementation of our proposed model is publicly available in <https://github.com/estefaniatalavera/Foodscenes.hierarchicalmodel>.

4.4.3 Dataset Split

In order to robustly generalize the proposed model and fairly test it, we assure that there are no images from the same scenes/events in both training and test sets. To this aim, we divide the dataset into events for the training and evaluation phases. Events are captured by sequentially recorded images that describe the same environment, and we obtain them by applying the SR-Clustering temporal segmentation

method introduced in (Dimiccoli et al., 2017). The division of the dataset into training, validation and test, aims to maintain a 70%, 10% and 20% distribution, respectively. As it can be observed in Fig. 4.3, *EgoFoodPlaces* presents highly unbalanced classes. In order to face this problem, we could either subsample classes with high representation, or add new samples to the ones with low representation. We decided not to discard any image due to the relatively small number of images within the dataset. Thus, we balanced the classes for the training phase by over-sampling the classes with fewer elements. The training process of the network learns from randomly crops of the given images, the over-sampling simply passes the same instances several times, until reaching the defined number of samples per class, which will correspond to the number of samples of the most frequent class. For all the experiments performed, the images used for the training phase are shuffled in order to give robustness to the network. Together with the *EgoFoodPlaces* dataset, the given labels, and the training, validation and test files are publicly available for further experimentation (<http://www.ub.edu/cvub/dataset/>).

4.4.4 Evaluation

We evaluate the performance of the proposed method and compare it with the baseline models by computing the accuracy, precision, recall and F_1 (F-score). We calculate them per each class, together with their ‘macro’ and ‘weighted’ mean. ‘Macro’ calculates metrics for each label, and find their unweighted mean, while ‘weighted’ takes into account the true instances for each label. We also compute the weighted accuracy. The use of weighted metrics aims to face the unbalanced of the dataset, and intuitively expresses the strength of our classifier. This metric normalizes based on the number of samples per class.

The F_1 score, *Precision* and *Recall* can be defined as:

$$F_1 = 2 \times \frac{Precision * Recall}{Precision + Recall}, \quad (4.5)$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}, \quad (4.6)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}. \quad (4.7)$$

Moreover, we qualitatively compare the given labels by our method and the best of the proposed baseline to sample images from the test set.

4.4.5 Results

We present the obtained classification accuracy at image level for the performed experiments in Table 4.1. As it can be observed, our proposed model achieves the highest accuracy and weighted average accuracy, with 75.46% and 63.20%, respectively, followed by the SVM and Random Forest for the accuracy and SVM and KNN for the weighted accuracy.

Our proposed hierarchical model has the capability of recognizing not only the 15 classes corresponding to the leaves of the tree in the semantic tree (see Fig.4.2), but also the meta-classes at the different semantic levels of depth. Thus, specialists can analyze the personal data and generate strategies for the improvement of the lifestyle of people by studying their food-related behaviour either from a broad perspective, such as when the person *eats* or *shops*, or into a more detailed one, like *if the person usually eats in a fast-food restaurant or at home*.

A logical question is if the model provides a robust classification of meta-classes as well. To this aim, we evaluate the classification performance at the different levels of the defined semantic tree. Note that since each class is related to a meta-class on a higher level, an alternative to our model would be to obtain the meta-classes accuracy from their sub-classes classification. We compare the accuracy of meta-classes from their classification by the proposed model vs inferring the accuracy from the classification of the subclasses samples for the set of baseline models. As one can observe in Table 4.2, our model achieves higher accuracy classifying meta-classes in all cases with 94.7%, 68.5%, 94.7% for Level 1 (L1), Level 2 (L2) and Level 3 (L3), respectively. This proves that it is a robust tool for the classification of food-related scenes classes and meta-classes.

If we observe the confusion matrix in Fig. 4.7, we can get insight about the miss-classified classes. We can see how our algorithm tends to confuse the classes belonging to the semantic level of *self-service* (acquiring) and *eating indoor* (eating). We believe that this is due to the unbalanced aspect of our data and the intrinsic similarity within the sub-categories of some of the branches of the semantic tree.

The classes with higher classification accuracy are *kitchen* and *supermarket*. We deduce that this is due to the very characteristic appearance of the environment that they involve and the number of different images of such classes in the dataset. On the contrary, *picnic area* is not recognized by any of the methods. The confusion matrix indicates that the class is embedded by the model into the class *restaurant*. This can be inferred by visually checking the images since in both classes a table and another person usually appear in front of the camera wearer. Moreover, from the obtained results, we can observe a relation between the previously computed Silhouette Score per class and the classification accuracy achieved by the classifiers.

Table 4.1: Food-related scene classification performance. We present the accuracy per class and model, and precision, recall and F1 score for all models. We rename the fine-tuning of the VGG365 as ‘FV’, and the later use of its output probabilities for the training of the State-of-the-Art models.

| | OurModel | FV | Tree+SVM | FV+RF | FV+SVM | FV+KNN | MACNet Sarker et al. (2018) | EnsembleCNNs |
|--------------------|----------|------|----------|-------|--------|--------|--------------------------------|--------------|
| bakery shop | 0.39 | 0.58 | 0.58 | 0.56 | 0.58 | 0.59 | 0.58 | 0.60 |
| bar | 0.31 | 0.13 | 0.15 | 0.11 | 0.11 | 0.17 | 0.17 | 0.15 |
| beer hall | 0.89 | 0.32 | 0.20 | 0.18 | 0.20 | 0.20 | 0.61 | 0.56 |
| pub indoor | 0.85 | 0.70 | 0.71 | 0.71 | 0.71 | 0.71 | 0.64 | 0.82 |
| cafeteria | 0.45 | 0.45 | 0.44 | 0.43 | 0.44 | 0.43 | 0.72 | 0.55 |
| coffee shop | 0.59 | 0.40 | 0.39 | 0.34 | 0.38 | 0.34 | 0.49 | 0.49 |
| dining room | 0.58 | 0.58 | 0.59 | 0.58 | 0.58 | 0.57 | 0.56 | 0.57 |
| food court | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| ice-cream parlor | 0.52 | 0.59 | 0.68 | 0.65 | 0.64 | 0.65 | 0.15 | 0.73 |
| kitchen | 0.89 | 0.87 | 0.89 | 0.87 | 0.86 | 0.86 | 0.88 | 0.90 |
| market indoor | 0.70 | 0.73 | 0.76 | 0.77 | 0.77 | 0.76 | 0.66 | 0.77 |
| market outdoor | 0.28 | 0.20 | 0.20 | 0.20 | 0.20 | 0.19 | 0.23 | 0.25 |
| picnic area | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| restaurant | 0.70 | 0.67 | 0.68 | 0.68 | 0.68 | 0.68 | 0.63 | 0.73 |
| supermarket | 0.85 | 0.81 | 0.81 | 0.79 | 0.81 | 0.80 | 0.75 | 0.84 |
| Macro Precision | 0.56 | 0.53 | 0.55 | 0.59 | 0.55 | 0.56 | 0.48 | 0.60 |
| Macro Recall | 0.53 | 0.48 | 0.47 | 0.44 | 0.46 | 0.45 | 0.49 | 0.52 |
| Macro F1 | 0.53 | 0.47 | 0.47 | 0.46 | 0.46 | 0.46 | 0.47 | 0.53 |
| Weighted Precision | 0.65 | 0.62 | 0.62 | 0.62 | 0.62 | 0.61 | 0.61 | 0.67 |
| Weighted Recall | 0.68 | 0.64 | 0.64 | 0.64 | 0.64 | 0.64 | 0.63 | 0.68 |
| Weighted F1 | 0.65 | 0.61 | 0.61 | 0.60 | 0.61 | 0.60 | 0.61 | 0.65 |
| Accuracy | 0.68 | 0.64 | 0.64 | 0.64 | 0.64 | 0.64 | 0.63 | 0.68 |
| Weighted Accuracy | 0.56 | 0.53 | 0.51 | 0.47 | 0.50 | 0.48 | 0.49 | 0.55 |

Table 4.2: Classification performance at different levels of the proposed semantic tree for food-related scenes categorization. We compute the achieved accuracy (Acc) per level and the weighted accuracy (W-Acc) where we consider the number of samples per class. The different semantic levels (L), Level 1 (L1), Level 2 (L2) and Level 3 (L3) are introduced in Fig. 4.2.

| | Our Method | | FV | | SVMTree | | FV+RF | | FV+SVM | | FV+KNN | | MACNet Sarker et al. (2018) | | EnsembleCNN | |
|----------------|--------------|--------------|-------|-------|---------|-------|-------|-------|--------|-------|--------|-------|--------------------------------|-------|-------------|--------------|
| | Acc | WAcc | Acc | WAcc | Acc | WAcc | Acc | WAcc | Acc | WAcc | Acc | WAcc | Acc | WAcc | Acc | WAcc |
| Level 1 (L1) | 0.944 | 0.947 | 0.927 | 0.919 | 0.934 | 0.931 | 0.928 | 0.922 | 0.927 | 0.924 | 0.927 | 0.910 | 0.884 | 0.865 | 0.923 | 0.913 |
| Level 2a (L2a) | 0.915 | 0.685 | 0.886 | 0.664 | 0.898 | 0.673 | 0.890 | 0.666 | 0.800 | 0.753 | 0.890 | 0.648 | 0.829 | 0.623 | 0.869 | 0.629 |
| Level 2b (L2b) | 0.893 | 0.947 | 0.890 | 0.940 | 0.890 | 0.944 | 0.885 | 0.945 | 0.897 | 0.935 | 0.885 | 0.927 | 0.860 | 0.906 | 0.856 | 0.955 |

Classes with high consistency are better classified, while classes such as *bar*, *bakery shop*, *picnic area*, or *market outdoor* have lower classification performance.

The achieved results are rather quantitatively similar. Therefore, we perform the *t-test* to evaluate the statistical significance of the differences in performance. Our proposed model outperforms FV, SVMtree, FV+RF, FK+KNN, FV+SVM, MacNet, and ensembleCNNs with statistical significance ($p=0.038 * 10^{-16}$, $p=0.042 * 10^{-12}$, $p=0.087 * 10^{-14}$, $p=0.057 * 10^{-13}$, $p=0.079 * 10^{-16}$, $p=0.087 * 10^{-19}$, and $pvalue=3.24 * 10^{-1}$ for paired t-test). The smaller the p value, the higher the statistical significance.

From the results, we can discuss that the performance by the ensemble of CNNs is similar to the proposed model. This happens when it is evaluated at the level of image classification. We can see in Table II how the proposed hierarchy outperforms the baseline methods when classifying at the different levels of the taxonomy tree.

Qualitatively, in Fig. 4.8 we illustrate some correct and wrong classifications by our proposed model and the trained SVM (FV-SVM). We highlight the ground-truth class of the images in boldface. Even though the performance of the different tested models does not differ much, the proposed model has the ability to better generalize, as its weighted average accuracy indicates.

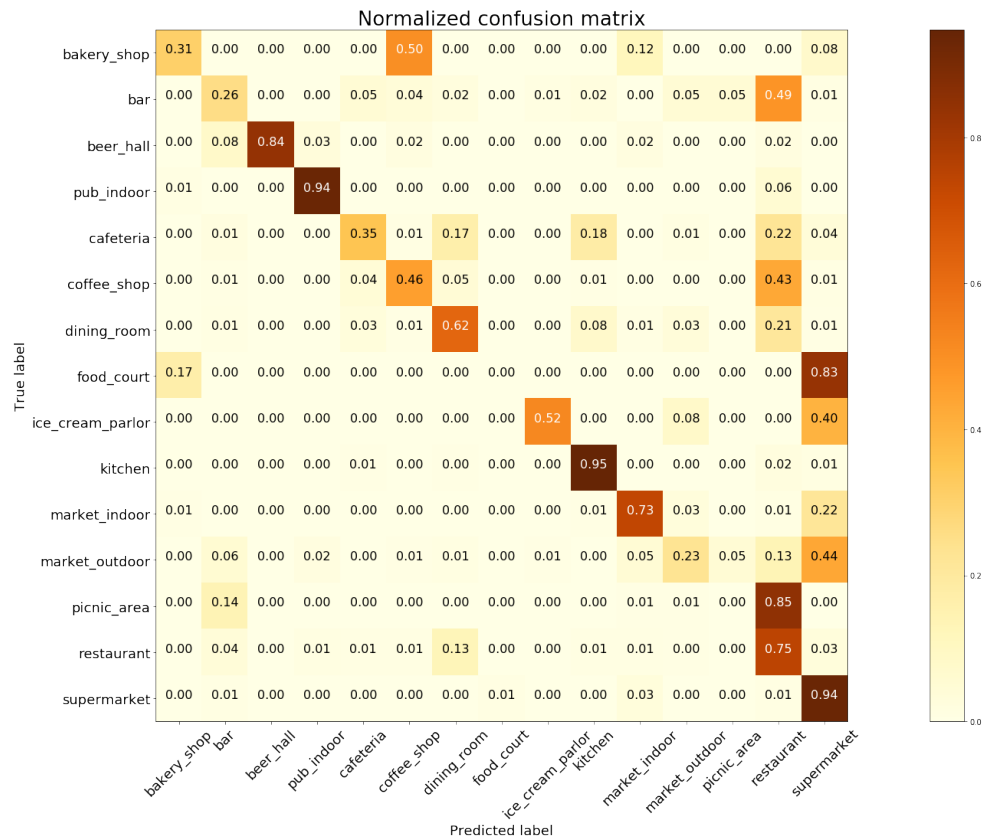


Figure 4.7: Confusion matrix with the classification performance of the proposed hierarchical classification model.

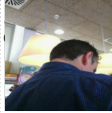
| Our approach | FV-SVM | Our approach | FV-SVM | Our approach | FV-SVM |
|--|---|--|---|---|---|
|  | dining room restaurant coffee shop bar kitchen |  | dining room kitchen restaurant bar kitchen |  | supermarket restaurant kitchen cafeteria dining room |
|  | market indoor supermarket picnic area restaurant kitchen |  | dining room kitchen restaurant coffee shop supermarket |  | bar restaurant supermarket coffee shop pub indoor |
|  | bar supermarket market outdoor ice cream shop |  | bakery shop market outdoor supermarket kitchen coffee shop |  | restaurant dining room bar cafeteria coffee shop (picnic area) |
|  | bar coffee shop market outdoor restaurant kitchen |  | supermarket restaurant ice cream shop market outdoor coffee shop |  | supermarket restaurant supermarket coffee shop market outdoor bakery shop |

Figure 4.8: Examples of top 5 classes for the images in the test set. We show the results obtained by the proposed model, and compare them with the obtained ones by the trained SVM classifier. The class in bold corresponds to the true label of the given image.

4.5 Discussions

The proposed dataset is composed of manually selected images from recorded day photo-streams. These extracted images belong to food-related events, described as groups of sequential images representing the same scene. We find important to highlight that for the performed experiments, images belonging to the same event stayed together for either training or testing phase. Even though the classification of such scenes could have been events rather than images, we do not dispose of a higher number of events for the training phase in the case of event-based scene classification. The creation of a bigger egocentric dataset is a recurrent ongoing work. Next lines of work will address the analysis of events in order to study if they are connected and time-dependent.

Recorded egocentric images can be highly informative about the lifestyle, behaviour and habits of a person. In this work, we focus on the implementation of computer vision algorithms for data extraction from images. More specifically, on characterizing food-related scenes related to an individual for future assistance in controlling obesity and other eating disorders being of high importance for society.

Next steps could involve the analysis of other information e.g. the duration and regularity of nutritional activities. Based on extracted information regarding individuals, their daily habits can be extracted and characterized. The daily habits of

people can be correlated to their personality since people's routine affects them differently. Moreover, within this context social relations and their relevance can be studied: the number of people a person sees per day, the length and frequency of their meetings and activities, etc and how social context influence people. All this information extracted from egocentric images is still to be studied in depth leading to powerful tools for an objective, long-term monitoring and characterization of the behaviour of people for better and longer life.

The introduced model can be easily extrapolated and implemented to other classification problems with semantically correlated classes. Organizing classes in a semantic hierarchy and embedding a classifier to each node of the hierarchy allow considering the estimated intermediate probabilities for the final classification.

The proposed model computes the final classification probability based on the aggregation of the probabilities of the different classification levels. The random probability of a given class is $1/|C|$, where $|C|$ is the number of children the parent class of that node has. Hence, having a high number of sub-classes (children nodes) for a specific node would tend to lower probability. There is a risk that a 'wrong class node' gets higher final classification probability if it has few brother-sin the tree compared to the 'correct class node'.

Application to recorded days characterization

Food-related scenes recognition is very useful to get understanding of the patterns of behaviour of people. The presence of people at certain food-related places is of importance when describing their lifestyle and nutrition. While in this work we focus on the classification of such places, we use the labels given to the photo-streams to characterize the camera wearer's 'lived experiences' related to food. The characterization is given by the proposed model allows us to address the scene detection at different semantic levels. Thus, by using high-level information we increase the robustness and the level of the output information of the model.

In Fig. 4.9, we illustrate a realistic case where each row represents a recorded day by the camera wearer. As we have previously highlighted, our proposed model focuses on the classification of food-related scenes in egocentric photo-streams. However, the previous classification step would be the differentiating among Food and Non-food related images. In (Cartas et al., 2017) the authors addressed activity recognition in egocentric images. Thus, we apply their network and focus on images labelled as 'shopping' and 'eating or drinking', to later apply our proposed hierarchical model. In Fig. 4.9 we can observe how not all labels are represented in the recorded days since it will depend on the life of the person. We can also monitor when the camera wearer goes for lunch to the *cafeteria*, and conclude that s/he

goes almost every day at the same time. We can recognize how *restaurant* always occurs in the evening. With this visualization, we aim to show the consistency of the proposed tool for the monitoring of the time spent by the user at food-related scenes. The automatic and objective discovered information can be used for the improvement of the health of the user.

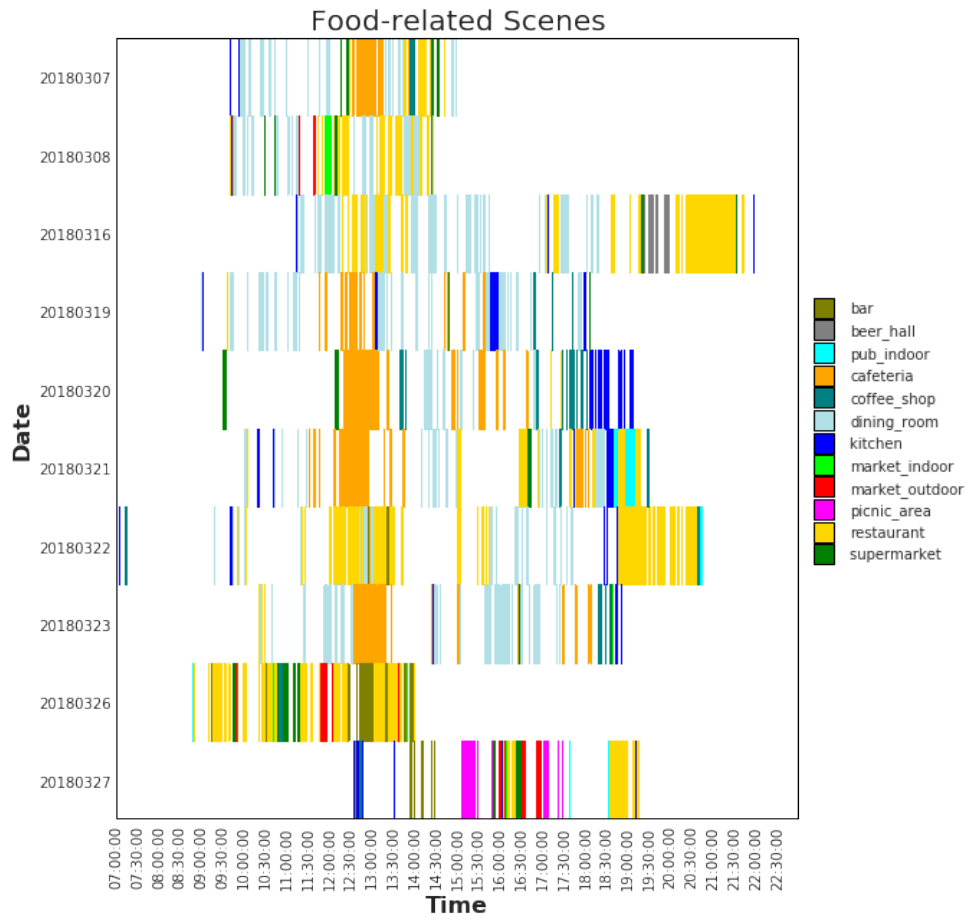


Figure 4.9: Illustration of detected food-related events in egocentric photo-streams recorded during several days by the camera wearer.

4.6 Conclusions

In this paper, we introduced a multi-class hierarchical classification approach, for the classification of food-related scenes in egocentric photo-streams. The contributions of our presented work are three-fold:

- A taxonomy of food-related environments that considers the main activities related to food (eating, cooking, buying, etc.). This semantic hierarchy aims to analyse the food-related activity at different levels of definition. This will allow a better understanding of the behaviour of the user.
- We propose a hierarchical model based on the combination of different layers of deep neural network, mirroring a given taxonomy for food-related scenes classification. This model is easily adapted to other classification problems and implemented on top of other different CNNs and traditional classifiers. The final classification of a given image is computed by combining the intermediate probabilities for the different levels of classification. Moreover, it showed its ability to classify images into meta-classes with high accuracy. This ensures that the final classification label, if not correct, will belong to a similar class.
- A new dataset that we make publicly available. FoodEgoPlaces is composed of more than 33000 egocentric images describing 15 categories of food-related scenes of 11 camera wearers. We publish the data set as a benchmark in order to allow other scientists evaluating their algorithms and comparing their results with ours and with each other. We hope that future research addresses what we believe as a relevant topic: nutritional behaviour analysis in an automatic and objective way, by analysing the user's daily habits from a first-person point of view.

The performance of the proposed architecture is compared with several built baseline methods. We use a pre-trained network on top of which we train our food-related scenes classifiers. However, transfer learning has shown its good performance when addressing problems where the lack of huge amounts of data is a problem. By building on top of pre-trained networks, we achieve results that outperform traditional techniques on the classification of egocentric images into challenging food-related scenes. Moreover and as an incentive, the proposed model has the ability of end-to-end automatically classifying different semantic levels of depth. Thus, specialists can analyze the nutritional habits of people and generate recommendations for improvement of their lifestyle by studying their food-related behaviour either from a broad perspective, such as when the person *eats* or *shops*, or into a more detailed one, like *when the person is eating in a fast-food restaurant*.

The analysis of the eating-routine of a person within its context/environment can help to control his/her diet better. For instance, someone could be interested in knowing the number of times per month that s/he goes to eat somewhere (last layer of the taxonomy). Moreover, our system can help to quantify the time spent at fast-food restaurants, that have shown to negatively affect adolescents health (Jeffery et al., 2006). In a different clinical aspect, the capacities for preparing meal or shopping are considered as one of the main instrumental daily activities to evaluate cognitive decline (Morrow, 1999). Our model allows analysing the custodian activities related to food-scenes represented in the first layer of the taxonomy. Hence, our proposed model integrates a set of food-related scenes and activities, that can boost numerous applications with very different clinical or social goals.

As future work, we plan to explore how to enrich our data using domain adaptation techniques. Domain adaptation allows the adaptation of the distribution of data to other target data distribution. Egocentric datasets tend to be relatively small due to the low-frequency rate of the recording cameras. We believe that by combining techniques of transfer learning, we will be able to explore how the collected dataset can be extrapolated to already available data, sets such as Places2. We expect that the combination of data distributions will improve the achieved classification performance. Therefore, further analysis of this line will allow us to get a better understanding of people's lifestyle, which will give insight into their health and daily habits.

Section 5.3 is taken from:

E. Talavera, N. Strisciuglio, N. Petkov, P. Radeva, "Sentiment Recognition in Egocentric Photostreams," Proceedings of the 9th Iberian Conference on Pattern Recognition and Image Analysis (IBPRIA), Pattern Recognition and Image Analysis, Chapter Springer Verlag, pp. 471-479, 2017.

Sections 5.2 and 5.4 is taken from:

E. Talavera, P. Radeva, N. Petkov, "Towards Egocentric Sentiment Analysis," Proceedings of the 16th International Conference, (EUROCAST), Part II, Lecture Notes in Computer Science, Vol. LNCS 10672, Springer International Publishing, pp. 297-305, 2018.

Chapter 5

Recognition of Induced Sentiment when Reviewing Personal Egocentric Photos

Abstract

Lifelogging is a process of collecting a rich source of information about the daily life of people. The availability and use of egocentric data are rapidly increasing due to the growing use of wearable cameras. In this work, we introduce the problem of sentiment analysis in egocentric events focusing on the moments that compose the images recalling positive, neutral or negative feelings to the observer. Given egocentric photostreams capturing the wearer's days, we propose a method for the classification of the sentiments in egocentric pictures based on global and semantic image features extracted by Convolutional Neural Networks. Such moments can be candidates to retrieve according to their possibility of representing a positive experience for the camera's wearer. We carried out experiments on an egocentric dataset, which we organized in 3 classes on the basis of the sentiment that is recalled to the user (positive, negative or neutral). Our model makes a step forward opening the door to sentiment recognition in egocentric photostreams.

5.1 Introduction

Mental imagery is the process in which a feeling of an experience is invoked by a person in the absence of external stimuli. Therapists assume that it is directly related with emotions (Holmes and et al., 2006), leading to some questions about the effect of images that depict past moments: *Can an image make the process of mental imagery easier?* or *Can specific images help us to invoke feelings and moods?*

Although our mood is influenced by the environment and social context that surrounds us, egocentric data do not always catch our attention or induce the same emotion. We consider that the creation of an electronic diary of positive moments will help to improve the perception of the user of his/her own life. Usually, users are interested in keeping special moments, images with sentiments that will allow them in the future to re-live the personal moments captured by the camera. An automatic tool for sentiment analysis of egocentric images is of high interest to make possible the processing of the big collection of lifelogging data and keeping out just the images of interest i.e. of high charge of positive sentiments.

We approach this problem in this work from two different perspectives. On one hand, we propose to analyse the relation of semantic concepts extracted from images that belong to the same scene. To this end, we defined a classification model where one-vs-all SVM classifiers were trained and evaluated with the features describing semantic and global information from images. On the other hand, we propose to combine semantic concepts, given that they have associated sentiment values (Borth et al., 2013), with general visual features extracted by a CNN (Krizhevsky, Sulskever and Hinton, 2012). Semantic concepts extracted from images represent a finite subset of what is present in the image, not covering the whole image content. Visual features extracted by CNNs can help to summarize the whole image content at an intermediate level.

Our contribution is an analytic tool for positive emotion retrieval seeking events that best represent a pleasant moment to be invoked within the whole set of a day photo-stream. We focus on the event's sentiment description where we are observers without inner information about the event, i.e. from an objective point of view of the moment under analysis.

5.2 Related works

Automatic sentiment image analysis is a complicated task since there is no consensus between the different sentiment ontologies presented in the literature. Table 5.1 illustrates the ambiguity of the problem, reporting several sentiments ontology related to images. The first group (Machajdik and Hanbury, 2010; You et al., 2016; Yi

et al., 2014) assigns 8 main sentiments as excitement, awe or sadness to the images with assigned discrete positive (1) and negative (-1) sentiment value. The second group (Dan-Glauser and Scherer, 2011; Lang et al., 1997) defines a different set of sentiments as valence or arousal and discrete positive (1), neutral (0), or negative (-1) values assigned to the images according to the sentiments. In contrast, the third group (Nojavanasghar and et al., 2016) assigns up to 17 sentiments (6 basics and 9 complex) and each image of the dataset is assigned a continuous value in a scale from 1 to 4. Given the ambiguity of the semantic sentiment assignment, with labels difficult to classify into positive or negative sentiments, the last group (Borth et al., 2013) defines up to 3244 Adjective Noun Pairs (ANP) (e.g. 'beautiful.girl') and assigns a continuous sentiment value in a range of [-2,2] to them. The main idea is that the same object according to its appearance has positive or negative sentiment value like '*angry.dog*' (-1.55) and '*adorable.dog*' (+1.45). A natural question is until which extent the 3244 ANPs represent a scene captured by the image, taking into account the difficulty to detect them automatically (Mean average accuracy $\sim 25\%$).

| DataSets | Source | Images | Semantic sentiment labels | Sentiment Values |
|--|--------------------|---------------|--|-------------------------------|
| Abstract & Artphoto (Machajdik and Hanbury, 2010) | | 280 & 806 | positive: contentment, amusement, excitement, awe, negative: sadness, fear, disgust, and anger | {1,-1} |
| You's Dataset (You et al., 2016) | Flickr Instagr | 23000 | positive: contentment, amusement, excitement, awe, negative: sadness, fear, disgust, and anger | {1,-1} |
| CASIA-WebFace (Yi et al., 2014) | | 494k | anger, disgust, fear happy, neutral, sad, surprise | [1,0,-1] |
| IAPS(Lang et al., 1997) | | 1182 | valence, arousal, and dominance | [1,7] |
| GAPED (Dan-Glauser and Scherer, 2011) | | 732 | valence, arousal, and normative significance | {1,0,-1} |
| EmoReact (Nojavanasghar and et al., 2016) | Youtube | 1102 clips | 17 sentiments: 6 basic emotions (positive: happiness, surprise, negative: sadness, fear, disgust, and anger), and 9 complex emotions: (curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, frustration). | [1,4] |
| VSO + TwitterIm (Borth et al., 2013) | Flickr Twitter | 0.5M | Not, but Adjective Noun Pairs (3244) | Flickr[-2,2] Twitter[-1,1] |
| You_RobustSet (You and et al., 2015) | Twitter | 1269 | Non-semantic labels: Positive and Negative | {1,-1} |
| UBRUG- EgoSenti* | Wearable Camera | 12088 | Non-semantic labels: Positive, Neutral and Negative | {1,0,-1} |

Table 5.1: Different image sentiment ontologies.

Given the difficulty of image sentiment determination, ambiguity and lack of consensus in the bibliography, added by the difficulty given by the egocentric im-

ages, we focus on the image sentiment as a discrete value expressing a ternary sentiment value (positive (1), negative (-1) or neutral (0) value) similar to (You and et al., 2015). Egocentric data is of special difficulty, since we do not observe the wearer and his/her, i.e. from facial or corporal expressions, but rather from the perspective of what the user sees. Moreover, in real life, fortunately, negative emotions have much less prevalence than neutral and positive, that makes very difficult to have enough examples of negative egocentric images and events. Thus, the problem we address in this article is what effect an egocentric image or event has on an observer (positive, neutral or negative) (see Fig.5.1), instead of attempting to specify an explicit semantic image sentiment like sadness; and how to develop automatic tool for sentiment value detection (positive, vs. neutral vs. negative) and egocentric dataset in order to validate its results. Going further, in contrast to the published work, we claim to automatically analyse the sentiment value of egocentric events i.e. a group of sequential images that represents the same scene. In the case of egocentric images, the probability that a single image describes an event is low; there are a lot of images that just capture wall, sky, ground or partially objects. For this reason, we are interested to automatically discover how the event captured by the camera influences the observer, that is to automatically determine the ternary sentiment values of the events, which are richer in information and involve the whole moment's experience. For example, an event being in a dark and narrow, grey space would influence negatively, a routine scene like working in the wearer's office could influence the observer neutrally and an event where the wearer has spent some time with friends in a nice outdoor space could influence positively to the observer.

Automatic sentiment analysis from images is a recent research field. In the literature, sentiment recognition in conventional images has been approached by computing and combining visual, textual, or audio features (Nojavanasghar and et al., 2016; Poria and et al., 2014; Wang et al., 2014; You and Et, 2016). Other characteristics, such as facial expressions have also been used for sentiment prediction (Yuan and et al., 2013). The combination of visual and textual features extracted from images is possible due to the wide use of online social media and microblogs, where images are posted accompanied by short comments. Therefore, multimodal approaches were proposed, where both sources of information are merged (Wang et al., 2014; You and Et, 2016) for automatic sentiment value detection.

Recently, with the outstanding performance of Convolutional Neural Networks (CNN), several approaches to sentiment analysis relied on deep learning techniques for classification and/or features extraction combined with other networks or methods (Campos and et al., 2015; Levi and Hassner, 2015; You et al., 2016; Yu et al., 2016). The work in (You et al., 2016) applies fine-tuning on the AlexNet to classify the 8 emotions: sadness, angry, content, etc. In contrast, in (Campos and et al., 2015)

they propose to fine-tuned CaffeNet with oversampling to classify into Positive or Negative sentiments. In (Levi and Hassner, 2015) a novel transformations of image intensities to 3D spaces is proposed to reduce the amount of data required to effectively train deep CNN models. In (Yu et al., 2016) the authors use logistic regression to classify into 3 sentiments using CNN features. In (Chen et al., 2014), the authors perform a fine-tuning on a CNN model and modify the last layer to classify 2089 ANPs. However, no work has addressed the sentiment image and event analysis in egocentric datasets.

5.3 Sentiment detection by global features analysis

In this section, we describe the proposed method for sentiment recognition from egocentric photo-streams, which is based on visual (extracted by CNN) and semantic (in terms of ANPs) features extracted from the images. An architectural overview of the proposed system is depicted in Fig. 5.2.

a) Temporal Segmentation:

Given that egocentric images have a smaller field of view and thus do not capture entirely the context of the event, we need to detect the events of the days. To this aim, we apply the SR-Clustering algorithm for temporal segmentation of photo-streams (Dimiccoli et al., 2017). The clustering procedure is performed on an image representation that combines visual features extracted by a CNN with semantic features in terms of visual concepts extracted by Imagga's auto-tagging technology (<http://www.imagga.com/solutions/auto-tagging.html>).

b) Features Extraction:

For the computation of the semantic features in terms of the ANPs, we use the DeepSentiBank Network (Chen et al., 2014). Given an image, the DeepSentiBank network considers the 2089 best performing ANPs. Applying the DeepSentiBank on them gives a 2089-D feature vector, where the feature values correspond to the ANPs likelihood in the image. These values are multiplied by the sentiment value associated with the concepts. Note that each ANP has a positive or negative sentiment value assigned, but not 0 for neutral sentiment.

However, the 2089 ANPs not necessarily have the power to explain the "richness" of any scene in an image. Hence, we integrate the ANPs feature vector with a feature descriptor provided by the penultimate layer of a CNN (Krizhevsky, Sutskever and Hinton, 2012) that summarizes the whole context of the image. The resulting



Figure 5.1: Examples of Positive (green), Negative (red) and Neutral (yellow) images.

feature vector is composed of 4096 features. We combine the ANPs and the CNN feature vectors into a 6185-D feature vector, in order to construct a more reliable and rich image representation that relates image semantics expressed by the ANPs with clear sentiment value with the CNN cues as an intermediate image representation. We apply the Signed Root Normalization (SRN) to transform the CNN feature vectors to a more uniformly distributed space followed by a l_2 -normalization (Zheng et al., 2014).

c) Classification:

We use the proposed feature vectors to train a multi-class SVM classifier due to its high generalization capability (Joachims, 2000). This is ensured by the SVM learning algorithm that finds a separation hyperplane that maximizes the separation margin

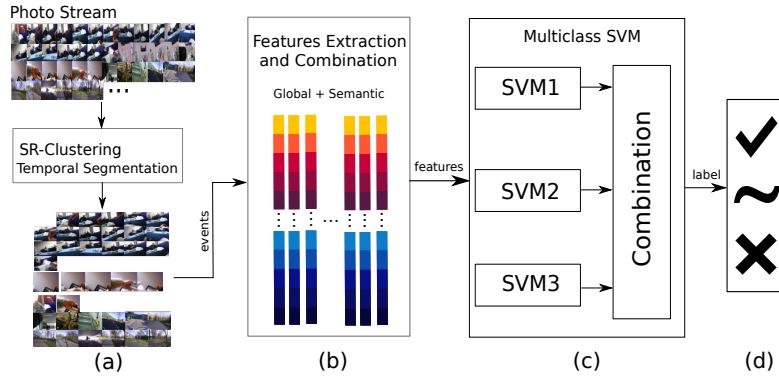


Figure 5.2: Architecture of the proposed method. (a) Temporal segmentation of the photo-stream into events. (b) CNN and ANPs features are extracted from the images and (c) used as input to the trained multi-class SVM model. (d) The model labels the input image as Positive, Neutral or Negative.

between the classes. We employ a 1-vs-all design for the multi-class problem, as suggested in (Foggia et al., 2015). The cardinality of the classes in the proposed dataset is not balanced, which affects the computation of the training error cost on the corresponding positive and negative samples. We set the cost of the training error on the positive and negative class according to their cardinality for each SVM of the pool of classifiers. In the implementation of the SMVs, we set the training error costs according to the ratio $r = n^-/n^+$, where n^+ and n^- are the number of positive and negative examples, respectively. At this stage, the decision of the classifier is taken at image level. To classify an event, we use a majority vote on the image level classification output.

5.3.1 Experimental Setup

Data set

We collected a dataset of 12471 egocentric pictures, which we call UBRUG-EgoSenti. The users were asked to wear a Narrative Clip Camera, which takes a picture every 30 seconds, hence each day around 1500 images are collected for processing. The images have a resolution of 5MP and JPG format.

We organize the images into events according to the output of the SR-clustering algorithm (Dimiccoli et al., 2017). From the originally recorded data, we discarded those events that are composed of less than 6 images, so obtaining a dataset composed of 12088 images grouped in a total of 233 events, with an average of 51.87

images per event and std of 52.19. We manually labelled the events following how the user felt while reviewing them by assigning *Positive*, *Negative* and *Neutral* values to them, some examples of which are given in Fig 5.1. The dataset, for which the details are in Table 5.2, is publicly available and can be downloaded from: <http://www.ub.edu/cvub/dataset/>.

| Class | Images | #Events | Mean Im Event | Std Im Event |
|----------|--------|---------|---------------|--------------|
| Positive | 4737 | 83 | 57.07 | 52.34 |
| Neutral | 6169 | 107 | 57.65 | 57.18 |
| Negative | 1182 | 43 | 27.49 | 26.44 |
| Total | 12088 | 233 | 51.88 | 52.19 |

Table 5.2: Description of the UBRUG-EgoSenti dataset.

Experiments and Results

We carried out 10-fold cross-validation. Events from different classes are uniformly distributed among the various folds, which are thus independent of each other. We evaluated the performance of the proposed system on single images and at event level. For the UBRUG-Senti dataset, the ground truth labels are given at event level. All the images that compose a certain event, are considered as having the same label of such event. Given an event composed of M images, we aggregate the M classification decisions by majority vote. We measure the performance results of our method by computing the average accuracy.

| | Image Classification | | | | | Event Classification | | | | |
|-----------------------|----------------------|------|------|-------|-------|----------------------|------|------|-------|-------|
| | Pos | Neg | Neu | All | | Pos | Neg | Neu | All | |
| | mean | | | mean | std | mean | | | mean | std |
| Semantic Features | 59.2 | 42.4 | 44.4 | 48.67 | 22.87 | 71.2 | 42 | 47.3 | 53.50 | 30.77 |
| CNN Features | 70 | 61.3 | 45.7 | 59.00 | 22.80 | 80.8 | 71 | 48.9 | 66.90 | 27.67 |
| Semantic+CNN Features | 72 | 60.8 | 46 | 59.60 | 23.17 | 82.1 | 73.5 | 48.9 | 68.17 | 30.07 |

Table 5.3: Performance results achieved at image and event level.

In Table 5.3, we report the results achieved by the proposed methods at image and event level. We achieved an average image classification rate of 59.60% with a standard deviation of 23.17, when we apply the proposed method. The average event classification rate is 68%, when the proposed features are employed, which corresponds to 82%, 73.5% and 49% for positive, negative and neutral events, respectively. Up to our knowledge, unfortunately, there is no work in the literature on egocentric image sentiment recognition neither event sentiment recognition to compared with. Even the works on image sentiment analysis in conventional images

(Campos and et al., 2015; Levi and Hassner, 2015; You et al., 2016; Yu et al., 2016) use different datasets and objectives (8 semantic sentiments vs. binary or ternary sentiment values) that make difficult their direct comparison. Fig. 5.3 shows some example results. As can be seen, the algorithm learns to classify events with the presence of routine objects into *neutral* events. Events wrongly classified as *neutral* are shown in Fig. 5.3(left) and Fig. 5.3(middle). As an example, the last row of Fig. 5.3(left) is classified as *neutral*, probably due to the presence of the *pc* in the image, while it was manually labelled as *positive*, because it shows social interactions. As for Fig. 5.3(left) and Fig. 5.3(right), events were mislabelled as *negative* probably due to the “homogeneity” and “greyness” of the images within the events, e.g. events were considered as *negative* when most of the information in the image corresponded to the asphalt of the road.

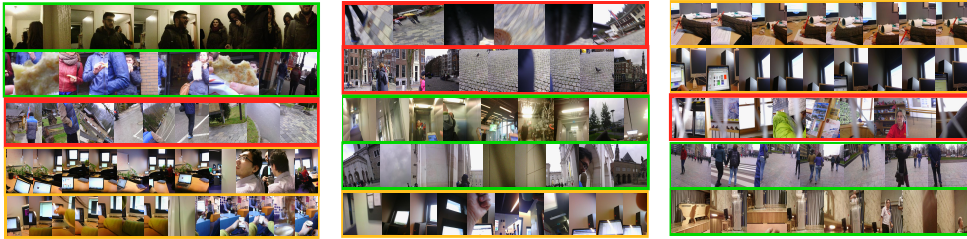


Figure 5.3: Examples of the automatic event sentiment classification. The events are grouped based on the sentiment defined by the user: (right) Positive, (middle) Negative, and (left) Neutral. The events frame colour corresponds to the label given by the model: Positive (green), Negative (red) and Neutral (yellow).

5.4 Sentiment detection by semantic concepts analysis

Given an egocentric photo-stream, we propose scene emotion analysis seeking for events that represent and can retrieve a positive feeling from the user. We apply event-based analysis since single egocentric images cannot capture the whole essence of the situation. By combining information from several images that represent the same scene, we get closer to a better understanding of the event.

a) Temporal segmentation:

We apply temporal segmentation on the egocentric photo-streams using the proposed method in (Dimiccoli et al., 2017). The clustering procedure is performed on an image representation that combines visual features extracted by a CNN with

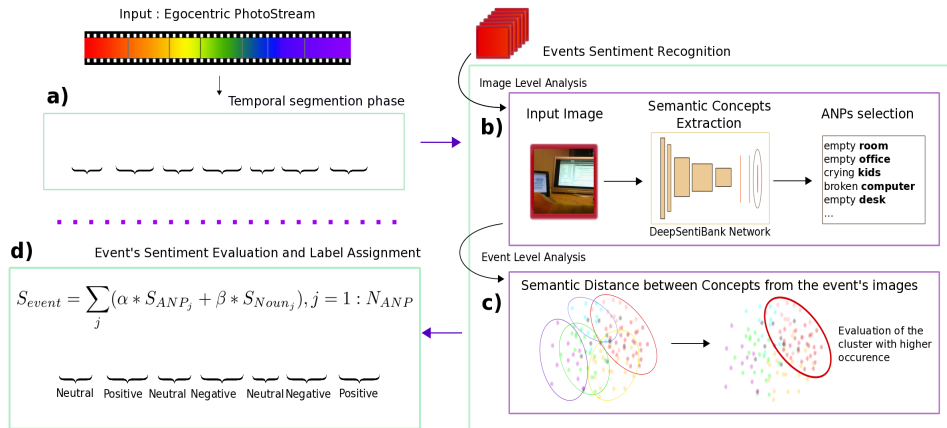


Figure 5.4: Sketch of the proposed method. First, a temporal segmentation is applied over the egocentric photo-stream (a). Later, semantic concepts are extracted from the images using the DeepSentiBank (Chen et al., 2014) (b). The semantic concepts with higher occurrence are selected as event descriptors (c). Finally, the ternary output is obtained by merging the sentiment values associated to the event’s semantic concepts (d).

semantic features in terms of visual concepts extracted by Imagga’s auto-tagging technology¹. In Fig. 5.3 we present some examples of events extracted from the dataset, we introduce below.

b) Event’s sentiment recognition:

The model relies on semantic concepts extracted from the images to infer the event sentiment associated. However, it relies not only on the semantic concepts extracted by the net with their associated sentiment, but also on how those semantic concepts can be interpreted by the user. We apply the DeepSentiBank Convolutional Neural Network(Chen et al., 2014) to extract the images semantic information since it is the only introduced model that extract semantic concepts (ANPs) with sentiment values associated. Given an image, the output of the network is a 2089-D feature vector, where the values correspond to the ANPs likelihood in the image.

Besides taking into account the sentiment associated with the ANPs, the influence of the common concepts within an event are also analysed. We categorize the noun into Positive, Neutral or Negative. There is a wide range of semantic concepts within the ontology, but many of them seem to repeat concepts that even from the

¹<http://www.imagga.com/solutions/auto-tagging.html>

user perspective would be difficult to differentiate when looking at an image; such as 'girl' from 'woman' or 'lady'.

When facing our egocentric images challenge, the VSO presents several drawbacks. On one hand, this tool is trained to recognise up to 2089 concepts, which can not describe all possible scenarios. On the other hand, despite including that big amount of concepts, many of them categorize objects into categories difficult to visually interpret or differ by the human eye. Examples can be the distinction between 'child', 'children', 'boy', or 'kid' from an image. To overcome this problem, we generate a parallel ontology with what we consider an egocentric view of the concepts, i.e., we cluster the concepts a person would merge based on their semantic.

Egocentric analysis of the VSO: We cluster the semantic concepts based on the similarities between the noun components of the ANPs, which are computed using the wordNet tool². Following what would be considered as similar from an egocentric point of view, we manually refine the resulted clusters into 44 categories. We label the clusters as Positive, Neutral or Negative. In Table 5.4 we present some of the ego-semantic clusters.

| Positive | | | Neutral | | | Negative | | |
|----------|--------------|---------|-------------|-----------|-----------|-----------|-----------|-----------|
| petals | christmas | award | car | study | bible | tumb | bug | nightmare |
| rose | winter | present | cars | science | book | tombstone | bugs | accident |
| flora | snow | honor | machine | history | card | monument | insect | shadows |
| park | santa | gift | vehicle | economy | stiletto | grave | worm | noise |
| yard | sketch | heroes | rally | market | sins | memorial | cockroach | scream |
| plant | cartoon | dolls | train | industry | record | stone | decay | night |
| garden | drawing | dolls | competition | statue | paper | graveyard | garbage | darkness |
| | comics | toy | race | sculpture | poem | cemetery | trash | shadow |
| | illustration | toys | control | museum | interview | grief | shit | |
| | humor | lego | metal | | | pain | | |

Table 5.4: Examples of clustered concepts based on their semantic similarity, initially grouped following the distance computed by the WordNet tool.

5.4.1 Sentiment Model

Given an event, the event's sentiment analysis model (see Fig. 5.4) performs as follows;

1. Given the ego-photo-stream we apply the temporal segmentation, analyse events with a minimum of 6 images, i.e. that last for at least 3 minutes.
2. Extract the ANPs of each event frame and rank them by their probability ($Prob_{ANP_i}$) of describing an image.

²<http://wordnet.princeton.edu>

3. Select the top-5 ANPs per image, since we consider that those are the concepts with higher relevance, thus better capturing the image’s information. After this step, the model ends up with a total of M semantic concepts per event, where $\{M = \text{Number of images} \times 5\}$.
4. Cluster the M semantic concepts based on their Wordnet-based nouns semantic distances. As a result, we have clusters of concepts with semantic similarity. For the event sentiment computation (S_{event}), focus on the largest cluster.
5. Finally, fuse the sentiment associated with the ANPs and noun’s cluster following the eq. (5.1):

$$S_{event} = \sum_j (\alpha * S_{ANP_j} + \beta * S_{Noun_j}), j = 1 : N_{ANP}, \quad (5.1)$$

where $S_{ANP_j} = (S_{ANP_j}^{VSO} * Prob_{ANP_j})$, $S_{ANP_j}^{VSO}$ is the ANP’s sentiment given by the VSO and S_{Noun} is the label of the noun, α and β are the contributions (%) of the ANPs and the nouns. Take into account the probability associated to the ANPs aiming to penalize the ANPs with low relation to the image content.

5.4.2 Experimental Setup

Data set

We collected a dataset of 4495 egocentric pictures, which we call UBRUG-Senti. The user was asked to wear the Narrative Clip Camera³ fixed to his/her chest during several hours every day and was asked to continue with his/her normal life. Since the camera is attached to the chest, the frames vary following the user’s movement and describe the user’s view of his/her daily indoor/outdoor activities. It involves challenging backgrounds due to the scene variation, handled objects appearing and disappearing during images sequences, and the movement of the user. The camera takes a picture every 30 seconds, hence each day around 1500 images are collected for processing. The images have a resolution of 5MP and JPG format.

After the temporal clustering (Dimiccoli et al., 2017), we obtained a dataset composed of 4495 images grouped in a total of 98 events. The events were manually labelled based on how the user felt while reviewing them. The labels assigned were *Positive* (36), *Negative* (43) and *Neutral* (19). Some examples are given in Fig 5.3.

³<http://getnarrative.com/>

Experiments and Results

During the experimental phase, we evaluated the contributions of ANPs and nouns by defining different combinations of α and β . We performed a balanced 5-fold cross-validation. For each of the folds, we used 80% of the total of events per label of our dataset and compute the best pair of α and β values. This is a parameter selection process that is later re-evaluated in a test phase with a different set of events.

Validation: To evaluate the effectiveness of the scene detection approach, we use the *Accuracy*, as the rate of correct results, and the *F-Score* (F1). The F1 is defined as : $F1 = 2(RP)/(R + P)$, where P is the precision ($P = TP/(TP + FP)$), R is the recall ($R = TP/(TP + FN)$) and TP , FP and FN respectively are the number of true positives, false positives and false negatives of the event’s sentiment label correctly identified.

Results: Tables 5.5 and 5.6 present the results achieved by the proposed method at image and event level, respectively. The model achieves an average training accuracy of $73\pm 3.8\%$ and F-score of $59\pm 5.4\%$ and test accuracy of $75\pm 8.2\%$ and F-score of $61\pm 13.2\%$, when $\alpha = 0.8$ and $\beta = 0.2$, i.e. when the ANP information is considered; although the major contribution comes from the noun sentiment associated. As expected, neutral events are the most challenging ones to classify.

| | Accuracy | | | F-Score | | |
|----------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| | beta = 0.2 alpha = 0.8 | beta = 0.5 alpha = 0.5 | beta = 0.8 alpha = 0.2 | beta = 0.2 alpha = 0.8 | beta = 0.5 alpha = 0.5 | beta = 0.8 alpha = 0.2 |
| Ours | 0.60 | 0.63 | 0.73 | 0.35 | 0.43 | 0.59 |
| Evaluating 3 Clusters | 0.68 | 0.66 | 0.68 | 0.48 | 0.45 | 0.48 |
| Evaluating with weights | 0.65 | 0.65 | 0.66 | 0.41 | 0.43 | 0.47 |

Table 5.5: Parameter-selection results

In order to contextualize our results, we fine-tune the well-known *GoogleNet* deep convolutional neural network (Ma et al., 2016) to classify into Positive, Neutral and Negative. We use 80%, 10% and 10% of the dataset for training, validation and testing respectively. The network achieves an accuracy of 55%.

From the results we can conclude that the application of the DeepSentiBank presents drawbacks when applied to egocentric photo-streams. To begin with and as commented before, the 2089 ANPs not necessarily have the power to represent what the image captured about the scene, taking into account the difficulty to detect them automatically (Mean average accuracy of the net $\sim 25\%$). Moreover, the ANPs

| | Accuracy | F-Score |
|----------------------------|---------------------------|------------------|
| | beta = 0.8 alpha = 0.2 | |
| Ours | 0.75±0.08 | 0.60±0.13 |
| Evaluating 3 Clusters | 0.69±0.1 | 0.50±0.15 |
| Evaluating with weights | 0.74±0.1 | 0.58±0.15 |

Table 5.6: Test set results

present the limitation that they are classified strictly into Negative or Positive concepts. Thus, moments from our daily routine, which are often considered as neutral, are difficult to recognize.

Sentiments recognition from an image or a collection of images is a difficult process due to its ambiguity. A challenge in the model construction for sentiment recognition consists in taking into account the bias due to the subjective interpretation of images by different users. Furthermore, the boundaries between neutral/positive and neutral/negative sentiments are not clearly defined. A *neutral* feeling is difficult to interpret. From the results, we observe that *neutral* events are the most challenging ones to classify. Another challenging aspect concerns the grouping of image sentiments into event sentiment, since events can have non-uniform sentiments.

A further step towards better understanding of the image and sentiment analysis is needed, due to the subjectivity of what an image can recall to different persons. To this aim, having annotations by different persons is critical to evaluate the inter- and intra-observer variability.

From the results, the intuition that we get is that non-routine events and specially when moments are social, have a higher probability of being positive. In contrast, routine events will most probably be considered as neutral. Negative events as accidents have low prevalence to be learned. Yet, hostile and empty environments could lead to negative sentiments too. Future works will address the study of emotional events and their relation to daily routine.

5.5 Discussion and conclusions

In this work, we proposed for the first time models and a dataset for sentiment recognition from egocentric images and events, recorded by a wearable camera.

Sentiment recognition from an image or a collection of images is a difficult process due to the subjectivity of the task. A challenge in the model construction for

sentiment recognition consists in taking into account the bias due to the subjective interpretation of images by different users. Furthermore, the boundaries between neutral/positive and neutral/negative sentiments are not clearly defined. A *neutral* feeling is difficult to interpret. From the results, we observe that *neutral* events are the most challenging ones to classify. Another challenging aspect is the fact that events are represented by groups of images that do not necessarily share the same associated sentiment. Thus, by giving a sentiment label to an event, we extrapolate it to the images that compose it, being aware that this might imply some errors.

In (Talavera, Radeva and Petkov, 2017) we first introduced a labelled dataset composed of 98 events. Later, in (Talavera, Strisciuglio, Petkov and Radeva, 2017) we extended it to 233 events, grouping 12088 images, from 20 days recorded by 3 different users.

The first proposed approach is based on the extraction of CNN and semantic features with associated sentiment value. It analyses semantic concepts called Adjective-Noun-Pairs (ANPs) extracted from the images, which have an associated sentiment value and describe the appearance of concepts in the images. The sentiment prediction tool is based on new semantic distance of ANPs and fusion of ANPs and nouns sentiments extracted from egocentric photo-streams. This model obtained a classification accuracy of 75% on the test set, with a deviation of 8% over the first version of the dataset. The second proposed approach is based on a classification model where one-vs-all SVM classifiers were trained and evaluated with the features describing semantic and global information from the images. Using the proposed method we obtained average events and image sentiment accuracy of 68.17% and 58.60%, with a standard deviation of 30.07% and 23.17%, respectively.

Analysing the obtained results, we conclude that the polarity of the ANPs makes it difficult to classify '*Neutral*' events. However, most of our daily life is composed of such events, which can be considered as routine. Furthermore, we get the intuition that non-routine events have a higher probability of being positive, especially when moments are social. In contrast, routine events will most probably be considered as neutral. Negative events as accidents have low prevalence to be learned. Yet, hostile and empty environments could lead to negative sentiments too. Future works will address the study of emotional events and their relation to daily routines.

A further step towards a better understanding of the image and sentiment analysis is needed, due to the subjectivity of what an image can invoke to different persons. To this aim, having annotations by different persons is critical to evaluate the inter- and intra-observer variability. Moreover, future experiments will address the generalization of the model over datasets collected by other wearable cameras, as well as recorded by different users.

Published as:

E. Talavera, A. Cola, N. Petkov, P. Radeva, "Towards Egocentric Person Re-identification and Social Pattern Analysis", 1st Conference on Applications of Intelligent Systems (APPIS), pp. 203-211, published in the proceedings in the series Frontiers in AI and Applications (IOS Press), 2018.

Chapter 6

Towards Egocentric Person Re-identification and Social Pattern Analysis

Abstract

Wearable cameras capture a first-person view of the daily activities of the camera wearer, offering a visual diary of the user behaviour. Detection of the appearance of people the camera user interacts with for social interactions analysis is of high interest. Generally speaking, social events, life-style and health are highly correlated, but there is a lack of tools to monitor and analyse them. We consider that egocentric vision provides a tool to obtain information and understand users social interactions. We propose a model that enables us to evaluate and visualize social traits obtained by analysing social interactions appearance within egocentric photo-streams. Given sets of egocentric images, we detect the appearance of faces within the days of the camera wearer, and rely on clustering algorithms to group their feature descriptors in order to re-identify persons. Recurrence of detected faces within photo-streams allows us to shape an idea of the social pattern of behaviour of the user. We validated our model over several weeks recorded by different camera wearers. Our findings indicate that social profiles are potentially useful for social behaviour interpretation.

6.1 Introduction

Human social behaviour involves how people influence and interact with others, and how they are affected by others. This behaviour varies depending on the person, and is influenced by ethics, attitudes, or culture (Allport, 1985). Understanding the behaviour of an individual is of high interest in social psychology. House et al. addressed the problem of how social relationships affect health (House et al., 1988) and demonstrated that social isolation leads to major risk factors for mortality. Moreover, Yang et al (Yang et al., 2016) observed that lack of social connections is associated with health risks in specific life stages, such as the risk of inflammation in adolescence, or hypertension in old age. Also, as in Kawachi et al. (Kawachi and Berkman, 2001) was highlighted, social ties have a beneficial effect in order to maintain psychological well-being.

Considering the importance of the matter, automatic discovery and understanding of the social interactions are of high importance to the scientists, as they remove the need for manual labour. On the other hand, egocentric cameras are useful tools as they offer the opportunity to obtain images of the daily activities of users from their own perspective. Therefore, providing a tool for automatic detection and characterization of social interactions through the visual recorded visual data can lead to personalized social pattern discoveries.

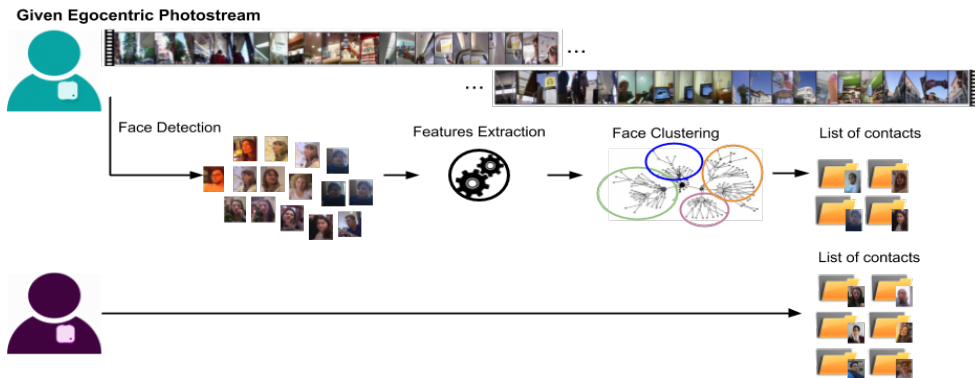


Figure 6.1: Architecture of the proposed model. First, we apply the Viola&Jones algorithm (Viola et al., 2001) to detect appearing faces in the photostreams. Later, we apply the OpenFace tool to convert the faces to feature vectors. We propose to define the re-identification problem as a clustering problem with a later analysis of the grouped faces occurrence.

The rest of the chapter is organized as follows: in Section 6.2, we present an overview of the recent person re-identification approaches. In Section 6.3, we define the Social Pattern Analysis approach, by performing face detection, faces clustering

and evaluation of their occurrence. In Section 6.4, we describe the experimental setup. Finally, in Section 6.5 we discuss our findings and draw conclusions and future research lines.

6.2 Related works

Social patterns analysis is commonly addressed as a re-identification problem. Re-identification usually considers information about detected faces in order to facilitate future detection.

People re-identification is of high importance in the area of video surveillance, often identifying pedestrians recorded by different cameras (Bak and Carr, 2017; Chen et al., 2017; Fan et al., 2017; Li et al., 2017; Zhao et al., 2013). Security cameras generally capture the appearance of people, offering information such as clothes and pose. Challenges within this area of research include the changes in views, illumination and human body poses.

Zhao et al. (Zhao et al., 2013) proposed to incorporate saliency maps and patch matching for unsupervised person re-identification. In (Li et al., 2017), a deep neural network was introduced to learn and localize pedestrians, to later learn features from the body and parts of the body of the recorded people. Munawar et al. introduced a Convolutional Neural Network (CNN) to simultaneously register and represent faces (Hayat et al., 2017). Chen et al. addressed the problem of cross-camera variations and proposed a hashing based approach (Chen et al., 2017). They proposed to transform high-dimensional feature vector through a binary coding scheme to compact binary codes that preserve identity.

The methods introduced above evaluate images recorded by several surveillance cameras. These recorded images usually capture the body of the people and the local context. However, person re-identification from egocentric photo-streams is different from person re-identification from datasets recorded by security cameras. On the other hand, egocentric images are commonly recorded by chest-worn cameras, they show social interactions from a closer perspective. Therefore, the face of the opposite person of the wearer commonly represents the main source of information about the person.

Chenyou et al. targeted to establish person-level correspondence across first- and third-person videos (Fan et al., 2017). To this end, they proposed a semi-Siamese CNN architecture. In (Aghaei et al., 2017) the authors addressed social signal analysis from egocentric photo-streams. To this end, they proposed to reach a characterization of the social pattern of a wearable photo-camera user through first, detection of social interactions, and later categorization of detected social interaction into ei-

ther a formal or informal meeting. Through applying a face clustering algorithm, people forming the social environment of the user are localized throughout the social events of the user. This allows the social pattern characterization of the user through quantifying the density, diversity and social trends of the user.

The work we present goes one step forward and proposes to compare social patterns of different individuals by analysing their social behaviour from various aspects. After applying clustering of the detected faces, our proposed model quantifies the occurrence and duration of the interactions of the user with different individuals. Therefore, in this work, we do not consider information about previously detected faces, but we rely on clustering techniques over the camera users data, in order to find the occurrence of the appearance of people around them. We consider that for our problem, this occurrence can be considered as re-identifying people along recorded days of the user.

6.3 Social Patterns Characterization

People tend to interact with others along their day. By using wearable cameras, an egocentric perspective of those specific activities is captured. However, in order to address the analysis of social patterns of individuals, we first need to detect the appearance of people with whom the user interacts. Therefore, our approach towards social pattern analysis proposes to combine face detector and face clustering algorithms.

6.3.1 Person Re-Identification

Face detection is performed by applying the well-known Viola-Jones algorithm (Viola et al., 2001). Our model translates the re-identification problem to a clustering problem. We rely on average linkage Agglomerative Hierarchical Clustering (AHC) to find the relation of the feature vectors extracted from the detected faces.

The consistency of the clusters is important for the later analysis of the occurrence of people within photostreams. We propose to estimate the robustness of the clusters by computing the Pearson's correlation coefficient among their samples. It is a measure of the linear relationship between two quantitative random variables, x and y , that in our problem represent the two feature vectors describing faces. Unlike covariance, Pearson's correlation r_{xy} is independent of the measurement scale of the variables being determined as:

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1) s_x s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

where n is the sample size, \bar{x} and \bar{y} are the samples mean, and x_i and y_i are the single samples indexed with i .

The coefficient values have normalised values from -1 to 1, representing no similarity and identity, respectively. We discard an image within a cluster if the similarity coefficient does not reach a minimum value. The cluster consistency is checked when the mean of the distance of the images feature descriptors is between 0.4 and 0.8. These empirical values are selected after running several experimental tests. A mean value higher than 0.8 is considered as robust. Inconsistent clusters were removed if the mean similarity value among their elements was less than 0.4. We defined a minimum similarity coefficient of 0.70. Clusters or images not following this consistency constraint are discarded and not later evaluated as social interactions.

When the clusters are found, and their consistency checked, we consider that people re-appear in the photostream when their faces belong to the same cluster. The temporal appearance of the people surrounding the camera wearer throughout photostreams describes the social pattern of behaviour of the wearable camera user. Therefore, after applying the clustering algorithm, the resulted groups can be analysed to find the occurrence and duration of relations.

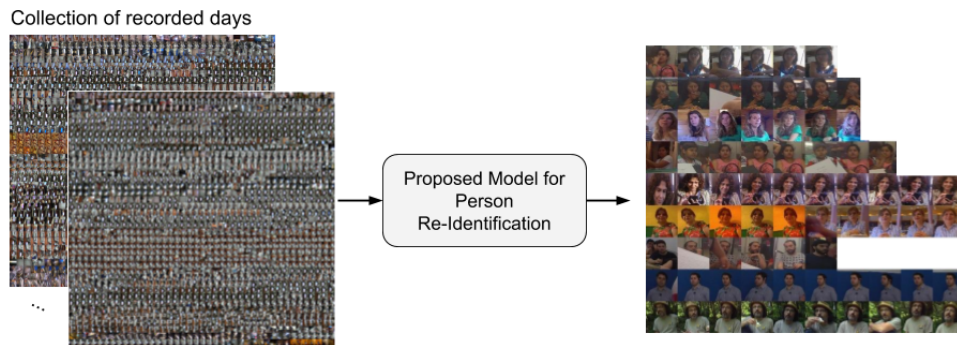


Figure 6.2: Samples of the clusters obtained by applying our model for person re-identification on a set of recorded days by a camera wearer.

We define several constraints when analysing the obtained clusterings due to the problem characteristics. On one hand, we seek sequential images with the appearance of a person. The wearable camera used for this experiment takes pictures every 20-30 second. We describe an event as a group of consecutive images of a duration of a minimum of 3 minutes. Therefore, a cluster composed of images describing an event shorter than that time is not considered as an event, and consequently nor as social interaction of interest. On the other hand, the movement of the camera

follows the movement of the user. This leads to challenging pose changes of the person, the camera wearer is interacting with, which makes faces detection and description challenging. Therefore, the appearance of faces can be either missed by the face detector algorithm, or not recognized by the neural network that extracts the face descriptors. Hence, we introduce a 15 minutes time-frame gap between temporal sequential images when calculating the duration of a social interaction.

6.3.2 Social Profiles Comparison

A visual description of the extracted information is helpful when the aim is to interpret and compare social profiles of several users. This allows us to easily gain understanding of social behavioural differences among people. Since social patterns describe the camera wearer social interactions, a social pattern profile relates to a single individual, describing a personalized social behaviour. Information such as *time spent interacting with others*, or on the contrary, the *time the user spent alone*, allow us to get insight into the social behaviour of a person. Although camera users do not wear the camera 24h per day, they wear it along hours when most of the activities of their daily living occur. In this work, we propose the estimation of the following basic social characteristics through the analysis of the recorded hours of the camera users:

- *Num p/day*: Average number of persons with whom the user interacts per days.
- *Inter/day*: Average number of social interactions per day.
- *T/P*: Average number of minutes spent with every person.
- *T/Inter*: Average number of minutes spent per interaction.
- *T/A*: Average number of minutes spent alone per day.

6.4 Experiments

6.4.1 Dataset

We collected data from 4 subjects who were asked to wear Narrative Clip camera (<http://getnarrative.com/>) fixed to their chest. This camera has a resolution of 2fpm and a normal lens. They captured information about their social daily routine, taking pictures of the people with whom they interacted. Since there is no training involved in this approach, the whole dataset is analysed by the proposed model. The dataset is organized as follows:

- *User 1*: A set of 8766 images, composed by 1627, 1384, 395, 1490, 643, 1376, 1851 images per day, respectively.
- *User 2*: A set of 10916 images, composed by 1395, 1587, 1926, 1615, 1643, 1371, 1379 images per day, respectively.
- *User 3*: A set of 9053 images, composed by 729, 1601, 1055, 1753, 1302, 1434, 1179 images per day, respectively.
- *User 4*: A set of 5343 images, composed by 780, 665, 747, 844, 809, 760, 792 images per day, respectively.

6.4.2 Experimental setup

We tested the face appearance detector over a test-set of 317 images, where 20 different people appear, and with a total of 377 faces. The best balance between performance and recognition accuracy rate was achieved configuring the *scale factor* and *minimum number of neighbours* parameters to 1.2 and 5, respectively. The algorithm achieved averages of 82.8%, 56.6% and 65% rates for Precision, Recall, and F-score, respectively. Later, we use the OpenFace tool (Amos et al., 2016), a trained deep neural network that extracts 128-D feature vectors from the detected faces.

After detecting the appearance of faces in the egocentric photostreams and extracting their descriptors, the clustering method is applied to find their relation and temporal occurrence. Our proposed clustering algorithm is evaluated over an extended version of the test-set, composed by 4280 egocentric images. We compared its performance to the state-of-the-art MeanShift (Comaniciu and Meer, 2002) and Spectral Clustering (Ng et al., 2002) techniques as well as with other configuration of agglomerative clusterings, with different dissimilarity metrics. The robustness for the obtained clusters is considered for all the methods evaluated. We evaluate the obtained results with Precision, Recall, and F-Measure metrics, see Table 6.1.

Table 6.1: Average Precision, Recall, and F-Measure result for each of the tested methods on the extended test-set composed by egocentric images.

| | Proposed clustering model | | | Mean Shift | | | Spectral Clustering | | |
|-------------------|---------------------------|--------|-----------|------------|--------|-----------|---------------------|--------|-----------|
| | Precision | Recall | F-Measure | Precision | Recall | F-Measure | Precision | Recall | F-Measure |
| Extended Test-set | 81,66 | 33,48 | 44,14 | 37,69 | 15,47 | 21,70 | 93,50 | 29,74 | 42,47 |

We expected that the clusters relate to different people that the wearer interacted with. Finally, information about the individual interactions is derived through the evaluation of the faces occurrence along the day.

6.4.3 Results

The obtained results are reported both, numerically in Table 6.2 and visually, through social profiles in Fig. 6.3.

Table 6.2: This table shows the social behavioural traits obtained from the detected social interactions for the different camera wearer.

| | Social Behavioural Traits | | | | |
|--------|---------------------------|-------------|-----------------|---------------|-----------------|
| | Num p/day | Avg int/day | Avg t/int (min) | Avg t/p (min) | Avg t/alone (h) |
| User 1 | 9 | 12 | 12 | 12 | 8h 23m |
| User 2 | 7 | 10 | 17 | 17 | 15h 52min |
| User 3 | 3 | 4 | 21 | 21 | 11h 39min |
| User 4 | 5 | 6 | 15 | 15 | 7h 42min |

In Fig. 6.3, we can observe how social patterns differ among individuals. For instance, User 2 interacts with a lower number of people per day than User1, but the duration of those interactions is higher. On the other hand, User 3 is the individual that spends a higher average time per person, even though he interacts with the lower number of people per day. We can infer from that that his social interactions are with on a small group of people.

Qualitative results suggest that the presented automatic model for social patterns analysis rapidly obtains social behaviour understanding of the individuals. Therefore, further interpretations of social profiles are on social specialists hands. Moreover, on the weeks of egocentric sets, a total of 182 clusters were obtained: 61, 63, 16, and 42, for User1, User2, User3 and User4, respectively.

6.5 Conclusions

In this work, we proposed a model to automatically analyse social behaviour of wearable cameras users, through quantifying and visualizing the occurrence of their social interactions. This is of high interest for social psychology, since it removes the need for manual labour when analyzing the social behaviour of people. To this end, we rely on clustering algorithms to find the relation of the detected appearance of people throughout the egocentric photostreams. We propose to obtain five social characteristics of the detected social interactions. These social traits are used to create a social profile through their visualization. Social profiles allow scientists to obtain information and compare different social behaviour of individuals, for its later study and understanding. Currently, this model and the obtained results are under discussion with social psychologists. Their comments about relevant traits to be obtained from the recorded photo-streams will infer a more robust social profile. We

plan to explore the understanding of the kind of relation the camera wearers share with the people they interact with. An application will be to use the recognized contacts of the user for cognitive training of patients suffering from Alzheimer disease.



Figure 6.3: Obtained social profiles as a result of applying our method to egocentric photo-streams recorded by different users. For a better interpretation, we present a social profile per user, and a common one where the social profiles are overlapping. The last one offers a clearer view of differences among individuals.

Chapter 7

Summary and Outlook

7.1 Work Summary

This dissertation provides with several solutions for the understanding of the lifestyle and behavioural patterns from egocentric photo-streams collected with wearable cameras. Five main applications have guided this work: the temporal segmentation of egocentric photo-streams; the discovery of Routine and Non-routine related days; the classification of food-scenes in egocentric images; the recognition of induced sentiment when reviewing own collected photo-streams; and the characterization of social interactions.

Deep learning has had a huge impact in the computer vision community for the classification and description of images. In particular, in this thesis, we have relied on the use of these techniques for the above-mentioned applications. Due to the limited amount of collected data, we use *transfer learning* theory in the different classification problems that we have addressed. Moreover, we made use of detected objects, places, and faces for the semantic description of the images. This obtained information allowed us to built on top of pre-trained models for the understanding of the lifestyle and behavioural patterns of the camera wearer.

Egocentric photo-streams describe a first-person view of the life of the camera wearer. These images are collected with a wearable camera that usually has a low frame-rate and describe where users spend time by following their body movements. This fact leads to consecutive highly visual different images when the user walks, as well as to similar ones when the user stays static doing some activity, such as *watching tv*. For the summarization and analysis of specific activities or time frames, temporal segmentation of egocentric photo-streams is a useful tool. In this work, we introduced a novel temporal segmentation model based on the hierarchical clustering method which computed the relation among images based on global and semantic features extracted from them. The obtained segmentation was coherent among several manual segmentations provided by different people, showing its robustness for further applications within the egocentric vision field.

The analysis of behavioural patterns relates to the description of the routine. We proposed a model for the discovery of behavioural patterns relying on topic modelling for the automatic finding of abstract topics within the recorded days. Topic modelling is a method for natural language analysis. Therefore, we translate the collected days into documents, which are composed of the detected objects in the images that constitute the photo-stream. Discovered abstract topics are evaluated when analyzing the documents collected by all users, or just at a personal level. We also evaluate the performance of the found topics for the task of classifying the collected days into Routine or Non-Routine related. We also test the performance of the model when evaluating time slots of different duration. The performed experiments establish a robust baseline showing that it is feasible to get insight into the behavioural patterns of people.

The places where people spend their time describe their lifestyle. More specifically, the food-related scenes showed to have an impact on their health. Therefore, the identification of food-related scenes in the collected photo-streams is critical for a better understanding of the lifestyle of a person. We proposed a hierarchical model for the classification of egocentric images into 15 different food-related scenes. Food-related scenes are visually and semantically related. Therefore, we have introduced a taxonomy describing the relationship between the studied classes. This taxonomy allows the analysis of the collected photo-streams through activity and location label. The proposed model adapts to the proposed taxonomy. The first stage of classification is between three food-related activities: eating, preparing and acquiring food. The final classification differentiates among the 15 proposed classes: bakery, bar, beer hall, cafeteria, coffee shop, dining room, food court, ice cream parlour, kitchen, market indoor, market outdoor, picnic area, pub indoor, restaurant, and supermarket. In order to give a robust classification output, the final classification probabilities for a given image are computed as the multiplication of prior probabilities of the given classification tree. The proposed model has shown to be a powerful tool for the later characterization of the nutritional habits of the camera wearer.

The analysis of inferred sentiment by reviewing past experiences through the collected photo-streams started with a collaboration with psychologists who worked with people suffering from depression. We focus on the classification of the images into three main classes: positive, negative, or neutral. The proposed model bases its classification on the semantic features obtained from the images. These semantic features were obtained from an existent model that detected objects with an associated sentiment value, such as: *beautiful view*, *lonely chair*, or *damaged building* with associated values of 1.69, -0.44, and -1.42, respectively. Our hypothesis was that the detected semantics would allow us to describe the feeling the images irradiate to the owner in the reviewing process. However, the available tool was not able to cor-

rectly detect such concepts in our images. Moreover, the final classification ended into: negative images as the non-informative ones; neutral images as the ones that describe work-related scenes; and the positive images were those with scenes related to social interactions, eating, or walking outside. This research was the starting point of the analysis of the routine of a person.

Appearing faces in the collected photo-streams show the social interactions of the camera wearers throughout their days. The detection and analysis of these faces in images is possible with existent and publicly available tools, such as OpenCV and OpenFace. In this work, we made use of such tools for the identification of social interactions and build a model to characterize them. Our proposed model performs person re-identification throughout the sequences for the identification of familiar people. Due to the lack of baseline works in this specific field of research, we proposed several metrics related to the occurrence and duration of the detected social interactions for their quantification.

Benchmarking: Together with the approaches described, we have introduced two novel and home-made datasets: *EgoFoodPlaces* and *EgoRoutine*. The first one is composed of more than 33,000 images and describes 15 different food-related scenes. It is further described in Chapter 4. The latest was collected by 7 different users for periods of at least two weeks. It is composed of 103 days, with a total of more than 100,000 images. This dataset is described in Chapter 3. Both datasets are available on the website of our research group: <http://www.ub.edu/cvub/dataset/>. This will encourage and allow other researchers to evaluate their algorithms with ours. If feasible, it would represent a significant step forward for the automatic and personalized characterization of a person's social life. One could argue that the final usability and applicability of this technology is not ensured. However, in (Gelonch et al., 2019), psychologists discuss that older adults have a high level of acceptance, concluding that the benefits for memory overcome previous privacy concerns.

7.2 Outlook

In this section, we describe ideas and open directions on how the work presented in this dissertation can be extended for future studies.

Fig. 7.1 illustrates the future lines of research that will further develop the proposed applications pipelines in this manuscript for the understanding of human behaviour from collected egocentric photo-streams.

First, we discuss the classification of food-related scenes in collected photo-streams. The improvement of the performance of the proposed classifiers will allow a better characterization of the nutritional routine of people leading to an improvement of

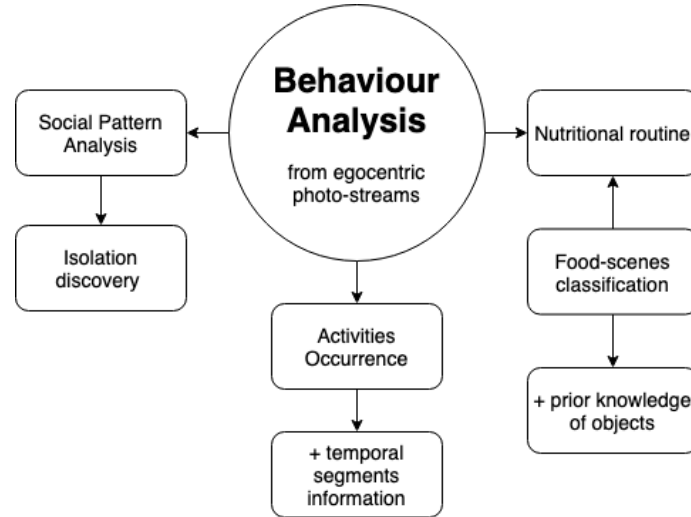


Figure 7.1: An overview of directions for further development for the understanding of behavioural patterns from egocentric photo-streams. These directions include various aspects: the characterization of the social interactions of the camera wearer, with the possibility of detecting isolation; the analysis of the nutritional routine relying on the analysis of occurrence of food-related scenes; and the inclusion of prior knowledge about appearing objects in certain scenes for the improvement of food-scenes recognition; the inclusion of information about temporal boundaries of scenes within the collected days for the analysis of activities in the frame of behavioural analysis.

their healthy lifestyle. Further work will analyze the classification of food-related scenes based on detected concepts in the images. We plan to study how the inclusion of priors of appearing objects in the target scenes affect the final classification. As an example, the final classification of a given image will be modified if it is labelled by the proposed model as *market outdoor* and at the same time, objects like *television* are found with a relatively high probability. We believe that this will help us improve the classification accuracy by avoiding “non-common sense” associations.

The process of identifying behavioural patterns through the discovery of abstract topics showed its potential for the characterization of the lifestyle of the camera wearer. Our proposed model relies on the application of the statistical process of topic modelling to the semantic features obtained from the image. Following this up, we believe that further research should go on the lines of nutritional behaviour and social patterns analysis. The classification of images into food-related scenes is the first step for the analysis of nutritional habits of the camera wearer. A more

personalized advice from specialists can be given if details about the occurrence of food-related scenes in the life of a person are automatically and objectively obtained. This shows the importance of developing tools for the analysis of the nutritional routine of a person. Future work will address the analysis of how days are related based on the food-related activities: eating, preparing and acquiring food. We believe this will help in describing the lifestyle of people for the later improvement of their health.

Second, we propose several research lines following the work done on analysis of behavioural patterns. The field of behaviour analysis is a wide area since different studies can focus on different and specific aspects of the lifestyle of a person. In Chapter 3, we concluded that activity patterns give relevant information that allows us to better perform a distinction among similar days. However, that was a general approach for the classification of days based on detected objects that gives an overview of the behaviour of a person. Following that line, we foresee that the characterization of a person's behavioural patterns can be addressed by the quantification of the performed activities throughout his or her collected days in the form of photo-streams.

In Chapter 6, we addressed the characterization of social patterns of behaviour by quantifying the social relations of the camera wearer. We will explore in future works the analysis of social habits and social activities occurrence through the collected days. We believe that this analysis will allow a better understanding of the habits of the person, helping to automatically detect isolation or certain behaviours related to specific disorders such as a depression.

Finally, dividing days into time-slots has helped us to characterize them more accurately and to when classify them into Routine and Non-routine compare them by maintaining temporal information. Future work will evaluate the performance of the proposed model for behaviour analysis when including information about temporal boundaries of events happening through the day. This information might be useful for the definition of time-slots to be compared, making the comparison more flexible with respect to activities duration. Methods such as the one proposed in Chapter 2 (Dimiccoli et al., 2017) can be used for the detection of boundaries within the collected days.

Bibliography

- Aghaei, M., Dimiccoli, M. and Radeva, P.: 2015, Towards social interaction detection in egocentric photo-streams, *International Conference on Machine Vision*.
- Aghaei, M., Dimiccoli, M. and Radeva, P.: 2016a, Multi-face tracking by extended bag-of-tracklets in egocentric videos, *Computer Vision and Image Understanding, Special Issue on Assistive Computer Vision and Robotics* **149**, 146–156.
- Aghaei, M., Dimiccoli, M. and Radeva, P.: 2016b, With whom do i interact? detecting social interactions in egocentric photo-streams, *Proceedings of the International Conference on Pattern Recognition, IEEE*, pp. 2959–2964.
- Aghaei, M., Dimiccoli, M. and Radeva, P.: 2017, All the people around me: face discovery in egocentric photo-streams.
- Alletto, S., Serra, G., Calderara, S. and Cucchiara, R.: 2015, Understanding social relationships in egocentric vision, *Pattern Recognition* **48**(12), 4082–4096.
- Allport, G. W.: 1985, The historical background of social psychology, *The Handbook of Social Psychology*.
- Altman, N. S.: 1992, An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician* **46**(3), 175–185.
- Alvera-Azcárate, A., Sirjacobs, D., Barth, A. and Beckers, J.-M.: 2012, Outlier detection in satellite data using spatial coherence, *Remote Sensing of Environment* **119**, 84–91.
- Amos, B., Ludwiczuk, B., Satyanarayanan, M. et al.: 2016, Openface: A general-purpose face recognition library with mobile applications, *CMU School of Computer Science* **6**.
- Andersen, C. K., Wittrup-Jensen, K. U., Lolk, A., Andersen, K. and Kragh-Sørensen, P.: 2004, Ability to perform activities of daily living is the main factor affecting quality of life in patients with dementia, *Health and quality of life outcomes* **2**(1), 52.

- Bak, S. and Carr, P.: 2017, One-shot metric learning for person re-identification, *IEEE Conference on Computer Vision and Pattern Recognition*.
- Biagioni, J. and Krumm, J.: 2013, Days of our lives: Assessing day similarity from location traces, *International Conference on User Modeling, Adaptation, and Personalization* pp. 89–101.
- Bifet, A. and Gavaldà, R.: 2007, Learning from time-changing data with adaptive windowing, *Proceedings of the 2007 SIAM international conference on data mining*, SIAM, pp. 443–448.
- Blei, D. M., Ng, A. Y. and Jordan, M. I.: 2003, Latent dirichlet allocation, *Journal of machine Learning research* **3**(Jan), 993–1022.
- Bolaños, M., Dimiccoli, M. and Radeva, P.: 2017, Toward storytelling from visual lifelogging: An overview, *IEEE Transactions on Human-Machine Systems* **47**, 77–90.
- Bolaños, M., Garolera, M. and Radeva, P.: 2014, Video segmentation of life-logging videos, *Articulated Motion and Deformable Objects*, Springer-Verlag, pp. 1–9.
- Bolaños, M., Mestre, R., Talavera, E., Giró-i Nieto, X. and Radeva, P.: 2015, Visual summary of egocentric photostreams by representative keyframes, *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, pp. 1–6.
- Borth, D., Ji, R., Chen, T., Breuel, T. and Chang, S.-F.: 2013, Large-scale visual sentiment ontology and detectors using adjective noun pairs, *Proceedings of the 21st ACM international conference on Multimedia*, ACM, pp. 223–232.
- Boykov, Y., Veksler, O. and Zabih, R.: 2001, Fast approximate energy minimization via graph cuts, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(11), 1222–1239.
- Cadmus-Bertram, L., Marcus, B. H., Patterson, R. E., Parker, B. A. and Morey, B. L.: 2015, Use of the fitbit to measure adherence to a physical activity intervention among overweight or obese, postmenopausal women: self-monitoring trajectory during 16 weeks, *JMIR mHealth and uHealth* **3**(4), e96.
- Campos, V. and et al.: 2015, Diving Deep into Sentiment: Understanding Fine-tuned CNNs for Visual Sentiment Prediction, *ASM* pp. 57–62.
- Cartas, A., Dimiccoli, M. and Radeva, P.: 2017, Batch-based activity recognition from egocentric photo-streams, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2347–2354.
- Castro, D., Hickson, S., Bettadapura, V., Thomaz, E., Abowd, G., Christensen, H. and Essa, I.: 2015, Predicting daily activities from egocentric images using deep learning, *proceedings of the 2015 ACM International symposium on Wearable Computers*, ACM, pp. 75–82.
- Chen, J., Wang, Y., Qin, J., Liu, L. and Shao, L.: 2017, Fast person re-identification via cross-camera semantic binary transformation, *IEEE Conference on Computer Vision and Pattern Recognition*.

- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L.: 2018, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE transactions on pattern analysis and machine intelligence* **40(4)**, 834–848.
- Chen, T., Borth, D., Darrell, T. and Chang, S.-F.: 2014, DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks, *arXiv preprint arXiv:1410.8586* p. 7.
- Chin, S. T. S., Anantharaman, R. and Tong, D. Y. K.: 2011, Emotional intelligence and organisational citizenship behaviour of manufacturing sector employees: An analysis., *Management* **6(2)**.
- Chollet, F.: 2017, Xception: Deep learning with depthwise separable convolutions, *IEEE Conference on Computer Vision and Pattern Recognition* pp. 1800–1807.
- Comaniciu, D. and Meer, P.: 2002, Mean shift: a robust approach toward feature space analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 603 – 619.
- Cortes, C. and Vapnik, V.: 1995, Support-vector networks, *Machine learning* **20(3)**, 273–297.
- Dan-Glauser, E. S. and Scherer, K. R.: 2011, The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance, *Behavior research methods* **43(2)**, 468–77.
- de Haan, E., Van Oppen, P., Van Balkom, A., Spinhoven, P., Hoogduin, K. and Van Dyck, R.: 1997, Prediction of outcome and early vs. late improvement in ocd patients treated with cognitive behaviour therapy and pharmacotherapy, *Acta Psychiatrica Scandinavica* **96(5)**, 354–361.
- de Wijk, R. A., Polet, I. A., Boek, W., Coenraad, S. and Bult, J. H.: 2012, Food aroma affects bite size, *BioMed Central* .
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: 2009, Imagenet: A large-scale hierarchical image database, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.
- Dimicoli, M., Bolaños, M., Talavera, E., Aghaei, M., Nikolov, S. G. and Radeva, P.: 2017, Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation, *Computer Vision and Image Understanding* **155**, 55–69.
- Ding, Z. and Fei, M.: 2013, An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window, *International Federation of Automatic Control* **46(20)**, 12–17.
- Doherty, A. R., Hodges, S. E., King, A. C. and et al.: 2013, Wearable cameras in health: the state of the art and future possibilities, *American journal of preventive medicine*, Vol. **44(3)**, Springer, pp. 320–323.

- Doherty, A. R. and Smeaton, A. F.: 2008, Automatically segmenting lifelog data into events, *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 20–23.
- Donini, L. M., Savina, C. and Cannella, C.: 2003, Eating habits and appetite control in the elderly: the anorexia of aging, *International psychogeriatrics* **15**(1), 73–87.
- Drozdal, M., Vitrià, J., Seguí, S., Malagelada, C., Azpiroz, F. and Radeva, P.: 2014, Intestinal event segmentation for endoluminal video analysis, *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp. 3592–3596.
- Eagle, N. and Pentland, A.: 2006, Reality mining: Sensing complex social systems, *Personal Ubiquitous Comput.* **10**(4), 255–268.
- Ermes, M., Parkka, J., Mantyjarvi, J. and Korhonen, I.: 2008, Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions, *IEEE Transactions on Information Technology in Biomedicine* **12**(1), 20–26.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al.: 1996, A density-based algorithm for discovering clusters in large spatial databases with noise., *ACM Transactions on Knowledge Discovery from Data* **96**(34), 226–231.
- Falomir, Z.: 2012, Qualitative distances and qualitative description of images for indoor scene description and recognition in robotics, *AI Communications* **25**(4), 387–389.
- Fan, C., Lee, J., Xu, M., Kumar Singh, K., Jae Lee, Y., Crandall, D. J. and Ryoo, M. S.: 2017, Identifying first-person camera wearers in third-person videos, *IEEE Conference on Computer Vision and Pattern Recognition*.
- Farrahi, K. and Gatica, D.: 2011, Discovering routines from large-scale human locations using probabilistic topic models, *ACM Transactions on Intelligent Systems and Technology* **2**(1), 3.
- Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N. and Vento, M.: 2015, Reliable detection of audio events in highly noisy environments, *Pattern Recognition Letters* **65**(1), 22–28.
- Fontana, J. M., Farooq, M. and Sazonov, E.: 2014, Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior, *IEEE Transactions on Biomedical Engineering* **61**(6), 1772–1779.
- Furnari, A., Farinella, G. and Battiato, S.: 2016, Temporal Segmentation of Egocentric Videos to Highlight Personal Locations of Interest, pp. 474–489.
- Furnari, A., Farinella, G. and Battiato, S.: 2017, Recognizing Personal Locations From Egocentric Videos, *IEEE Transactions on Human-Machine Systems* **47**(1), 1–13.
- Furnari, A., Farinella, G. M. and Battiato, S.: 2015, Recognizing personal contexts from egocentric images, *IEEE International Conference on Computer Vision Workshop* pp. 393–401.

- Garofolo, J. and et al.: 1993, TIMIT Acoustic-Phonetic Continuous Speech Corpus, *Philadelphia: Linguistic Data Consortium* .
- Gelonch, O., Ribera, M., Codern-Bove, N., Ramos, Silvia, Q., Maria, Chico, G., Cerulla, N., Lafarga, P., Radeva, P. and Garolera, M.: 2019, Acceptability of a lifelogging wearable camera in older adults with mild cognitive impairment: a mixed-method study, *BMC Geriatrics* .
- Ghosh, S. and Reilly, D. L.: 1994, Credit card fraud detection with a neural-network, *27th Hawaii International Conference on System Sciences* **3**, 621–630.
- Goldman, D. B., Curless, B., Salesin, D. and Seitz, S. M.: 2006, Schematic storyboarding for video visualization and editing, *ACM Trans. Graph.* **25**(3), 862–871.
- Habibian, A. and Snoek, C.: 2014, Recommendations for recognizing video events by concept vocabularies, *Computer Vision and Image Understanding* **124**, 110–122.
- Hayat, M., Khan, S. H., Werghi, N. and Goecke, R.: 2017, Joint registration and representation learning for unconstrained face identification, *The IEEE Conference on Computer Vision and Pattern Recognition*.
- He, K., Zhang, X., Ren, S. and Sun, J.: 2016, Deep residual learning for image recognition, *IEEE Conference on Computer Vision and Pattern Recognition* pp. 770–778.
- Herranz, L., Jiang, S. and Li, X.: 2016, Scene Recognition With CNNs: Objects, Scales and Dataset Bias, *Conference on Computer Vision and Pattern Recognition* pp. 571–579.
- Higgs, S. and Thomas, J.: 2016, Social influences on eating, *Current Opinion in Behavioral Sciences* **9**, 1–6.
- Higuchi, M. and Yokota, S.: 2011, Imaging environment recognition device. US Patent 7,983,447.
- Ho, T. K.: 1995, Random decision forests, *Proc. of the Third International Conf. on Document Analysis and Recognition Vol.1* pp. 278–282.
- Hodge, V. and Austin, J.: 2004, A survey of outlier detection methodologies, *Artificial intelligence review* **22**(2), 85–126.
- Hoeffding, W.: 1963, Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* **58**(301), pp. 13–30.
- Hoffman, J., Sergio, S., Tzeng, E. S., Hu, R., J. Donahue, R. G., Darrell, T. and Saenko, K.: 2014, Lsda: Large scale detection through adaptation, *Advances in Neural Information Processing Systems*, pp. 3536–3544.
- Holmes, E. A. and et al.: 2006, Positive Interpretation Training: Effects of Mental Imagery Versus Verbal Training on Positive Mood, *Behavior Therapy* **37**(3), 237–247.

- Hopkinson, J. B., Wright, D. N., McDonald, J. W. and Corner, J. L.: 2006, The prevalence of concern about weight loss and change in eating habits in people with advanced cancer, *Journal of pain and symptom management* **32**(4), 322–331.
- House, J. S., Landis, K. R. and Umberson, D.: 1988, Social relationships and health, *Science* **241**(4865), 540–545.
- Jeffery, R. W., Baxter, J., McGuire, M. and Linde, J.: 2006, Are fast food restaurants an environmental risk factor for obesity?, *International Journal of Behavioral Nutrition and Physical Activity* **3**(1), 2.
- Jia, Y.: 2013, Caffe: An open source convolutional architecture for fast feature embedding, <http://caffe.berkeleyvision.org/>.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T. and Eecs, U. C. B.: 2014, Caffe: Convolutional Architecture for Fast Feature Embedding.
- Joachims, T.: 2000, Estimating the Generalization Performance of a SVM efficiently, *International Conference on Machine Learning* pp. 431–438.
- Jojic, N., Perina, A. and Murino, V.: 2010, Structural epitome: a way to summarize one's visual experience, pp. 1027–1035.
- Kawachi, I. and Berkman, L. F.: 2001, Social ties and mental health, **73**, 458–467.
- Kelly, P., Marshall, S., Badland, H., Kerr, J., Oliver, M., Doherty, A. and Foster, C.: 2013, An ethical framework for automated, wearable cameras in health behavior research, *American journal of preventive medicine* **44**(3), 314–319.
- Kemps, E., Tiggemann, M. and Hollitt, S.: 2014, Exposure to television food advertising primes food-related cognitions and triggers motivation to eat, *Psychology & Health* **29**(10), 1192.
- Keogh, E. J. and Pazzani, M. J.: 2001, Derivative dynamic time warping, *Proceedings of the 2001 SIAM international conference on data mining* pp. 1–11.
- Koskela, M. and Laaksonen, J.: n.d., Convolutional Network Features for Scene Recognition, pp. 1–4.
- Krizhevsky, A., Sutskever, I. and Hinton, G. E.: 2012, ImageNet Classification with Deep Convolutional Neural Networks, *NIPS* pp. 1–9.
- Krizhevsky, A., Sutskever, I. and Hinton, G. E.: 2012, Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems* **25** pp. 1097–1105.
- Lang, P., Bradley, M. and Cuthbert, B.: 1997, International Affective Picture System (IAPS): Technical Manual and Affective Ratings, *NIMH* pp. 39–58.

- Larson, N., Story, M. and J, M.: 2009, A review of environmental influences on food choices, *Annals of Behavioural Medicine* **38**, 56–73.
- Laska, M., Hearst, M., Lust, K., Lytle, L. and Story, M.: 2015, How we eat what we eat: identifying meal routines and practices most strongly associated with healthy and unhealthy dietary factors among young adults, *Public Health Nutrition* **18(12)**, 2135–2145.
- Lazebnik, S., Schmid, C. and Ponce, J.: 2006, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2**, 2169–2178.
- LeCun, Y., L. Bottou, L., Bengio, Y. and Haffner, P.: 1998, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86(11)**, 2278–2324.
- Lee, M. L. and Dey, A. K.: 2008, Lifelogging memory appliance for people with episodic memory impairment, *UbiComp* .
URL: <http://portal.acm.org/citation.cfm?doid=1409635.1409643>
- Lee, Y. and Grauman, K.: 2015, Predicting important objects for egocentric video summarization, *International Journal of Computer Vision* **114(1)**, 38–55.
- Levi, G. and Hassner, T.: 2015, Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns, *ICMI* pp. 503–510.
- Lewis, D. D.: n.d., Reuters-21578.
- Li, C., Cheung, W. K. and Liu, J.: 2015, Elderly mobility and daily routine analysis based on behavior-aware flow graph modeling, *2015 International Conference on Healthcare Informatics*, pp. 427–436.
- Li, D., Chen, X., Zhang, Z. and Huang, K.: 2017, Learning deep context-aware features over body and latent parts for person re-identification, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Li, Z., Wei, Z., Jia, W. and Sun, M.: 2013, Daily life event segmentation for lifestyle evaluation based on multi-sensor data recorded by a wearable device, *Proceedings of Engineering in Medicine and Biology Society*, pp. 2858–2861.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: 2014, Microsoft coco: Common objects in context, *European conference on computer vision* pp. 740–755.
- Lin, W.-H. and Hauptmann, A.: 2006, Structuring continuous video recording of everyday life using time-constrained clustering, *Proceedings of SPIE, Multimedia Content Analysis, Management, and Retrieval* **959**.
- Liu, F. T., Ting, K. M. and Zhou, Z.-H.: 2008, Isolation forest, *8th IEEE International Conference on Data Mining* pp. 413–422.

- Liu, J., Johns, E., Atallah, L., Pettitt, C., Lo, B., Frost, G. and Yang, G.-Z.: 2012, An intelligent food-intake monitoring system using wearable sensors, *9th International Conference on Wearable and Implantable Body Sensor Networks* pp. 154–160.
- Lu, Z. and Grauman, K.: 2013, Story-driven summarization for egocentric video., *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2714–2721.
- Ma, M., Fan, H. and Kitani, K. M.: 2016, Going deeper into first-person activity recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1894–1903.
- Maaten, L. v. d. and Hinton, G.: 2008, Visualizing data using t-sne, *Journal of machine learning research* **9**, 2579–2605.
- Machajdik, J. and Hanbury, A.: 2010, Affective image classification using features inspired by psychology and art theory, *Proceedings of the 18th ACM international conference on Multimedia*, ACM, pp. 83–92.
- Makris, D. and Ellis, T.: 2005, Learning semantic scene models from observing activity in visual surveillance, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **35**(3), 397–408.
- Martin, A., Brouwers, P., Lalonde, F., Cox, C., Foster, N. L. and Chase, T. N.: 1986, Towards a behavioral typology of alzheimer’s patients, *Journal of clinical and experimental neuropsychology* **8**(5), 594–610.
- Martin, D., Fowlkes, C., Tal, D. and Malik, J.: 2001, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, *Proceedings of 8th International Conference on Computer Vision*, pp. 416–423.
- Miller, G. A.: 1995, Wordnet: a lexical database for english, *Communications of the ACM* **38**(11), 39–41.
- Morrow, S.: 1999, Instrumental activities of daily living scale, *AJN The American Journal of Nursing* **99**(1), 24CC.
- Nam, J. and Tewfik, A.: 1999, Dynamic video summarization and visualization., in J. F. Buford and S. M. Stevens (eds), *ACM Multimedia*, pp. 53–56.
- Neisser, U.: 1988, Five kinds of self-knowledge, *Philosophical psychology* **1**(1), 35–59.
- Ng, A. Y., Jordan, M. I. and Weiss, Y.: 2002, On spectral clustering: Analysis and an algorithm, in T. G. Dietterich, S. Becker and Z. Ghahramani (eds), *Advances in Neural Information Processing Systems 14*, pp. 849–856.
- Nojavanasghar, B. and et al.: 2016, EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children, *International Conference on Multimodal Interfaces* pp. 137–144.

- Oliveira-Barra, G., Bolaños, M., Talavera, E., Dueñas, A., Gelonch, O. and Garolera, M.: 2017, Serious games application for memory training using egocentric images, *International Conference on Image Analysis and Processing* pp. 120–130.
- Parzen, E.: 1962, On estimation of a probability density function and mode, *The annals of mathematical statistics* pp. 1065–1076.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: 2011, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825–2830.
- Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G. and Kokmen, E.: 1999, Mild cognitive impairment: clinical characterization and outcome, *Archives of neurology* **56**(3), 303–308.
- Platt, J. et al.: 1999, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, *Advances in large margin classifiers* **10**(3), 61–74.
- Poleg, Y., Arora, C. and Peleg, S.: 2014, Temporal segmentation of egocentric videos, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2537–2544.
- Poleg, Y., Ephrat, A., Peleg, S. and Arora, C.: 2016, Compact cnn for indexing egocentric videos, *IEEE Winter Conference on Applications of Computer Vision* pp. 1–9.
- Poria, S. and et al.: 2014, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, *Neurocomputing* **174**, 50–59.
- Pujol, O., Radeva, P. and Vitria, J.: 2006, Discriminant ecoc: A heuristic method for application dependent design of error correcting output codes, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(6), 1007–1012.
- Quattoni, A. and Torralba, A.: 2009, Recognizing indoor scenes., *Computer Vision and Pattern Recognition* pp. 413–420.
- Ravi, D., Lo, B. and Yang, G.-Z.: 2015, Real-time food intake classification and energy expenditure estimation on a mobile device, *12th International Conference on Wearable and Implantable Body Sensor Networks* pp. 1–6.
- Redmon, Joseph & Farhadi, A.: 2018, Yolov3: An incremental improvement, *arXiv preprint arXiv:1804.02767* .
- Reeder, B. and David, A.: 2016, Health at hand: a systematic review of smart watch uses for health and wellness, *Journal of biomedical informatics* **63**, 269–276.
- Rokach, L. and Maimon, O.: 2005, Clustering methods, *Data mining and knowledge discovery handbook* pp. 321–352.

- Rousseeuw, P. J. and Driessen, K. V.: 1999, A fast algorithm for the minimum covariance determinant estimator, *Technometrics* **41**(3), 212–223.
- Ryff, C. D.: 1995, Psychological well-being in adult life, *Current directions in psychological science* **4**(4), 99–104.
- Salvador, S. and Chan, P.: 2007, Toward accurate dynamic time warping in linear time and space, *Intell. Data Analysis* **11**(5), 561–580.
- Sanlier, N. and Seren Karakus, S.: 2010, Evaluation of food purchasing behaviour of consumers from supermarkets, *British Food Journal* **112**(2), 140–150.
- Sarker, M., Kamal, M., Rashwan, H. A., Talavera, E., Banu, S. F., Radeva, P. and Puig, D.: 2018, Macnet: Multi-scale atrous convolution networks for food places classification in egocentric photo-streams, *Proceedings of the European Conference on Computer Vision* .
- Schmand, B., Walstra, G., Lindeboom, J., Teunisse, S. and Jonker, C.: 2000, Early detection of alzheimer’s disease using the cambridge cognitive examination, *Psychological Medicine* **30**(3), 619–627.
- Seiter, J., Derungs, A., Schuster-Amft, C., Amft, O. and Tröster, G.: 2015, Daily life activity routine discovery in hemiparetic rehabilitation patients using topic models, *Methods of information in medicine* **54**(03), 248–255.
- Sellen, A., Fogg, A., Aitken, M., Hodges, S., Rother, C. and Wood, K. R.: 2007, Do life-logging technologies support memory for the past?: an experimental study using sensecam, *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, pp. 81–90.
- Sevtsuk, A. and Ratti, C.: 2010, Does Urban Mobility Have a Daily Routine? Learning from the Aggregate Data of Mobile Networks, *Journal of Urban Technology* **1**(17), 41–60.
- Silvia, P. J. and Gendolla, G. H.: 2001, On introspection and self-perception: Does self-focused attention enable accurate self-knowledge?, *Review of General Psychology* **5**(3), 241–269.
- Simonyan, K. and Zisserman, A.: 2015, Very Deep Convolutional Networks for Large-Scale Image Recognition, *International Conference on Learning Representations (ICRL)* pp. 1–14.
- Smith, M. A.: 1997, Video skimming and characterization through the combination of image and language understanding techniques, *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, pp. 775–781.
- Society for Personality and Social Psychology: 2014, How we form habits, change existing ones, *ScienceDaily* .
- Spiliopoulou, M., Faulstich, L. C. and Winkler, K.: 1999, A data miner analyzing the navigational behaviour of web users, *Proceedings of the Workshop on Machine Learning in User Modelling* .

- Stalonas, P. M. and Kirschenbaum, D. S.: 1985, Behavioral treatments for obesity: Eating habits revisited, *Behavior Therapy* **16**(1), 1–14.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: 2015, Going Deeper with Convolutions, *Computer Vision and Pattern Recognition*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: 2016, Rethinking the inception architecture for computer vision, *Conf. on Computer Vision and Pattern Recognition* pp. 2818–2826.
- Tal, A. and Wansink, B.: 2013, Fattening Fasting: Hungry Grocery Shoppers Buy More Calories, Not More Food, *JAMA internal medicine* **173**(12), 1146–1148.
- Talavera, E., Cola, A., Petkov, N. and Radeva, P.: 2018, Towards Egocentric Person Re-identification and Social Pattern Analysis, *1st Conference on Applications of Intelligent Systems (APPIS)*.
- Talavera, E., Dimiccoli, M., Bolanos, M., Aghaei, M. and Radeva, P.: 2015, R-clustering for egocentric video segmentation, *Lecture Notes in Computer Science*, Vol. 9117, Springer Verlag, pp. 327–336.
- Talavera, E., M. Sarker, M., Puig, D., Petkov, N. and Radeva, P.: 2014, Hierarchical approach to classify food scenes in egocentric photo-streams, *Journal Biomedical and Health Informatics*.
- Talavera, E., Petkov, N. and Radeva, P.: 2019, Unsupervised routine discovery in egocentric photo-streams, *8th Conference on Computer Analysis of Images and Patterns*.
- Talavera, E., Radeva, P. and Petkov, N.: 2017, Towards egocentric sentiment analysis, pp. 297–305.
- Talavera, E., Strisciuglio, N., Petkov, N. and Radeva, P.: 2017, Sentiment recognition in egocentric photostreams, pp. 471–479.
- Tan, P. N., Steinbach, M. and Kumar, V.: 2005, *Introduction to Data Mining, (First Edition)*, Addison-Wesley Longman Publishing Co., Inc.
- Trickler, C.: 2013, An overview of self-monitoring systems.
- Viola, P., Jones, M. et al.: 2001, Rapid object detection using a boosted cascade of simple features, *IEEE Conference on Computer Vision and Pattern Recognition* **1**(511-518), 3.
- Wang, L., Wang, Z. and Du, W.: 2015, Object-Scene Convolutional Neural Networks for Event Recognition in Images, pp. 1–6.
- Wang, M., Cao, D., Li, L., Li, S. and Ji, R.: 2014, Microblog Sentiment Analysis Based on Cross-media Bag-of-words Model, *International Conference on Internet Multimedia Computing and Service* pp. 76–80.

- Wei, J., Hollin, I. and Kachnowski, S.: 2011, A review of the use of mobile phone text messaging in clinical and healthy behaviour interventions, *Journal of telemedicine and telecare* **17**(1), 41–48.
- Wiles, R., Prosser, J., Bagnoli, A., Clark, A., Davies, K., Holland, S. and Renold, E.: 2008, Visual ethics: Ethical issues in visual research.
- Wood, W., Quinn, J. and Kashy, D.: 2002, Habits in everyday life: Thought, emotion, and action, *Journal of Personality and Social Psychology* **83**(6), 1281–1297.
- Xu, Y. and Damen, D.: 2018, Human routine change detection using bayesian modelling, *2018 24th International Conference on Pattern Recognition* pp. 1833–1838.
- Yang, Y. C., Boen, C., Gerken, K., Li, T., Schorpp, K. and Harris, K. M.: 2016, Social relationships and physiological determinants of longevity across the human life span, *Proceedings of the National Academy of Sciences* **113**(3), 578–583.
- Yesavage, J. A.: 1983, Bipolar illness: correlates of dangerous inpatient behaviour, *The British Journal of Psychiatry* **143**(6), 554–557.
- Yi, D., Lei, Z., Liao, S. and Li, S. Z.: 2014, Learning Face Representation from Scratch, *arXiv* .
- You, Q. and Et, A.: 2016, Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia, *ACM International WSDM Conference* pp. 13–22.
- You, Q. and et al.: 2015, Robust Image Sentiment Analysis using Progressively Trained and Domain Transferred Deep Networks, *AAAI Conference on Artificial Intelligence* pp. 381–388.
- You, Q., Luo, J., Jin, H. and Yang, J.: 2016, Building a large scale dataset for image emotion recognition: The fine print and the benchmark, *AAAI Conference on Artificial Intelligence* .
- Yu, F., Zhang, Y., Song, S., Seff, A. and Xiao, J.: n.d., Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop.
- Yu, S. X. and Shi, J.: 2003, Multiclass spectral clustering, *Proceedings of the 9th IEEE International Conference on Computer Vision* p. 313.
- Yu, Y., Lin, H., Meng, J. and Zhao, Z.: 2016, Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks, *Algorithms* **9**(2), 41.
- Yuan, J. and et al.: 2013, Stribute: Image Sentiment Analysis from a Mid-level Perspective Categories and Subject Descriptors, *International Workshop on Issues of Sentiment Discovery and Opinion Mining* pp. 101–108.
- Yürüten, O., Zhang, J. and Pu, P.: 2014, Decomposing activities of daily living to discover routine clusters, *28th Conference on Artificial Intelligence* .
- Zhao, R., Ouyang, W. and Wang, X.: 2013, Unsupervised salience learning for person re-identification, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3586–3593.

-
- Zheng, L., Wang, S., He, F. and Tian, Q.: 2014, Seeing the Big Picture: Deep Embedding with Contextual Evidences, p. 10.
- Zhou, B., Khosla, A., Lapedriza, A., Torralba, A. and Oliva, A.: 2016, Places: An Image Database for Deep Scene Understanding, *ArXiv* pp. 1–12.
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. and Torralba, A.: 2017, Places: A 10 million image database for scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* .
- Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A.: 2014, Learning Deep Features for Scene Recognition using Places Database, *Advances in Neural Information Processing Systems* 27 pp. 487–495.

Summary

Describing people's lives has become a hot topic in several disciplines. Lifelogging appeared in the 1960s as the process of recording and tracking personal activity data generated by the daily behaviour of a person. The development of new wearable technologies allows to automatically record data from our daily living. Wearable devices are light-ware and affordable, which shows potential for the increase of their use by our society. Egocentric images are recorded by wearable cameras and show a first-person view of the life of the camera wearer. These collected images show an objective view of the daily life of a person and thus are a rich source of information about her or his habits. However, there is lack of tools for the analysis of collections of egocentric photo-sequences and thus room for progress.

This thesis investigates the development of automatic tools for the analysis of egocentric images with the ultimate goal of getting understanding of the lifestyle of the camera wearer. This work addresses five main topics in the field of egocentric vision:

1. *Temporal photo-sequences segmentation*: We introduce an automatic model for the definition of temporal boundaries for the division of egocentric photo-sequences into moments, which are sequences of images describing the same environment. The model is based on global and semantic features and achieves a 66% F-score over the EDUB-Seg dataset.
2. *Routine discovery*: We propose an automatic tool for the discovery of routine-related days and the visualization of patterns of behaviour, based on the use of topic modelling over semantic concepts extracted from the photo-sequences. The introduction of the EgoRoutine dataset composed of a total of 104 days is part of this work. The model is able to classify days into routine and non-routine related with an accuracy of 80%.
3. *Food-related scenes recognition*: We introduce a hierarchical classifier for the recognition of visually highly similar food-related images into 15 different classes that describe

daily activities related to food consumption, acquisition, and preparation. We introduce the EgoFoodScenes dataset, which our model is able to classify into the 15 categories with an accuracy of 68%.

4. *Sentiment retrieval*: We explore the sentiment associated with images by classifying them into Positive, Neutral, and Negative. Our model is based on the analysis of global features and obtained semantic concepts with associated sentiment. We obtain an accuracy of 75%. Results show that positive images relate to outdoor environments or with social interactions, neutral to work-related environments, and negative to non-informative or visually not clear images .
5. *Social pattern characterization*: We propose a model that characterizes the social behaviour of the camera wearer based on the occurrence of people that the camera wearer meets throughout her/his data collection. The proposed social parameters allow the definition of a radar chart that shows its potential for the comparison of social patterns among individuals.

The introduced and made publicly available egocentric datasets and the obtained results in the different performed experiments indicate that behaviour can be identified and studied. We conclude that the developed automatic algorithms for the analysis of egocentric images allow a better understanding of the lifestyle of the camera wearer. Applications based on the analysis of this data can lead to the improvement of the quality of life of people and therefore, are worth to continue exploring.

Samenvatting

Het beschrijven van het leven van mensen is in verschillende disciplines een hot topic geworden. Lifelogging is ontstaan in de jaren zestig van de vorige eeuw als het proces van het vastleggen en volgen van het dagelijkse gedrag van een persoon. De ontwikkeling van nieuwe draagbare technologieën maakt het mogelijk om automatisch gegevens uit ons dagelijks leven vast te leggen. Draagbare apparaten zijn licht en betaalbaar en zijn dus zeer interessant voor gebruik in onze samenleving. Persoonlijke beelden vanuit een eerste-persoonsperspectief worden opgenomen door draagbare camera's en geven een objectief beeld van het dagelijks leven van een persoon. Daarmee is deze verzameling beelden een rijke bron van informatie over haar of zijn gewoonten. Er is echter een gebrek aan hulpmiddelen voor de analyse van verzamelingen egocentrische fotoreeksen en dus is er ruimte voor vooruitgang.

Dit proefschrift onderzoekt de ontwikkeling van automatische hulpmiddelen voor de analyse van egocentrische beelden met als uiteindelijk doel inzicht te verkrijgen in de levensstijl van de cameradrager. Dit werk behandelt vijf hoofdonderwerpen op het gebied van egocentrische visie:

1. *Tijdelijke fotoreekssegmentatie*: We introduceren een automatisch model voor het definiëren van tijdsgrenzen om egocentrische foto-sequenties in momenten te verdelen die dezelfde omgeving beschrijven. Het model is gebaseerd op globale en semantische functies en behaalt een 66 % F-score met de EDUB-Seg dataset.
2. *Routine-ontdekking*: We stellen een automatische tool voor die routine-gerelateerde dagen en de visualisatie van gedragspatronen ontdekt en die is gebaseerd op het gebruik van topic modelling over semantische concepten uit de fotoreeksen. De introductie van de EgoRoutine-dataset bestaande uit een totaal van 104 dagen maakt deel uit van dit werk. Het model is in staat om dagen in te delen in routine- en niet-routine-gerelateerde dagen met een nauwkeurigheid van 80%.

3. *Voedselgerelateerde scèneherkenning*: We gebruiken een hiërarchische classificeerder voor de herkenning van visueel zeer gelijkwaardige voedsel-gerelateerde beelden in 15 verschillende klassen die de dagelijkse activiteiten met betrekking tot voedselconsumptie, -verwerking en -bereiding beschrijven. We gebruiken de EgoFoodScenes-dataset die ons model kan indelen in 15 categorieën met een nauwkeurigheid van 68%.
4. *Sentiment retrieval*: We onderzoeken het sentiment dat gepaard gaat met beelden door ze te classificeren in Positief, Neutraal en Negatief. Ons model is gebaseerd op de analyse van globale kenmerken en verkregen semantische concepten met bijbehorend sentiment. Met het model wordt een nauwkeurigheid van 75 % verkregen. De resultaten tonen aan dat positieve beelden betrekking hebben op buitenomgevingen of op sociale interacties, neutraal op werkgerelateerde omgevingen, en negatief op niet-informatieve of visueel onduidelijke beelden.
5. *Karakterisering van sociale patronen*: We stellen een model voor dat het sociale gedrag van de cameradrager karakteriseert op basis van het aantal mensen dat de cameradrager ontmoet tijdens haar of zijn gegevensverzameling. De voorgestelde sociale parameters maken het mogelijk om een radark kaart te definiëren die potentieel mogelijk maakt om sociale patronen tussen individuen te vergelijken.

De geïntroduceerde en openbaar gemaakte egocentrische datasets en de verkregen resultaten in de verschillende uitgevoerde experimenten geven aan dat gedrag kan worden geïdentificeerd en onderzocht. We concluderen dat de ontwikkelde automatische algoritmen voor de analyse van egocentrische beelden een beter begrip mogelijk maken van de levensstijl van de cameradrager. Toepassingen gebaseerd op de analyse van deze gegevens kunnen leiden tot verbetering van de levenskwaliteit van personen en zijn daarom de moeite waard om verder te verkennen.

Resumen

Describir la vida de las personas se ha convertido en un tema candente en varias disciplinas. Lifelogging apareció en la década de los 60 como el proceso de registrar y rastrear datos de actividad personal generados por el comportamiento diario de una persona. El desarrollo de nuevas tecnologías portátiles permite almacenar automáticamente datos de nuestra vida diaria. Dichos dispositivos son livianos y asequibles, lo que muestra potencial para su uso por parte de nuestra sociedad. Las imágenes egocéntricas son grabadas por cámaras portátiles y muestran una vista en primera persona de la vida del usuario. Esta recopilación de imágenes muestra una visión objetiva de la vida diaria de una persona y, por lo tanto, son una rica fuente de información sobre sus hábitos. Sin embargo, faltan herramientas hoy en día no hay herramintetas para el análisis de colecciones de fotosecuencias egocéntricas y, por lo que hay espacio para el progreso.

Esta tesis investiga el desarrollo de herramientas automáticas para el análisis de imágenes egocéntricas con el objetivo final de comprender el estilo de vida del usuario de la cámara. Este trabajo aborda cinco temas principales en el campo de la visión egocéntrica:

1. *Segmentación temporal de secuencias de imágenes:* Introducimos un modelo automático para la definición de límites temporales con el objetivo de dividir secuencias de imágenes egocéntricas en momentos. Entendemos como momentos secuencias de imágenes que describen el mismo entorno. El modelo se basa en características globales y semánticas y logra un F-score del 66% sobre el conjunto de datos EDUB-Seg.
2. *Descubrimiento de la rutina:* Proponemos una herramienta automática para el descubrimiento de días relacionados con la rutina y la visualización de patrones de comportamiento. La introducción del conjunto de datos EgoRoutine compuesto por un total de 104 días es parte de este trabajo. El modelo puede clasificar los días en rutinarios y no rutinarios con una precisión del 80%.

3. *Reconocimiento de escenas relacionadas con la comida:* Presentamos un clasificador jerárquico para el reconocimiento de 15 clases diferentes de escenas relacionadas con los alimentos, que son visualmente muy similares y que describen actividades diarias relacionadas con el consumo, la adquisición y la preparación de alimentos. Además, presentamos el conjunto de datos EgoFoodScenes, el cual nuestro modelo puede clasificar en las 15 categorías con una precisión del 68%.
4. *Entender el sentimiento evocado:* Exploramos el sentimiento asociado con las imágenes clasificándolas en Positivo, Neutro y Negativo. Nuestro modelo se basa en el análisis de características globales y conceptos semánticos obtenidos con sentimientos asociados. Obtenemos una precisión del 75%. Los resultados muestran que las imágenes positivas se relacionan con ambientes al aire libre o con interacciones sociales, las neutrales con ambientes laborales y las negativas con imágenes no informativas o visualmente no claras.
5. *Caracterización del patrón social:* Proponemos un modelo que caracteriza el comportamiento social del usuario de la cámara basándose en la ocurrencia de personas que el usuario de la cámara se encuentra a lo largo de su recopilación de datos. Los parámetros sociales propuestos permiten la definición de un gráfico de radar que muestra su potencial para la comparación de patrones sociales entre individuos.

Los conjuntos de datos egocéntricos introducidos y puestos a disposición del público junto con los resultados obtenidos en los diferentes experimentos realizados indican que el comportamiento puede identificarse y estudiarse. Concluimos que los algoritmos automáticos desarrollados para el análisis de imágenes egocéntricas permiten una mejor comprensión del estilo de vida del usuario. Las aplicaciones basadas en el análisis de estos datos pueden conducir a la mejora de la calidad de vida de las personas y, por lo tanto, vale la pena continuar estudiándolas.

Acknowledgements

This PhD journey ends with these lines. I would like to start by thanking my promoters Prof. Petia Radeva and Prof. Nicolai Petkov. You gave me the opportunity to grow both as a person and as a researcher by your side. The most precious gift you can give someone is your attention and time, so thank you for yours.

Thanks to the reading committee Prof Michael Biehl, Prof. C. N. Schizas, Prof. J. Vitrià, and Prof. G. M. Farinella for reviewing this manuscript. A special thank to the secretaries at Bernouilliborg, especially to the enthusiastic Ineke, you made my life easier at RUG.

Doing a PhD is not taking the easy path. However, I would choose this path all over again, not just because of all that I have learned - that is quite a lot - but the experiences that I have lived and the people I have met. I have introduced myself as a Sandwich PhD, most of the times causing some laughs. But yes, I used to say I was the 'ham and cheese' between the universities of Groningen and Barcelona. This type of position pushed me to grow fast, living in two different countries with very different cultures. I enjoyed it.

I want to thank my paranympths, Laura and Ahmad, not just for being by my side on such a relevant day, but also for being such good friends from the first day, despite the distance, and throughout the process. My bella Fiorini, we arrived to Groningen in the same week and I keep enjoying when you share your ideas with me, you convey warmth and happiness. Ahmad, I am glad I met you - discussing all types of topics with you made my day in countless times. I wish you both success in life, and if possible, with not too much distance from me.

Charmaine and George, I still remember the first time I met you, that dark and cold night on January 2015, when I first arrived in Groningen. You two have always supported me and I will always be grateful for that - I love the beautiful family you two created. Jiapan, living with you and Astone for one year made me get to know and love you even more. People still smile when I refer to you as 'my Chinese', but I truly feel it! Ours will be a life-long relation.

Our old and now extended *Intelligent Party* group, with whom we made a great and fun team: Ahmed, Laura, Nicola, Andreas, Manuel, Kitty, Ugo, Sreejita, Chenyu (Astone), Jiapan, Laura, Sara, Maria, Godliver, Sofia, Daniel, and Renata. The already PhDs for a while, M. Biehl and M. Wilkinson were always there with good advice, food, and fun - Thanks!

Barcelona, a beautiful city that offers everything where I did my master's degree and two years of PhD. I thank all my research group colleagues for sharing their knowledge and skills - we made a good working team and I learned a lot from them. Maya, you were the first person I met in UB, I hope that we live again in the same *somewhere else*. Marc, if I could choose, I would always like to work on a desk next to yours! Together with Edu, Bea, Pedro, Mariella, Axel, Juan Luis, Alejandro, Eduardo, and Gabriel, we made UB life fun and had many Graniers and 'Risais'. But Barcelona was not just UB. Mireia and Maite, I know you since the first week I moved to Barcelona, back in 2012. You supported me throughout these years and became an important piece of my daily life. Thanks for your unconditional friendship - I really miss you. Collaborations sometimes bring friendships. I also thank Señorita, from the University of Otago, who became a good friend after many Skype meetings.

In Mallorca I had my family and lifelong friends, Patricia, Vicky, Pau, Marga, Lida, Jose, and Francesc. It is always great to catch up when I go back home. I also really enjoy this new condition of being the guest at my sister's and Ismael's home - I expect more visits and road trip together in the near future.

PhD life in Groningen is vivid. GOPHER introduced me to the city from a different perspective and to people who touched my heart. Antonija and Eric, you were the highlight. While writing this, nice memories come to mind from our sweet moments in Barcelona, Girona, Mallorca, and Ameland. In the Spring of 2016, I also joined the PhD Day program committee team. It was a great experience to meet and work together with people from different disciplines. Monique, Mustapha, Ionela, Steven, Marleen, Xu, and Kumar, I enjoyed our meetings and movie nights. Monique, we made and make a good team. Hugs for Daniela and Emilia too.

Maik, you always enthusiastically believed in me and in my project. Thanks for supporting me throughout this journey. Eres genial!

And finally, the most important acknowledgement goes to my beloved family who has supported me in all stages of my life. Lidia, my witty and intelligent sister, I wish you success on everything you face, you are the most capable person I know. I am lucky to have you as partner in life. My biggest thanks go to my parents, mamá y papá, siempre habéis creído que podía hacer lo que me propusiese, y me apoyasteis en todas mis decisiones. Si he llegado a este punto, y a ser como soy, es gracias a vosotros. La hermana y yo nunca podremos devolver tanto como nos habéis dado. Este logro es vuestro también. Os quiero.

I see many of the people I have met during this PhD journey as part of my extended family - because of this, I consider myself a very lucky person.

Thank you all, bedankt iedereen, gracias a todos!

Estefanía Talavera Martínez
Groningen
December 1, 2019

Research Activities

Journal Papers

- **E. Talavera**, C. Wuerich, N. Petkov, P. Radeva, “*Topic Modelling for Routine Discovery from Egocentric Photo-streams*”, (Submitted - Under Review).
- **E. Talavera**, M. Leyva-Vallina, Md M. Sarker, D. Puig, N. Petkov, P. Radeva, “*Hierarchical approach to classify food scenes in egocentric photo-streams*”, Journal Biomedical and Health Informatics (JBHI), IF 4.217, Q1, 2019.
- Md. M. Kamal Sarker, H. A. Rashwan, F. Akram, **E. Talavera**, S. F. Banu, P. Radeva, D. Puig, “*Recognizing Food Places in Egocentric Photo-streams using Multi-scale Atrous Convolutional Networks and Self-Attention Mechanism*”, IEEE Access, Pages 39069-39082, Vol. 7, IF 4.098, Q1, 2019.
- M. Dimiccoli, M. Bolaños, **E. Talavera**, M. Aghaei, G. Stavri, P. Radeva, “*SR-Clustering: Semantic Regularized Clustering for Egocentric Photo Streams Segmentation*”, International Journal Computer Vision and Image Understanding (CVIU), pp. 55-69, Vol. 155, IF 2.645, Q1, 2016.
- S. John, R. Butson, **E. Talavera**, R. Spronken-Smith, P. Radeva, “*Beyond perceptions: exploring Reality Mining to research student experience*”, (Submitted - Under Review).
- S. John, **E. Talavera**, A. Cartas, R. Butson, R. Spronken-Smith, P. Radeva, “*Re-framing our understanding of student experience: the use of photographs to capture activity*”, (Submitted), .

Book Chapters

- G. Oliveira-Barra, M. Bolaños, **E. Talavera**, O. Gelonch, M. Gardera, P. Radeva, “*Lifelog Retrieval for Memory Stimulation of People with Memory Impairments*”, Book Chapter *Multimodal behavior analysis in the wild*, 2017

- **E. Talavera**, N. Petkov, P. Radeva, “Egocentric vision for behavioural understanding”, Book Chapter *Wearable Sensors: Fundamentals, Implementation and Applications*, (Submitted)

Conference Proceedings

- **E. Talavera**, N. Petkov, P. Radeva, “Unsupervised routine discovery in egocentric photo-streams”, 18th Conference on Computer Analysis of Images and Patterns, published in proceedings as Chapter Springer Verlag, 2019.
- M. Kamal, H. Rashwan, **E. Talavera**, S. Furruka, P. Radeva, D. Puig, “MACNet: Multi-scale Atrous Convolution Networks for Food Places Classification in Egocentric Photo-streams”, 3rd Workshop on Egocentric Perception, Interaction and Computing (EPIC), published in the proceedings, 2018.
- A. Cartas, M. Dimiccoli, **E. Talavera**, P. Radeva, “On the Role of Event Boundaries in Egocentric Activity Recognition from Photostreams”, 3rd Workshop on Egocentric Perception, Interaction and Computing (EPIC), extended Abstract, 2018.
- **E. Talavera**, A. Cola, N. Petkov, P. Radeva, “Towards Egocentric Person Re-identification and Social Pattern Analysis”, 1st Applications of Intelligent Systems (APPIS), pp. 203-211, published in the proceedings in the series *Frontiers in AI and Applications* (IOS Press), 2018.
- G. Oliveira-Barra, M. Bolaños, **E. Talavera**, A. Dueñas, O. Gelonch, M. Gardera, “Serious Games Application for Memory Training Using Egocentric Images”, ICIAP, published in proceedings as Chapter Springer Verlag, 2017.
- **E. Talavera**, N. Strisciuglio, N. Petkov, P. Radeva, “Sentiment Recognition in Egocentric Photostreams,” 9th Iberian Conference on Pattern Recognition and Image Analysis (IBPRIA), pp. 471-479, Pattern Recognition and Image Analysis, published in proceedings as Chapter Springer Verlag, 2017
- **E. Talavera**, P. Radeva, N. Petkov, “Towards Egocentric Sentiment Analysis,” 6th International Conference on Computer Aided Systems Theory (EUROCAST), pp 297-305, published in proceedings as Chapter Springer Verlag, 2018.
- **E. Talavera**, N. Petkov, P. Radeva “Towards Unsupervised Familiar Scene Recognition in Egocentric Videos,” In 8th GI Conference on Autonomous Systems, pp. 80-91, published in proceedings as Chapter VDI Verlag, 2015.
- M. Bolaños, R. Mestre, **E. Talavera**, X. Giro-i-Nieto, P. Radeva, “Visual Summary of Egocentric Photostreams by Representative Keyframes”, In International Workshop on Wearable and Ego-vision Systems for Augmented Experience (WEsAX), pp. 1-6, published in the proceedings, 2015.
- **E. Talavera**, M. Dimiccoli, M. Bolaños, M. Aghaei, P. Radeva, “R-Clustering for Egocentric Video Segmentation,” 7th Iberian Conference on Pattern Recognition and Image Analysis (IBPRIA), pp. 327-336, Pattern Recognition and Image Analysis, Chapter Springer Verlag, 2015.

Research Fund

- APIF Predoctoral Scholarship from University of Barcelona - led by Prof. Petia Radeva, Spain. Term: from July 2018 to March 2019.
- ICREA Predoctoral Scholarship from University of Barcelona - led by Prof. Petia Radeva, Spain. Term: from March 2017 to July 2018.
- Promovendus PhD Scholarship from University of Groningen - led by Prof. Dr. Nicolai Petkov. Term: from February 2015 to February 2017.
- Collaboration Grant within the project “Internacionalització de projectes d’investigació AR000312 HORIZON 2020” - led by the Prof. Petia Radeva, Spain. Term: from September 2014 to January 2015.

Summer Schools

- ICVSS, International Computer Vision Summer School, Siracusa, Sicily, 11-16th July 2015.

Talks

- “Deep Learning and applications to activity recognition from Egocentric Photostreams”, Tutorial at the 1st International Conference on Applications of Intelligent Systems, AP-PIS 2018, together with Prof. Petia Radeva and MSc. Marc Bolaños (Las Palmas, Spain).
- Oral presentation in the 1st 3 Minutes Thesis Competition organized by the University of Groningen, March 2018.

Organized Seminars

- Member of the Program Committee for the PhD Day of 2016 at the University of Groningen.
- Organization member as volunteer at CAIP 2015, in Valletta, Malta.
- Organization member as volunteer at APPIS 2017, in Gran Canarias, Spain.

Followed Courses

- University Teaching Skills, duration of 70h, from the University of Groningen, 2019.
- Supervising thesis students/Begeleiden van thesisstudenten, from the University of Groningen, 2019.

Teaching duties

- Co-lecturer in the course Introduction to Intelligent Systems, in the bachelor of Computer Science, from the University of Groningen, Sept - Nov 2019.
- Main lecturer in the course Software Engineering, in the bachelor of Computer Science, from the University of Groningen, Feb - Jun 2019.
- Teacher Assistant in the course Artificial Vision, in the bachelor of Computer Science, from the University of Barcelona, fall semester 2017-2018 and 2018-2019

About the Author



Estefanía Talavera Martínez was born on September 21st, in Torreperogil, Jaén, within the region of Andalucía (Spain). When she was 9 she moved to Mallorca with her family.

For her undergraduate studies she joined the Degree in Industrial Engineering, specialized in Industrial Electronics, from the University of the Balearic Islands (UIB). The subject *Industrial Vision* dragged her attention to the computer vision world. In 2012, she moved to Barcelona and joined the M.Sc. in Biomedical Engineering, from Polytechnical University of Catalunya (UPC) and University of Barcelona (UB). It was there when she met Prof. Petia Radeva, with whom she made her first steps into the egocentric vision topic. She finished her master thesis "Towards unsupervised lifelogging video segmentation" with a qualification of 9.5/10.

In a hot summer day in Mallorca, August 2014, she received an email from Prof. Nicolai, her application for a 4 years joint PhD with the University of Groningen had been accepted. From February 2015 she started her PhD journey under the supervision of Prof. Nicolai Petkov (RUG) and Prof. Petia Radeva (UB), through the Ubbo Emmius program.

In 2016, she joined the Program Committee for the organization of the PhD Day 2016, a conference organized by and for PhD students of the University of Groningen. This experience allowed her to improve her organization skills.

Her research interests are in the field of image analysis, more specifically egocentric vision and medical imaging. In her studies she proposed several techniques for egocentric images analysis, such as inferred sentiment computation from visual and semantic features extracted from the images, and behavioral patterns analysis by describing routines, understood as the repetition of activities.

She balances her life by dancing salsa, hanging out with friends, visiting family in Mallorca, and traveling around the world.