UNIVERSITAT DE BARCELONA

Social Signal Processing from Egocentric Photo-Streams

Maedeh Aghaei

Social Signal Processing from Egocentric Photo-Streams

Maedeh Aghaei

Social Signal Processing from Egocentric Photo-Streams

Thesis submitted to
University of Barcelona
Department of Mathematics and Computer Science
for the award of the degree

of

Doctor of Philosophy

by

Maedeh Aghaei

UNIVERSITAT DE
BARCELONA

Department of Mathematics and Computer Science
University of Barcelona

May 2018

Director:	Dr. Petia Radeva
Co-director:	Dr. Mariella Dimiccoli
Thesis committee:	Dr. Dima Damen
	Dr. David Masip
	Dr. Lluis Fuentemilla

Acknowledgment

Reaching the final line of a book is always a bittersweet farewell. Now that this book came to its end, I would like to gratefully mention those who helped me in reaching the final line. Today, I am a better version of myself and I owe you all many of it. I can only but thank for every minute shared.

First and foremost, I would like to thank my supervisors, Dr. Petia Radeva and Dr. Mariella Dimiccoli, for their invaluable guidance, support, motivation, and encouragement. Thank you Petia for believing in me since my Master thesis and for giving me the opportunity to realize the PhD studies. Thank you Mariella for walking this path with me.

My sincere gratitude also goes to Prof. Marco Cristani, for providing me the opportunity of joining their team in the University of Verona, for his passionate interest, and precious guidance. Thank you Marco.

My greatest appreciation goes to my family, for sheltering me with their love along the good, and not-so-good moments over all these years.

I truly want to name each and every friend that I made in the MAIA lab. To Estefania, for her most sincere friendship, energy, care, and support. To Bea, for her pure kindness and lovely companionship. To Nadia, for all the chats, laughter, and teas. To Marc for all the memories, experiences, travels, and funs. To Edu, for all the Risas and Gambas Rebozadas. To Alejandro, for keeping my company in the empty UB during the sleepless deadlines. To Julio, to Gabriel, to Eduardo, and to Juan Luis for their invaluable companionship over these years. Thank you guys, without you MAIA is not any VIP. I also would like to mention my lab mates in the University of Verona. To Marco Godi, for the great amount of fun involved in working with you. Thank you Marco Carletti, Irina, Irtiza, Francesco, Pietro, Davide, and Mateo for your hospitality and all the nice moments we shared.

I cannot leave unmentioned my supportive and understanding friends out of university. Cheers Aida, to all the solutions we figured for the world's problems over our wine nights! Thanks Mohammad, for your unconditional and heart-warming friendship. Thanks Cherry, for the unique being you are. Gracias Cris, y gracias Belen, que nada compara con un día de la playa, una excursión en Verona y una practica de yoga con vosotras.

This work was mainly supported by APIF grant from University of Barcelona, and Cooperint grant from University of Verona. I would like to thank all the involved administrative staff in both universities who made my path easier to ride.

Last, but definitely not least, my most sincere appreciation to Pasquale, for making the last miles of this marathon the loveliest to me.

Thank you. Gracias. Grazie.

Abstract

Wearable photo-cameras offer a hands-free way to record images from the camera-wearer perspective of daily experiences as they are lived, without the necessity to interrupt recording due to the device battery or storage limitations. This stream of images, known as egocentric photo-streams, contains important visual data about the living of the user, where social events among them are of special interest. Social interactions are proven to be a key to longevity and having too few interactions equates the same risk factor as smoking regularly. Considering the importance of the matter, there is no wonder that automatic analysis of social interactions is largely attracting the interest of the scientific community.

Analysis of unconstrained photo-streams however, imposes novel challenges to the social signal processing problem with respect to conventional videos. Due to the free motion of the camera and to its low temporal resolution, abrupt changes in the field of view, in illumination condition and in the target location are highly frequent. Also, since images are acquired under real-world conditions, occlusions occur regularly and appearance of the people undergoes intensive variations from one event to another.

Given a user wearing a photo-camera during a determined period, this thesis, driven by the social signal processing paradigm presents a framework for comprehensive social pattern characterization of the user. In social signal processing, the second step after recording the scene is to track the appearance of multiple people who are involved in the social events. Hence, our proposal begins by introducing a multi-face tracking which holds certain characteristics to deal with challenges imposed by the egocentric photo-streams. Next step forward in social signal processing, is to extract the so-called social signals from the tracked people. In this step, besides the conventionally studied social signals, clothing as a novel social signal is proposed for further studies within the social signal processing. Finally, the last step is social signal analysis, itself. In this thesis, social signal analysis is essentially defined as reaching an understanding of social patterns of a wearable photo-camera user by reviewing captured photos by the worn camera over a period of time. Our proposal for social signal analysis is comprised of first, to detect

social interactions of the user where the impact of several social signals on the task is explored. The detected social events are inspected in the second step for categorization into different social meetings. The last step of the framework is to characterize social patterns of the user. Our goal is to quantify the duration, the diversity and the frequency of the user social relations in various social situations. This goal is achieved by the discovery of recurrences of the same people across the whole set of social events related to the user.

Each step of our proposed pipeline is validated over relevant datasets, and the obtained results are reported quantitatively and qualitatively. For each section of the pipeline, a comparison with related state-of-the-art models is provided. A discussion section over the obtained results is also given which is dedicated to highlighting the advantages, shortcomings, and differences of the proposed models, and with regards to the state-of-the-art.

Keywords: Social Signal Processing, Egocentric Vision, Low frame-rate wearable cameras, Photo-Streams, Multi-Face Tracking, Social Interaction Detection, Social Interaction Categorization, Face Clustering, Social Pattern Characterization, Clothing social signal.

Resumen

Las cámaras portables ofrecen una forma de capturar imágenes de experiencias diarias vividas por el usuario, desde su propia perspectiva y sin la intervención de este, sin la necesidad de interrumpir la grabación debido a la batería del dispositivo o las limitaciones de almacenamiento. Este conjunto de imágenes, conocidas como *secuencias de fotos egocéntricas*, contiene datos visuales importantes sobre la vida del usuario, donde entre ellos los eventos sociales son de especial interés. Las interacciones sociales han demostrado ser clave para la longevidad, el tener pocas interacciones equivale al mismo factor de riesgo que fumar regularmente. Teniendo en cuenta la importancia del asunto, no es de extrañar que el análisis automático de las interacciones sociales atraiga en gran medida el interés de la comunidad científica.

Sin embargo, el análisis de secuencias de fotos impone nuevos desafíos al problema del *procesamiento de las señales sociales* con respecto a los videos convencionales. Debido al movimiento libre de la cámara y a su baja resolución temporal, los cambios abruptos en el campo de visión, en la iluminación y en la ubicación del objeto son frecuentes. Además, dado que las imágenes se adquieren en condiciones reales, las oclusiones ocurren con regularidad y la apariencia de las personas varía de un evento a otro.

Dado que un individuo usa una cámara fotográfica durante un período determinado, esta tesis, impulsada por el paradigma del procesamiento de señales sociales, presenta un marco para la caracterización integral del patrón social de dicho individuo. En el procesamiento de señales sociales, el segundo paso después de grabar la escena es rastrear la apariencia de varias personas involucradas en los eventos sociales. Por lo tanto, nuestra propuesta comienza con la introducción de un seguimiento de multiples caras que posee ciertas características para hacer frente a los desafíos impuestos por las secuencias de fotos egocéntricas. El siguiente paso en el procesamiento de señales sociales es extraer las señales sociales de las personas bajo análisis. En este paso, además de las señales sociales estudiadas convencionalmente, en esta tesis se propone *la vestimenta* como una nueva señal social para estudios posteriores dentro del procesamiento de señales sociales. Finalmente, el último paso, es el análisis de señales sociales. En esta tesis, el análisis

de señales sociales se define esencialmente como la comprensión de los patrones sociales de un usuario de cámara portable, mediante la revisión de fotos capturadas por la cámara llevada durante un período de tiempo. Nuestra propuesta para el análisis de señales sociales se compone de diferentes pasos. En primer lugar, detectar las interacciones sociales del usuario donde se explora el impacto de varias señales sociales en la tarea. Los eventos sociales detectados se inspeccionan en el segundo paso para la categorización en diferentes reuniones sociales. El último paso de la propuesta es caracterizar los patrones sociales del usuario. Nuestro objetivo es cuantificar la duración, la diversidad y la frecuencia de las relaciones sociales del usuario en diversas situaciones sociales. Este objetivo se logra mediante el descubrimiento de apariciones recurrentes de personas en todo el conjunto de eventos sociales relacionados con el usuario.

Cada paso de nuestro método propuesto se valida sobre conjuntos de datos relevantes, y los resultados obtenidos se evalúan cuantitativa y cualitativamente. Cada etapa del modelo, se compara con los trabajos relacionados más recientes. También, se presenta una sección de discusión sobre los resultados obtenidos, que se centra en resaltar las ventajas, limitaciones y diferencias de los modelos propuestos, y de estos con respecto al estado del arte.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

We are living in interesting times, where Artificial Intelligence is hard wired with every aspect of our living. Scientists continuously develop better *imitation games* for computers to more effectively mimic functions of the human brain. Today, smart machines are intended to exhibit intelligence in all aspects of the human intelligence, as it is not only restricted to IQ, but is interlaced with a wide range of cognitive modalities [5]. Theory of multiplicity of intelligence was first characterized by Howard Gardner in his book *Frames of Mind: The Theory of Multiple Intelligences*. Gardner argues that one important modality of intelligence is *interpersonal* or *social intelligence* and explores how human beings react to the world and interact with it and each other.

The importance of social intelligence is indisputable. It is constantly considered as an invaluable factor to determine the quality of life and is consistently associated with better outcomes across the lifespan, ranging from academic achievement and substance use in adolescence to mental and physical health and longevity in adulthood [6, 7]. Empirical pieces of evidence from research studies have repeatedly shown that stronger social relationships are associated with feeling happier, better coping with daily and major life stressors, and consequently living a longer life [8, 9]. As a result, automatic recognition of social interactions from images and videos has increasingly drawn scientific interest [10]. Recognition of social

(a) Surveillance perspective (b) Sousveillance perspective

Figure 1.1: Figure of a social interaction from surveillance perspective (a) and sousveillance perspective (b).

interactions, relying solely on visual cues is a valuable task from the computer vision perspective that confines the analysis to the visual information, eliminating the need for acquiring additional information and major privacy concerns. The emergence of Social Signal Processing (SSP) domain is a consequence of realizing the importance of the matter by the researchers. In fact, the pursued goal by SSP community is to provide machines with a naturalistic social intelligence similar to human social behavior.

Early works on SSP were motivated mainly by video surveillance applications [11, 12]. Surveillance cameras, however, capture the environment from the fixed and external third-person perspective and fail in capturing real involvement in social interactions at the personal level. In contrary to surveillance cameras, wearable cameras offer the possibility of capturing social cues from a more intimate perspective, known as ego-vision. Wearable cameras allow capturing natural photos of the daily interactions of camera-wearers, where they naturally attempt to reach a clear view of whom they are engaged in during a social interaction (see Fig. 1.1).

Vinciarelli et al. [13] suggest to formalize SSP in a four-step pipeline in which, after having recorded the scene and detected humans (step 1 and 2), in step 3, feature extraction has to be performed, where features are behavioral cues whose interpretation brings to individuate social signals. In step 4, social signals have to be grounded in the scene context, in order to understand social interactions. This thesis goes in the direction of SSP in egocentric photo-streams, where the scene

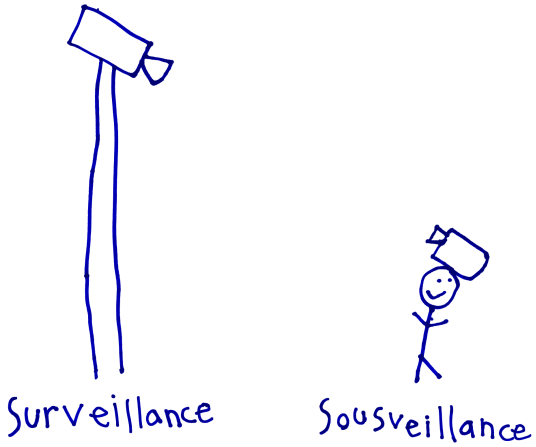


Figure 1.2: A drawing by Steve Mann's six-year-old daughter, illustrating surveillance versus sousveillance. Both words are French, which means "to watch" (veillance), "from above" (sur), or "from below" (sous). Figure is adapted from [1].

is recorded by a Narrative photo-camera[1], people are detected and tracked within our multi-face tracking proposal, and social signals are extracted and analyzed for each tracked person in various levels to reach a comprehensive understanding of the social pattern of the wearable photo-camera user.

SSP in egocentric photo-streams despite being in its initial phase, has already attracted the attention of the technological community. The MIT Fifth Sense Project [2] is an example. For people with impaired vision system, performing activities of daily living that sighted people typically perform without additional effort is critical. Within this project, researchers use machine learning methods to detect and model social interactions of a wearable camera user or a robot, aiming at providing visually-impaired users with better awareness of their social context.

1.2 Background

The concept of using a wearable camera as a monitoring modality dates back to the WearComp work of Mann in 1998 [14, 15]. However, it was not until the introduction of Microsoft SenseCam in 2004 that researchers began experimenting with large scale egocentric recordings of human life (lifelogging) and its health applications [16]. Since then the topic received some attention until 2012, when

[1] http://getnarrative.com/
[2] http://people.csail.mit.edu/teller/misc/bocelli.html

Kanade and Hebert [17] argued that the egocentric perspective is an inverse to the traditional surveillance perspective and that it "senses the environment and the subjects activities from a wearable sensor, is more advantageous [than surveillance] with images about the subjects environment as taken from his/her viewpoints". It was in the same year that TIME magazine reported that wearable cameras "will transform society because they introduce a two-sided surveillance and sousveillance"[3]. From 2012 to date, the topic is receiving an exponentially ascending attention from the science and technology communities [18].

Recording of an activity by the performer of the activity is referred to as sousveillance. For the facility, sousveillance is typically performed by a small wearable or portable camera [19]. The literal translation of the term "sousveillance" from French is "observation from below", where *below* can be either interpreted physically (mounting cameras on people rather than on height), or hierarchically (crowd doing the watching, rather than higher authorities) [20]. In this regard, a subset of sousveillance is defined as *inverse surveillance* aiming at performing a watchful vigilance from the perspective of a participant in a society [21]. Although this subset of sousveillance has its proper applications [22, 23], sousveillance typically involves recording by ordinary people from first-person perspectives.

Personal sousveillance happens normally out of different purposes such as art, science, or technology. An example is Alberto Frigo, the conceptual media artist which is perhaps the most extreme example of lifelogger which under the *2004-2040* project is continuously documenting 36 aspects of his life to understand himself at his 60 years of age[4]. Another example is the company of *Nestle* which in a project conceptually designed to *reverse the male gaze*, used a bra as a point-of-view for a camera[5]. In science and technology, Kanade and Hebert for the first time described a prototypical sousveillance in 2012 that is composed by three basic components: a localization component to estimate the surrounding, a recognition component to identify object and people, and an activity recognition component to provide information about the current activity of the user. Together, these three components provide a complete situational awareness of the user. Following the proposed idea by Kanade and Hebert, the first computational techniques for egocentric analysis focused on hand-related activity recognition [24] and social interaction analysis [25]. Also, given the unconstrained nature of the video and the

[3]http://techland.time.com/2012/11/02/eye-am-a-camera-surveillance-and-sousveillance-in-the-glassage/

[4]http://www.2004-2040.com

[5]http://time.com/3449830/nestle-bra-cam-breasts/

huge amount of data generated, temporal segmentation [26] and summarization [27] were among the first addressed problems. To date, researchers have explored the use of egocentric vision for diverse topics including: activity recognition [28], social saliency estimation [29], multi-agent egocentric vision systems [30], privacy preserving techniques [31], attention-based activity analysis [32], hand pose analysis [33], Ego graphical User Interfaces (EUI) [34], and understanding social dynamics, and attention [35].

Today, the wearable cameras are small, lightweight, and thus convenient to use for digital recording of visual information from daily interactions of the camera-wearer without the need for the user intervention (see Fig. 1.3). These cameras depending on the frame-rate commonly can be classified as photo-cameras and video-cameras. The former (e.g., Narrative Clip and Microsoft SenseCam), are usually worn on the chest, and are characterized by a low frame-rate (up to 2 fpm) that allows capturing images over a long period of time without the need of recharging the battery. The sequence of images captured by egocentric photo-cameras are commonly referred to by *photo-streams*. Consequently, these cameras offer considerable potential for inferring knowledge about behavior patterns and lifestyle of the user. However, due to the low frame-rate and the free motion of the camera, temporally adjacent images typically present abrupt appearance changes so that motion features cannot be reliably estimated. The latter (e.g., Google Glass, GoPro), are commonly mounted on the head, and due to their high frame-rate (around 35 fps) allow to capture fine temporal details of interactions. Consequently, they offer the potential for in-depth analysis of special activities of the camera-wearer. However, since the camera is moving with the wearer's head, global motion estimation of the wearer becomes unfeasible and images can result blurred frequently due to abrupt movements of the head.

This thesis goes in the direction of SSP in the domain of egocentric photo-streams to achieve a broad understanding of the social patterns of a wearable photo-camera user. In SSP, social signals are known as a bunch of non-verbal behavioral cues that occur over short-time intervals and people usually use them to express themselves when engaged in a social situation. The term behavioral cue is typically used to describe a set of temporal changes in neuromuscular and physiological activities that one exhibits in a certain social situation, but the definition can include a broader context. According to Vinciarelli et al. [13], observable behaviors can be classified into four main categories as physical appearance (beauty, attractiveness, etc.), gesture and posture (reading of sign language, human affect present in body parts, etc.), face and eye behavior (smile, frown, pain, etc.), and

(a) WearComp - 1988 (b) Narrative - 2017

Figure 1.3: (a) Steve Mann's Visual Filter for continuous live webcast as well as viewing, (b) Narrative photo-camera.

space and environment (physical proximity, seating arrangements, etc.). To provide a sensible example, when people get involved in social interactions, they tend to stand in determined close positions to other people to avoid occlusions (space and environment social signal) and organize orientations to naturally place the focus on the subjects of interest (face and eye behavior). This phenomenon was first studied in sociology by Kendon in the theory of F-formation [36]. F-formation is defined as a pattern that people instinctively maintain when interacting and can be measured based on the mutual distances and orientations of the individuals in the scene.

Adoption of the F-formation theory by the computer vision community was a foot-stone in formalizing the problem [37, 38]. As defined by Kendon, "an F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access". F-formation comprises of 3 spaces: the people involved in an interaction stand in the *p-space*, where they all look inwards to a common empty space surrounded by the *p-space* that forms the *o-space*. External people who do not belong to this interaction are not accepted in the *p-space* and they belong to any space outside of the *p-space* known as the *r-space* (see Fig.

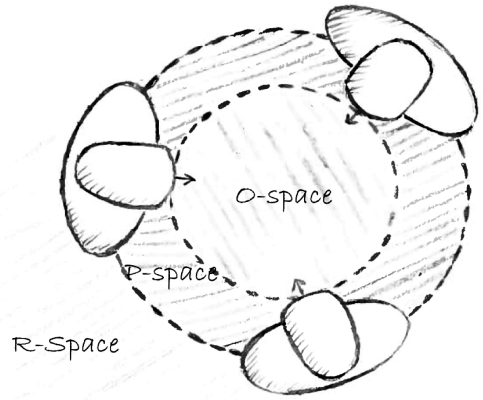


Figure 1.4: Example of the F-formation: people involved in the social interaction stand in p-space. Common empty space surrounded by the p-space, where every participant is looking into it, forms the o-space. Any space outside of these two space is r-space, that is out of the scope of this social interaction.

1.4).

Traditionally, the studied social signals in computer vision were one or more among the aforementioned observable behavioral cues for further analysis of a social event. This thesis tackles the problem of social pattern characterization of a wearable photo-camera user by adopting SSP related technologies in the domain of egocentric photo-streams. Specifically, due to the lack of a public dataset, the EgoSocialStyle dataset is recorded to enable design and development of SSP related techniques for addressing the problem. After recording of the scene which is considered as the first step in SSP, in the second phase, a multi-face tracking model is introduced which is designed to cope with the particularly induced challenges by the domain. The third phase is to extract relevant social signals from the tracked people and the recorded scene, which enable analysis of the problem in the following phase. In addition to the conventionally studied social signals in SSP, the role of some social signals which traditionally received less attention, such as 3D head pose, facial expression, and clothing of the individuals are also explored. As in the last step in SSP, social signals are inspected within various modules, to reach an understanding of the characteristics of social patterns of the user.

SSP in egocentric vision introduces novel advantages as well as novel challenges with regards to third-person vision (see Sec. 1.3). Following, we summarize some

of the offered capabilities by the egocentric vision which classically have been considered by the community to tackle related problems in this domain:

- Ego-vision provides a continuous paradigm for sampling the visual world of the camera-wearer. Egocentric videos represent what the user daily sees, while third-person videos only capture a fraction of the life of the user which fits within the static and limited camera field of view. In this manner, ego-vision offers important information about social living of the user as lived.

- An egocentric video is personalized and corresponds to what a particular person sees. As a result, the algorithms and techniques can be personalized to the characteristics and preferences of that person.

- Ego-vision actively captures moments that a person might be passively living. It records important social life moments of the person where the user naturally turns to have a clear view of other involved people and objects in the interaction.

- In ego-vision, the context of the action is firmly present. Perception benefits from the context and it plays an important role in recognizing the category of social activities. For example, if an interaction takes place in a conference room, that interaction is most likely related to a formal meeting. Time of the day is another source of contextual information. Often individuals follow a daily routine of interactions in their daily activities where more probably their informal gatherings occur during the evenings.

With this historical context in mind, throughout this thesis, we study that in the egocentric photo-stream setting:

- Continuous recording aids in detection, categorization, and characterization of social interactions of the user.

- The personalized nature of the ego-vision enables personalized characterization of social interactions of a camera-wearer.

- Personalization also provides the possibility of predicting social routine of the user.

- It is possible to extract the context of a social interaction, and social context is indeed an important factor in characterizing social routines of the camera-wearer.

Figure 1.5: Example of a sequence captured by Narrative clip camera inside a train. Two people are seated in front of the camera-wearer, which one of them the camera-wearer is interacting with?

1.3 Research Issues

Wearable cameras are small and lightweight. They acquire first-person images and videos automatically, without the user intervention, with different resolutions and frame-rates. However, ego-vision besides providing various opportunities that make SSP tasks simpler, it introduces new challenges (see Fig. 1.5):

- The wearable camera regardless of being head mounted or chest mounted, is worn in a naturalistic setting. This leads to huge variability in visual data in terms of background variation, illumination conditions, and object appearance.
- The camera-wearer is not visible in the image and what he/she is doing has to be inferred from the information in the visual field of the camera, implying that important information about the wearer, such for instance as pose or facial expression estimation, is not available.
- In the case of photo-cameras with low frame-rate, it is common to experience drastic visual changes in even temporally adjacent photos. This issue leads to unavailability of the information which traditionally could be extracted from the temporal coherency among the frames, such as optical flow.
- The field of view of the wearable cameras due to proximity to the scene and objects is not specifically wide. Therefore, many conventional detection and recognition algorithms such as those that require seeing the limbs and the joints of the subject, in practice are not applicable to ego-vision.
- Egocentric field of view is incapable of capturing socializing moments that do not involve face-to-face interaction (as when walking together).

Besides the aforementioned issues which are consequences of SSP in ego-vision, there are some issues that are directed to SSP independently from the perspective of the camera.

- Despite all the recent advances in SSP, design, and development of automated systems for unveiling the conveyed meaning by some behavioral cues such as blinks, smiles, crossed arms, etc. in a social situation remains unsolved.

- SSP is not standalone and fragmentation of the task over several scientific communities including those in psychology, computer vision, and signal processing, makes it specifically difficult.

- In SSP many issues are still open. For instance, there is a lack of proper machine vision models to detect and analyze human social behavior in different scales and the appropriate psychological and cognitive theories that can provide useful concepts for them.

- The legal, ethical, and policy issues surrounding the ego-vision are still arguably unsolved and leave a large room for further exploration.

1.4 Research Contributions

This thesis argues that SSP can be effectively formalized in the domain of egocentric photo-streams. Throughout this thesis, we demonstrate that SSP is dividable into a set of sub-tasks and each sub-task becomes resolvable through leveraging social signal extraction techniques and learning descriptive models considering the continuity of social signals along the photo-streams. These models are generic models which can be easily adapted to specific scenarios for personalized social interaction analysis.

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

- We introduce a new model for multi-face tracking in the domain of egocentric photo-streams (see chapter 3). In chapter 3, we explain the associated features of our extended-bag-of-tracklets proposal for multi-face tracking that helps to effectively overcome the drastic visual variations of faces and discontinuity that is imposed by the photo-streams. Our proposed method relies on the deep-matching approach for finding the face correspondences along a sequence, which is robust against visual variations imposed by the wearable camera and its low temporal resolution. Our proposed model is robust in detecting face occlusions and is able to localize them. Within our proposed model for multi-face tracking, we also proposed to assign to each tracklet

a confidence value which determines how likely correct is the final tracking results. This is an important factor in designing trackers as it facilitates further analysis of tracklets. The main results of chapter 3 are published in a conference in 2014 [39] and in the CVIU journal in 2016 [2].

- We present new models for social interaction detection in egocentric photo-stream setting (see chapter 4). In this chapter, we present two models: one is based on frame-level analysis of social interactions, and the other is based on event-level analysis. In the development of both of our proposed models, we adapted the sociological notion of F-formation into machine vision analysis. In our proposed analysis, we studied the role of considered relevant features in the psychological studies for detection of social interactions. Specifically, we studied the role of *facial expressions* in automatic detection of social interaction. In this chapter, we prove the importance of the 3D head pose of individuals in addition to the previously studied *yaw* head pose in the social interaction detection. Moreover, we present a comparative discussion over the obtained results by each frame-level and sequence-level analysis and report the robustness of the sequence-level analysis over the frame-level analysis. The main results given in this chapter are published in two conferences [40, 41], and partially in the CVIU journal 2018 [42].

- We present a new pipeline for categorization of social interactions into two broad categories of formal and informal meetings (see chapter 5). Our proposed method based on an extensive body of literature suggests to study high-level features describing the environment where the social interaction takes place as the most relevant feature in this analysis. In our proposed model, we also demonstrate the role of *facial expressions* of the involved people in the interaction in the categorization task. For the analysis of features, we propose a frame-level as well as an event-level method and demonstrate the advantages of the event-level analysis. We present comparative results with the state-of-the-art models and report the superiority of our proposed model. The main results given in this chapter are published in the CVIU journal 2018 [42].

- We propose a new model for face clustering in the domain of egocentric photo-streams (see chapter 6). Our proposed method is built upon a multi-face tracking model and is designed to calculate the similarities between face-sets instead of face-examples relying on the similarity score obtained

by applying the deep-matching approach. Upon calculating the similarities among face-sets and employing both inner-track as well as inter-track constraints, the agglomerative clustering with a previously learned threshold is applied to decide on the final cluster members where each cluster ideally belongs to the face appearance of one person across the dataset. In the same chapter, we also provide a wide comparison with the baseline models to emphasize the importance of each component of our proposed model. Also, to prove the robustness of our proposed model, a comparison with a relevant state-of-the-art model is provided. The main results of this chapter are previously published in a conference [3].

- We propose a new pipeline which aggregates together the findings in the previous chapters to draw a comprehensive image of the social pattern of a wearable photo-camera user (see chapter 7). In this chapter, we formally define the frequency, diversity, social trend, and duration of a social interaction and demonstrate that social interactions of a user can be characterized according to these four terms, generally and specifically with a certain person. We prove our claim quantitatively and qualitatively and draw a sensible conclusion out of the temporal map of the social interactions of the user. In addition to demonstrating the obtained results over our proposed dataset, EgoSocialStyle, we also report the result of our proposed model over the public dataset, EGO-GROUP. The main results of this chapter are reported in the CVIU journal 2018 [42].

- In addition to study conventional social signals, in the format of a position paper, we outline the main steps of the first systematic analysis towards relieving the relationship between clothing and social signals, from the SSP perspective (see chapter 8). In this chapter, in a *question answering* format, we propose a framework within the scope of computer vision to measure the effect of clothing in SSP, as a sender or receiver of the social signals. Our study is built on top of reviewing a vast amount of related literature in sociology, psychology, and computer vision. In this chapter, we also mention that our future goal is to reach an understanding of the human personalities through observing their clothing patterns and report the results of the first steps taken by us towards this analysis. The main findings of this chapter are reported in a conference [43].

1.5 Thesis Organization

This thesis begins by providing an insight into the background of social signal analysis in computer vision in the next chapter. Chapter 3 introduces tracking, the second step towards social interaction analysis in SSP. Chapter 4 is devoted to social interaction detection and chapter 5 details the proposed approach for social interaction categorization. Chapter 6 is dedicated to the problem of face clustering in egocentric photo-streams. Details about the characterization of social interactions in egocentric photo-streams are discussed in chapter 7. Chapter 8 covers details about the relations between clothing and social signals and, chapter 9 highlights the main conclusions and holds discussions about the possible future paths.

CHAPTER 2

Previous Work

2.1 Multi-Face Tracking

Despite the importance of *tracking* in the analysis of social interaction, this problem received less attention in ego-vision than the same problem in third-person vision [44]. Tracking in ego-vision is a different problem from the tracking in conventional videos in several aspects. Conventional tracking facilitates itself with the assumption of temporal coherence among visual information present in the video frames, while temporal coherence does not hold for egocentric photo-streams. Moreover, in egocentric photo-streams, the appearance of the target, as well as its position, may change drastically from frame to frame. In addition, due to changes in the camera field of view caused by body movement of the camera-wearer, background modeling becomes a more challenging issue (see Fig. 2.1).

When reviewing the state-of-the-art trackers, two main categories of conventional trackers can be found: offline trackers and online trackers. The former assumes that object detection in all frames has already been performed and trajectory construction is achieved by linking different detections and tracks in offline mode [45, 46, 47]. This property of offline trackers allows for global optimization of the path and thus, makes them potentially suitable for dealing with large visual variations of the objects among the frames of photo-streams. As an example, Berclaz et al. [45] reformulate the linking step between detections and trajectories as a constrained flow optimization approach, which results in a convex problem

Figure 2.1: Example of a sequence captured by Narrative clip camera during an interaction. Changes in the target location and appearance due to the changes in the camera movement can be appreciated.

that can be solved using the k-shortest paths algorithm. In order to overcome the noisy probabilities of candidates that may be produced by the object detector, the authors arranged a set of assumptions including the limited motion of the target. Zamir et al. [47] solve the data association problem for one object at a time, while implicitly incorporating the rest of the objects using global association by employing *Generalized Minimum Clique Graphs* (GMCP). GMCP incorporates both motion and appearance model over the whole temporal span for optimization. In the development of aforementioned trackers, the authors assume a rather fixed or predictable position for targets in the adjacent frames of the video. Although this assumption is generally applicable in conventional videos, it does not hold in the egocentric photo-streams setting.

In comparison with offline trackers, for online trackers, the target position is provided in the initial frame and the tracker needs to establish the state of the target in the following frames of the video. Among state-of-the-art online trackers, those that are relatively tolerant to occlusion and drastic appearance changes, are more suitable for egocentric photo-streams [48, 49, 50, 51]. Kalal et al. presented a *Tracking, Learning, Detection* (TLD) framework [52], which works by training a discriminative classifier over labeled and unlabeled examples. This method performs well in handling short-term occlusions but strongly relies on optical flow, which cannot be applied in low temporal resolution sequences. *Compressive Tracking* (CT) [48], uses an appearance model based on features extracted in a compressed sensing domain. This method is relatively robust to changes in appearance and performs favorably in challenging datasets, outperforming TLD. However, CT is not robust to large displacements of the target, which are very frequent in egocentric sequences. In *Locally Orderless Tracking* (LOT) [50], target and candidates in the new frame are segmented first into superpixels and among the set of candidates, the one which has the least distance to the target is selected as the target in the new frame. LOT tracker offers adaptation to object appearance variations by matching with flexible rigidity through measuring the

distance between superpixels. Similar to LOT, *SuperPixel Tracker* (SPT) [49] extracts superpixels of the target. SPT extracts the color histograms of the superpixels from the first 4 frames and based on these features, clusters superpixels by using mean-shift. A confidence value is assigned to each cluster, from which the superpixels confidence of all pixels of the cluster is derived. In the next frame, the candidate window with the highest confidence summed over all superpixels in the window is selected as the new target. Mei et al. presented L1O [51] as a tracker which explicitly detects occlusions. In L1O, the candidate windows with a reconstruction error above a threshold are selected for L1-minimization. When a certain number of the pixels of the candidate window are occluded, L1O detects an occlusion, which disables the model updating.

Conventional online trackers usually search for the target in the new frame, around its previous position in the current frame. These trackers are mostly dependent on the object appearance in the very first frames and generally require the feature patches in neighboring frames to be close to each other. However, under specific conditions of egocentric photo-streams, such presumptions will result in the gradual departure of the estimated target from the true target state, which eventually leads to tracking loss.

The trackers in *Low Frame Rate* (LFR) videos are the most similar to ours [53, 54]. Li et al. presented a temporal probabilistic combination of discriminative models of different learning and service period, known as their *lifespan* [53]. Each model is learned from different ranges of samples, with different subsets of features, to achieve varying levels of discriminative power. Different models are fused by a cascade particle filter, to achieve multiple stages of importance sampling. However, this work falls into the pre-trained tracking class that its performance also depends on the training data; an issue that we try to avoid, due to the peculiarity of our dataset that presents a relatively small number of images in each trackable segment. A recent work about LFR tracking was presented by Zhou et al. [54]. The authors proposed a *Nearest Neighbor Field* (NNF) driven stochastic sampling framework for abrupt motion tracking. In this work, NNF provides candidate regions, where the target may exist. Smoothing Stochastic Approximate Monte Carlo (SSAMC) sampling scheme predicts the state of the target more effectively. Finally, the method refined the result with a sparse representation based template matching technique.

Although the body of literature regarding tracking is huge, most existing approaches cannot be directly applied to egocentric photo-streams, either because of the unpredictability of motion or because of drastic appearance changes that char-

acterize this data. Furthermore, most of the methods are not able to track multiple targets simultaneously or require the manual specification of the initial position of the target. To this end, we proposed the *Bag-of-Tracklets* (BoT) [39] for tracking in egocentric photo-streams acquired by Sensecam camera (3 fpm). The underlying key idea of our approach is that detection and tracking can be integrated to achieve strong discriminative power. This approach belongs to the offline class of trackers, that allows for general optimization of tracklets. Optimization consists of generating a tracklet for each detected target and categorizing similar tracklets into groups, that should correspond to different persons. This approach simply allows for the rejection of unreliable bag-of-tracklets, and eventually extracts a single prototype for each reliable bag-of-tracklets. The detailed explanation of our proposed pipeline for multi-face tracking in egocentric photo-stream setting is given in chapter 3 of the thesis.

2.2 Social Interaction Detection

2.2.1 Social interaction in computer vision

Microsociology, or social interaction, as defined by Erving Goffman [55] is a process by which people act and react to those around them. The importance of automatic analysis of visual data for the purposes of detection of social interactions has been recognized by the computer vision community within several studies [56, 13]. Most of the previous studies in social interaction computing were focused on finding potential groups of interacting people, also known as Free-standing Conversational Groups (FCG) in conventional still images or videos. In this regard, Groh et al. [57] proposed to use the relative distance and shoulder orientations between each pair of people to measure social interactions on small temporal and spatial scales. This has been done through training a probabilistic classifier which can then be used for characterizing the social context.

In sociology, the introduction of the F-formation theory by Kendon [36] was a foot-stone to formalizing social interaction settings. F-formation is defined as a geometrical pattern that interacting people tend to follow by adjusting their location and orientation towards each other in the space to avoid mutual occlusion. The computer vision community later adopted the F-formation theory to detect groups of interacting people from images and videos [37, 58, 59, 38, 11]. Cristani et al. [37] proposed to solve the task using a Hough-Voting F-Formation (HVFF) strategy to find the common area of interaction by accumulating the density of

the overlapping votes of each interacting person. Built upon a multi-scale Hough-Voting policy, Setti et al. [59] modeled small FCG as well as large groups of people, relying on different voting sessions.

The problem of finding F-formations has also been formulated as finding dominant sets and using proxemics by employing the graph clustering algorithm [58], the graph-cuts framework for clustering individuals [11], heat-map based feature representation of interacting people [38], and defining an intermediate representation of how people interact [60].

2.2.2 Social interaction in ego-vision

The boom of interest in ego-vision during the past few years [18], naturally led to the exploration of social interaction analysis in this setting. For social interaction analysis in an egocentric scenario, the most exploited features are the face location and the pattern of attention of the visible individuals, as well as the head movements of the first-person when the camera is worn on the head.

Fathi et al. [25], proposed a Markov Random Field model to infer the 3D location to which a person is looking at during a social interaction, that relies on the camera intrinsic parameters. They further used this information to classify social interactions into three classes, namely *discussion*, *dialogue*, and *monologue*, depending on the active role played by the participants in the interaction. To the best of our knowledge, this is the only previously introduced work about egocentric social interaction categorization.

Later, Alletto et al. [61] proposed a method for identifying multiple social groups from egocentric videos, that do not rely on the camera intrinsic parameters for 3D projection; hence, the method is applicable to any head-mounted wearable camera. Park and Shi [62] introduced the concept of *social saliency* defined as the likelihood of joint attention from a spatial distribution of social members. A social formation is modeled as an electric dipole moment allowing to encode a spatial distribution of social members using a social formation feature. Recently, [63] proposed to model the dynamics of micro-actions and reactions between two camera-wearer engaged in a dyadic interaction to reach a deeper understanding of the ongoing social interaction between them. In this work, the authors demonstrate that the integration of the first-person perspective of both parties in a dyadic interaction fosters micro-action recognition task in this setting. In another recent attempt, [64] offered to analyze social interaction sequences and detect them applying a Hidden Markov - Support Vector Machine (HM-SVM). Their focus was

on modeling what they called *interaction features*, mainly physical information of head and body.

All the aforementioned works share three main common characteristics. First, the high temporal resolution of videos (30-60 fps), which allows relying on the temporal coherence among video frames to robustly estimate the head pose of appearing people and modeling the foreground. Second, the head-mounted cameras, which permits the modeling of head movements and attention patterns of the user. Third, the common goal by them, that is restricted to finding potential social groups of people in the scene, with exception of [25], that goes deeper into the categorization of social interactions, but strongly relies on the head motion for that.

2.3 Social Interaction Categorization

An important factor in social pattern characterization of a user is *diversity* of social interactions which highlights the density of participation of individuals in various categories of social interactions, i.e. formal or informal category of meetings [65, 66, 67, 68]. *Meetings* are defined as gatherings at which humans communicate, convince, cajole, conspire, and collaborate [69]. In general sociology, a formal meeting is defined as a pre-planned event where two or more people come together at a pre-planned place at a particular time to discuss specific matters for the purposes of achieving a specific goal [69]. An informal meeting is more casual, requires less planning, and usually can take place at any casual space from a park to a dinning hall.

Statistical analysis of the social interactions diversity has been considered as a helpful tool to optimize workspace [67], to minimize the cost of meetings [70], and to maximize the effectiveness of interactions among of group members and in the social structure of a broader organization [71]. However, these studies are carried out in non-automated manners by visually reviewing the images and other involved signals of interest such as sound.

2.4 Face Clustering

Face clustering is a largely unconstrained problem and rich body of work in the literature has focused on finding how to exploit characteristics of the dataset or of the particular application to constrain it. The most common applications are

interactive tagging of photo albums [72, 73, 74] and video organization [75, 76]. In the context of face discovery in photo albums, Lee et al. [72] introduced a new constraint known as *social context of co-occurring people*, following which people of the same social context often appear together. For example, faces of the family members usually tend to co-occur even in different photos. The system first trains a separate detector for each individual and later, uses the detector to discover novel face clusters by taking advantage of co-occurrence constraints. In the same scenario, Zhu et al. [73] presented a Rank-Order distance to measure the dissimilarity between two faces. This work exploits the fact that faces of the same person usually form close sub-clusters in the feature space. A similar idea is proposed by Xia et al. [74], who exploited two constraints: an individual only may appear once in a picture, and the number of instances of the same person must be lower than the total number of pictures. The problem is then formulated as a constrained K-Means, which is solved through Minimum Cost Flow linear network optimization strategy. Imposing constraints to achieve more accurate clustering is observed in several other works attempting to cluster faces in videos. Xiao et al. [75] proposed a Weighted Block-Sparse Low Rank Representation (WBSLRR) which learns a low rank data representation, while considering two defined prior constraints. First, the inner-track constraint states that any two faces in the same face track belong to the same person. Therefore clustering is first performed over face-tracks instead of individual faces. Second, the inter-track constraint that states face-tracks belonging to faces that appear in the same frame, does not belong to the same person. A similar idea has been employed by Cinbis et al. [77], to learn a distance metric for face identification in videos that pulls close together faces in an inner-track relation, and pushes away those in inter-track relation. More recently, as in many other computer vision tasks, deep features proved their efficiency in data representation for face clustering [78, 76]. However, deep learning based approaches are supervised and hence require a previous learning stage involving identity-labeled faces. Therefore, they are most suited for face re-identification.

2.5 Social Interaction Characterization

The crucial role of personalized characterization of the social pattern of a user has been recognized particularly in the medical domain. Related works thoroughly investigate the feasibility of using a wearable camera for personalized health monitoring that leads to increasing the number of positive clinical outcomes. In this

line, Aung et al. [79] and Chow et al. [80] pinpoint how mobile technologies through continuous monitoring, allow precise assessments of human behavior and ultimately individual mental health. In the same path, Hodges et al. [81] and Berry et al. [82] suggested to use wearable cameras for detecting relapse in people affected by depression and Granholm et al. [83] proposed it for ecological momentary assessment of social functioning in schizophrenia. In the context of memory training of people affected by mild cognitive impairment, pictures of social interactions are specially treated to trigger autobiographical memory [84]. Recently, Dhand et al. [85] used wearable cameras for monitoring the lifestyle of stroke survivors and Brown et al. [86] discussed the advantages and disadvantages of incorporating wearable cameras into social psychological research and reported data variation on different social situations. In all the aforementioned studies, the key component is to track social interactions of the user in terms of duration and frequency and to monitor their possible variation over time. Indeed, the importance of *duration* and *frequency* of social interactions in the study of social patterns is well recognized in the literature [87, 88]. In chapter 7, we will introduce a pipeline for automatic analysis of duration, type, frequency, and diversity of social interactions in the context of social pattern characterization from egocentric photo-streams.

2.6 Clothing: A Social Signal Processing Perspective

Despite the rich body of work on the role of behavioral cues in non-verbal social signal processing [89], deeper understanding of social signals requires further cues discovery. In this respect, some visual behavioral cues such as gesture, posture, gaze, physical appearance, and proxemics have received high attention, while other possible features such as clothing have been traditionally little studied [13].

This is an important lack in the social signal processing literature since clothing affects behavioral responses in the form of impression formation or person self-perception. Several past studies in the social sciences aimed to assess this influence, showing that formality of the clothing influences impression of others towards a person [90] as well as the self-perception of people towards themselves [91, 92]. More recently, the influence of clothing on the decision making of individuals has been investigated [93]. A few studies also have shown that clothing may be an indicator of ethnicity, culture, socioeconomic status [94, 95], and even surroundings

of the people [96].

Other studies show that clothing correlates with the personality traits of people in a way that people with formal clothing perceive actions and objects, the inter-relationship and the intra-relationship between them in a more meaningful manner [97]. Clothing can make a person feel comfortable or not in a social situation [98] and can be considered as a determinant of how long it takes for strangers to trust one and how much they may trust them [99]. Aforementioned studies are pieces of evidence of the importance of clothing in social signaling. Arguably, clothing can be considered as the most evident blueprint of individuals, which is completely dependent on their conscious choices, is not as transient as a gesture, and is more evident than any micro-signals such as a sarcastic smile among the facial expressions.

Various experiments have been previously performed to measure the clothing effect on human behavior. According to the critical review of Johnson et al. [100], the effect of clothing on human behavior usually is measured in combination with other variables. However, despite the rich body of work, to the best of our knowledge, all of the previous experimental studies were performed and analyzed manually.

CHAPTER 3

Multi-Face Tracking in Egocentric Photo-Streams

3.1 Introduction

People during a full day may often engage in various social events. In a social event, i.e. a coffee break in a conference, a user by wearing a photo-camera captures the moments that might be of interest for later retrieval. However, the first step towards social event retrieval from these photo-streams is to find and track the appearing people in them. Precisely, people who get engaged in a social event with the user appear in a number of consecutive frames while irrelevant people to the camera-wearer only appear occasionally in the photos-streams and normally do not stay in front of the user for a long time. Hence, by incorporating additional information about the tracked people, their involvement in the social interaction with the user can be analyzed.

Extracting relevant information from egocentric photo-streams is not a trivial task. Indeed, a massive number of unconstrained images can be gathered even over a relatively limited period of time (up to 3000 images per day using the Narrative Clip). Moreover, given the unpredictability of the camera motion and the low temporal resolution of the camera, abrupt changes of the scene occur frequently. By considering the introduced limitations for people tracking in egocentric photo-streams, in this chapter, we present the *extended-Bag-of-Tracklets*

(eBoT) approach by introducing several features that help in increasing tracking robustness in this setting.

The rest of the chapter is organized as follows: in Sec. 3.2, we define the Confidence-based eBoT for multi-face tracking, by performing seed and tracklet generation, grouping tracklets, prototypes extraction and occlusion treatment. In Sec. 3.3, we introduce our experimental setup and discuss comparative results and finally, in Sec. 3.4, we close the chapter by drawing conclusions and sketching future work.

3.2 Methodology

Our approach to track multiple faces in egocentric photo-streams consists of four main steps: *seed and tracklet generation*, *grouping tracklets into Bag-of-tracklets*, *prototype extraction*, and *occlusion treatment*.

3.2.1 Seed and tracklet

Prior to any computation, the first step of the proposed method is to organize the long and unconstrained egocentric photo-streams into homogeneous temporal segments. To this end, we apply SR-clustering [101], an unsupervised temporal segmentation method, specifically formulated for egocentric photo-streams. SR-clustering consists in a Graph-Cut algorithm that finds a trade-off between the under-segmentation produced by a concept drift detector, and the over-segmentation resulting from agglomerative clustering. The clustering is performed over both semantic visual concepts and global image features to group temporally adjacent images into semantically homogeneous segments.

Among the set of created segments from the temporal segmentation step, those that contain trackable persons are of particular interest for our purpose. To determine if a segment contains trackable persons, the ratio between the number of frames with detected faces and the number of frames of the segment is measured. If the ratio is higher than a predefined threshold (0.5 in this work), then the segment is considered as a segment containing trackable persons. Hereafter, we refer to a potentially social segment of a photo-stream as a *sequence*. As the output of this phase, a set of bounding boxes surrounding the face of each person throughout the sequence is collected. The collected bounding boxes, hereafter called *seeds* are shown in red in Fig. 3.1.

3.2 Methodology

Due to the nature of the egocentric setting, to detect the visible faces, an *in the wild* face detector [102] that substantially outperforms state-of-the-art face detectors [103] is applied on each frame of the sequence. The detector is based on a mixture of trees with a shared pool of parts, where, every facial landmark is defined as a *part* and a global mixture is used to model topological changes due to the viewpoint. Different mixtures share part templates that allow modeling a large number of views with low complexity. Moreover, as shown by the authors, tree-structured models perform effectively at capturing global elastic deformations, while being easy to optimize using dynamic programming. The global mixture can also be used to capture large deformation changes for a single viewpoint, such as changes in expression. Despite the relatively good performance of the detector, it sometimes produces some false positives or false negatives due to the blurring effect that happens frequently in egocentric photos-streams.

Figure 3.1: Detected faces (seeds) are shown by red bounding boxes in a sequence. An example of false negatives can be observed in frames 8 and 9. Only a subsample of the original sequence is shown.

For each seed, a set of correspondences to it along the sequence is generated, called a *tracklet*. The tracklet is generated by propagating the seed in the sequence forward and backward using a similarity measure to be detailed below. As a result, a tracklet $T^i = \{t^i_b, ..., t^i_s, ..., t^i_e\}$ associated to the seed i found at time s, begins in a time b, where the backward tracking ends (first frame in the sequence), and ends at time e, where the forward tracking ends (last frame in the sequence). In the rest of this chapter, we will keep the convention of using the variable t to refer to the bounding box surrounding the faces, the upper-index to identify the tracklet, and the sub-index to identify the frame. Note that theoretically, the number of generated tracklets should be of the order of the number of found seeds. For example, in the ideal case where face detector does not fail, two persons appearing in all the 100 frames of a sequence, would generate 200 tracklets, each one of length 100 frames.

To backward and forward propagation of a seed found in the frame s of a sequence, every other frame of the sequence is inspected to find the most similar

region in it. In order to deal with the abrupt displacements of the target, the whole area of the new frame is inspected employing the sliding window approach where the size of the sliding window corresponds to the size of the seed. This leads to the generation of a set of sample regions of the same size of the seed in each new frame. In addition to the sample regions generated by the sliding window, all the previously detected seeds in that frame are also considered among the set of sample regions. The intention is to take into consideration the possible face size variation due to its distance variation from the camera in a different frame.

After generation of the set of sample regions, the next step is to find the most similar region to the seed. However, to reduce the computational complexity, a less complex criteria is applied firstly to prune out irrelevant sample regions. The criteria is to reject those samples whose similarity to the seed in the HSV color space is lower than a pre-defined threshold. Later, the similarity between the seed and every remaining sample in a frame of the sequence is measured by its average deep-matching score [104]. The deep matching is conceived as a 2D-warping, that is able to deal with various kinds of object-induced or camera-induced image deformations, including scaling factors and rotations. Instead of using SIFT patches as descriptors, each SIFT patch is split into four so-called quadrants and, assuming independent motion (to some extent) of each of the four quadrants, the similarity is computed to optimize the positions of the four quadrants of the target descriptor.

Figure 3.2: An example of a tracklet generated by deep matching. The red bounding box corresponds to the seed that the tracklet is generated from it. The green box in each frame corresponds to the sample with the highest deep matching score to the seed.

For simplicity, let us consider two sequences of R-dimensional descriptors in a 1D warping case: the *reference*, that corresponds to the seed, say $P_s = \{p_{s,i}\}_{i=0}^{R-1}$, and the *target*, say $P_t = \{p_{t,i}\}_{i=0}^{R-1}$, that corresponds to a sample in a frame. The optimal warping between them is defined by the function $w^* : \{0, \ldots, R-1\} \rightarrow \{0, \ldots, R-1\}$ that maximizes the average value of similarities between their ele-

ments:

$$\Lambda(w^*) = \max_{w \in W} S(w) = \max_{w \in W} M_i\{sim(P_s(i), P_t(w(i)))\}_{i=0,\ldots,R-1} \tag{3.1}$$

where $w(i)$ returns the position of element i in P_t, M_i is the average value of the set of similarity values generate by varying i and sim is the non-negative cosine similarity between pixel gradients. The deep matching algorithm is built upon a multi-stage architecture that interleaves convolutions and max-pooling at three different scales among the feasible warpings between descriptors. The set of *feasible warpings* W is defined recursively so that finding the optimal warping w^* can be done efficiently by a dynamic programming strategy. Fig. 3.2 illustrates an example of a generated tracklet based on deep matching for one of the seeds in the sequence shown in Fig. 3.1. The seed is depicted by a red bounding box and the green bounding boxes correspond to the samples with the highest deep matching score to the seed in every frame. As can be seen, the tracklet corresponds to the same person who generated the seed.

3.2.2 Extended-Bag-of-Tracklets

The tracklets generated by the seeds belonging to the same person in a sequence are likely to be similar to each other; we aim to group them into one eBoTs, where there is no intersection between eBoTs by definition. Let us consider an eBoT, say $\mathbb{T}$, as a set containing a tracklet, $\mathbb{T} = \{T^i\}$, where T^i does not belong to any other eBoT. Also, let us consider another tracklet T^j that has not been assigned to any eBoT yet. Let t_k^i and t_k^j be the bounding boxes, where the person is detected (by the face detector or by the tracker) at frame k for tracklets T^i and T^j, respectively.

We define the similarity between two tracklets T^i and T^j as the average of the area of the intersection between t_k^i and t_k^j divided by the area of their union:

$$\mathcal{S}(T^j, T^i) = \frac{1}{|T^i|} \sum_{k=1}^{|T^i|} \frac{|t_k^j \bigcap t_k^i|}{|t_k^j \bigcup t_k^i|}.$$

Given a tracklet T^j, it will be added to the eBoT $\mathbb{T}$, if the similarity between T^j and all tracklets in $\mathbb{T}$ is high enough. In this work, we experimentally found that the threshold 0.2 to include a tracklet in an eBoT leads to the best results. Before adding tracklets to an eBoT, we sort them based on their similarity to the first tracklet in the eBoT. Since the next tracklets need to be compared to the existing tracklets in an eBoT, sorting tracklets prior to other computations, helps

to avoid the aggregation of biased tracklets in the eBoT.

The similarity of tracklet, T^j to the eBoT, $\mathbb{T}$ is defined as the average of the similarities to all its tracklets:

$$\widetilde{\mathcal{S}}(T^j, \mathbb{T}) = \frac{1}{|\mathbb{T}|} \sum_{T^i \in \mathbb{T}, T^i \neq T^j} \mathcal{S}(T^j, T^i) \tag{3.2}$$

where $|\mathbb{T}|$ is the number of tracklets in the eBoT. After grouping by similarity, all tracklets in an eBoT are very likely to correspond to the same person.

However, not all the tracklets in an eBoT are equally reliable. In addition, some eBoTs may correspond to seeds that are false positive detections. While the first issue is related to the prototype extraction and will be addressed in the next subsection, here we detail how to remove unreliable eBoTs that do not correspond to any person in the video. To this end, we define the *density* of an eBoT as $d(\mathbb{T}) = \frac{|\mathbb{T}|}{|T|}$, where $|\mathbb{T}|$ is the number of its tracklets and $|T|$ is the length of the sequence.

Ideally, the density value is equal to 1 and corresponds to a situation where there are as many tracklets in the eBoT, as the number of frames the person persisted in the sequence and the person appears in every frame of the sequence. In practice, since the face detection algorithm, as well as the matching algorithm, holds some errors, the eBoT is looking for the consensus between the different tracklets to obtain the right tracking outcome. As expected, reliable eBoTs show different behavior from unreliable ones, the latter having low density. Based on this observation, those eBoTs that have a density lower than a predefined threshold are detected as unreliable eBoTs and are discarded. In this work, we empirically found that a threshold of 0.2 gives good results. By excluding unreliable eBoTs, we obtain a number of eBoTs as the number of persons in that sequence (see Fig. 3.3).

3.2.3 Prototype extraction

A prototype extracted from an eBoT $\mathbb{T}$, should represent all the tracklets in the eBoT. Note that the detection of the target in a given frame of the sequence varies depending on the seed that generated the tracklet. Thus, a prototype should return the most common location of the face among the tracklets of an eBoT in every frame. To this end, the bounding box of a prototype in a frame is chosen as the one which has the biggest intersection with the rest of the bounding boxes of the

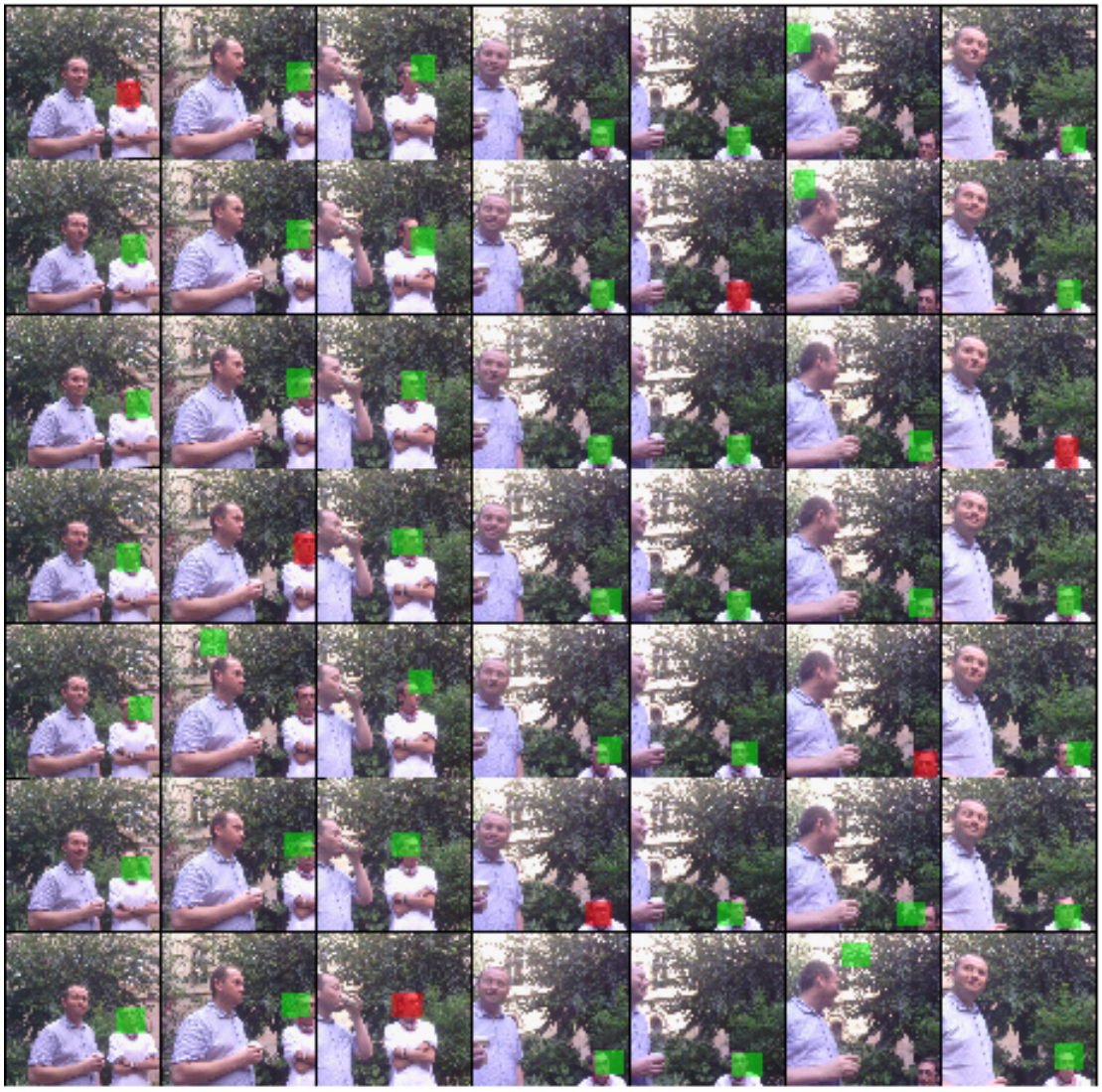


Figure 3.3: Example of a reliable eBoT -after excluding unreliable eBoT- extracted from the sequence in Fig. 3.1. Each row shows a tracklet in the eBoT which in total consists of 7 tracklets. The red bounding box in each row indicates the seed of that tracklet and green bounding boxes are the samples with the highest average deep matching score to their corresponding seed. As can be appreciated, all tracklets in the eBoT correspond to the same person.

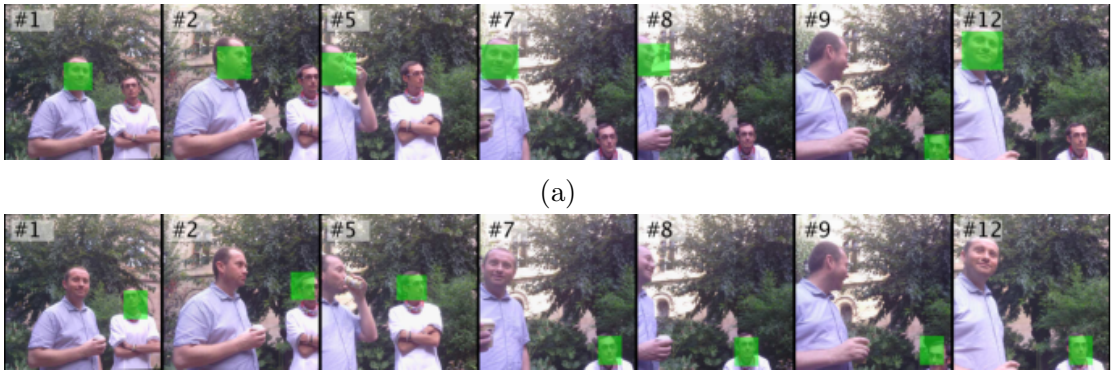


Figure 3.4: Two Prototypes extracted for the two persons in the sequence.

other tracklets in the same eBoT in that frame, namely:

$$\hat{T} = \{\hat{t_b}, \ldots, \hat{t_k}, \ldots, \hat{t_e}\}, \text{ so that } \hat{t_k} = \underset{i=1,\ldots|\mathbb{T}|}{\arg\max} \sum_{j=1,\ldots,|\mathbb{T}|, j\neq i} t_k^i \bigcap t_k^j,$$

where $|\mathbb{T}|$ is the number of tracklets in the eBoT, $(t_k^i, t_k^j)_{i\neq j}$ are the bounding boxes of detected faces in the k-th frame of tracklets T^i and T^j from the eBoT $\mathbb{T}$, respectively.

Fig. 3.4 shows two prototypes, each of them extracted from a separate eBoT where only one of them is shown in Fig. 3.3. Note that the prototypes correctly tracked the persons although the face detector missed the person in several frames. Missed detections can be seen in Fig. 3.1.

3.2.4 Occlusion treatment

Besides optimizing the localization of the target, a good prototype should also indicate occlusions or unreliable detections. In order to increase the accuracy of the proposed method, in the final prototype, those frames where the target is fully or partially occluded or there is an unreliable detection are detected and removed. To this goal, a function $\Lambda(t_s^i, t_k^i)$ is defined that associates to each bounding box t_k^i of a tracklet T^i the value of the deep matching score to its seed t_s^i. The *frame confidence* is defined as the average of the normalized deep matching scores of its bounding boxes of all the tracklets of the same eBoT, in that frame, that is:

$$\mathcal{C}_k = \frac{1}{|\mathbb{T}|} \sum_{i=1}^{|\mathbb{T}|} \Lambda(t_s^i, t_k^i), \tag{3.3}$$

In Eq. 3.3, $\mathcal{C}_k$ is the frame confidence, $|\mathbb{T}|$ is the number of tracklets in the eBoT, t_s^i is the seed of the i-th tracklet of the eBoT and t_k^i is the bounding box of frame k of the i-th eBoT tracklet. The deep matching scores between bounding boxes in the eBoT have been normalized between zero and one.

When there is a severe or partial occlusion of the face, or the target is missing, the confidence of the eBoT on that frame $\mathcal{C}_k$ experiences a drop. This phenomenon can be observed in Fig. 3.6, where, due to partial occlusion of faces in frames 5 and 6 in Fig. 3.6 (a) and frames 6 in Fig. 3.6 (b), the confidence value in these frames has a minimum and lies under the pre-defined threshold for occlusion estimation. In all the cases of occlusions that are shown in Fig. 3.6 (a) and (b), the face of the person is only partially occluded. This fact shows the robustness of the method in estimating large changes in facial appearance.

The value of the threshold for estimating occlusions, say L, is calculated over a subset of 15 sequences that constitute the training dataset. Fig. 3.5 shows the normalized confidence value calculated using Eq. 3.3, for frames where the target is occluded (left) and for frames where the target is not occluded (right). For non-occluded frames we used the ground-truth tracklet to compute the confidence values, whereas for occluded frames we generate a fake-tracklet by randomly defining a bounding box where there is not a face. As a tracklet is generated for each seed, in Fig. 3.5 we plot on the left the median value and the mean value of deep matching score over all the generated fake-tracklets and on the right the median value and the mean value of deep matching score over all the ground-truth tracklets over a sequence. The threshold L (black line), emerges from the median of all the median confidence values over occluded frames. We obtained this value as $L = 0.12$.

After estimating occlusions, we refine the frame confidence presented in Eq. 3.3, considering it zero for occluded frames, that is:

$$\mathcal{C}_k = \begin{cases} \frac{1}{|\mathbb{T}|} \sum_{i=1}^{|\mathbb{T}|} \Lambda(t_s^i, t_k^i), & \text{if } \frac{1}{|\mathbb{T}|} \sum_{i=1}^{|\mathbb{T}|} \Lambda(t_s^i, t_k^i) \geq L \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

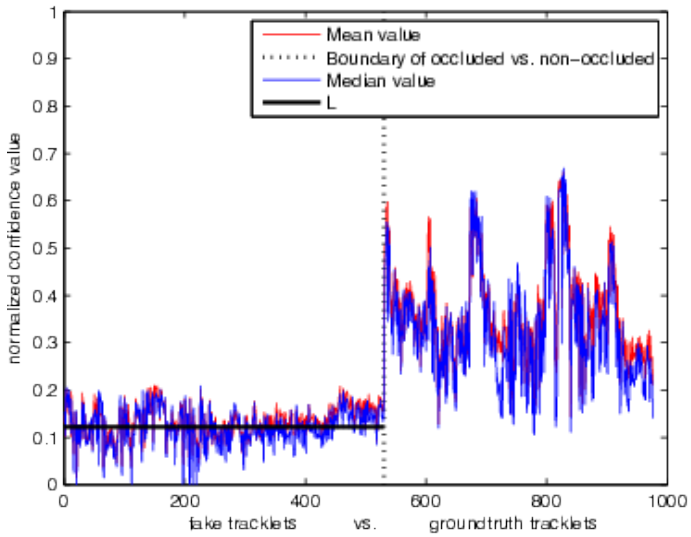


Figure 3.5: Normalized confidence value for fake tracklets generated from an occluded target (left) and for ground-truth tracklets (right). The threshold L that is used to estimate occlusions, is depicted in black.

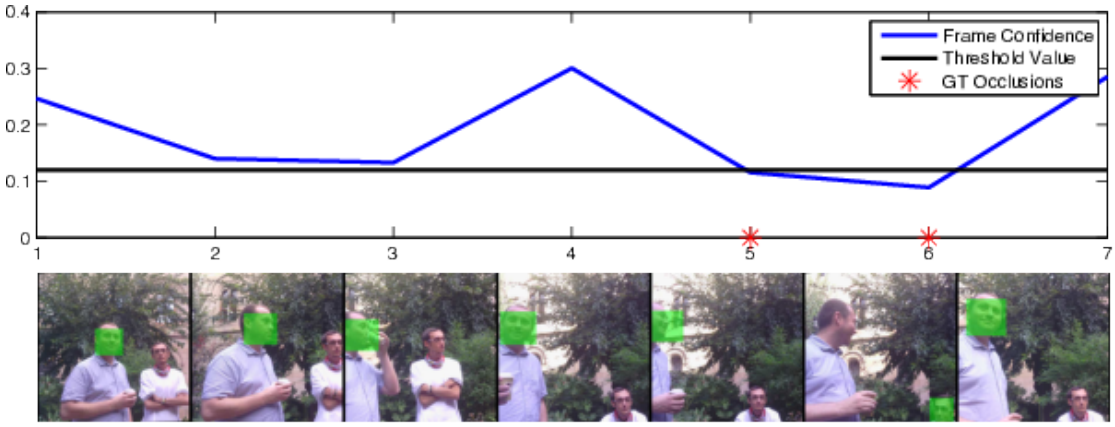
3.2.5 Confidence of prototypes

A prototype can be very useful as a basis for applications, such as finding the type of a social interaction and social roles. Thus, confidence estimation of an extracted prototype is a valuable task. We define the *prototype confidence* as the mean confidence over all its frames that do not undergo occlusion weighted by a term that penalizes occlusions, that is:

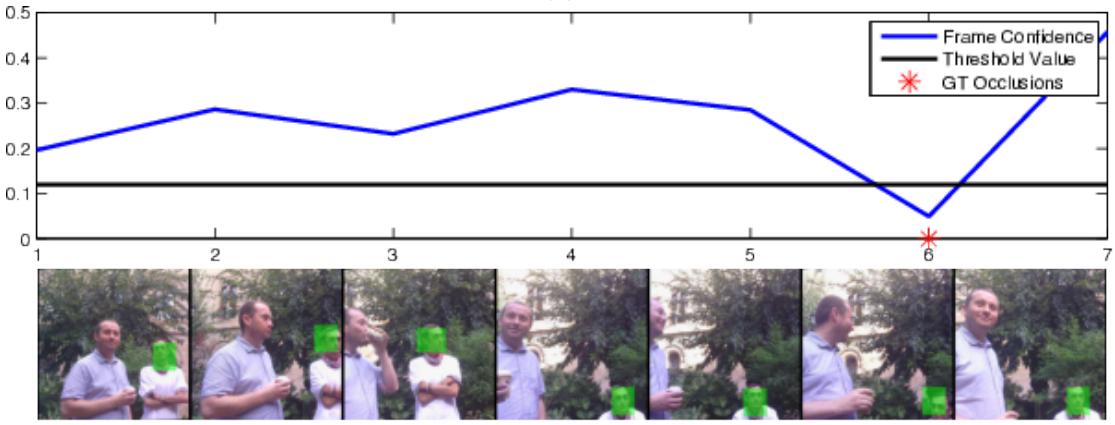
$$\mathcal{C}(\hat{T}) = \frac{1}{|\hat{T}|} \sum_{k=1,\dots,|\hat{T}|} \mathcal{C}(\hat{t_k}) \times max((1 + \beta \log((|\hat{T}| - z)/|\hat{T}|)), 0) \tag{3.5}$$

where $|\hat{T}|$ is the length of the prototype, z is the number of frames, where the face is occluded or missing, and β is a control parameter that depends on the performance of the detector (we found that $\beta = 1$ gives reasonable results). Note that, in the absence of occlusion, the confidence from Eq. 3.4 and Eq. 3.5 are the same.

Eq. 3.5 is inspired from the definition of tracklet confidence given by Bae and Yoon in *Multi-Object Tracking based on Tracklet Confidence* [105]. The first term is related to the coherence in the appearance of the target along the tracklet: a more coherent appearance in a tracklet increases the confidence of the tracklet.



(a)



(b)

Figure 3.6: Frame confidence of two prototypes shown in Fig. 3.4, as defined in Eq. 3.3. The occurrence of occlusion for every person in the sequence in the ground-truth is shown by red stars in the plot. The black line corresponds to L, the threshold determined to estimate occlusions. As can be seen, the occurrence of the face occlusion indicated in the ground-truth, highly coincides with the calculated confidence drop of the face in that frame.

The second term is related to the continuity of the tracklet: it decreases for occluded tracklets. Therefore, the final prototype should have a larger confidence than all the tracklets in an eBoT. After estimating occlusions for the prototypes, we associate a confidence value to each tracklet of the eBoT by using Eq. 3.5, and verify that the confidence of the prototype is higher than the highest tracklet confidence in the eBoT. After evaluation, the average confidence value of all prototypes in our test set has a value of 0.54, which is higher than the average of the confidence value of all the tracklets in all eBoTs, being 0.32.

3.3 Validation

3.3.1 Dataset

Currently, there is no dataset for person tracking with ground-truth information in egocentric photo-streams. Hence, to measure the performance of the proposed model, we created a dataset acquired by the Narrative Clip camera. We manually annotated the sequences that contain trackable people and localized the position of their faces. The dataset has been acquired by five users of different ages. Each user wore the camera for a number of non-consecutive days over an 80 days period, collecting ∼20.000 images. The images have been acquired in diverse environments with different illumination conditions while the user was performing varying activities. Our dataset contains a total number of 108 different trackable persons along 80 sequences of the average length of 25 frames. Table 3.1 provides further details of the proposed dataset.

Table 3.1: Detailed breakdown of our dataset made of ∼20.000 images captured by 5 users

User	**Days**	**Total frames**	**Total frames with person(s)**	**Total frames with occlusion**	**Average daily duration**
1	30	6478	680	53	8h
2	5	1228	125	17	8h
3	10	3428	220	27	8h
4	28	6894	850	96	8h
5	7	2178	425	22	6h

3.3.2 Experimental setup

After partitioning a photo-stream captured by the Narrative Clip into segments, a face detector is applied to exclude non-trackable segments and generate possible seeds for trackable segments, called sequences. Then, a tracklet is generated for each seed in a sequence. Finally, the tracklets are grouped into eBoTs and a final prototype with estimated occlusion is extracted from each reliable eBoT. These prototypes constitute the final output of our method. In the next section, quantitative and qualitative comparison between our approach and other tracking approaches is provided.

We measured the performance of our method using CLEAR MOT [106] on the resulting prototypes (with and without occlusion estimation). Additionally, we compared its performance with six other state-of-the-art methods. CLEAR MOT consists of multiple metrics as follows. The Multiple Object Tracking Precision (MOTP) evaluates the intersection area over the union area of the bounding boxes:

$$MOTP = \frac{1}{|M_s|} \sum_{k \in M_s} \frac{|t_k \bigcap gt_k|}{|t_k \bigcup gt_k|},$$

where M_s is the set of frames in a sequence in which the tracked bounding box t_k intersects the ground-truth bounding box gt_k, and $|M_s|$ is the cardinality of M_s. MOTP quantifies the accuracy of the tracker by estimating the precise location of the object, regardless of its ability in keeping consistent trajectories.

On the other side, the Multiple Object Tracking Accuracy (MOTA) estimates the accuracy of the results by penalizing False Negatives (FN), False Positives (FP) and IDentity Switching (IDS), namely:

$$MOTA = 1 - \frac{\sum_{k=1}^{l}(FN_k + FP_k + IDS_k)}{\sum_{k=1}^{l} GT_k},$$

where k refers to the frame number, l is the length of the sequence, and GT_k states for the number of faces in the ground-truth to be tracked at frame k. FN_k and FP_k donate the number of false negatives and false positives in a frame k, respectively. IDS_k is equal to 1 when the detection does not overlap with its corresponding ground-truth face target, but with another face.

Both metrics intuitively express the overall strength of each tracker and are suitable for general performance evaluations. Furthermore, the qualitative comparative results are also shown over four different sequences in the next section.

3.3.3 Discussion

Quantitative evaluation: To the best of our knowledge, the only work which was exclusively introduced for person tracking in egocentric photo-streams is BoT [39]. Most of the available tracking techniques are not directly applicable to egocentric photo-streams, since they follow assumptions such as temporal consistency between frames or smooth variation in target and background appearance, that do not hold for egocentric photo-streams. Still, we compared our approach to six different state-of-the-art algorithms that are applicable to egocentric photo-streams since they do not rely on motion information nor background modeling.

The selected trackers are designed for tracking one object at a time, but in our dataset, more than one person appears in the sequence. Thus, we applied the trackers separately for each person to adapt them to our scenario. In this case, the tracking problem reduces to one object tracking and therefore for evaluation measurements we do not consider the IDS metric for these methods as proposed by Smeulders et al. in [44]. In Table 3.2, we show the percentage of MOTP, MOTA, FP, FN and IDS on the results of AMT [54], BoT [39], CT [48], LOT [50], L1O [51], and SPT [49]. We also show how the estimation of occlusions improves the performance of the proposed method in most of the metrics.

Table 3.2: Performance comparison

Methods	**MOTP↑**	**MOTA↑**	**FP↓**	**FN↓**	**IDS↓**
AMT (Abrupt Motion Tracking)	60.99%	59.65%	16.70%	23.65%	-
BoT (Bag of Tracklets)	48.39%	43.44%	22.9%	20.17%	14.30%
CT (Compressive Tracking)	35.05%	15.32%	33.07%	51.61%	-
LOT (Locally Orderless Tracking)	42.27%	15.57%	33.12%	51.13%	-
L1O (L1 Tracker with Occlusion Detection)	37.25%	25.87%	31.81%	42.32%	-
SPT (SuperPixel Tracking)	40.75%	39.31%	23.56%	37.13%	-
eBoT (prototype, occlusions not excluded)	68.32%	72.08%	15.19%	**10.60%**	**2.13%**
eBoT (prototype, occlusions excluded)	**70.27%**	**80.23%**	**5.12%**	12.51%	**2.13%**

As can be observed, the difference among CT, LOT, L1O, and SPT in terms of precision (MOTP) is small, where CT has the smallest value. This can happen since this tracker does not change the scale of the bounding box, while other methods have a relatively good mechanism for scale adaptation. BoT and AMT have higher precision than other methods, with AMT outperforming BoT. This can be justified because in AMT the true object is introduced for the tracker in the initial frame of the sequence, whereas BoT is fully automatic.

In terms of accuracy (MOTA), CT and LOT perform much the same as each

other. This might be a consequence of regular appearance model updates for both trackers which leads to object loss when they encounter a large variation between frames. However, L1O and SPT perform slightly better, since they are able to estimate occlusions, leading to lower amount of FP. SPT and LOT use superpixel representation, which is more suited for bigger objects. Thus, they perform better when the face is closer to the camera and looks bigger. On the other hand, AMT is designed for tracking on low frame-rate videos and performs quite well on our dataset, being able to outperform BoT. However, it can easily miscalculate the position of the target, when there is more than one face in the frame. The miscalculation may happen due to use of a color-based likelihood model that can easily get confused by finding a region with similar colors to the target.

As it can be seen in the lower part of Table 3.2, the proposed method in this paper outperforms the state-of-the-art. The seventh and the eighth lines in the Table 3.2 show evaluation metrics obtained before and after estimating occlusions. The estimation of occlusions reduces FP, while slightly increases the FN rate due to wrongly eliminating some true detections in the final prototypes. The proposed method for prototype extraction drastically reduces FP, FN, and IDS, since it optimizes the localization of the detection.

From this evaluation, we can state that the proposed model is able to robustly track multiple peple faces under challenging conditions. Moreover, this improvement is achieved without relying on any strong assumptions and without the need of a cumbersome training stage.

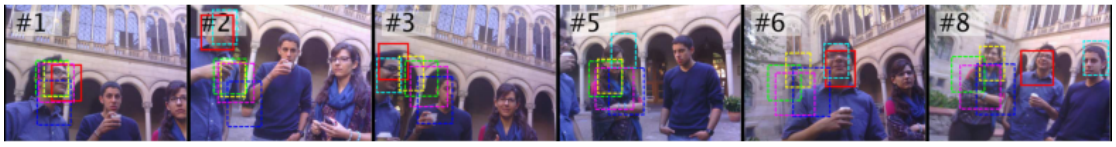
Qualitative evaluation: The tracking results of the proposed approach together with the results of the previously introduced trackers are shown over four different sequences in Fig. 3.7, Fig. 3.9, Fig. 3.10, and Fig. 3.8. Every sequence contains multiple persons and tracking result of each tracker is shown by a specific color in every frame of the sequences. The result of the proposed approach is shown by a red bounding box around the face of the person. In the frame, where our method detects an occlusion, no bounding box is shown. For the sake of visualization, if a sequence contains more than one person, the tracking result for each person is shown in a separate line. Fig. 3.7 shows the final prototypes with estimated occlusions of the prototypes shown in Fig. 3.4. Fig. 3.9 and Fig. 3.10 show the result for a sequence of two different persons and Fig. 3.8 shows them for a sequence of three different persons.

Among the state-of-the-art methods, AMT has the best performance on our dataset, because it was designed to cope with abrupt motion changes. However, it can easily produce FP in the presence of multiple persons for not being a multi-

(a)

(b)

Figure 3.7: Results of applying different methods on an egocentric photo-stream. Different bounding boxes show the tracking results of the CT, LOT, AMT, SPT, L1O and our proposed approach. Occlusions can be observed in frame #9 of 3.7a and frames #4 and #9 of 3.7b.

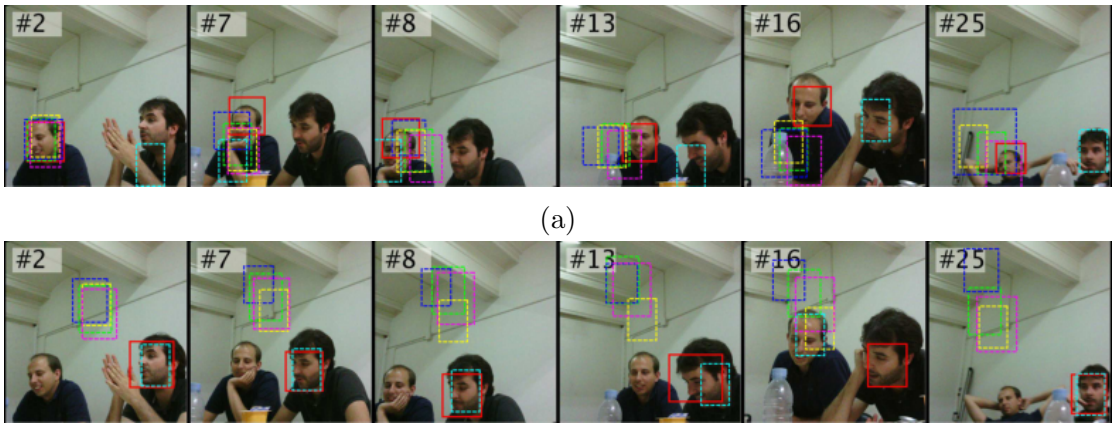


(a)

(b)

(c)

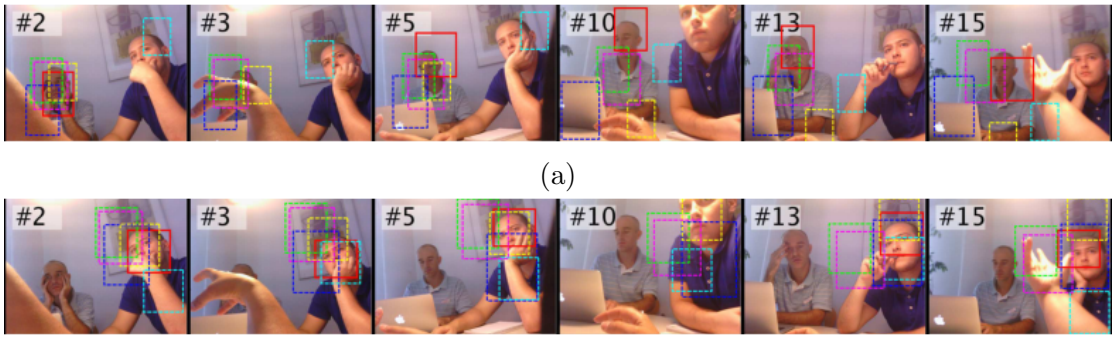
Figure 3.8: Results of applying different methods on an egocentric photo-stream. Different bounding boxes show the tracking results of the CT, LOT, AMT, SPT, L1O and our proposed approach. Occlusions can be observed in frame #5 of 3.8a and frame #6 of 3.8c.



(a)

(b)

Figure 3.9: Results of applying different methods on an egocentric photo-stream. Different bounding boxes show the tracking results of the CT, LOT, AMT, SPT, L1O and our proposed approach.



(a)

(b)

Figure 3.10: Results of applying different methods on an egocentric photo-stream. Different bounding boxes show the tracking results of the CT, LOT, AMT, SPT, L1O and our proposed approach. Occlusions can be observed in frame #3 of 3.10a and frame #10 of 3.10b.

tracking method. As can be observed, CT, LOT, L1O, and SPT are unable to find the target when its location varies largely. In addition, a common drawback among the AMT, BoT, CT, and LOT is that they are unable to localize target occlusion. As expected, it can be seen that the tracking results of the proposed approach closely match the person's face. However, the method assigns a wrong region to the track, when a person's face is occluded, causing the occurrence of FP or IDS. Still, our method is able to precisely estimate occlusions or wrongly assigned detection.

Another observation taken from these experiments is that our proposed method works better when the people are closer to the camera. As the distance of the people from the camera increases, the resolution of the image on their face region decays. This phenomenon leads to the generation of fewer seeds by the face detector and to unreliable matches by the deep matching approach. The illumination condition is another important factor as well. eBoTs is quite robust to illumination changes, although it performs better when the images are not too dark.

3.3.4 Complexity analysis

Regarding the complexity of our proposed algorithm, one can easily see that the most expensive part is the construction of the tracklets, where the deep matching is applied with a sliding window approach to all the windows with a similar color to the seed in the HSV color space. The most expensive part of the deep matching algorithm lies in the computation of the first level convolutions. However, the computational burden would be mitigated by using a GPU or a faster matching algorithm [107], that achieves comparable performances. Finding the optimal matching score among all feasible non-rigid warpings for all square patches at different scales, from the first image at all locations in the second image can be done with complexity $O(PP')$, where P and P' are the numbers of pixels of both images. Usually, the size of the seed image is between 5000 and 6000 pixels and the number of samples to be considered is about 2000. On a CPU Intel i5 - 2.53 GHz, with operating system Windows 7 - 64 bit, 4G of RAM, it takes on average about 1 minute per each pair of images to find the similar candidate to the seed. It is easy to see that the complexity of the rest of algorithms to construct the eBoT and extract the prototype is $O(M * N^2)$, where M is the number of faces appearing in the sequence and N is the length of the sequence, taking less than a minute in the aforementioned computer.

3.4 Summary

In this chapter, we introduced our novel method to track multiple-faces in egocentric photo-streams that substantially outperforms state-of-the-art. In the following chapters, the importance of multiple-face tracking in the analysis of social events in egocentric vision will be thoroughly explained. Extended-bag-of-tracklets approach, to deal with various types of object-induced or camera-induced image deformations, tracklets are computed by using the average deep-matching score between the seed and each sample in different frames. Moreover, in order to extract the final prototype, eBoT introduces a useful measure of confidence to estimate and discard occlusions and missed detections. A quantitative comparison between our model and other six state-of-the-art methods over a dataset of 20,000 images showed the advantage of the proposed model under drastic changes of poses, scales and object appearances.

In this chapter, we presented the *extended-Bag-of-Tracklets* (eBoT) approach by introducing several features that help in increasing tracking robustness even in photo-streams acquired by cameras with lower frame-rates (2 fpm) and narrower fields of view:

The advantages of eBoT approach can be summarized as follows:

- To manage the close appearance of people to the camera, eBoT reliably detects people characterizing them by their face instead of their body.
- To improve the control over target deformations and scale variations, eBoT employs a new approach for finding correspondences based on an average deep matching score instead of the sparse representation of features.
- eBoT presents a more robust way to compute the prototype of the bag-of-tracklets by trying to extract the most reliable bounding box frame-wise, instead of tracklet-wise.
- eBoT is tolerant of face occlusions and is able to explicitly localize them which leads to the higher precision of results.
- eBoT introduces a confidence term to measure the reliability of the prototypes which facilitates further analysis of final tracklets.

CHAPTER 4

Social Interaction Detection in Egocentric Photo-Streams

4.1 Introduction

Distinct past efforts have been carried out in computer vision to solve the social interaction detection problem in both conventional and egocentric vision employing different visual features as discussed in chapter 2. In this chapter, our focus slightly different from the aforementioned methods is to tackle the problem of social interaction detection in egocentric photo-streams. We are interested in the automatic detection of the social events occurred in the real-world setting, such as coffee breaks, casual work meetings or a sudden encounter in a park. The complex visual appearance of natural scenes makes the task especially challenging.

In this chapter, we propose two approaches to solve this problem: frame-level and event-level social interaction detection. In the frame-level approach, social interaction status of each individual in the scene with the camera-wearer is established in every frame of the photo-stream. The presence of a social interaction is decided in every single frame separately and eventually, if the number of found interactions with regards to the full length of the sequence is higher than a predefined threshold, then that specific person is considered as socially interactive. In the event-level analysis, we make use of the temporal evolution of social signals along a potential social event. We aim to describe how the evolution of the social

(a)

(b)

Figure 4.1: Examples of two sub-sampled sequences in EgoSocialStyle test set. In 4.1a the user is involved in a social interaction while 4.1b demonstrates a sequence where although the user is among the crowd, he is not specifically interacting.

signals inferred from human behavior, can be employed to decide if the appearing people in a sequence are interacting with the camera-wearer or not.

This chapter is organized as follows: Sec. 4.2 is devoted to provide details about both the frame-level (4.2.2) and event-level (4.2.3) models for social interaction analysis in egocentric photo-streams. In Sec. 4.3 we present the employed methodology for validation of both models and discuss the obtained results. In the same section, we also compare both frame-level and event-level models for social interaction detection from a different perspective. Finally, main characteristics of both models and their performance on the proposed dataset are summarized in Sec. 4.4.

4.2 Methodology

We, as humans are naturally able to recognize if two or more people are interacting by simply looking at a sequence of images (see Fig. 4.1). However, this is not as trivial for a computer program. Let us define a *sequence* as a potential social segment of a photo-stream extracted by applying the video segmentation tool of Dimiccoli et al. [108]. Given a sequence, social signals are first extracted at every frame and later are analyzed either in frame-level or their evolution is analyzed over time at sequence-level to detect social interactions.

4.2.1 Social signal extraction

Tracking the appearance of people is generally considered as the first step prior to any social behavior analysis in machine vision. In this work for tracking, we employed the extended-Bag-of-Tracklets (eBoT) [2] as a multi-person tracking algorithm in the egocentric photo-stream setting. The set of bounding boxes corresponding to the same face in a sequence, resulting from eBoT, is called a *prototype*, where the number of prototypes in a sequence is equal to the number of tracked people in it as more than one individual may appear in a single sequence.

Both approaches presented in this chapter rely on the F-formation formalization for social interaction detection. As the F-formation model assumes a bird's-eye view of the scene, we represent each bounding box in a prototype by a (x, d, o) triplet, so that x denotes the position of the person in the horizontal axis of the image and with regards to the user, d denotes its distance, and o its head orientation. The tracking process directly provides us with the x position of a face. Both parameters, d and o should be calculated for all the participants in the social interaction, being the user and the visible people in a sequence. However, in our egocentric setting, x is not a reliable feature to be considered as it constantly undergoes large variations due to the unpredictable movements of the camera and its low frame-rate (see Fig. 4.1a). Moreover, when it comes to interaction with the user, the x position of the visible people as far as they do not occlude each other, does not play a crucial role. Therefore, we only consider the (d, o) pair to analyze the F-formation.

Distance: In the egocentric setting, the user is obviously located at no distance from the camera, O. The distance $d(O, p_j)$ of the j-th tracked person (p_j) from the camera is estimated based on the camera-pinhole model by learning its relation with the vertical face height of the person [109]. According to our observations, the relation between the face height of individuals and their distance from the camera is best modeled as a second degree polynomial of the face height of the person [40].

For training the polynomial regression function, we used the height of the face of 3 different individuals measured in all the following set of distances:

$$\{30, 50, 70, 100, 150, 200, 250\}\, cm.$$

The distance feature is represented by:

$$\varphi_d(p_j) = d(O, p_j) \in \mathcal{R}.$$

Without loss of generality, in the feature vector, we will omit the reference to the person p_j and the wearable camera O.

Generally, the distance among the interacting people strictly depends on the type of the interaction and the number of people involved in the interaction as defined by Hall [110] and has the range [0,120] for casual relationships and no relationship can occur in a distance further than 350 cm. In our setting, we empirically discovered that the distance $d \leq 150cm$ is where social interaction normally takes place.

Orientation: The head orientation of each individual gives a rough estimation of where the person is looking at. *Yaw* head orientation is the most commonly studied social signal in social interaction detection. However, in this work, in addition to the *yaw* (ω_z) head orientation, *pitch* (ω_y), and *roll* (ω_x) head orientations are also studied. Hence, the orientation feature is given by:

$$\varphi_o(p_j) = (\omega_x(p_j), \omega_y(p_j), \omega_z(p_j)) \in \mathcal{R}^3,$$

where each of ω_x, ω_y, and ω_z has a value between [-90°,90°]. As the camera is basically worn on the chest of the user, we assume the user possibly looks at anywhere in the space, but with a higher probability of looking at other engaged people in the interaction.

Facial expression: During a social interaction, people exhibit a large number of non-verbal communication cues including facial expressions. Facial expressions as stated by Hess et al. [111], are often referred to as automatic demonstrations of affective internal states used as communicative means in interaction with others. The overlooked importance of facial expressions for social interaction detection is mostly noticed within the scenes recorded in crowded places where people often stand in close proximity to strangers with whom they do not necessarily interact (see Fig. 4.1b). In this situation, relying solely on distance and orientation of the individuals for social interaction detection may lead to disputable predictions (see Fig. 4.2). Our observation on real social situations led us to intuitively explore the role of facial expression in social interaction detection as an additional feature besides the pure geometrical features imposed by the F-formation.

In this work, facial expressions and face orientation are extracted by making use of Microsoft Cognitive Service[1]. Facial expression is presented as a predicted vector of probabilities for each of 8 different facial expressions consistently associated to emotions in the occidental culture, being *neutral*, *happiness*, *surprise*, *sadness*,

[1]https://azure.microsoft.com/en-us/services/cognitive-services/face/

(a) (b)

Figure 4.2: A same person is shown in two different social events where facial expression probabilities of the person are also presented. When the person is not interacting with the user (4.2b), her dominant facial expression is *Neutral*, while when interacting (4.2a) her dominant facial expression varies to *Happiness*.

anger, *disgust*, *fear*, and *contempt* [112]. For a given person p_j, we proposed to consider the index of the dominant facial expression that is a discrete value between 1 (*neutral*) and 8 (*contempt*):

$$\varphi_e(p_j) = \arg \max_{k \in 1,\dots,8} e_k(p_j).$$

4.2.2 Frame-level analysis

In the frame-level analysis of social signals, our focus is on finding the *o-space* and the individuals who are forming it within the F-formation formalization in every frame of a sequence. To this end, we adopt the Hough-Voting F-Formation (HVFF) strategy for social interaction discovery in videos recorded by surveillance cameras introduced by Cristani et al. [37] to the egocentric photo-stream scenario. HVFF aims at detecting the o-space, taking as input the position of the people and their head orientations. This algorithm is based on a Hough-Voting strategy, where weighted local features vote for a location in the image plane and the generalized Hough procedure does not require the local features to be in a fixed number. This approach provides the estimation of the circular o-spaces and the index of the people that form them, thus, enables social interaction detection solely from visual cues.

4.2.2.1 Hough-Voting F-formation

In the surveillance setting, to specify the o-space in the image plane, the (yaw) head orientation ω_z and the position of the people in the scene (x, y) is required. Let us assume a special case where only two individuals, p_1 and p_2 are interacting together. These two people are located at positions (x_1, y_1) and (x_2, y_2) with head orientations ω_{z_1} and ω_{z_2}, respectively. Let us assume that the two individuals are exactly facing each other and are located at a distance where the interaction can happen ($d \leq 150cm$). Given these (hard) constraints, each j-th subject votes for a candidate center $C(j)$ of the o-space, with coordinates $x_{C(j)}, y_{C(j)}$

$$C(j) = \left[x_{C(j)}, y_{C(j)}\right] = \left[x_j + r \cdot cos(\omega_{z_j}), y_j + r \cdot sin(\omega_{z_j})\right], j = 1, \ldots, J \quad (4.1)$$

where the radius $r = d/2 = 0.75$. However, the condition where people face each other at a position that enables them to vote in exactly one point is rather rare. In order to deal with this problem, some uncertainty is injected in the voting procedure of the proposed HVFF. To this end, primarily, uncertainty in the positions and the head orientation of the different subjects is modeled by random Gaussian variables, i.e.,

$$\left[x_j, y_j, \omega_{z_j}\right]^T \sim \mathcal{N}(\mu_j, \Sigma_j)$$

where $\mu_j = \left[x_j, y_j, \omega_{z_j}\right]^T$ and $\Sigma_j = diag(\sigma_x^2, \sigma_y^2, \sigma_{\omega_z}^2)$. This uncertainty is transferred in the voting approach by drawing S-1 samples (being μ_j the S-th sample).

Each generated sample of the j-th subject has associated a weight, which is the likelihood of being extracted from its generating distribution. Each sample votes for a candidate position in the same way of Eq. 4.1. The votes given by different samples from different individuals accumulate in an intensity accumulation space with an index which associates them to the individuals who generated them. In this way, the o-space can be localized by finding the maximum values in the intensity accumulation matrix, and the associated subjects to it can be identified by inferring the index of individuals who vote for them.

4.2.2.2 Hough-Voting F-formation in egocentric photo-streams

To adapt the Hough-Voting model for social interaction detection to the egocentric setting, namely ego-HVFF, a set of adaptation is required. Although the x position of individuals is used to depict them in the scene, the y position is omitted and

instead the distance d of the individuals is considered. Hence, the parameters σ_x^2 and σ_d^2 are used to project the position of the people in the range of $3\sigma_{x(d)}$. In other words, these values allow being flexible about the classes of relations taken into account by the distance parameter. σ is set to $\sigma_x^2 = \sigma_d^2 = 400cm$. For the appearing people in the scene, the value of $\sigma_{\omega_z}^2$ depends on the quantization of the head orientation. As we employed 7 head orientations, we kept $\sigma_{\omega_z}^2 = 0.005$. The head orientation of the camera-wearer is set to 0, with $\sigma_{\omega_z}^2 = 0.1$ to cover approximately 180 degrees in front of him as his head orientation is not extractable from the chest-worn wearable camera. The parameter S is empirically chosen as $S = 800$ for the visible people and as double for the camera-wearer. These parameters are chosen empirically over the training set of EgoSocialStyle.

4.2.3 Event-level analysis

Despite most existing methods which make little use of the evolution of the features over time, in this work we employed Long Short-Term Memory Recurrent Neural Network (LSTM) which is adapted for learning over sequential data. The proposed method aims to describe how the evolution of the social clues characterizing the F-formation theory which inferred from human behavior, can be employed to decide if the appearing people in a sequence are interacting with the camera-wearer or not. To the best of our knowledge, this work is first to detect social interactions with the camera-wearer at sequence level instead of frame-level information in the domain of egocentric photo-streams.

4.2.3.1 Temporal representation of social signals

In this setting, the problem of social interaction detection is formulated as a binary time-series classification, where the time-series dimension corresponds to the number of selected social signals for the analysis as explained in Sec. 4.2.1. The complete setting is a $5-$dimensional time-series representing the time-evolution of the k-th interaction feature for each prototype. The task is to classify each time-series as interacting with the user or not. All the aforementioned interaction features are extracted in every frame of the sequence at time step τ to build the time-series representation of a prototype:

$$\varphi_{detection}(\tau, p_j) = (\varphi_d(\tau, p_j), \varphi_o(\tau, p_j), \varphi_e(\tau, p_j)) \in \mathcal{R}^5,\ \tau = 1, 2, \ldots.$$

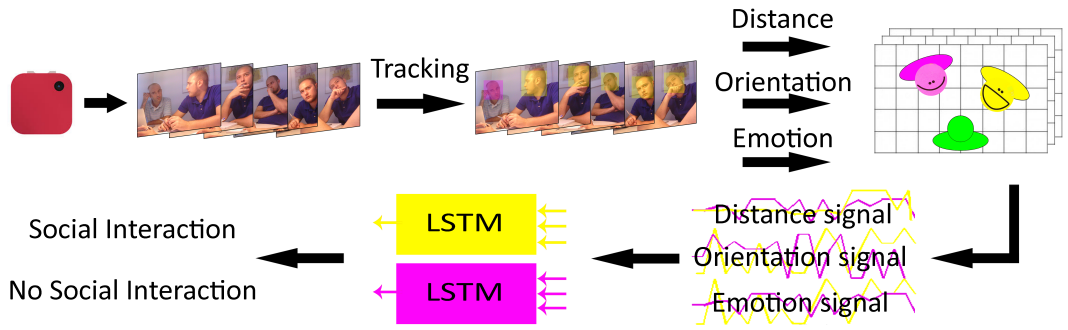


Figure 4.3: Pipeline of the proposed model for event-level social interaction detection.

4.2.3.2 Time-series classification by LSTM

Time-series classification is a predictive modeling problem and what makes this problem difficult is that the original sequences can vary in length, be comprised of a very large vocabulary of input symbols and may require the model to learn the long-term context or dependencies between symbols in the input time-series. In this context, RNNs with LSTMs showed great promise to learn the information hidden among steps of a sequence [113, 114]. LSTM owes its ability to its incorporated memory cells that use logistic and linear units with multiplicative interactions with input and output gates. In this way, it overcomes the exponential error decay problem of RNN and increasing complexity of HMM for learning long-term dependencies.

For egocentric sequence binary classification purpose, in this paper, we propose to train an LSTM network by introducing to it the time-series from each sequence as presented in the previous subsection at each time step. All the aforementioned features for each sequence are introduced to the network as input. The system must learn to classify sequences of different lengths to interacting or not by analyzing the feature vectors associated to each sequence. Hence, the system needs to learn to protect memory cell contents against even minor internal state drift. The scheme of the proposed model for the social interaction detection is depicted in Fig. 4.3.

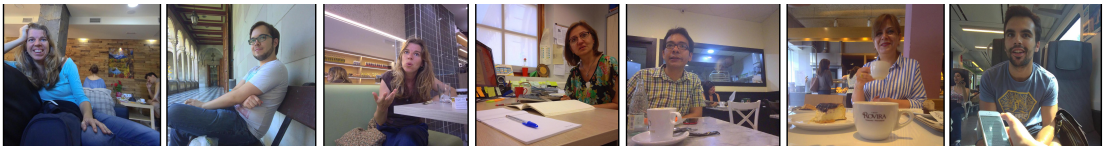


Figure 4.4: Examples of images of social interactions from EgoSocialStyle. EgoSocialStyle is the proposed dataset in this work and is captured by 9 different users in different social contexts using the Narrative Clip camera. In this work, EgoSocialStyle is employed to evaluate the proposed framework for the purpose of social pattern characterization of a user.

4.3 Validation

In this section, we introduce our dataset for social interaction analysis in egocentric photo-streams, namely EgoSocialStyle and describe the proposed experimental setup to validate our proposed approaches. A comprehensive discussion over the obtained results is also given in this section.

4.3.1 Dataset

EgoSocialStyle has been acquired by 9 users wearing a Narrative clip camera during the participation in gathering the dataset while they were living their daily life without any constraints. The camera was set to automatically capture a photo every 30 seconds once being worn. The participants who gathered the dataset had different ages and profiles and wore the camera in different and random days and times of the week. Sequences in our dataset have different lengths, varying from 20 to 60 frames (10 to 30 minutes of interactions).

The training set of EgoSocialStyle is acquired by 8 users; each user wore the camera for a number of non-consecutive days over a total of 100 days period, collecting over 100,000 images in total, wherein 3,000 images among them a total number of 62 different persons appear.

The test set is acquired by a single user, who did not participate in acquiring the training set as we aimed to study the generalization ability of our model for social pattern characterization of a person. The user wore the camera for 30 consecutive days collecting 25,200 images, where 2,639 of which correspond to social events. There are 35 sequences with more than one person appearing in them over 113, in total. 40 different trackable persons appear in the test set.

Face annotations in the whole dataset are attained using the Microsoft face annotation tool [112]. Participants were asked to provide a label (interacting/not

Table 4.1: EgoSocialStyle dataset consists of train set and test set captured by 9 different users. The details about each set is provided in this table.

#	Users	Days	Images	Social Images	People	Sequences	Prototypes	Interacting	Formal
Train	8	100	100,000	3,000	62	106	132	102	42
Test	1	30	25,200	2,639	40	113	172	130	25

interacting, formal/informal) for their own sequences. Table 4.1 provides further details of the proposed dataset.

4.3.2 Frame-level analysis

The efficiency of ego-HVFF on finding the social interactions in EgoSocialStyle is measured by calculating its accuracy in truly detecting individuals who actually interacted with the camera-wearer during a sequence. The ground-truth in this case for each sequence is a binary number for each individual who appears along the sequence. One is assigned to interacting and zero to non-interacting individuals with the camera-wearer. From the interaction probability map of each individual, we decide they interacted with camera-wearer if they are detected as being interacting in more than half of the frames of a sequence.

Note that ego-HVFF inherently only considers the yaw head orientation, and make no use of pitch and roll head orientation. Therefore, in the frame-level analysis employing ego-HVFF, pitch and roll head orientation, as well as facial expression, are not considered.

4.3.3 Event-level analysis

4.3.3.1 Data augmentation

A large amount of data for better training of deep models is a well-recognized necessity. However, the required time to acquire and label real data for this purpose is not negligible and is where artificial data augmentation could have an impact. A proper data augmentation is one which provides a reasonable set of data in addition and similar to the already existing data in the training set, but also slightly different from them to reduce overfitting of the model in learning a task [115]. Besides the impact of data augmentation in the production of additional data, it is also considered a helpful tool to provide balance to unbalanced data.

This especially is of interest in our case where to acquire sequences without any social interaction is more difficult than sequences with social interaction.

To augment the data at hand, we employed the proposed idea by Krizhevsky et al. [116]. The principle idea consists of augmenting signals by adding slight variations to them, which can be done by adding eigen-features on top of each different feature in a sequence. This has been achieved through applying PCA and then adding multiples of the found principal components to each sequence. The magnitude of the principal components is proportional to the corresponding eigenvalues times a random variable drawn from a Gaussian with mean zero and small standard deviation (0.01, in this work). This scheme generates more data in addition to the original training data by applying label-preserving transformations to them.

Let $\Phi = (\varphi_{1,n}(\tau), \varphi_{2,n}(\tau), \ldots, \varphi_{K,n}(\tau))$, $n = 1, \ldots, N$ is the set of all the N time series in our training set where $\tau = 1, \ldots, \mathcal{T}$, is the length of the sequences and consequently the time-series and, $k = 1, \ldots, K$, is the dimension of the time-series. In the social interaction detection task, N is equal to the total number of prototypes in the training set.

The augmentation of Φ from N to $\hat{N}$, with $\hat{N} = \Delta N$, is achieved through adding the vector $\hat{\Phi}_n(\tau) = (\phi_{1,n}(\tau), \phi_{2,n,}(\tau), \ldots, \phi_{K,n}(\tau))$ to the frame τ of the n-th time-series in Δ number of attempts. $\hat{\Phi}_n(\tau)$ is obtained as:

$$\hat{\Phi}_n(\tau) = [P_1, P_2, \ldots, P_K][\theta_{1,n}(\tau)\lambda_1, \theta_{2,n}(\tau)\lambda_2, \ldots, \theta_{K,n}(\tau)\lambda_K]^T,$$

where P_k and λ_k are the k-th eigenvector and eigenvalue of the $K \times K$ covariance matrix of feature values, respectively, and $\theta_{k,n}(\tau)$ is the aforementioned random variable. It is worth to mention that in the social interaction detection task, $K = 4$. In the social interaction detection, since the facial expression is a variable with discrete values, we did not consider to alter it in the data augmentation. Instead, when we generated new samples of time-series from an original time-series, we only repeated the facial expression signal of the original time-series in the augmented time-series.

4.3.3.2 Network structure and hyper-parameter optimization

In this work, we used the most commonly used version of LSTM in literature, known as vanilla LSTM [117] for time-series classification. This architecture is a three layer network consisting of the input layer, the LSTM hidden layer and a sigmoid output layer, where the input layer has forward connections to all units

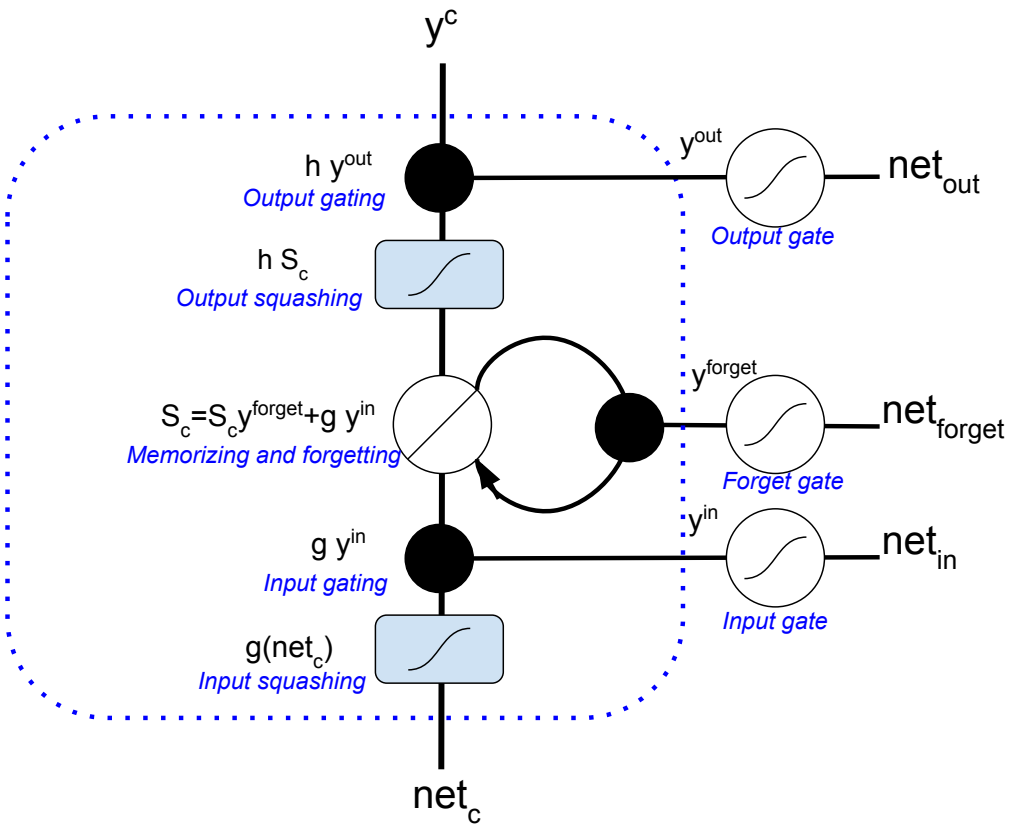


Figure 4.5: Architecture of a LSTM cell. net_c is the combination of present input and past cell state which gets fed not only to the cell itself, but also to each of its three gates. The black dots are the gates themselves, which determine respectively whether to let new input in, erase the present cell state, and/or let that state impact the networks output at the present time step. S_c is the current state of the memory cell, and gy^{in} is the current input to it.

in the hidden layer and each LSTM is composed of various numbers of memory cells. We added a dropout layer between the hidden layer and the output layer to mitigate the overfitting problem. Vanilla LSTM in contrary to the first introduced version of LSTM [118], features *forget gate* in addition to *input gate* and *output gate*. It also incorporates peephole connections and uses full Backpropagation Through Time (full-BPTT) instead of truncated gradient training. An example of the LSTM cell is shown in Fig. 4.5.

In vanilla LSTM, the output of the LSTM block is recurrently connected back

Table 4.2: Best performing hyperparameters for each setting of social interaction detection analysis.

	Learning rate	Momentum	Dropout rate	Batch size	Epoch	#Cells
SID1	0.001	0.7	0.0	20	50	30
SID2	0.01	0.8	0.0	30	50	35
SID3	0.001	0.7	0.5	50	100	30
SID4	0.001	0.5	0.0	20	100	100

to the block input and all of the gates, but it does not use *full gate recurrence* as in the initial version of LSTM. Full gate recurrence means that all the gates receive recurrent inputs from all gates at the previous time-step which greatly increases the number of parameters that has been discouraged in the literature [119]. Stochastic Gradient Decent method (SGD) is used for optimization in full-BPTT training. As the task at hand is a binary classification problem, we used an output layer with a single neuron and a sigmoid function to make 0 or 1 predictions and a log loss as the loss function. Due to the higher computational complexity of the gate specific dropout techniques in the hidden layer, we did not use any of them.

We are interested in the best performance that can be achieved for different settings of features. For this reason, we chose to tune the hyperparameters for each setting separately. Grid search with 3-fold cross-validation on the training set has been used in order to obtain best performing hyperparameters. The studied parameters for the grid-search are learning rate, momentum, dropout rate, batch size, number of epochs, and number of LSTM blocks per hidden layer. We made log-uniform sampling over the following interval of hyper-parameters: [0.0001,0.1] learning rate, [0.1,0.9] momentum, [0.0,0.9] dropout rate, [100,1000] batch size, [10,100] epochs, and [10,200] number of LSTM blocks. The best performing hyperparameters per each setting are given in Table 4.2.

4.3.4 Experimental results and discussion

As mentioned earlier, each dimension of a time-series is a variation of a unique feature along the sequence. In this section, to prove the importance of each feature and to discover the optimal combination of features, we train and test individual networks by introducing time-series composed of a different combination of features.

Table 4.3: Social interaction detection results. The best results in terms of precision, recall, and accuracy are achieved through training and testing the model on the SID4 setting.

	ego-HVFF	**SID1**	**SID2**	**SID3**	**SID4**
Precision	82.75%	80.76%	88.49%	88.59%	**91.66%**
Recall	55.81%	64.61%	76.92%	77.69%	**84.61%**
Accuracy	58.38%	61.62%	75.00%	75.58%	**82.55%**

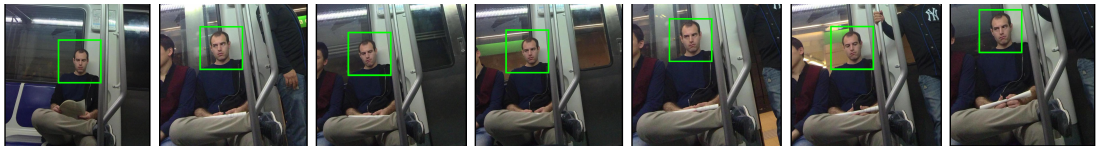
In this task, four set of settings are explored as:

- **SID1:** Distance + Yaw
- **SID2:** Distance + Yaw + Pitch + Roll
- **SID3:** Distance + Yaw + Facial expression
- **SID4:** Distance + Yaw + Pitch + Roll + Facial expressions

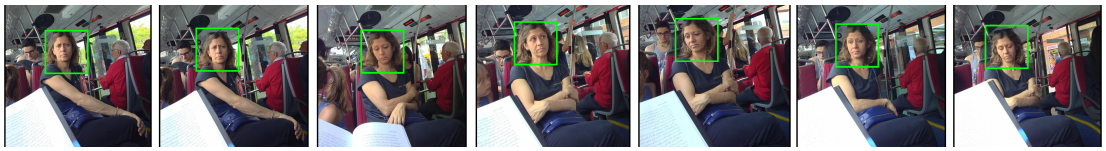
SID1 is the baseline setting in which only pure geometrical features implied by F-formation are studied in event-level. In SID2, pitch and roll in addition to yaw as the main indicator of face orientation in previous works are studied. SID3 follows the same pattern as SID1, but includes facial expression features to observe the effect of facial expressions in addition to commonly studied features for social interaction detection. Finally, SID4 includes all the discussed features for social interaction detection analysis. It is important to note that the data augmentation is only performed once for the complete 4-dimensional setting (SID4) and data in other settings is formed by selecting the required dimensions from the complete setting.

In Table 4.3, we report the obtained precision, recall and accuracy values for each of the above settings, as well as the obtained results for the frame-level social interaction detection employing ego-HVFF.

The best obtained results, in all terms of precision, recall and accuracy belong to the SID4 setting containing all the proposed features (distance, yaw, pitch, roll, facial expressions) for social interaction detection. Comparing SID1 with each of SID2 and SID3 shows that the incorporation of each of the other head orientation information and facial expression in the analysis leads to more robust social interaction detection, while facial expression shows to have a slightly stronger impact (SID3) than additional head orientations (SID2). Ego-HVFF only considers



(a) Correctly detected as no-social interaction employing SID3 and SID4, incorrectly detected as social interaction employing SID1

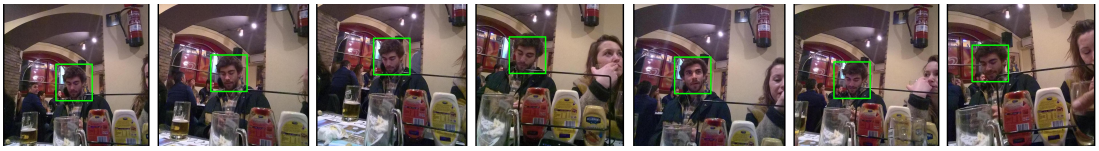


(b) Correctly detected as no-social interaction employing SID3 and SID4, incorrectly detected as social interaction employing SID1

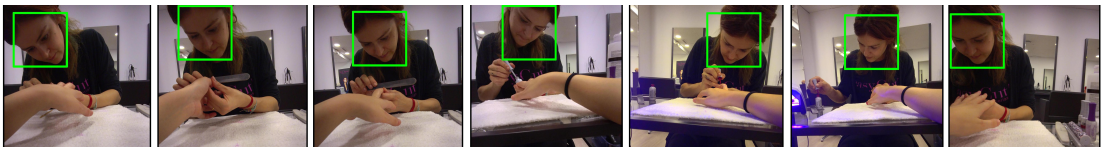
Figure 4.6: Two examples to highlight the role of facial expression. We assume the invariant *Neutral* facial expression of the individual led to classification success employing both SID3 and SID4 settings, and classification failure employing SID1 setting which does not include facial expression information. For better observability in the cluttered scene, face examples of the individuals are shown by a green bounding boxed around them.

distance and yaw orientation (SID1) for social interactions detection. However as expected, temporal analysis of SID1 in sequence-level leads to more accurate social interaction detection than frame-level analysis of the sequences as it has been achieved by ego-HVFF. Our reasoning is that since in this task all the social signals originate from the face appearance of the third-person, face occlusions due to movements of the camera or the user itself, lead to social signals discontinuity. Therefore, analysis of the sequences in frame-level results in the direct exclusion of occluded frames from the analysis while sequence-level analysis in the format of time-series mitigates the social signals fragmentation impact by considering the relation among the rest of the frames of a sequence.

Fig. 4.6 and Fig. 4.7 are visual demonstrations of how facial expressions and additional head orientations aid in more robust social interaction detection. In Fig. 4.6a and Fig. 4.6b, although the subjects are oriented towards the user and they are in relatively close proximity to the camera, we assume their invariant *neutral* facial expressions were a determinant factor in helping the model to correctly classify them as not interacting with the user. Another scenario can be observed in Fig. 4.7a and Fig. 4.7b. In Fig. 4.7b, despite the close proximity of the subject to the user and although her yaw orientation is inclined towards the user, we assume the uncommon pitch orientation of her head aided the model to correctly



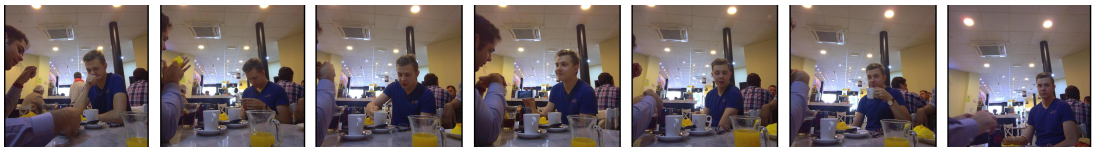
(a) Correctly detected as social interaction employing SID2 and SID4



(b) Correctly detected as no-social interaction employing SID2 and SID4

Figure 4.7: Two examples to emphasize the role of pitch and roll head orientation in social interaction detection. Sequences are correctly classified employing both SID2 and SID4 settings, and incorrectly classified employing SID1 setting which lacks pitch and roll head orientation information.

(a) Incorrectly detected as no-social interaction employing any of the settings

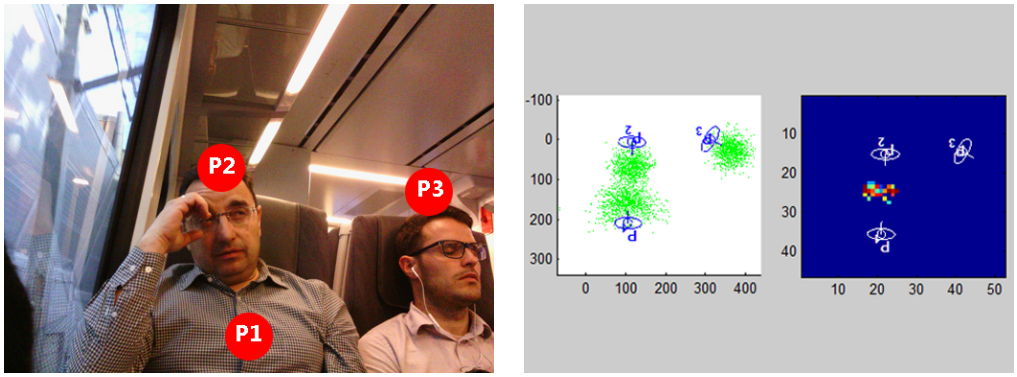


(b) Incorrectly detected as no-social interaction employing any of the settings

Figure 4.8: Examples of two sub-sampled sequences in our dataset, where sequences could not be correctly detected as interacting employing any of the settings. The uncommon head pose of the individuals in both sequences led to the model failure.

classify the sequence as not interacting with the user. Two failure cases of the detection model can be observed in Fig. 4.8. This could happen due to the uncommon head pose of the interacting people and their dominant *neutral* facial expression. Indeed in none of the examples, the interacting people are looking towards the user.

The observations from the experiments with ego-HVFF reveal that the method generally performs comparably to event-level analysis within SID1 setting. It basically fails on the frames where the interacting people are standing farther than



(a) (b)

Figure 4.9: (a) A frame of a social interaction captured by Narrative clip camera. The camera-wearer (P1) is indeed interacting with the P2, but not with P3. (b) the votes given by each individual can be seen by green clouds in front of each individual. The area of intersection between the clouds of camera-wearer and P2 can be seen in the right most plot. The colorful pixels are indication of the discovered F-formation by the ego-HVFF among P1 and P2.

a pre-defined distance for forming an F-formation ($150cm$) or when they do not look at the camera-wearer during the interaction. The latter is obserevd among interacting people with autism disorder or those who look at somewhere else when they are thinking. An example of the ego-HVFF and the found F-formation for one frame of the sequence of Fig. 1.5 is shown in Fig. 4.9.

4.4 Summary

In this chapter, we introduced two methods for social interaction detection in the domain of egocentric photo-streams: frame-level and event-level social interaction detection. The former approach is intended to decide whether there is a social interaction between camera-wearer and the other visible people in the scene, separately in every frame of a sequence. Eventually, the algorithm decides whether there is a social interaction between the camera-wearer and each individual in the scene if the number of found interacting frames between them exceeds half of the length of the sequence. In the latter approach, a set of social signals is extracted for every person in the scene and its evolution over the sequence is modeled as a multi-dimensional time-series. LSTM is employed for social interaction detection through time-series analysis.

In this chapter, we empirically demonstrated that:

- Each of the social signals studied in this work (distance among the interacting people, their head orientation, and facial expression) is important in the detection of social interactions.
- Aggregation of all the social signals to form the time-series, leads to the highest social interaction detection accuracy rate.
- The overlooked importance of *facial expressions* among the non-verbal social signals in SSP is brought into consideration. Facial expression is an important factor in augmenting the social interaction detection accuracy rate.
- Sequence-level analysis of social signals is preferred over frame-level analysis of social events. In fact, frame-level analysis has two major drawbacks. First, the classification precision is highly dependent on the selected threshold for the task. Second, the interdependency between frames and evolution of features over the time is not considered.

CHAPTER 5

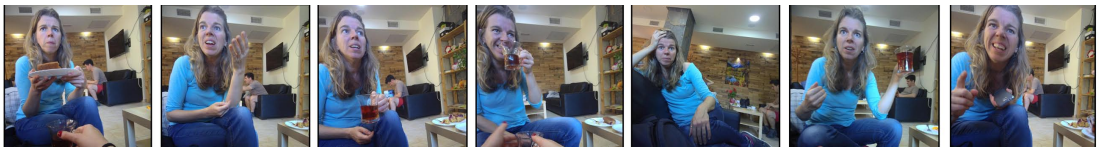
Social Interaction Categorization in Egocentric Photo-Streams

5.1 Introduction

Social interaction categorization is the task of typifying an occurred social interaction. Typification can occur within various scopes: a social interaction can be internal or external, friendly or hostile, familiar or institutional, formal or informal, etc. In the literature, three major elements have been typically exploited for social interaction categorization: the physical setting or place, the social environment, and the activities surrounding the interaction [120].

In this chapter, following Xiong et al. [69] we propose to categorize the detected social interactions into two broad categories of meetings as *formal meetings* and *informal meetings*, also known as *informal gatherings*. In this regard, we propose a model for categorization of social interactions in the domain of egocentric photo-streams. We explore the appropriate features as suggested by the relevant studies in the social science and empirically demonstrate their importance for modeling the task at hand. We employ a frame-level as well as an event-level analysis of features and report their differences, advantages, and shortcomings. For the event-level analysis of features, we made use of LSTMs and also HMMs and provide a discussion over the obtained results by each of them.

(a) Formal meeting



(b) Informal meeting

Figure 5.1: Example of two sub-sampled sequences, demonstrating the engagement of the user in different categories of social interactions; a formal meeting (5.1a), and an informal meeting (5.1b). The variations in the environment as well as facial expressions of the person in different events can be appreciated.

This chapter is organized as follows: Sec. 5.2 is dedicated to introducing the image features extracted for this task and the proposed methods for social interaction categorization in egocentric photo-streams. In Sec. 5.3 we explain the pipeline for validation of the proposed method and discuss the obtained results by different validation settings, in Sec. 5.4 we summarize the proposed method and pinpoint the key findings of this study.

5.2 Methodology

Looking closely from the computer vision perspective at the definition of each meeting as given in chapter 2, environmental features show sign of discriminative power. Therefore, for social interaction categorization, we base our approach on the use of environmental features. In addition, we also attempt to study the impact of the facial expressions of involved individuals in the interaction on defining the category of a social interaction. Fig. 5.1 shows an example of each meeting, where the differences between the environment and facial expression between two settings can be easily observed. Our approach takes into account the temporal evolution of both environmental and facial expression features by modeling them as multi-dimensional time-series and relies on the classification power of LSTM for binary classification of each time-series into either a formal or an informal meeting.

5.2.1 Feature extraction

Global features: As explained earlier, the surrounding environment of an interaction is considered among the main indicators for categorizing a meeting. Among different features for image representation, CNN features showed exceptional results for global representation of the context in images [121]. In this work, we represent each image with a feature vector extracted by taking the output of the last fully connected layer of the VGGNet (VGG16) [122] pre-trained on the Imagenet dataset [123]. However, since the image feature vector consists of thousands of variables, the computational cost becomes significant when it comes to further processing. In addition, the Hughes phenomenon [124] is inevitable when it comes to learning a high-dimensional feature space with a limited number of training samples in machine learning in general and in RNNs, specifically [125].

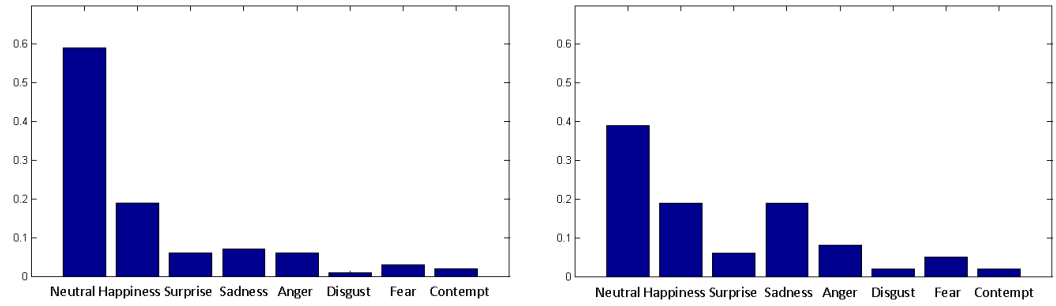
In this work, to resolve the curse of dimensionality of CNN features we propose first to apply quantization and then to apply PCA to keep the most important components of the quantization result. To quantize the CNN features, we propose to re-write them as discrete words as proposed by Amato et al. [126]. This method takes advantage of the inverted-index approach to deal with the sparsity of the CNN features to associate each component of the feature vector with a unique alphanumeric keyword. This conversion leads to a sparser textual representation of the CNN features in which the relative term is proportionally related to the feature intensity. This method showed great promises in retrieval applications.

CNN feature to word conversion essentially represents each component of the L2-normalized CNN feature vector, $f_k, k = 1, ..., 4096$, as a word:

$$w_k = \lfloor Q f_k \rfloor,$$

where $\lfloor \rfloor$ denotes the floor function, and Q is an integer positive quantification factor being $Q > 1$. For instance, if we fix $Q = 2$, for $f_k < 0.5$, then $w_k = 0$, while for $f_k \geq 0.5$, $w_k = 1$. The factor Q has a regulator effect on the features for further processing. The smaller the Q the sparser is the new feature vector and it represents less details about the original feature vector. In this work, $Q = 15$ is used which results in highly sparse feature vector representation of integer values: $(w_1, w_2, \ldots, w_{4096})$.

Given the high sparsity of the obtained word representation, a PCA is applied over the so obtained feature vectors extracted from all the images of the dataset and from the emerging representation, 95% of the most important information are kept. This process results in a 35-dimensional feature vector, $\varphi_{CNN} \in \mathcal{R}^{35}$,



(a) (b)

Figure 5.2: Bar-plot of facial expression variations over 10 randomly selected sequences for each of 5.2a formal and 5.2b informal meetings from the training set in EgoSocialStyle. Each sub-figure shows the mean of the observed facial expressions for each detected face in all the frames of 10 randomly selected sequences. Within informal meetings, people seem to express more freely their emotions as more variation can be observed.

while keeping the most important environmental features of the image. Note that applying PCA on the raw CNN features without conversion to word representation, does not result in a feature vector dimension smaller than hundreds. We are interested in keeping the dimensionality of features in the order of tens.

Facial expression: Following our hypothesis that formal and informal meetings can be characterized by the environmental characteristic as well as the facial expression of participants, integration of both features is required. A proof for this hypothesis is illustrated in Fig. 5.2 that shows the bar-plot of eight facial expressions for both formal and informal meetings. These bar-plots obtained using ground-truth information, suggest that people express more freely their emotions in informal meetings. Facial expression features in this task are extracted as the mean of facial expressions of the total number of J people detected in each frame of a sequence:

$$\varphi_{e,k} = \frac{1}{J}\sum_{j=1}^{J} e_k(p_j), k = 1, \ldots, 8.$$

5.2.2 Temporal analysis of representative features

To achieve joint effect of global image features representing the environment and facial expression features of individuals on social interaction categorization, the 8-dimensional vector of facial expression probabilities ($\varphi_e(\tau)$) is directly concate-

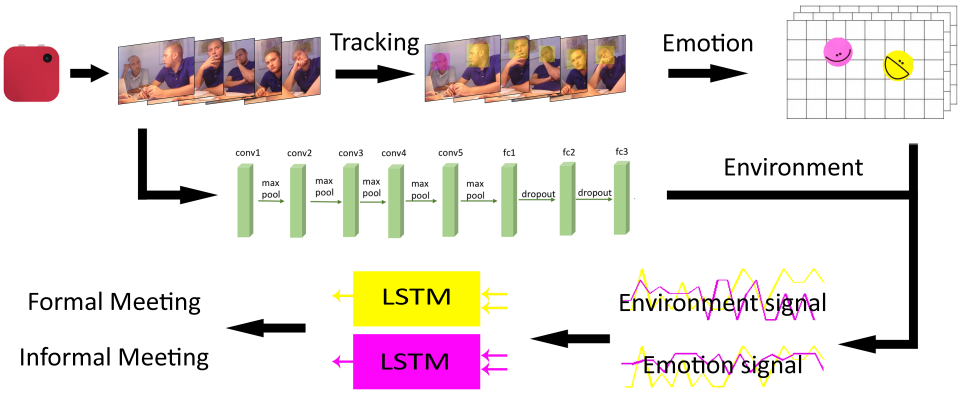


Figure 5.3: Pipeline of the proposed model for event-level social interaction categorization.

nated to the environmental features represented by global image characteristics of the event ($\varphi_{CNN}(\tau)$). Given a sequence, the time-series of interaction sequences are constructed as follows for the social interaction categorization:

$$\varphi_{categorization}(\tau) = (\varphi_{CNN}(\tau), \varphi_e(\tau)) \in \mathcal{R}^{43}, \ \tau = 1, 2, \ldots$$

The scheme of the proposed model for the social interaction detection is depicted in Fig. 5.3.

5.3 Validation

We validated our proposed method for social interaction categorization on the EgoSocialStyle. Time-series classification task into either a formal or an informal meeting is reached relying on the LSTM power for time-series classification. The same idea and approach as in *social interaction detection* task is applied for data augmentation in this task as well, with the only difference that the total number of original time-series in this task (N) is equal to the number of sequences in the training set, and dimension of the time-series is $K = 32$. For data augmentation, we did not consider to alter the facial expression signal neither in the social interaction categorization task, since the facial expression feature vector originally

Table 5.1: Best performing hyperparameters for each setting of social interaction categorization analysis.

	Learning rate	Momentum	Dropout rate	Batch size	Epoch	#Cells
SIC1	0.001	0.8	0.0	50	50	200
SIC2	0.001	0.9	0.0	50	20	150
SIC3	0.01	0.8	0.5	100	50	200

contains values of probabilities which must sum to 1 and altering them leads to a change in their essence. Instead, we only repeated the facial expression signal of the original time-series in the augmented time-series. The best performing hyperparameters per each setting are given in Table 5.1.

We kindly refer the readers to Sec. 4.3.1 for more details about the EgoSocialStyle dataset, and Sec. 4.3.3 for a comprehensive review on the data augmentation process and network structure for time-series classification.

5.3.1 Experimental results and discussion

The following settings of features are considered for the temporal analysis of social interaction categorization task:

- **SIC1:** Environmental (VGG)
- **SIC2:** Environmental (VGG-finetuned)
- **SIC3:** Environmental (VGG-finetuned) + Facial expressions

We assume that global features of an event, namely environmental features, have the largest impact on the categorization of it. Therefore in this section, the first setting (SIC1) studies only environmental features which are extracted from the last fully connected layer of the VGGNet trained over the Imagenet and preprocessed as explained in Sec. 5.2.1. VGGNet trained on the Imagenet is highly capable of grasping the general semantics in an image. However, fine-tuning the network for a specific task over relevant data for that task adapts the pre-trained network to that specific purpose. Therefore, we assume the extracted features from the fine-tuned network ideally lead to better representation of the desired classification task. In SIC2, the environmental features are extracted in the same manner as SIC1, but from the fine-tuned VGGNet over the training set of the

Table 5.2: Social interaction categorization results. The best results in terms of precision, recall, and accuracy are achieved through training and testing the model on the SIC3 setting.

	HM-SVM	VGG-FT	SIC1	SIC2	SIC3
Precision	76.82%	86.81%	87.91%	89.01%	**91.48%**
Recall	63.65%	89.77%	90.90%	92.04%	**97.72%**
Accuracy	64.87%	82.30%	83.18%	84.95%	**91.15%**

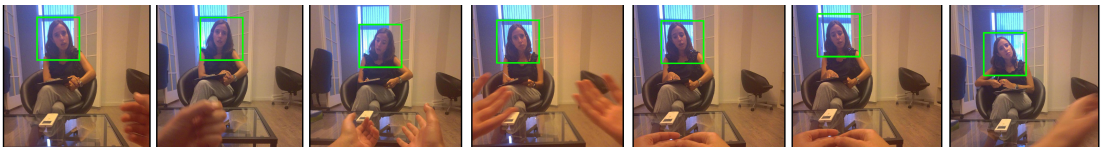
EgoSocialStyle. The features are preprocessed in the same manner as explained in Sec. 5.2.1. Fine-tuning the network is achieved through the instantiation of the convolutional part of the model up to the fully-connected layers and then training fully-connected layers on the photos of the training set. The last setting to be studied is SIC3, which explores jointly the effect of facial expressions as well as the environmental features in social interaction categorization.

In this task, we used VGGNet pre-trained over Imagenet for feature extraction, while any other CNN architecture suitable for image feature extraction could be employed and finding the optimal CNN architecture was out of the scope of this work. Also, due to the narrow field of view of the Narrative camera, the Imagenet dataset was preferred to a seemingly more relevant dataset such as Places [127]. In the images captured by Narrative, a scene is better observed by the set of visible objects in it rather than the wide view of the scene.

In Table 5.2, we report the precision, recall and accuracy values obtained for each setting of the aforementioned settings. Additionally, we compared our obtained results with HM-SVM [64] which is an applicable state-of-the-art method to our setting as this model similar to ours makes use of extracted features in the egocentric setting and analyzes them in sequence-level but different to our proposed model employs an HMM to model interaction sequences according to features to categorize them. To apply HM-SVM, the HMM is trained using our training set where features follow the SIC3 setting. The HM-SVM is later employed to label the interaction state. We also report achieved results by a baseline method, VGG-FT, in which the fine-tuned VGG network on the photos of the training set in EgoSocialStyle is tested over the pool of photos in the EgoSocialStyle test set. Thus, it is considered a frame-level modeling of the problem.

The obtained results suggest that temporal analysis of environmental features extracted from fine-tuned VGGNet in SIC2 setting outperforms temporal anal-

(a) Correctly detected as informal meeting employing SIC3



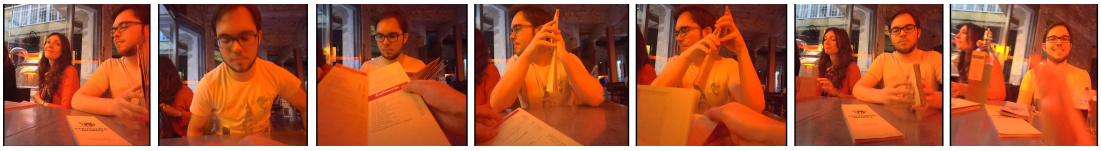
(b) Correctly detected as formal meeting employing SIC3

Figure 5.4: Two successful examples employing SIC3 setting, emphasizing on the role of facial expressions in social interaction categorizations. The method trained over mere general features employing SIC2 setting did not lead to the right categorization of each of the sequences.

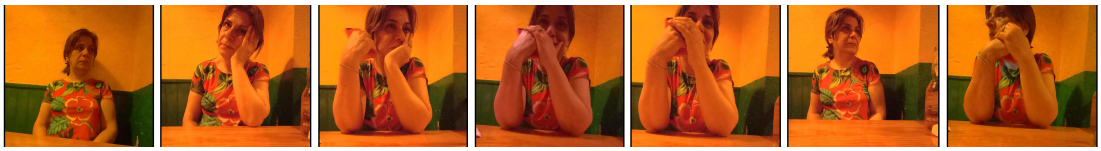
ysis of environmental features extracted from VGGNet before fine-tuning in the SIC1 setting. Temporal analysis of fine-tuned features also outperforms frame-level analysis of fine-tuned features in VGG-FT which is also an indication of the importance of temporal analysis of features in this task. The combination of environmental features extracted through fine-tuned VGG network and feature vector of facial expressions probabilities leads to the highest performance of the model. HM-SVM is trained and tested with features in the SIC3 setting. However, the obtained results suggest that the LSTM demonstrates more power in modeling the problem at hand than the HMM.

It is worth to note that, due to the extensive amount of data that end-to-end models need for training (few million data) and to our limited number of image sequences in the dataset, we did not consider to design our proposed model in an end-to-end fashion. Indeed, making use of pre-trained networks, like emotion, makes a more effective use of the resources when the available data is small compared with the amount of data needed to train the individual sub-networks.

In Fig. 5.4, two sequences are shown in which the aggregation of facial expressions with the general environmental features employing SIC3 leads to the correct categorization of them. In Fig. 5.4a, although the environment is the indicator of a formal meeting, we assume the variant facial expressions of the subject aids the model to correctly classify it as an informal meeting. On the contrary, in Fig. 5.4b despite the scene not implying a formal meeting, we assume the dominant *neutral* facial expression of the subject leads to the correct categorization of



(a) Incorrectly detected as formal meeting employing SIC3



(b) Incorrectly detected as formal meeting employing SIC3

Figure 5.5: Two failure examples of the model trained on any of the social interaction categorizations settings. We assume misleading environmental features in 5.5a and invariant *neutral* facial expressions of the subject in 5.5b led to these failure cases.

the sequence as a formal meeting. Fig. 5.5 shows two cases where the model fails to correctly categorize social interactions due to misleading features transmitted from the scene. Both Fig. 5.5a and 5.5b are informal gatherings which are classified incorrectly as formal meetings. Our assumption is that in Fig. 5.5a the model confuses the menu with a piece of paper which is an important characteristic of a formal meeting. We also assume in Fig. 5.5b the invariant *neutral* facial expression of the person leads the model to fail.

5.4 Summary

In this chapter, we introduced a novel method for social interaction categorization into formal or informal meetings. Based on an extensive body of sociological literature, the importance of a set of high-level image features in automatic analysis of this task is investigated for the first time. We have proposed different models as well as different settings of features for this analysis, including frame-level and event-level analysis of features for this analysis. For the event-level analysis of features, our proposed method for multi-dimensional time-series classification using LSTM is compared to HMM for time-series classification. Following the validation of the models, a discussion over the obtain results is provided to highlight the advantages of LSTM over HMM in addressing this problem and differences among various settings are underlined.

In this chapter, we empirically demonstrated that:

- Setting and characteristics of the social environment are important factors for social interaction categorization.
- Facial expression is an important factor in augmenting the social interaction categorization accuracy rate.
- Sequence-level analysis is preferred over frame-level analysis of social events.
- LSTM is preferred over HMM for sequence-level analysis of time-series in this task.

CHAPTER 6

Face Clustering in Egocentric Photo-Streams

6.1 Introduction

Face clustering, also known as face discovery, is the task of grouping face images in a dataset into either known or unknown number of disjoint groups. Face clustering is a suitable task for cases where the identity of people in the dataset is not available. Face clustering, has vast number of applications, such as interactive photo album tagging [72, 73, 74], social media [128], and medical purposes [129, 130, 131, 85, 81, 82]. Face clustering is also useful to unveil less noticed matters about the social life of the user: with whom does the user interact the most? how many times did the user meet a friend last month?

In this chapter, we propose a fully unsupervised approach for face clustering from egocentric photo-streams collected over a long period of time. In this context, face is the most discriminating feature of a person since, depending mostly on the clothing, a person appearance may change drastically in different days or even at different times of the day. To cope with the extreme intra-class variability of faces, we propose to first track the appearance of multiple faces into the same event using [2], and then considering both the *inner-track* and *inter-track* constraints, to cluster similar faces across the events into an unknown number of groups.

Figure 6.1: Each row is the resulting prototype of tracking by eBoT [2] over a sequence of two people.

This chapter is organized as follows: In Sec. 6.2 we detail our proposed approach for face clustering in egocentric photo-streams. In Sec. 6.3 we introduce the dataset used in this paper as well as the experimental setting and we discuss the experimental results. Finally, in Sec. 6.4, we summarize the content and the contributions of this approach.

6.2 Methodology

Given a large and unconstrained photo-stream captured by a wearable camera, we propose a face clustering approach by leveraging inner-class and inter-class constraints derived from the face tracking of people across the photo-stream.

6.2.1 Face-example vs. Face-set

To overcome the challenges imposed by the free motion of the camera and by its low temporal resolution, eBoT multi-face tracking [2] is applied on resulted sequences from segmentation of photo-streams, to extract prototypes (tracklets) of the appearing people in them. A final prototype keeps the bounding boxes of face occurrences of one individual along that sequence, so in the case that two persons appear in a sequence, eBoT outputs two separate prototypes that localize face occurrences of each individual, separately (see Fig. 6.1).

Due to the characteristics of the camera, faces appear in a variety of views and in different ambient conditions even within a sequence. We treat all the observed occurring variations of the same face in a sequence as a unique representation of the same face for face discovery in the whole dataset. Hereafter, we refer to each bounding box in a prototype as a *face-example* and define all the bounding boxes of a prototype as a *face-set*.

6.2.2 Face discovery in egocentric photo-streams

Unlike the majority of face discovery frameworks that solely rely on pair-wise comparison of face-examples at the time to find face matches, we propose a model which is built upon a tracking framework that provides us with a set of correct variations of the same face in one sequence. In this way, the proposed algorithm reshapes the face discovery challenge from *face-example-pair* comparison, to *face-set-pair* comparison. In our approach, the deterministic factor in deciding whether two different face-sets belong to the same person is defined through a measure of dissimilarity. We first calculate the dissimilarity between all the possible pairs of face-sets, and then, based on these measurements, we employ a hierarchical clustering technique to discover the most similar face-sets.

6.2.2.1 Dissimilarity between two face-sets

For simplicity, let us suppose that given two face-sets, say R and T, we want to measure the dissimilarity between *target*, T, and the *reference*, R. Let $l(R)$ and $l(T)$ be the lengths of R and T, respectively. Let $r_i \in R$ be the $i-$th face example in the R, where $i = 1, \ldots, l(R)$ and $t_j \in T$ be the $j-$th face example in the T, where $j = 1, \ldots, l(T)$. To compute the dissimilarity between R and T, we first define two similarity matrices: S^R representing the similarity between all possible pairs of face-examples in R, and S^T representing the similarity between face-examples in R and face-examples in T. We compute S^R as to build a baseline about how similar faces inside a face-set are.

The similarity between two face-examples is measured by their average deep-matching score [104]. The deep-matching is a descriptor matching algorithm, built upon a multi-stage architecture with interleaving convolutions and max-pooling layers and uses dense sampling to retrieve correspondences with deformable patches. More specifically, instead of using SIFT patches as descriptors, each SIFT patch is split into four quadrants and, assuming independent motion of each of the four quadrants, the similarity is computed to optimize the quadrant positions of the

target descriptor. As a consequence, the descriptor is able to deal with various kinds of image deformations, including scaling factors and rotations. Denoting by $\Delta(x, y)$ the value of the deep-matching between x and y, the elements of S^R are defined as $s^R_{i,k} = \Delta(r_i, r_k), i, k = 1, \ldots, l(R)$ and the elements of S^T as $s^T_{i,m} = \Delta(r_i, t_m)$, with $i = 1, \ldots, l(R), m = 1, \ldots, l(T)$. Finally, the dissimilarity $\delta(R, T)$ between R and T is calculated as the absolute difference between the median value of S^R and S^T, say φ^R and φ^T, respectively:

$$\delta(R, T) = \left|\varphi^R - \varphi^T\right| \tag{6.1}$$

6.2.2.2 Clustering of face-sets

To cluster face-sets based on their dissimilarity, we used agglomerative clustering, a hierarchical bottom-up approach that repeatedly merges pairs of clusters based on a measure of dissimilarity to form larger clusters. In this work, the initial clusters are face-sets and Eq. (6.1) is used to measure the dissimilarity between face-set-pairs. All dissimilarity relations between face-set-pairs are encoded by the matrix $\mathcal{D} \in \mathcal{R}^{N \times N}$, where N is the total number of face-sets. To take into account the fact that face-sets extracted from the same sequence should belong to different subjects, we force the dissimilarity between these face-sets to be maximal. Specifically, we introduce the constraint matrix $\mathcal{C} \in \mathcal{R}^{N \times N}$, where its elements $c_{m,n} = 1$ if the face-sets $r_m \in R$ and $t_n \in T$ were extracted from the same sequence and $c_{m,n} = 0$, otherwise. We then multiply each element of $\mathcal{D}$, say $d_{m,n}$, by the weight $w_{m,n} = c_{m,n} + 1$.

To determine the cut-off threshold, that is, when to stop merging clusters at a selected precision, we measured $\delta(R, T)$ of various face-sets in two manner: first, $\delta_s(R, T)$, where R and T are different face-sets of the same person, and second, $\delta_d(R, T)$, where R and T belong to two different people. The cut-off threshold θ is taken as the median value of all the values of $\delta_s(R, T)$. These calculations are performed over a training dataset consisting of 100 face-sets. Fig. 6.2 shows the $\delta_s(R, T)$ values on the left and $\delta_d(R, T)$, values on the right, over the training dataset, where the vertical is the separating line between them and the horizontal lines are the median of $\delta(R, T)$ values in each section.

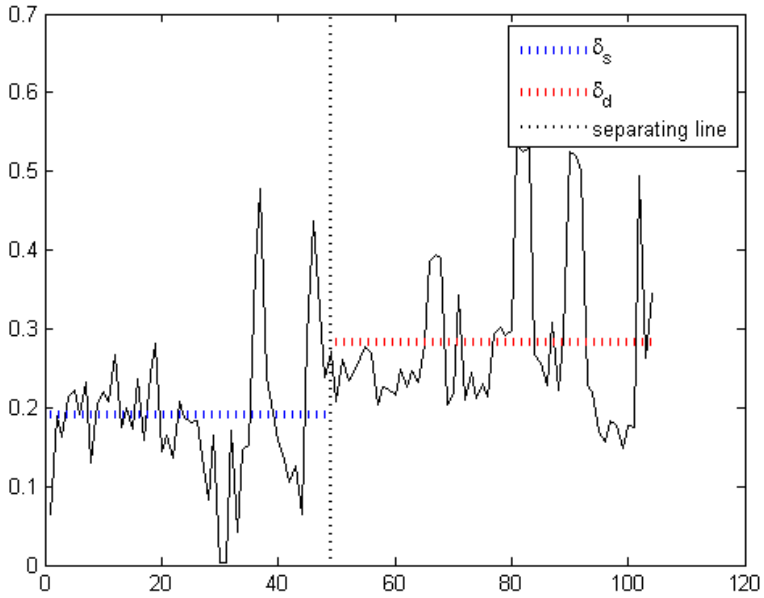


Figure 6.2: Threshold estimation on a separate training set, different from EgoSocialStyle: left side of the separating line shows $\delta_s(R,T)$ values, and right side of it shows $\delta_d(R,T)$ values. The horizontal lines are the median of $\delta(R,T)$ values in each section.

6.3 Validation

6.3.1 Dataset

We validated the proposed model for face clustering on the test set of the EgoSocialStyle. Some additional information to those given in Sec. 4.3.1 are that the 25,200 images are captured by one user wearing the Narrative Clip camera during 30 days. It contains 2,033 faces belonging to 40 persons, whose bounding boxes have been manually annotated in ground-truth. In average, each person appears in 6 sequences and 3 days. There are 35 sequences with more than one person appearing on them over 113, in total.

As mentioned earlier, a separate dataset is used to select a cutoff value discussed in the previous subsection. It is acquired by 8 users; each user wore the Narrative clip for a number of non-consecutive days over a total of 100 days period, collecting 30,000 images, where 3,000 images of them contain a total number of 100 different trackable persons. Sequences in both datasets have different lengths, varying from 10 to 40 frames and they have been acquired in real-world conditions, including inside and outside scenes.

6.3.2 Baselines

In this section, we evaluate state-of-the-art methods with different settings over our dataset. The following is a brief description of each setting.

- **M1 (WBSLRR):** The proposed method by Xiao et al. [75] is applied on the face-sets obtained by applying eBoT. This method similar to ours considers the inner-track and inter-track constraints introduced in Sec. 6.2.2.2 to learn a more discriminative low rank data representation.

- **M2 (Spectral, Open-face, Face-pairs):** An implementation of face analysis tool with deep neural networks based on the work proposed by Schroff et al. [78], known as OpenFace [132] is employed. First, faces are detected using a pre-trained model for face detection. Second, they are transformed in an attempt to make the eyes and bottom lip appear approximately in the same location on each image. Third, a deep neural network is used to embed the face on a 128-dimensional unit hypersphere. Finally and forth, the spectral clustering method is used to group faces into groups corresponding to different subjects.

  As the first detection method is common among our proposed model (in tracking using eBoT) as well as baseline models, we performed it once using Openface and kept it intact for all the models. In this way, we avoid a possible bias in the analysis that could be imposed by the differences among different face detectors.

- **M3 (Agglomerative, Open-face+Euclidean, Face-pairs):** The same setting as M2 is employed, despite variations in the forth step. In this setting, Agglomerative clustering is applied over the pair-to-pair Euclidean distance between 128-dimensional face features.

- **M4 (Spectral, Open-face, Face-set-pairs):** As an attempt to validate the effect of employing the inner-track and inter-track constraints, we used as the initial clusters the face-sets resulting from applying eBoT. A unique 128-dimensional feature vector as the mean value of all the 128-dimensional faces feature vectors is representing each face-set.

- **M5 (Agglomerative, Open-face+Euclidean, Face-set-pairs):** A similar setting to M4 for face-set representation is employed. Agglomerative clustering is then employed to cluster face-sets based on their Euclidean distance from each other.

Table 6.1: Percentage of NMI and ARI values for different baseline settings (M1-M5), and our proposed model (M6-M7)

	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
NMI	24.31	21.18	58.35	68.95	19.21	78.79	**83.68**
ARI	00.59	00.21	31.66	01.49	00.42	23.44	**33.84**

6.3.3 Evaluation measurements

To compare our proposed method with the baseline models, we used two distinct widely known measurement techniques for clustering evaluation with known true-labels: Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI). Both measurements have values ranging from 0 to 1, with 1 indicating that the clustering result perfectly matches the ground-truth.

NMI measures the mutual information between the labels predicted by the classifier and the true-labels, ignoring permutations. It is calculated based on entropy and does not need to predefine cluster numbers. NMI utilizes inner-cluster distinctness and intra-cluster agglomeration to measure clustering results and needs to relate the labels indicating the clusters acquired by the clustering algorithm to the labels predefined by the user.

ARI on the other hand measures how similar the labels predicted by the classifier are to the true-labels. Mathematically, ARI is related to the accuracy. It evaluates on a pairwise-basis if two sets of labels are incorrectly grouped so its value is representative of the true clustering result. ARI evaluates how well the algorithm splits input data into different clusters by looking at the relationship between clusters and not between clusters and the given labels.

6.3.4 Discussion

Although NMI and ARI validate the results in distinct ways, both follow the same trend as it can be observed for different methods in Table 6.1. M1 to M5 are the baselines introduced previously, M6 is the proposed model without considering the inter-track constraints, and M7 is the complete pipeline of the proposed method, as described in Sec. 6.2.2.

Open-face is a robust method for extraction of facial features. However, as it can be observed, the proposed method employing the deep-matching approach can grasp a more robust idea of the similarity between face-example pairs which is

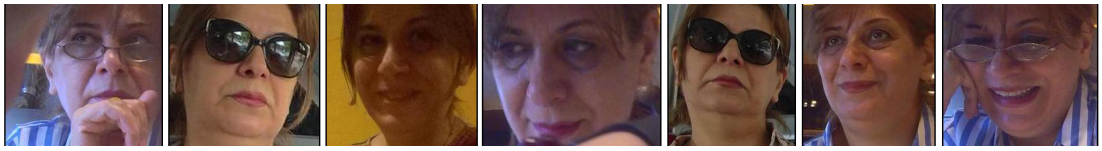


Figure 6.3: A few examples of faces belonging to one cluster obtained by applying our proposed model on the EgoSocialStyle test set. The visual variation among face examples can be appreciated.

proved by higher NMI and ARI values comparing M2 to M5 with M6 and M7. Additionally, WEBSLRR, despite its robust pipeline for face clustering in controlled environments, performs poorly on our dataset. We consider that this is due to using only pixel intensities for the image representation. Also, as expected in our proposed method, constraining the problem by exploiting inter-track constraints (M7) allows to improve the accuracy up to 10% considering ARI with respect to the same approach without inter-track constraints (M6).

The experimental results unveil the challenges involved in clustering of appearing faces in photo-streams captured by a wearable photo-camera under free-living conditions. Indeed, face appearances may change even along the same event since people take out or put on their accessories such as glasses, a hat, makeup, and so forth, making the problem very challenging (see Fig 6.3). However, our conclusion from the experiments is that the unbalanced number of images per individual in a face-set is the most challenging problem for face clustering. In this work, we aimed to study only facial attributes, disregarding any additional information. This analysis is important when the additional features are not either available, because of the nature of the applications or they are costly to provide.

6.4 Summary

In this chapter, we addressed the face clustering problem in the challenging domain of egocentric photo-streams. The problem at hand is complex to solve as we rely solely on face attributes in an image set captured under free-living conditions. The proposed model, through employing a deep-matching technique grasps robust representation of the face similarities. Moreover, by applying two inner-track and inter-track constraints, the proposed model achieves a relatively high performance while outperforming the state-of-the-art methods and baselines.

Main characteristic and advantages of our proposed model can be summarized as follows:

6.4 Summary

- The proposed method can be considered the first complete pipeline for face clustering in the domain of egocentric-photo-streams.
- The proposed model relies on the deep-matching method to find various appearances of the same face that undergoes drastic variations in the egocentric setting during a long time capturing of images.
- To deal with the aforementioned challenges, our proposed model is built upon a multi-face tracking to incorporate both the inner-track and inter-track constraints to improving the robustness of the results.
- A strategy is designed and validated to learn the cut-off threshold of the agglomerative clustering over a training set.
- Capability of our proposed model, as well as the baselines, is measured based on two broadly used metrics for calculating the clustering performance with known labels.

CHAPTER 7

Social Pattern Characterization in Egocentric Photo-Streams

7.1 Introduction

Building upon the previous chapters, this chapter goes beyond event-level social interaction analysis in egocentric photo-streams, relying on the long-term observation and analysis of social interactions of a user. Characterizing social patterns of a camera-wearer requires its identification through quantifying the frequency, the diversity, and the type of social interactions during the observation period. This is accessible when social events of the user are previously localized and their type is categorized. In sociology, interaction frequency is the total number of social interactions per unit time and interaction diversity indicates how diverse is one's social interaction considering two types of formal and informal social interactions. This is accessible as our proposed model is a hierarchical model that initiates by segmenting social interactions and indexing them sequentially in time as individual events.

A visual overview of the proposed pipeline is given in Fig. 7.1. As the first step, face tracking is employed over the potential social events to localize the position of an interacting person with the user along it. Later, from the bird's-eye view representation of the scene, social signals (social distance, face orientation, facial expression), as well as environmental features, are extracted for each frame

and used to represent each sequence as a time-series. An LSTM is employed to classify each time-series according to the task at hand: social interaction detection or categorization. On the other side, face clustering enables determination of the diversity and the frequency of social interactions. Finally, social pattern characterization requires the integration of all tasks. By leveraging the proposed framework, we seek to address the following key questions: *How often does the user engage in social interactions? With whom does the user interact most often? Are the interactions with this person mostly formal or informal? With how many people does the user interact during a month? How often does the user see a specific person?*

In this chapter, we formalize the common terms used for characterizing a social interaction and report them numerically over the EgoSocialStyle dataset. We quantitatively and qualitatively demonstrate that our proposal for social pattern characterization leads to sensible understandings of social patterns of a wearable-camera user. To the best of our knowledge, this work is the first comprehensive social pattern characterization study from a first-person perspective. Social pattern characterization is the ultimate step of our proposed framework for social signal processing in egocentric photo-streams.

This chapter is organized as follows: Sec. 7.2 is devoted to bringing into details the proposed methodology for generic as well as person-specific social pattern characterization in the domain of egocentric photo-streams. In the same section, we also explain the role of face-clustering in the analysis. In Sec. 7.3 we report the results of the proposed method on the EgoSocialStyle and the public EGO-GROUP dataset. Finally, in Sec. 7.4 we summarize the main contributions in this chapter.

7.2 Methodology

7.2.1 Generic social interaction characterization

Characterizing the social pattern of an individual implies the ability to define the nature of social interactions of the user from various temporal (frequency, duration, etc.) and social (type, identity, number of interaction people with the user, etc.) aspects. Providing a definition within the aforementioned contexts demands social interaction analysis of the user across several events during a long period of time. For this purpose, we define four concepts to characterize social interactions, namely *frequency*, *social trend*, *diversity*, and *duration*.

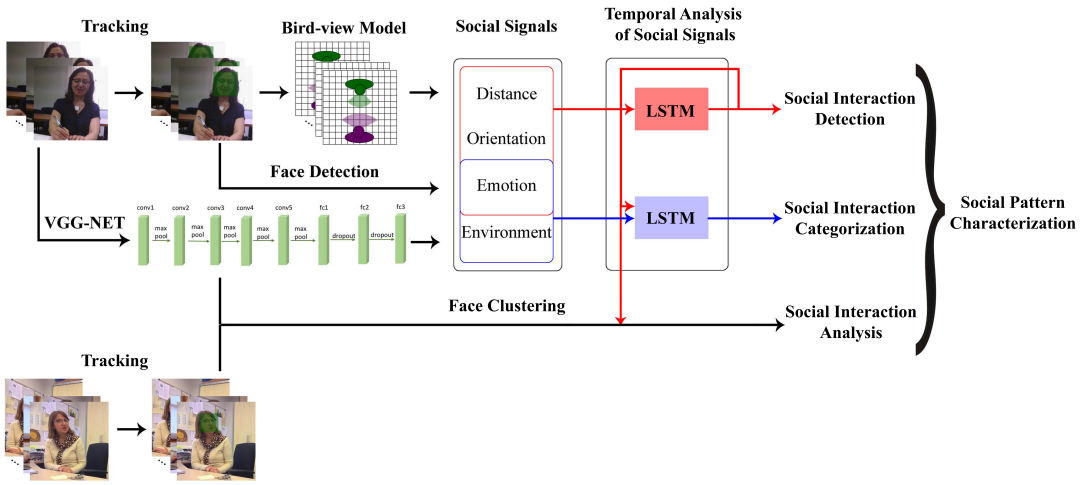


Figure 7.1: Complete pipeline of the proposed method. Face tracking is employed to localize the position of an interacting person with the user along a social event. From the bird's-eye view representation of the scene, social signals (social distance, face orientation, facial expression), as well as environmental features, are extracted for each frame and used to represent each sequence as a time-series. An LSTM is employed to classify each time-series according to the task at hand: social interaction detection or categorization. On the other side, face clustering enables determination of the diversity and the frequency of social interactions. Finally, social pattern characterization requires the integration of all tasks.

Frequency (F): Defines the normalized rate of formal (or informal) interactions of a person by the total number of observation days:

$$F_{formal(informal)} = \#formal(informal)\ interactions/\#days$$

Social trend (A): Indicates whether the majority of social interactions of a person are formal (or informal):

$$A_{formal(informal)} = \#formal(informal)\ interactions/\#all\ interactions$$

Diversity (D): Demonstrates how diverse are social interactions of a person. The term is defined as the exponential of the Shannon entropy calculated with natural

logarithms, namely:

$$D = 1/2 \exp\left(- \sum_{i \in \{formal, informal\}} A_i \ln(A_i) \right)$$

Note that when the person has the same number of formal and informal interactions (i.e. $A_{formal} = A_{informal} = 0.5$), $D = 1$.

Duration (L): Defines the longitude of each social interaction of the user. The duration is the longitude of the sequence corresponding to the i-th social interaction, say $L(i) = \mathcal{T}(i)r$, where $\mathcal{T}(i)$ is the number of frames of the i-th interaction and r is the frame-rate of the camera. Different statistics can be applied to the duration of interactions like average or median to characterize social interactions and extract the social pattern.

7.2.2 Person-specific social interaction characterization

In this subsection, we explore the aforementioned concepts for social interaction characterization of the user within the context of interaction with a specific person. This firstly requires that all the interactions of the user with a certain person to be localized. To this goal, the face clustering method introduced in the previous chapter (6) is employed to find various appearances of the same person among all the social events of the user. As mentioned in the previous chapter, the face clustering method is applied on the tracking step to cope with the extreme intra-class variability of faces. In a single event, tracking gathers a set of different appearances of the same face in that event, called a *face-set* in this context, which allows reshaping the face clustering task in different events to face-set clustering.

7.2.3 Face-cluster analysis

Let $\mathcal{C} = \{c_j\}$, $j = 1, \ldots, J$ be the set of clusters obtained by applying the face-set clustering method on the detected interacting prototypes, where J ideally corresponds to the total number of people who appeared in all social events of the user along the whole period of observation (e.g. a month). Each cluster, c_j, ideally contains all the different appearances of the person p_j across different social events, and $|c_j|$ is the cardinality of c_j which demonstrates the number of social interactions events of the user with the person p_j during the observation period.

As the clustering method and the proposed method for social interaction de-

Table 7.1: Social pattern characterization results, demonstrating the generic and person-specific frequency (F), social trend (A), diversity (D), and Duration (L) of the social interactions of the user.

	F-Formal	F-Informal	A-Formal	A-Informal	D	L	
Generic	0.83	2.50	0.25	0.75	0.87	25.19	1.32
Person-specific	0.25	1.00	0.20	0.80	0.59	18.80	0.96

tection and categorization act at event-level, inferring the interaction state of each sequence inside a cluster is straightforward. The frequency, the social trend, the diversity, and the duration of the interactions with a specific person, can be computed in the same manner as explained in 7.2.1, by restricting the considered interactions to the ones with the person of interest.

7.3 Validation

To illustrate the ability of the proposed framework for social pattern characterization of an individual, face clustering is applied on the test set. A total number of 83 clusters is obtained, which is almost double the size of the total number of prototypes in the test set. The largest cluster contains 77 number of faces from 5 number of sequences belonging to the same person in various social events.

The different statistics of the social interactions of the user, as well as those related to the most frequently interacted person, are given in Table 7.1. From our observation, it can be concluded that during the observation interval, the user most frequently interacts with a specific person 5 times, in 4 different days, and 4 times of which occurs during informal meetings. An interesting observation is that in a cluster containing different sequences, a sequence may belong to a formal or informal meeting which implies the user may have different types of interaction with the same person in various social events. According to the statistics reported in Table 7.1, generic diversity of social interaction of the user is relatively high (87%). Specifically, the user is three times more inclined towards having informal meetings than formal meetings (Generic A-Formal vs. A-Informal, 0.75 vs. 0.25) and thus, more frequently gets engaged in informal meetings as supported by the statistics. Interestingly, the generic social trend of the user is correlated to the person-specific one (0.05 difference in both formal and informal social trends). The above interpretation is expected when assuming an informal social interaction

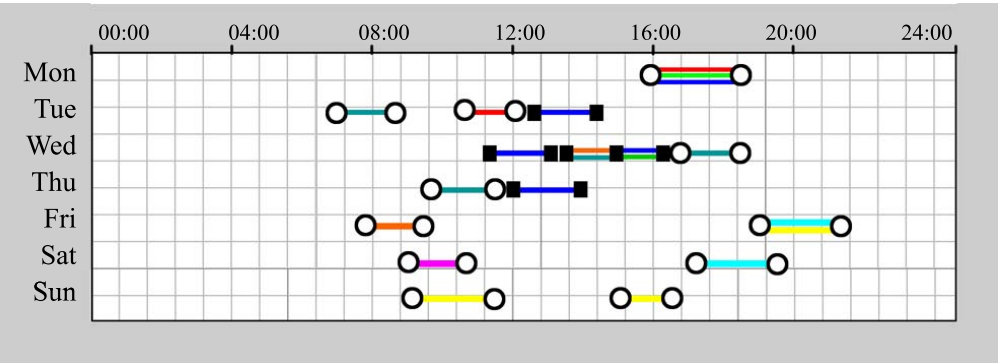


Figure 7.2: Temporal map of social interactions of the user during one week. The boundaries of an interaction are shown by circles for informal and squares for formal interactions. Different line colors are the index of interaction with different people and multiple lines within a boundary are indicative of the interaction with multiple people.

can occur at any time without any planning, while for formal social interactions normally planning is involved [69].

The social pattern of the user over one week according to the obtained results from clustering and inference to their types is visualized in Fig. 7.2. Social interactions are shown by horizontal colored lines, where the interaction boundaries are shown by circles for informal meetings and squares for formal meetings. Different colors correspond to different persons. Re-occurring people in one social event are shown with parallel lines within the same interval. As it can be observed in Fig. 7.2, informal social interactions of the user are happening at almost any time of the day and the formal social interactions are normally happening during the middle of the day.

7.3.1 Social pattern characterization on EGO-GROUP

Despite the lack of available datasets for the purpose of social pattern characterization in egocentric vision, to demonstrate the effectiveness of our proposed model, we applied the entire pipeline on EGO-GROUP [61], a most adaptable public dataset to our considered purpose in this work. Despite the fact that EGO-GROUP is not a designed dataset for computing the statistics of the social style of a user (social pattern characterization), it offers a benchmark that is directly suitable for social interaction detection and adaptable for social interaction categorization in the domain of egocentric vision.

EGO-GROUP is a social group detection dataset for egocentric vision, which consists of 18 videos collected in five different scenarios: laboratory, coffee break, conference room, outdoor, and party. The ground-truth data available with the dataset in addition to the type of each scenario provides interaction labels for each individual. To adapt the dataset to the definition of social interaction category in this work, we labeled the laboratory and the conference room videos as the formal meeting, and party, coffee break, and outdoor as informal meeting scenarios. As mentioned before, social pattern characterization purpose requires long-term monitoring of daily life of a person, while EGO-GROUP consists of single detached by scenario sequences that are captured under controlled, and not free-living conditions. For this reason, in this section, we report the obtained results for social interaction detection and categorization as well as face clustering.

For the sake of a fair comparison, we down-sample the videos captured in 15 fps to 1 fps photo-streams. Within the terminology used in this paper, we obtained 21 social events (sequences) and 76 prototypes. For social interaction detection, we followed the same proposal as explained in Chapter 4, with the only difference that the distance feature is calculated as it is proposed in the original paper [61]. For social pattern categorization, we used one event of each scenario for fine-tuning the network and used the new fine-tuned network to extract the word representation of training set for training the LSTM. Later, the appropriately trained LSTM is used for testing the model. For both of social interaction detection and categorization tasks, we only evaluate the models on the best-performing settings of features on EgoSocialStyle, being SID4 in the case of social interaction detection and SIC3 in the case of social interaction categorization. We report the obtained results on EGO-GROUP in terms of precision, recall, and accuracy in Table 7.2.

EGO-GROUP does not provide any clustering ground-truth to validate this task. However, as part of the entire framework, we also applied the clustering on this dataset. Examples of the face-examples in the biggest obtained cluster are shown in Fig. 7.3. This cluster contains 86 face-examples of the same person from several events across three different scenarios in EGO-GROUP.

7.4 Summary

Social pattern characterization of individuals requires long-term observation of their social interactions. For this purpose, wearable photo-cameras are specifically suitable as they allow long-term recording of the life of a user. In this chapter, we presented a formalized methodology for analysis of the social pattern of a user

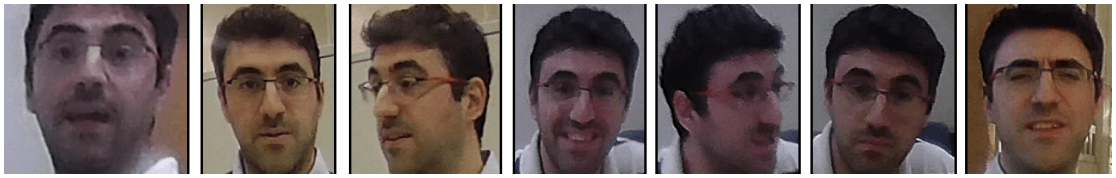


Figure 7.3: A few examples of faces belonging to the biggest cluster obtained by applying the face clustering clustering method [3] on the EGO-GROUP dataset. Face-examples in this clusters belong to three different scenarios of EGO-GROUP.

Table 7.2: The obtained results in terms of precision, recall, and accuracy on the best performing settings for both tasks of social interaction detection (SID4) and categorization (SIC3) on EGO-GROUP.

	Detection	**Categorization**
Precision	86.11%	90.00%
Recall	77.50%	75.00%
Accuracy	81.57%	76.47%

based on frequency, duration, and diversity of social interactions. To demonstrate the generalization ability of the proposed approach, we tested our proposed models on the test set of EgoSocialStyle which is acquired by a user who did not participate in acquiring the training set used for training them. Interpretable quantitative and qualitative results are a proof of claim. On the other side, we applied the entire model on EGO-GROUP, the most suitable public dataset for our purpose.

To summarize, in this chapter we demonstrated that:

- Social pattern characterization of a wearable photo-camera user is possible and can be achieved by closely monitoring the social behavior of the user through the visual analysis of its egocentric photos.
- A rough idea of the social pattern of an individual can be achieved by its characterization within four concepts: frequency, social trend, diversity, and duration of social interactions.
- In a similar way that social pattern of a wearable camera user can be characterized generally, it can be also characterized with one specific person within the social circle of the camera-wearer.
- Our proposed method for social pattern characterization can be effectively applied to any other dataset which holds certain characteristics for social

pattern characterization of a person, i.e. includes long-terms visual recordings of their livings.

CHAPTER 8

Clothing: A Social Signal Processing Perspective

8.1 Introduction

In our society and century, clothing is not anymore used only as a means for body protection. According to evidences studied within the social sciences, clothing brings a clear communicative message in terms of social signals, influencing the impression and behaviour of others towards a person. In fact, clothing correlates with personality traits, both in terms of self-assessment and assessments that unacquainted people give to an individual. The consequences of these facts are important: the influence of clothing on the decision making of individuals has been investigated in the literature, showing that it represents a discriminative factor to differentiate among diverse groups of people. Unfortunately, this has been observed after cumbersome and expensive manual annotations, on very restricted populations, limiting the scope of the resulting claims.

In this chapter, in the format of a position paper we want to sketch the main steps of the very first systematic analysis, driven by social signal processing techniques, of the relationship between clothing and social signals, both sent and perceived. Thanks to human parsing technologies, which exhibit high robustness owing to deep learning architectures, we are now capable to isolate visual patterns characterising a large types of garments. These algorithms will be used to capture

statistical relations on a large corpus of evidence to confirm the sociological findings and to go beyond the state of the art. In this chpater, we will sketch the future steps of the first systematic study on which social signals are conveyed by clothing, proposing a framework within the scope of computer vision to measure the clothing effect on the impression that we have on ourselves and that we trigger in the others. More precisely, in a first phase we will investigate the basic visual cues that could be associated to social signals; for example, checking how much tight shirts are associated to the social signal of attractiveness. In a second phase, we will perform a higher level analysis, investigating the types of behaviors that may result from a social interaction, in dependence on the type of garments worn by the interacting people; for example, analyzing interactions between formally versus casually suited individuals. All of this would be possible since computer vision technologies are now mature for a fine-grained analysis of the clothing, providing precise dense segmentations of outfits as results of human parsing algorithms, and automatically recognizing diverse clothing items [133, 134, 135] and styles [136].

The chapter is organized as follows: in Sec. 8.2, the literature on clothing in terms of social sciences is reviewed and clothing analysis in terms of human parsing approaches is reported. The core of the paper is also presented in this section, where we discuss our ideas related to the study of clothing under the social signal processing umbrella. The chapter ends with some final remarks in Sec. 8.4.

8.2 Methodology

8.2.1 Clothing and social semiotics

Semiotics, as originally defined by Ferdinand de Saussure, is "the science of the life of signs in society" [137]. Semiotics investigates *signs* and analyzes them to provide significance to a specific problem. There are three main elements in semiotics: the sign, what it refers to, and the people who use it. The people as social species and biological entities, instinctively evolved to survive better through facilitating living in a disciplined society by defining new signs and giving them an appropriate interpretation. Social semiotics is a subcategory of semiotics that studies how people design and interpret meanings and how these meanings are shaped by a specific social situation [138]. In social semiotics, the term *resource* is preferred over the term *sign* and represents a used signifier by the people to produce and to interpret communicative artifacts. In this respect, social semiotics is particularly useful in disclosing unnoticed significance and functionality of social resources and

each individual is a semiotician, since everybody constantly interprets the meaning of signs around them.

Humans signify specific social context through *resources* of all type, whether visual, verbal, or gestural. *Clothing* is a non-verbal resource that transfers meanings about individuals in the society. Cloths hold a symbolic and communicative role having the capacity to express style, identity, profession, social status, gender, or group affiliation of an individual. Although the symbolism that clothing carries on is not always clear, it evidently can be considered as the most desired personal image that one is willing to project to the society [139]. The study of how people use and interpret specific social context through dress is known as *clothing semiotics* or *fashion semiotics*, although some believe that clothing is distinct from fashion [140]. Within their definition, clothing is "any covering for the human body", while fashion is "the style of dress that is temporarily adopted by a discernible proportion of members of a social group, because that chosen style is perceived to be socially appropriate for the time and situation". Originally, clothing semiotics was studied from the fine arts perspective. Later, the perspective has been expanded and the study covered the human needs in this respect [141]. Subsequently, in the 1960s, the social and psychological implications of clothing began receiving more attention from scholars. Today, clothing remains a common topic of study in social psychology [142] as it conveys social meaning about an individual and groups of people. It is in this way that the semiotics of clothing can be linked to the social semiotics.

In spite of the fact that clothes have such large potential to convey a message, it must be noted that clothing semiotics understanding is complex. The social context affects the interpretation of clothing, thus, having a precise knowledge of the unconscious symbolism attached to forms, colors, textures, postures, and other expressive elements that affects the interpretation of clothing in a given culture is a desired quality in automatic analysis of this information.

8.2.2 Clothing and computer vision

Clothing style is commonly intended in computer vision as the set of visual attributes and category labels that describe an outfit [143, 134]. Examples of visual attributes are *colors* (red, green, etc.), *clothing patterns* (solid, striped, etc.), and more technical *qualitative expressions* (skin exposure, placket presence, etc.) [143]. However, it is worth noting that these attribute taxonomies cannot go beyond a certain level of details, i.e. fine grained details such as the type of hat are not

Figure 8.1: Parsing example. The input image (left) and the final output of parsing (right) employing the proposed model by Yamaguchi et al. [4].

listed in the list of attributes; in facts, only generic types of objects are available in existing dataset annotations. These visual attributes have been used to measure the similarity between outfits paving the road for the identification and analysis of visual trends in fashion [133]. The category labels are textual expressions that individuate a particular type of clothing item (shirt, sweater, etc.) [134]. In most of the cases, all these textual labels are accompanied by a pixel-wise segmentation of the outfit, in which each segment is associated to a category label, and to one [144] or more visual attributes [145] (see Fig. 8.1).

This segmentation is the output of an operation commonly referred as *human* [146, 147, 148], *clothing* [134] or *fashion parsing* [144]. Clothing style can be also modeled without referring to a particular outfit, but instead to a larger set of category labels and visual attributes [136].

Human parsing is usually performed by statistical classifiers, which operate after a training phase. The training data may consist of fully labeled data, which means images of individual outfits in which each of the pixels has a label indicating the category and/or the attribute [134]. This is the most reliable source of information to train a classifier, but it is extremely cumbersome to get: in facts, manual annotation is necessary, which requires 15-60 minutes to be carried out for each single image [149]. Alternatively, weak labeling can be provided, which means to have training data in which an entire image is associated to a set of textual labels (in other words, the textual labels are not localized over the pixels of the outfit image). This obviously reduces the human labor to get training samples, but at the same time is less expressive, leading to classifiers which are

not completely automatic: for example, the model introduced by Liu et al. [144] requires that the testing image too comes with textual labels that indicate what to look for in the image.

Most of the techniques for human parsing builds upon a preliminary operation, which is that of fitting a skeleton on the human body depicted in the input image. This operation is called pose estimation [150], and helps to introduce a structural prior for the parsing process, which individuate salient joints (ankles, knees, hips, shoulders, elbows, wrist, neck). These points are connected by sticks forming a skeleton, which in turn drive the parsing to align with it, providing anatomically plausible segmentations [134]. Unfortunately, pose estimation techniques are prone to errors in the case of missing data, due to occlusions or auto-occlusions; for this reason, images of single persons where the entire body is portrayed, are preferred. Images depicting parts of the body (as those ones captured via wearable sensors, where usually the whole body does not fit) represent a serious issue. In addition, pose estimation is weak in the case of large and long clothing, covering the structure of the body for what concerns some of the joints (a person wearing long dress has its knees completely covered). This issue has been recently faced by facing human parsing and pose estimation as two intertwined aspects of the same problem, introducing the concept of *semantic part* (such as leg, arm, head) [146]. A semantic part can be iteratively modeled with tools usually employed for human parsing (as the *Parselets* [135]) and as an ensemble of joints, taking from the pose estimation literature.

8.3 Validation

The four-step pipeline of Vinciarelli et al. for SSP [13] suggests that after having recorded the scene and detected humans (step 1 and 2), in the step 3, feature extraction has to be performed, and in the step 4, social signals have to be grounded with the scene context, in order to understand social interactions. In this chapter, we are interested specifically in the last two steps of the pipeline, since we assume that the scene has been already recorded and the individuals have been properly detected.

In the rest of this section, we will individuate the research questions (indicated with the letter **Q**) that can be inserted in these two steps, providing our intuitions about possible answers (letter **A**), driven by the literature of the human sciences and/or our speculations, together with the type of experiments we would like to carry out, to provide the community with deeper insight and novel tools for

clothing social signaling.

8.3.1 Clothing behavioral cues as an individual social signal

A-**Q1 - How much clothing-related cues are independent from other standard behavioral cues, in the determination of particular social signals?** The question essentially asks how the mapping from visual features related to clothing (for example, the type and appearance of a particular clothing item, e.g., a shirt) has to be carried out in dependence from other cues such as the ones reported in [13] (Table 1, pag.1745). In other words, this question is very preliminary and asks for a feasible and reliable protocol with which clothing-related cues can be analyzed without caring of the effects due to other features in determining social signals.

A-**A1 -** In social situations a clothing outfit comes with the body that wears it, so that other cues, in particular related to physical appearances, gesture and posture, face, emotional expressions and eyes behavior [13, 151] are obviously co-present and some cues may have different effects depending on the visible human body. For example, facial expression comes more into vision if only the upper body is visible. This could be the reason why online shops often present garments without the human body (Fig. 8.2). Thus, an analysis on these data seems reasonable and may help for answering A-Q1.

A simple yet important experiment would be that of checking whether the presence of different types of body appearances will change the nature of the social signal transmitted. In particular, our first step is to enrich the annotation of a clothing dataset, for example, the *Exact Street2Shop* dataset [152]. For a given garment, the dataset contains some "shop pictures", where the garment is usually located on a neutral background, without being worn by a human body. Together with this, the dataset offers a "street picture", where the same garment is worn by a subject among an undefined set of people. The idea is to first annotate standard semantic information about the people in the street photos (gender, expressions etc.). Next, different assessors will evaluate the street and the shop photos, defining the person wearing that particular garment in terms of social signals and personality traits. In the case of the street photos, the person in the picture is present and annotated, while in the case of the shop photos, persons are absent. The goal is to discover whether the presence of the person changes significantly the judgment of the assessors, and if this correlates with the semantic

Figure 8.2: The picture shows an example of clothing outfits typical of online shops.

information associated to the person.

A finer setup, given a particular person, could be that of isolating the most the cues related to clothing by masking behavioral cues coming from the face (blurring the face) or hiding the height (removing the background scene). The interrelation between behavioral cues and other features in terms of social signals has never been investigated in the literature.

A-**Q2 - How to evaluate the nature of a social signal generated by clothing behavioral cues?** For understanding this question, one may consider the *Brunswick's Lens* model. A simplified version of this model is adopted by Cristani et al. [153] (see Fig. 8.3). In few words, the model says that a social signal is not necessarily univocally intended. More in the detail, the model assumes that a social signal is sent by a sender, S, as a consequence of its internal state, μ_s, which is assumed to be measurable. For example, S feels himself extrovert (his internal state), and this awareness is measured by a self-assessment (for example, using the Big Five questionnaire [154]). S wears some clothing items and as we are assuming that clothing items are related in some way with the internal state of S, they can be assumed as an *externalization* of the internal state. The receiver R sees the clothing items worn by S, and infers about the internal state of S, which in this case is μ_r, to highlight that possibly is not equal to μ_s. This process, called *attribution*, which brings to a perceived state what can be measured itself. The Brunswick Lens model states that a social signal has high ecological validity ρ_{EV}, if there is a high correlation between the internal state of S and the features. Viceversa, a social signal has high representational validity ρ_{RV}, if the correlation between the features and the state inferred by R is high. Finally, if the internal state of S correlates with the inferred state of R, it means that the communication

through the social signal mediated by the features has high functional validity ρ_{FV}.

A-**A2 -** The Brunswick's Lens essentially states that the nature of a social signal should be measured considering the sender of the signal and the receiver. This opens up to diverse experiments, suggesting a protocol for each one of them. For example, in order to understand how a particular social signal built upon clothing behavioral cues is interpreted by a generic receiver, it is necessary to measure the perceived state of multiple assessors. If the correlation between the behavioral cues and the perceived state of the assessors is high, we may individuate the *implied* meaning of a particular outfit. In more practical terms, to assess whether an athletic outfit is a behavioral cue that communicates the social signal of extroversion, this can be asked to a set of assessors. If the features that characterize the athletic outfit correlate with the extroversion assessment, then this message can be understood that athletic outfit triggers a certain reaction in a generic audience in terms of social signals. In order to individuate the attributes that most consistently originate social signals, deep learning technologies will come into play. One of the most attractive features of deep architectures is that they can be "opened" and "visualized", allowing to easily interpret what is codified into the internal layers [155, 156, 157, 158]. Exploiting these strategies, once annotations of social signals have been extracted from garment images, the goal would be that of feeding them into deep architectures, capturing the most discriminative visual patterns. In this way, the generic semantic label of "athletic outfit" can be explained in terms of behavioral cues (in the sense of [13]), like shape, color and texture attributes.

A-**Q3 - Is there an agreement between one's self-image and the impression conveyed to others through his/her clothing style?** It has long been known that *clothing affects how other people perceive us as well as how we think about ourselves*. This question asks whether there is a consistency between self-perception of an individual and perception of other people towards him/her.

A-**A3 -** Often people choose what they wear as a means of self-expression. The individual measurement of the effect of clothing on self-perception and perception of others, has been studied previously by Heart [159]. However, the question of whether others perceive the desired message that the person wishes to communicate, has not been explicitly studied before. An experimental set up should first facilitate separate investigation of clothing effect over self-perception of an individual and perception of others over them and then study their correlation.

A-**Q4 - Which clothing behavioral cues are related to the social signal of the *attractiveness*?** This question asks if clothing style influences the

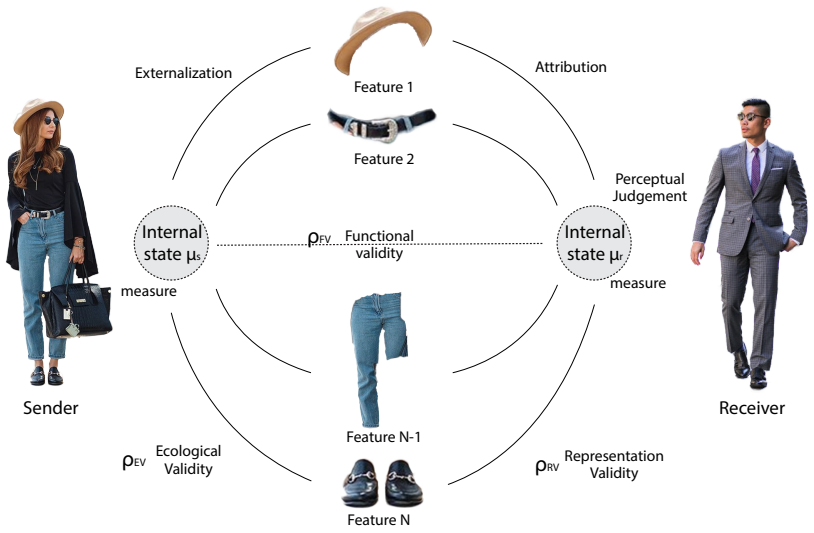


Figure 8.3: The picture shows a simplified version of the Brunswik Lens Model adapted to the transmission of a social signal between a Sender and a Receiver.

perception of attractiveness of others. Attractiveness is the main social signal associated to physical appearance [13] and attractive people tend to be considered as having high status, good personality, and being extrovert.

A-**A4 -** Clothing is considered as an indicator of socioeconomic status [94, 95] and personality traits [97]. According to Johnson et al. [100], the most investigated concepts using dress manipulations were *dress*, *status*, and *attractiveness* and listed the most experimented dress manipulation to study the effect of clothing on attractiveness concept as grooming, tidiness, makeup, and natural physical appearance such as hair and eye color, height and weight (see [100], Table 3). Clothing can be strongly related to arise perception of attractiveness in people towards a person. To prove this hypothesis, a possible experiment would be that of capturing the influence of single clothing items, or multiple clothing elements arranged in an outfit towards the definition of an attractive person. In the same line with A-Q1, other behavioural cues should be considered, selected or masked, so to avoid explaining away effects. Also in this case, deep learning architectures and tools to visualize them ([155, 156, 157, 158], see A-A2) will help in segregate and explicitly individuate visual attributes that convey the impression of attractiveness.

A-**Q5 - How much impact clothing have on the other individuals**

first impression? The importance of first impression comes more into attention in relation to its effect on the overall lasting impression. The last impression is what we will remember most about a social situation, however, one probably will not have a last impression if do not get the right first impression.

A-**A5 -** "You never get a second chance to make a first impression" as noted by Oscar Wilde. Although a large amount of cues aggregate together to form a first impression, we hypothesise clothing is a strong cue that eases the process for the people [160] to make a first impression. This quick judgment that happens in less than a minute [161], can lead us towards a set of assumptions about a set of personal traits of that person, such as attractiveness, likability, competence, and aggression [162]. Howlett et al. [163] studied the effect of clothing alone on the first impression and reported that clothing solely influences the first impression of the others even in limited exposure time. To detect the influence of individuals first impressions, the labeling of the *Exact Street2Shop* dataset (see A-A1) can be performed including the time dimension into play, enforcing the user to give a quick answer on the impression the clothing does trigger, explaining then by textual attributes the item(s) that leads him to such an answer.

8.3.2 Grounding clothing related social signals with scene context

B-**Q1 - How much clothing-related cues help in capturing the context of a social interaction?** The idea here is to study how clothing items worn by people involved in a focused or unfocused interaction [11] can tell about the interaction itself.

B-**A1 -** People wearing outfits of a very similar kind, different from that of the rest of the crowd, are connected with a high chance, and this in turn helps in individuating the nature of a social interaction. Sport players with the same attire and supporters with the same t-shirts in a spectator crowd could be considered as an example of this connection. In this case, simple counting algorithms, specialized to finding similar items in a scene, may be of a great help [164]. However, the problem becomes more challenging when it comes to other types of interactions, namely ordinary exchanges in generic scenarios (waiting in a bus stop, attending a conference, etc.).

An ideal solution is to develop models capable of first, assigning clothing visual attributes to social scenarios (learning the most expected garments on the beach, in a Starbucks, during a conference etc.), and second, individuating similarities

among outfits, from those very explicit (team outfits that are different only for the number depicted on the shoulders) to those more insidious to catch (individuating people that bring the same bag to individuate a social event). The interplay with social sciences lies in motivating the results from the learning stage of the model, that is, analyzing and interpreting the most emblematic garments for a particular interaction. Even in this case, deep learning technology will help, especially those architectures equipped with region proposal modules [165]. The novel idea here could be that of assuming the region proposal module as on-line evolving, drifting towards the detection of people exhibiting similar clothing.

B-**Q2 - How the clothing style drive people to socialize? Specifically, do people with the certain type of apparel socialize with similarly suited people?** This question asks whether our higher tendency to socialize with certain people is influenced by the clothing they wear, and if people tend to socialize more with similarly dressed people.

B-**A2 -** The approach towards answering this question is twofold. On one side, clothing in the same way as being considered as a flag to make visible a specific ideology, culture, or ethnicity, it also can be considered as a social catalysts among similarly dressed people. In an example, Nash [166] studied the influence of dressing on runners and stated that when two runners are dressed alike they engaged in an extended conversation as opposed to a short nonverbal greeting that occurred among runners that dressed differently from each other. On the other side, the effect of clothing on the people's self-perception, leads to variations in their social relations. As an example, feeling comfortable is an important factor in a social interaction and clothing has the power to make a person feel comfortable or not. Simply, when clothing can be used as a criterion for judgment, people may unconsciously feel judged and act according to it. The connection with pattern recognition here lies in the approaches for detecting gatherings of people, which are proven to be very robust and versatile to diverse types of scenarios [11, 167]. Applying pattern recognition techniques for correlating clothing types of interactants will unveil possible affinities which may facilitate social interactions.

8.3.3 Towards clothing style interpretation

A collage of typical photos during social exchange from EgoSocialStyle are reported in Fig. 8.4. As visible, people are having different types of interactions (eating together, discussing, looking at each other). In this setup, the analysis of the impressions triggered by the clothing can be carried out exploiting a rich set of

Figure 8.4: Montage of photos from EgoSocialStyle during social interactions. Fine details about the facial expression, body posture and hand gestures can be appreciated, but also the clothing can be observed at a fine grain.

fine grained features. In fact, details of the clothing can be observed in the second row of the figure. The analysis of this kind of details can help in the study of the relations between a given impression that a subject is producing and his/her garment. Unfortunately though, this type of analysis has never been performed so far.

In this regard, we performed an analysis to see whether deep CNN are in fact capable of providing us with a meaningful interpretation of clothing. More precisely, in the first step we are interested to answer *what makes a clothing to belong to a certain style?* This is important, since if components of an style are extractable individually or in relation to each other, then further analysis of that style can be formulated easier. For example, let us consider that *Hipster* style can be defined when a person is seen to be wearing sunglasses, a buttoned shirt with rolling sleeves, a hat, a pair of cotton pants with a leather belt, and possibly Converse shoes. In this way we are able to *describe* Hipster style with its i.e. six main characteristics provides an explanation to the style in the format of social signals, which simplifies further analysis of the style.

Various works towards extracting interpretable explanation of deep models have been previously introduced. An state-of-the-art is the model introduced by

Fong et al. [168]. This model presents an image saliency prediction paradigm by learning what part of an image if perturbated affects most the output score of the algorithm for classification task. We employed this model to make an idea over what are share common characteristics within a certain clothing style. Fig. 8.5 shows some primarily obtained results for four different styles introduced in Hipsterwars dataset [136]. These results are obtained by using trained GoogleNet [169] over Fashionista dataset and later fine-tuned over Hipsterwars. The order of training and fine-tuning is on one hand due to the large size of Fashionista with respect to shorter size of Hipsterwars. On the other side, Hipsterwars was the only available dataset with clothing style annotation until very recently.

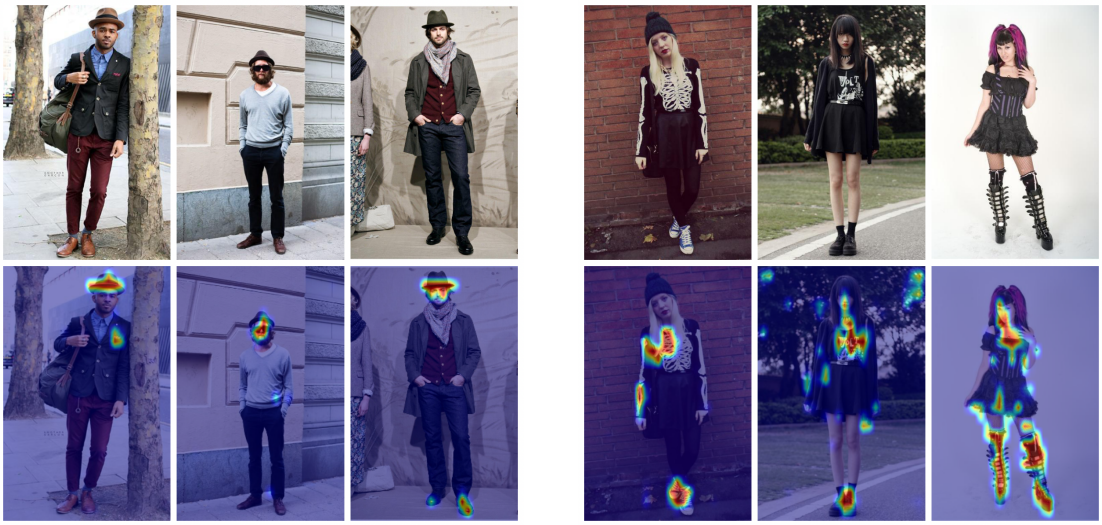
To the best of our knowledge, understanding over *what* makes each of these styles falls mainly within fashion studies and unfortunately there is no official definition for each of these styles yet available. However, as it can be observed in Fig. 8.5, salient points from each style share some common characteristics, while are different from characteristics of other styles. An example is the role of shoes in Pinup versus Bohemian style. In Pinup style, high-heels are the most determinant factor, while in recognition of Bohemian style, shoes almost have no importance.

8.4 Summary

In this chapter some research questions that are related to the investigation of the social signals associated to clothing are presented. The outcome of these questions, other than filling a gap in the social signal processing literature, may have important relapses. For example, it could facilitate the design of online personal stylists able of indicating, on the one side, the type of impressions one's outfit may trigger (associated to their particular body or posture), and on the other side, which are the most suitable outfits for attracting the attention of others. This has helpful applications in facilitating social interactions.

To summarize, in this chapter we presented:

- A systematic analysis, driven by social signal processing techniques, of the relationship between clothing and social signals, both sent and perceived.
- For the related analysis from the computer vision perspective, deep learning technologies seem to be the most convenient analysis tool. Other than being strongly effective as for classification and regression, one of the most attractive features of deep architectures is that they can be opened and visualized, allowing to easily interpret what is codified into the internal layers.



(a) Hipster

(b) Goth

(c) Pinup

(d) Bohemian

Figure 8.5: Saliency map for four different style examples employing [168]. As can be observed, salient areas for each style are different from other one, i.e. salient part in a Hipster style is hat and sunglasses, while salient part in the Pinup style are the high-heel shoes and reddish makeup.

- Despite experiments are in its very initial phase, we are confident that the intuition is good, that the technology is ready to purse our goals, and that the results would be of great impact not only in the fashion field, but in general in all the social signal processing area.

CHAPTER 9

Conclusions

In this thesis, a sequential pipeline for automated social signal processing in the domain of egocentric photo-streams is presented. The thesis begins by characterizing the benefits of using a wearable photo-camera and discusses the involved challenges in the task. Later, it highlights the importance of this new line of attack on classical SSP related computer vision problems. In continuation, it brings into details each component of the proposed pipeline in separated chapters and provides a discussion over the obtained results along the validation phase.

9.1 Findings

Our proposed model for multi-face tracking in egocentric photo-streams, extended-Bag-of-Tracklets, is designed to cope effectively with induced challenges by the domain, i.e. motion unpredictability, low frame-rate, and frequent occlusions. Due to the low temporal resolution of the camera, the model addresses the problem of tracking by treating it as a matching problem: the most similar patches to a face example are found by patch-matching analysis and gathered along a sequence. The set of gathered faces in a sequence which ideally belongs to one person forms a tracklet. In the next step, a voting strategy is applied to select the most common

face bounding box in every frame of the sequence. Unreliable bounding boxes are prone away by introducing and applying a new measure of confidence. In this chapter, we show that exclusion of occlusions is an important step in increasing the robustness of a tracking model. Bag-of-Tracklets, in this work, has been applied to reach a higher performance in tracking that could be achieved singularly by a simple forward-backward patch matching. However, this idea can be applied in any other similar scenario where a set of candidates for a patch are available and the task is to choose the most reliable patch among all.

We proposed two models for social interaction detection in chapter 4. One model is based on frame-level analysis of social signals, and the other one, based on their sequence-level analysis. Through comparing the social interaction detection performance achieved by each model, we discovered the importance of sequence-level analysis of social signals which leads to a higher performance. The studied social signals in each model according to F-formation formalization are the mutual distance between interacting people and their head orientation. However, as an additional social signal, we also studied the role of facial expressions of the interaction people on the social interaction detection performance and showed it boosts the social interaction detection robustness. Proving the importance of sequence-level analysis of social signals, as well as studying the role of facial expressions in social interaction detection tasks are two main contributions of this chapter. In addition, it should be noted that related analysis for social interaction detection given in this chapter is the first attempt at solving the task in the domain of egocentric photo-streams.

The importance of sequence-level analysis of related social cues for social interaction categorization into a formal or an informal meeting is also disclosed in chapter 5. In this chapter, we brought into machine vision analysis the proposed determinant factors studied in sociology -mainly environment of the meeting, for identifying its category. In this regard, we proposed to employ a model for extraction of compact features representing the meeting environment. We also studied the role of facial expression of the interacting people in determining the category of the social interaction. We consider this as the first attempt at automated analysis of social signals for categorization of social interactions in the domain of egocentric photo-streams.

In chapter 6, we proposed a model for face clustering in the domain of egocentric photo-streams. Our proposal is built upon the eBoT, our multi-face tracking model, to enable calculating the similarities between face-sets gathered along the sequences, instead of face-examples. In order to deal better with the drastic variation in face appearances during long time photo acquisition, the similarity score is obtained by applying the deep-matching approach. Upon calculating the similarities among face-sets and employing both inner-track as well as inter-track constraints, the agglomerative clustering with a previously learned threshold is applied to decide on the final cluster members where each cluster ideally belongs to the face appearance of one person in the dataset. We count our proposed model as the first work to tackle the problem of face clustering in the domain of egocentric photo-streams.

Our proposed pipeline for social pattern characterization of a wearable photo-camera user is wrapped up in chapter 7. In this chapter, we formally defined the principle terminologies to characterize the social style of a user, being frequency, diversity, social trend, and duration of a social interaction. We demonstrated that this task is possible by presenting quantitatively and qualitatively the results and drawing a sensible conclusion out of them. In addition to demonstrating the obtained results over our proposed dataset, EgoSocialStyle, we also reported the result of our proposed model over the public dataset, EGO-GROUP, which is a proof of the generalization ability of the proposed pipeline.

In this thesis, the majority of the experiments are held on our proposed dataset for social pattern characterization of a user, EgoSocialStyle. As mentioned in Sec. 4.3.1 of chapter 4, EgoSocialStyle comes with a vast amount of annotations, including interacting/not interacting flags per each prototype and Formal/informal meeting flag for each sequence, as well as face clustering annotations.

In chapter 8, we introduced a new branch of research in SSP and presented first steps towards its formalization. In this chapter, we proposed some research questions which we believe extrapolating them leads to solving the problem. For each question and within the SSP scope we suggested a coarse solution to them. In this chapter, we took an step towards clothing style understanding and reported the preliminary results of some of our own experiments towards answering it.

9.2 Future Lines

The contributions and limitations of each of presented methods within the pipeline for SSP in egocentric photo-streams are opening future research lines:

In this thesis, for SSP in egocentric photo-streams, a set of social signals are newly analyzed which conventionally were not subject to analysis. However, we believe there are other social signals which their role is still to be discovered in future studies. Some of these social signals such as *distance* among interacting people, and their *clothing style* as a clue to categorize the type of a social interaction, are previously mentioned in sociology, but has not arrived at automatic machine vision analysis. We assume further studies in sociology and psychology are required to reveal more related social signals for more effective social signal processing.

Moreover, in this thesis, we only expand social interaction analysis with their categorization into formal and informal meetings. However, we believe further studies can be taken place to study role of similar features, or other relevant feature to put social interactions into another set of categories, i.e. indoor and outdoor activities categorization. We theorize that assigning a different type labels to a social interaction can have important applications in automatic analysis of social patterns.

Our main concern about near future expansion of the work is regarding analysis of clothing from SSP perspective. Indeed, this is a novel area of research which can be tackled from various aspects, from clothing style analysis to its relation to personality traits. In this regard, the first-person perspective can also be useful for the analysis, since they can capture context together with the clothing and also can provide more visual details about the clothing of others due to their proximity to the subjects. This indeed is not an easy task and requires integration of fashion, sociology, psychology, and machine vision studies in one place. However, this problem once tackled, will activate many potential lines of research and applications.

Publications from this Thesis

Journals

- **Towards Social Pattern Characterization in Egocentric Photo-stream.**
  Maedeh Aghaei, Mariella Dimiccoli, Cristian Canton Ferrer, Petia Radeva.
  Journal of Computer Vision and Image Understanding (CVIU), 2018.

- **SR-Clustering: Semantic Regularized Clustering for Egocentric Photo Streams Segmentation.**
  Mariella Dimiccoli, Marc Bolanos, Estefania Talavera, **Maedeh Aghaei**, Stavri Nikolov, Petia Radeva.
  Journal of Computer Vision and Image Understanding (CVIU), 2016.

- **Multi-face Tracking by Extended Bag-of-tracklets in Egocentric Photo-streams.**
  Maedeh Aghaei, Mariella Dimiccoli, Petia Radeva.
  Journal of Computer Vision and Image Understanding (CVIU), 2016.

Conferences

- **Understanding Deep Architectures by Interpretable Visual Summaries.**
  Marco Carletti, Marco Godi, **Maedeh Aghaei**, Marco Cristani.
  The British Machine Vision Conference (BMVC), 2018. (*Submitted*)

- **All the People Around Me: Face Discovery in Egocentric Photo-streams.**

Maedeh Aghaei, Mariella Dimiccoli, Petia Radeva.
IEEE International Conference on Image Processing (ICIP), 2017.

- **Clothing and People - A Social Signal Processing Perspective.**
  Maedeh Aghaei, Federico Parezzan, Mariella Dimiccoli, Petia Radeva, Marco Cristani.
  IEEE Conference on Automatic Face and Gesture Recognition (FG), 2017.

- **With Whom Do I Interact? Detecting Social Interactions in Egocentric Photo-streams.**
  Maedeh Aghaei, Mariella Dimiccoli, Petia Radeva.
  International Conference on Pattern Recognition (ICPR), 2016.

- **Towards Social Interaction Analysis in Egocentric Photo-streams.**
  Maedeh Aghaei, Mariella Dimiccoli, Petia Radeva.
  International Conference on Machine Vision (ICMV), 2015.

- **R-Clustering for Egocentric Video Segmentation.**
  Estefania Talavera, Mariella Dimiccoli, Marc Bolanos, **Maedeh Aghaei**, Petia Radeva.
  Iberian Conference on Image Recognition and Pattern Analysis (IBPRIA), 2015.

- **Bag-of-Tracklets for Person Tracking in Life-Logging Data.**
  Maedeh Aghaei, Petia Radeva.
  Catalan Conference on Artificial Intelligence (CCIA), 2014.

Workshops

- **Social Style Characterization.**
  Maedeh Aghaei, Mariella Dimiccoli, Cristian Canton Ferrer, Petia Radeva.
  International Conference on Computer Vision (ICCV), Egocentric Perception, Interaction, and Computing Workshop (EPIC), 2017.

- **Wearable for Wearable: a Social Signal Processing Perspective for Clothing Analysis using Wearable Devices.**

Marco Godi, **Maedeh Aghaei**, Mariella Dimiccoli, Marco Cristani.
ACM International Conference on Multimedia Retrieval (ICMR), WEARMME workshop, 2017.

- **Behavior Analysis of Ants from Video Sequences.**
  Alejandro Cartas, **Maedeh Aghaei**, Christoph Gruter, Francis L. W. Ratnieks, and Constantino Carlos Reyes-Aldasoro.
  International Conference on Pattern Recognition (ICPR), Visual observation and analysis of Vertebrate And Insect Behavior Workshop (VAIB), 2016.

- **Towards Social Interaction Detection in Egocentric Photo-stream.**
  Maedeh Aghaei, Mariella Dimiccoli, Petia Radeva.
  International Conference on Computer Vision and Pattern Recognition (CVPR), First-Person Vision Workshop (FPV), 2016.

Bibliography

[1] S. Mann, K. M. Kitani, Y. J. Lee, M. Ryoo, and A. Fathi, "An introduction to the 3rd workshop on egocentric (first-person) vision," in *Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2014, pp. 827–832.

[2] M. Aghaei, M. Dimiccoli, and P. Radeva, "Multi-face tracking by extended bag-of-tracklets in egocentric photo-streams," *Computer Vision and Image Understanding*, vol. 149, pp. 146–156, 2016.

[3] ——, "All the people around me: face discovery in egocentric photo-streams," *International Conference on Image Processing*, 2017.

[4] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3570–3577.

[5] H. Gardner, "Frames of mind: The evidence to provide backing foe a theory of multiple semi-independent intelligence," 1983.

[6] J. Holt-Lunstad, T. B. Smith, M. Baker, T. Harris, and D. Stephenson, "Loneliness and social isolation as risk factors for mortality: a meta-analytic review," *Perspectives on Psychological Science*, vol. 10, no. 2, pp. 227–237, 2015.

[7] D. Umberson and J. Karas Montez, "Social relationships and health: A flashpoint for health policy," *Journal of health and social behavior*, vol. 51, no. 1_suppl, pp. S54–S66, 2010.

[8] S. Cacioppo, A. J. Grippo, S. London, L. Goossens, and J. T. Cacioppo, "Loneliness: Clinical import and interventions," *Perspectives on Psychological Science*, vol. 10, no. 2, pp. 238–249, 2015.

[9] R. J. Waldinger and M. S. Schulz, "What's love got to do with it? social functioning, perceived health, and daily happiness in married octogenarians." *Psychology and aging*, vol. 25, no. 2, p. 422, 2010.

[10] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and vision computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[11] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PloS one*, vol. 10, no. 5, p. e0123783, 2015.

[12] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[13] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[14] S. Mann, "Humanistic computing: "wearcomp" as a new framework and application for intelligent signal processing," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2123–2151, 1998.

[15] ——, "'wearcam'(the wearable camera): personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis," in *Wearable Computers, 1998. Digest of Papers. Second International Symposium on.* IEEE, 1998, pp. 124–131.

[16] A. R. Doherty, S. E. Hodges, A. C. King, A. F. Smeaton, E. Berry, C. J. Moulin, S. Lindley, P. Kelly, and C. Foster, "Wearable cameras in health: the state of the art and future possibilities," *American journal of preventive medicine*, vol. 44, no. 3, pp. 320–323, 2013.

[17] T. Kanade and M. Hebert, "First-person vision," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2442–2453, 2012.

[18] M. Bolanos, M. Dimiccoli, and P. Radeva, "Toward storytelling from visual lifelogging: An overview," *Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 77–90, 2017.

[19] S. Mann, "Sousveillance: inverse surveillance in multimedia imaging," in *ACM International Conference on Multimedia.* ACM, 2004, pp. 620–627.

[20] ——, "Sousveillance, not just surveillance, in response to terrorism," *Metal and Flesh*, vol. 6, no. 1, pp. 1–8, 2002.

[21] S. Mann, J. Nolan, and B. Wellman, "Sousveillance: Inventing and using wearable computing devices for data collection in surveillance environments." *Surveillance & society*, vol. 1, no. 3, pp. 331–355, 2002.

[22] E. A. Bradshaw, "This is what a police state looks like: sousveillance, direct action and the anti-corporate globalization movement," *Critical criminology*, vol. 21, no. 4, pp. 447–461, 2013.

[23] P. Reilly, "Every little helps? youtube, sousveillance and the anti-tescoriot in stokes croft," *New Media & Society*, vol. 17, no. 5, pp. 755–771, 2015.

[24] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *International Conference on Computer Vision.* IEEE, 2011, pp. 407–414.

[25] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Conference on Computer Vision and Pattern Recognition.* IEEE, 2012, pp. 1226–1233.

[26] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 2537–2544.

[27] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1346–1353.

[28] S. Z. Bokhari and K. M. Kitani, "Long-term activity forecasting using first-person vision," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 346–360.

[29] H. S. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," in *Advances in Neural Information Processing Systems*, 2012, pp. 422–430.

[30] I. R. Nourbakhsh, K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion, "Human-robot teaming for search and rescue," *Pervasive Computing*, vol. 4, no. 1, pp. 72–79, 2005.

[31] J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling, "Privaceye: Privacy-preserving first-person vision using image features and eye movement analysis," *arXiv preprint arXiv:1801.04457*, 2018.

[32] Y.-C. Su and K. Grauman, "Detecting engagement in egocentric video," in *European Conference on Computer Vision*. Springer, 2016, pp. 454–471.

[33] G. Rogez, J. S. Supančič, and D. Ramanan, "First-person pose recognition using egocentric workspaces," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 4325–4333.

[34] S. Mann, R. Janzen, T. Ai, S. N. Yasrebi, J. Kawwa, and M. A. Ali, "Toposculpting: Computational lightpainting and wearable computational photography for abakographic user interfaces," in *Canadian Conference on Electrical and Computer Engineering*. IEEE, 2014, pp. 1–10.

[35] V. Bettadapura, I. Essa, and C. Pantofaru, "Egocentric field-of-view localization using first-person point-of-view devices," in *Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 626–633.

[36] A. Kendon, "The f-formation system: The spatial organization of social encounters," *Man-Environment Systems*, vol. 6, pp. 291–296, 1976.

[37] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of f-formations." in *British Machine Vision Conference*, vol. 2, 2011, p. 4.

[38] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli, "Temporal encoded f-formation system for social interaction detection," in *ACM international conference on Multimedia*. ACM, 2013, pp. 937–946.

[39] M. Aghaei and P. Radeva, "Bag-of-tracklets for person tracking in life-logging data." in *Catalan Conference on Artificial Intelligence*, 2014, pp. 35–44.

[40] M. Aghaei, M. Dimiccoli, and P. Radeva, "Towards social interaction detection in egocentric photo-streams," in *Eighth International Conference on Machine Vision*. International Society for Optics and Photonics, 2015, pp. 987 514–987 519.

[41] ——, "With whom do I interact? detecting social interactions in egocentric photo-streams," in *International Conference on Pattern Recognition*. IEEE, 2016, pp. 2959–2964.

[42] M. Aghaei, M. Dimiccoli, C. C. Ferrer, and P. Radeva, "Towards social pattern characterization in egocentric photo-streams," *arXiv preprint arXiv:1709.01424*, 2017.

[43] M. Aghaei, F. Parezzan, M. Dimiccoli, P. Radeva, and M. Cristani, "Clothing and people-a social signal processing perspective," in *International Conference on Automatic Face & Gesture Recognition*. IEEE, 2017, pp. 532–537.

[44] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.

[45] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1806–1819, 2011.

[46] J. F. Henriques, R. Caseiro, and J. Batista, "Globally optimal solution to multi-object tracking with merged measurements," in *International Conference on Computer Vision*. IEEE, 2011, pp. 2470–2477.

[47] A. R. Zamir, A. Dehghan, and M. Shah, "Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs," in *European Conference on Computer Vision*. Springer, 2012, pp. 343–356.

[48] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *European Conference on Computer Vision*. Springer, 2012, pp. 864–877.

[49] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel tracking," in *International Conference on Computer Vision*. IEEE, 2011, pp. 1323–1330.

[50] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking," *International Journal of Computer Vision*, vol. 111, no. 2, pp. 213–228, 2015.

[51] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai, "Minimum error bounded efficient 1 tracker with occlusion detection," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 1257–1264.

[52] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[53] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 10, pp. 1728–1740, 2008.

[54] T. Zhou, Y. Lu, and H. Di, "Nearest neighbor field driven stochastic sampling for abrupt motion tracking," in *International Conference on Multimedia and Expo*. IEEE, 2014, pp. 1–6.

[55] E. Goffman, "The presentation of self in everyday life. 1959," *Garden City, NY*, 2002.

[56] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "Salsa: A novel dataset for multimodal group behavior analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1707–1720, 2016.

[57] G. Groh, A. Lehmann, J. Reimers, M. R. Frieß, and L. Schwarz, "Detecting social situations from interaction geometry," in *Social Computing, Second International Conference on*. IEEE, 2010, pp. 1–8.

[58] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," in *International Conference on Multimodal Interfaces*. ACM, 2011, pp. 231–238.

[59] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, "Multi-scale F-formation discovery for group detection," in *International Conference on Image Processing*. IEEE, 2013, pp. 3547–3551.

[60] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Discovering groups of people in images," in *European Conference on Computer Vision*. Springer, 2014, pp. 417–433.

[61] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara, "Understanding social relationships in egocentric vision," *Pattern Recognition*, vol. 48, no. 12, pp. 4082–4096, 2015.

[62] H. Park and J. Shi, "Social saliency prediction," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 4777–4785.

[63] R. Yonetani, K. M. Kitani, and Y. Sato, "Recognizing micro-actions and reactions from paired egocentric videos," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 2629–2638.

[64] J.-A. Yang, C.-H. Lee, S.-W. Yang, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "Wearable social camera: Egocentric video summarization for social interaction," in *International Conference on Multimedia & Expo Workshops*. IEEE, 2016, pp. 1–6.

[65] F. G. Mangrum, M. S. Fairley, and D. L. Wieder, "Informal problem solving in the technology-mediated work place," *The Journal of Business Communication (1973)*, vol. 38, no. 3, pp. 315–336, 2001.

[66] R. Muncy, "Disconnecting: Social and civic life in america since 1965," *Reviews in American History*, vol. 29, no. 1, pp. 141–149, 2001.

[67] M. Steinlin, "Knowledge management feng shui: designing knowledge sharing-friendly office space," *Knowledge Management for Development Journal*, vol. 1, no. 2, 2005.

[68] P. B. Hudson, S. M. Hudson, and R. F. Craig, "Distributing leadership for initiating university-community engagement," 2006.

[69] Y. Xiong and F. Quek, "Meeting room configuration and multiple camera calibration in meeting analysis," in *International Conference on Multimodal Interfaces*. ACM, 2005, pp. 37–44.

[70] M. Peter Valenzuela MD, "Applying creativity to health care: learning from innovative companies," *Physician executive*, vol. 38, no. 5, p. 34, 2012.

[71] H. Oh, M.-H. Chung, and G. Labianca, "Group social capital and group effectiveness: The role of informal socializing ties," *Academy of management journal*, vol. 47, no. 6, pp. 860–875, 2004.

[72] Y. J. Lee and K. Grauman, "Face discovery with social context." in *British Machine Vision Conference*, 2011, pp. 1–11.

[73] C. Zhu, F. Wen, and J. Sun, "A rank-order distance based clustering algorithm for face tagging," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 481–488.

[74] S. Xia, H. Pan, and A. K. Qin, "Face clustering in photo album," in *International Conference on Pattern Recognition*. IEEE, 2014, pp. 2844–2848.

[75] S. Xiao, M. Tan, and D. Xu, "Weighted block-sparse low rank representation for face clustering in videos," in *European Conference on Computer Vision*. Springer, 2014, pp. 123–138.

[76] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Joint face representation adaptation and clustering in videos," in *European Conference on Computer Vision*. Springer, 2016, pp. 236–251.

[77] R. G. Cinbis, J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in tv video," in *International Conference on Computer Vision*. IEEE, 2011, pp. 1559–1566.

[78] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 815–823.

[79] M. H. Aung, M. Matthews, and T. Choudhury, "Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies," *Depression and anxiety*, 2017.

[80] P. Chow, H. Xiong, K. Fua, W. Bonelli, B. A. Teachman, and L. E. Barnes, "Sad: Social anxiety and depression monitoring system for college students," 2016.

[81] S. Hodges, E. Berry, and K. Wood, "Sensecam: A wearable camera that stimulates and rehabilitates autobiographical memory," *Memory*, vol. 19, no. 7, pp. 685–696, 2011.

[82] E. Berry, N. Kapur, L. Williams, S. Hodges, P. Watson, G. Smyth, J. Srinivasan, R. Smith, B. Wilson, and K. Wood, "The use of a wearable camera, sensecam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: A preliminary report," *Neuropsychological Rehabilitation*, vol. 17, no. 4-5, pp. 582–601, 2007.

[83] E. Granholm, D. Ben-Zeev, D. Fulford, and J. Swendsen, "Ecological momentary assessment of social functioning in schizophrenia: impact of performance appraisals and affect on social interactions," *Schizophrenia research*, vol. 145, no. 1, pp. 120–124, 2013.

[84] E. Woodberry, G. Browne, S. Hodges, P. Watson, N. Kapur, and K. Woodberry, "The use of a wearable camera improves autobiographical memory in patients with alzheimer's disease," *Memory*, vol. 23, no. 3, pp. 340–349, 2015.

[85] A. Dhand, A. E. Dalton, D. A. Luke, B. F. Gage, and J.-M. Lee, "Accuracy of wearable cameras to track social interactions in stroke survivors," *Journal of Stroke and Cerebrovascular Diseases*, 2016.

[86] N. A. Brown, A. B. Blake, and R. A. Sherman, "A snapshot of the life as lived: wearable cameras in social and personality psychological science," *Social Psychological and Personality Science*, p. 1948550617703170, 2017.

[87] L. L. Carstensen, "Social and emotional patterns in adulthood: support for socioemotional selectivity theory." *Psychology and aging*, vol. 7, no. 3, p. 331, 1992.

[88] D. S. Berry and J. S. Hansen, "Positive affect, negative affect, and social interaction." *Journal of Personality and Social Psychology*, vol. 71, no. 4, p. 796, 1996.

[89] A. Vinciarelli and A. S. Pentland, "New social signals in a new interaction world: the next frontier for social signal processing," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 1, no. 2, pp. 10–17, 2015.

[90] J. H. Fortenberry, J. MacLean, P. Morris, and M. O'Connell, "Mode of dress as a perceptual cue to deference," *The Journal of Social Psychology*, vol. 104, no. 1, pp. 139–140, 1978.

[91] B. Hannover and U. Kühnen, "the clothing makes the self via knowledge activation1," *Journal of Applied Social Psychology*, vol. 32, no. 12, pp. 2513–2525, 2002.

[92] A. D. Adomaitis and K. K. Johnson, "Casual versus formal uniforms: flight attendants' self-perceptions and perceived appraisals by others," *Clothing and Textiles Research Journal*, vol. 23, no. 2, pp. 88–101, 2005.

[93] E. Stephan, N. Liberman, and Y. Trope, "The effects of time perspective and level of construal on social distance," *Journal of Experimental Social Psychology*, vol. 47, no. 2, pp. 397–402, 2011.

[94] R. Sybers and M. E. Roach, "Sociological-research-clothing and human-behavior," *Journal of Home Economics*, vol. 54, no. 3, pp. 184–187, 1962.

[95] J. Kerr, "Dress as a status symbol," *African American dress and adornment*, pp. 93–103, 1990.

[96] A. C. Murillo, I. S. Kwak, L. Bourdev, D. Kriegman, and S. Belongie, "Urban tribes: Analyzing group photos from a social perspective," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 28–35.

[97] M. L. Slepian, S. N. Ferber, J. M. Gold, and A. M. Rutchick, "The cognitive consequences of formal clothing," *Social Psychological and Personality Science*, vol. 6, no. 6, pp. 661–668, 2015.

[98] K. Hogan, "The secret language of business," *The Secret Language of Business: How to Read Anyone in 3 Seconds or Less*, pp. 1–16.

[99] A. R. Timming and D. Perrett, "Trust and mixed signals: A study of religion, tattoos and cognitive dissonance," *Personality and Individual Differences*, vol. 97, pp. 234–238, 2016.

[100] K. K. Johnson, J.-J. Yoo, M. Kim, and S. J. Lennon, "Dress and human behavior a review and critique," *Clothing and Textiles Research Journal*, vol. 26, no. 1, pp. 3–22, 2008.

[101] M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, S. G. Nikolov, and P. Radeva, "Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation," *Computer Vision and Image Understanding*, vol. 155, pp. 55–69, 2017.

[102] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2879–2886.

[103] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[104] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *International Conference on Computer Vision*. IEEE, 2013, pp. 1385–1392.

[105] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1218–1225.

[106] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *Journal on Image and Video Processing*, vol. 2008, p. 1, 2008.

[107] R. Timofte and L. Van Gool, "Sparse flow: Sparse matching for small to large displacement optical flow," in *Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 1100–1106.

[108] M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, S. G. Nikolov, and P. Radeva, "Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation," *Computer Vision and Image Understanding*, 2016.

[109] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara, "From ego to nos-vision: Detecting social relationships in first-person views," in *Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2014, pp. 580–585.

[110] E. T. Hall, "The hidden dimension," 1966.

[111] U. Hess and P. Bourgeois, "You smile–I smile: Emotion expression in social interaction," *Biological psychology*, vol. 84, no. 3, pp. 514–520, 2010.

[112] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," *ACM International Conference on Multimodal Interaction*, 2016.

[113] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 1942–1950.

[114] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *International Conference on Computer Vision*. IEEE, 2015, pp. 2407–2415.

[115] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" in *International Conference on Digital Image Computing: Techniques and Applications*. IEEE, 2016, pp. 1–6.

[116] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[117] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[118] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[119] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, 2017.

[120] E. Palispis, *Introduction to Sociology and Anthropology*. Manila: Rex Book Store, Inc, 2007.

[121] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 580–587.

[122] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[123] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[124] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *transactions on information theory*, vol. 14, no. 1, pp. 55–63, 1968.

[125] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks." *International Conference on Machine Learning*, vol. 28, pp. 1310–1318, 2013.

[126] G. Amato, F. Debole, F. Falchi, C. Gennaro, and F. Rabitti, "Large scale indexing and searching deep convolutional neural network features," in *International Conference on Big Data Analytics and Knowledge Discovery*. Springer, 2016, pp. 213–224.

[127] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, 2014, pp. 487–495.

[128] D. Wang, S. C. H. Hoi, and Y. He, "A unified learning framework for auto face annotation by mining web facial images," in *ACM International Conference on Information and Knowledge Management*. ACM, 2012, pp. 1392–1401.

[129] A. C. Holland and E. A. Kensinger, "Emotion and autobiographical memory," *Physics of life reviews*, vol. 7, no. 1, pp. 88–131, 2010.

[130] P. N. Lopes, M. A. Brackett, J. B. Nezlek, A. Schütz, I. Sellin, and P. Salovey, "Emotional intelligence and social interaction," *Personality and social psychology bulletin*, vol. 30, no. 8, pp. 1018–1034, 2004.

[131] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions." in *Advances in Neural Information Processing Systems*, vol. 2, 2012, p. 4.

[132] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[133] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Runway to realway: Visual analysis of fashion," in *Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 951–958.

[134] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving similar styles to parse clothing," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 5, pp. 1028–1040, 2015.

[135] J. Dong, Q. Chen, Z. Huang, J. Yang, and S. Yan, "Parsing based on parselets: A unified deformable mixture model for human parsing," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 88–101, 2016.

[136] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster wars: Discovering elements of fashion styles," in *European Conference on Computer Vision*. Springer, 2014, pp. 472–488.

[137] F. De Saussure, "The linguistic sign," 1985.

[138] R. Hodge and G. Kress, *Social semiotics*. Polity Press, 1988.

[139] W. T. Pauline, "The victorian era fashion history," 2006.

[140] G. B. Sproles and L. D. Burns, *Changing appearances: Understanding dress in contemporary society*. Fairchild Publications, 1994.

[141] R. B. Rubinstein, "Dress codes," *Boulder, CO: Westview*, 1995.

[142] Y. S. M. Owyong, "Clothing semiotics and the social construction of power relations," *Social Semiotics*, vol. 19, no. 2, pp. 191–211, 2009.

[143] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *European Conference on Computer Vision*. Springer, 2012, pp. 609–623.

[144] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *Transactions on Multimedia*, vol. 16, no. 1, pp. 253–265, 2014.

[145] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan, "Style finder: Fine-grained clothing style detection and retrieval," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 8–13.

[146] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan, "Towards unified human parsing and pose estimation," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 843–850.

[147] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan, "Matching-cnn meets knn: Quasi-parametric human parsing," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 1419–1427.

[148] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 12, pp. 2402–2414, 2015.

[149] P. Kohli, P. H. Torr *et al.*, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[150] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 1385–1392.

[151] C. Frith, "Role of facial expressions in social interactions," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1535, pp. 3453–3458, 2009.

[152] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *International Conference on Computer Vision*. IEEE, 2015, pp. 3343–3351.

[153] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis," in *ACM International Conference on Multimedia*, 2013, pp. 213–222.

[154] G. V. Caprara, C. Barbaranelli, L. Borgogni, and M. Perugini, "The big five questionnaire: A new questionnaire to assess the five factor model," *Personality and individual Differences*, vol. 15, no. 3, pp. 281–288, 1993.

[155] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[156] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 5188–5196.

[157] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[158] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. R. Mller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–14, 2016.

[159] A. Heart, "Apparel and its impact on self-perception," 2011.

[160] M. Barnard, *Fashion as communication*. Psychology Press, 2002.

[161] A. Todorov, M. Pakrashi, and N. N. Oosterhof, "Evaluating faces on trustworthiness after minimal time exposure," *Social Cognition*, vol. 27, no. 6, pp. 813–833, 2009.

[162] J. Willis and A. Todorov, "First impressions making up your mind after a 100-ms exposure to a face," *Psychological science*, vol. 17, no. 7, pp. 592–598, 2006.

[163] N. Howlett, K. Pine, I. Orakcioglu, and B. Fletcher, "The influence of clothing on first impressions: Rapid and positive responses to minor changes in male attire," *Journal of Fashion Marketing and Management: An International Journal*, vol. 17, no. 1, pp. 38–48, 2013.

[164] F. Setti, D. Conigliaro, M. Tobanelli, and M. Cristani, "Count on me: learning to count on a single image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[165] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[166] J. E. Nash, "Decoding the runner's wardrobe," *Conformity and Conflict, eds. James P. Spradley and David W. McCurdy, Boston: Little, Brown*, vol. 172, p. 185, 1977.

[167] J. Shao, C. Change Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 2219–2226.

[168] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," *arXiv preprint arXiv:1704.03296*, 2017.

[169] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2015.