Managing heterogeneous cues in social contexts : A holistic approach for social interactions analysis

Mahmoud Qodseya

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *05/10/2020* **par :**

Mahmoud QODSEYA

Managing Heterogeneous Cues in Social Contexts - A Holistic Approach for Social Interactions Analysis

JURY

Oscar PASTOR	Université polytechnique de Valence	Rapporteur
Christophe NICOLLE	Université de Bourgogne	Rapporteur
Jean-Pierre JESSEL	Université de Toulouse	Examinateur
Donatello CONTE	Université de Tours	Examinateur
Guillaume CABANAC	Université de Toulouse	Invité
Pierre DUFRESNE	Toulouse Tech Transfer	Invité
Florence SÈDES	Université de Toulouse	Directrice de Thèse
Jean-Pierre POULAIN	Université de Toulouse	Co-directeur de thèse

École doctorale et spécialité :
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse, Spécialité : Informatique et Télécommunications

Unité de Recherche :
IRIT : Institut de Recherche en Informatique de Toulouse (UMR 5505)

Directeur(s) de Thèse :
Florence SÈDES et *Jean-Pierre POULAIN*

Rapporteurs :
Oscar PASTOR et *Christophe NICOLLE*

Managing Heterogeneous Cues in Social Contexts - A Holistic Approach for Social Interactions Analysis

Mahmoud QODSEYA

octobre 2020

REMERCIEMENT

Je tiens à remercier tout particulièrement et à témoigner toute ma reconnaissance à ma directrice de thèse **Pr. Florence Sèdes** et mon codirecteur **Pr. Jean-Pierre Poulain**, pour l'expérience enrichissante, du fait de son interdisciplinarité, et pleine d'intérêt qu'ils m'ont permis de vivre durant ma thèse, pour leur implication, leurs précieux encadrements et surtout leur soutien, jusqu'à la concrétisation de mon contrat sur la plateforme OVALIE. L'humain est au cœur de cette thèse en Informatique, à l'image de leur humanité...

Pr. Jean-Pierre Jessel, aussi humain et l'esprit affûté, curieux et entrepreneur, s'est dévoué pour présider ce jury : merci à vous, M. le (Vice-)Président de notre Université.

Ma gratitude aux membres du jury pour avoir accepté d'évaluer ce travail :
Pr. Christophe Nicolle, d'après ce que m'a expliqué Pr. Sèdes, est très intéressé par l'objet de notre étude : l'interaction sociale au restaurant ☺. A-t-il enfin eu la réponse à sa question « qu'est-ce que l'intelligence artificielle va penser de moi ? » ?...

Pr. Oscar Pastor m'a convaincu, moi qui viens de la robotique et du monde de la vision, que « Using Conceptual Model Technologies for Understanding the Human ... » était un vrai verrou scientifique. L'analyse des interactions sociales ibériques nous tient à cœur pour de futures perspectives de collaboration.

M. Guillaume Cabanac, éminent « scientomètre », s'intéresse plutôt aux interactions scientifiques, mais après tout n'est-ce pas une autre forme de socialisation, dans ce monde (post-)(pré ?)confiné ? M. Cabanac, please, « T'AS PENSÉ À RETWEETER MON ARTICLE ? » ☺.

M. Donatello Conte, vous avez aiguisé mon appétit vec ☺ « Video Processing for Human Behavioral Analysis" Track@SAC 2020 et même si notre proposition fut refusée, le fait que de tels calls émergent prouve la pertinence en devenir de notre objet de recherche.

Je saisis cette occasion pour exprimer ma reconnaissance envers mes amis et collègues Mahdi Washha, Joël Courant, Franck Panta et Wafa Abdelghani qui m'ont apporté leur soutien moral et intellectuel tout au long de ma thèse, ainsi que l'ensemble des membres de l'IRIT qui m'ont consacré leur temps pour les enregistrements qui ont valorisé ma recherche dans notre mini-lab bricolé.

Un message à l'attention de **l'Université Fédérale Toulouse Midi-Pyrénées** et à la **Région Occitanie** sans qui rien n'aurait démarré: le financement de cette thèse dans la cadre du programme APR (Appel à Projets de Recherche) a matérialisé la collaboration entre les équipes de l'IRIT et du CERTOP, avec le soutien de l'ISTHIA et de Taylor's University (KL, Malaysia). Un grand pas pour l'interdisciplinarité et le partage des cultures...

J'adresse également mes remerciements à TTT (Toulouse Tech Transfer) pour avoir cru en notre projet SOIF (SOcial Interactions analysis Framework) dans le cadre du programme Proto'Pitch Challenge. **M. Pierre Dufresne**, son Président (CEO), est présent aujourd'hui et cette marque d'intérêt prouve le lien fort et nécessaire à la recherche que sont le transfert et la valorisation.

J 'adresse mes remerciements également à ma mère, à qui je dois la réussite, pour l'éducation qu'elle m'a prodiguée ; avec tous les moyens et au prix de tous les sacrifices qu'elle a consentis à mon égard, pour le sens du devoir qu'elle m'a enseigné depuis mon enfance, mon père, mes frères, et mes sœurs pour leurs encouragements. Je désire aussi remercier ma femme, Mounia, pour son soutien inconditionnel qui m'a donné la force et la patience d'accomplir ce travail. Enfin, je tiens à remercier ma fille Cataleya, qui est venue au monde il y a deux mois, de m'avoir donné un bonheur et un plaisir illimité.

Mahmoud Qodseya

The purpose of abstraction is not to be vague, but to create a new semantic level in which one can be absolutely precise.

— Edsger Dijkstra

ABSTRACT

Social interaction refers to any interaction between two or more individuals, in which information sharing is carried out without any mediating technology. This interaction is a significant part of individual socialization and experience gaining throughout one's lifetime. It is interesting for different disciplines (sociology, psychology, medicine, etc.). In the context of testing and observational studies, multiple mechanisms are used to study these interactions such as questionnaires, direct observation and analysis of events by human operators, or a posteriori observation and analysis of recorded events by specialists (psychologists, sociologists, doctors, etc.). However, such mechanisms are expensive in terms of processing time. They require a high level of attention to analyzing several cues simultaneously. They are dependent on the operator (subjectivity of the analysis) and can only target one side of the interaction. In order to face the aforementioned issues, the need to automatize the social interaction analysis process is highlighted. So, it is a question of bridging the gap between human-based and machine-based social interaction analysis processes.
Therefore, we propose a holistic approach that integrates multimodal heterogeneous cues and contextual information (complementary "exogenous" data) dynamically and optionally according to their availability or not. Such an approach allows the analysis of multi "signals" in parallel (where humans are able only to focus on one). This analysis can be further enriched from data related to the context of the scene (location, date, type of music, event description, etc.) or related to individuals (name, age, gender, data extracted from their social networks, etc.). The contextual information enriches the modeling of extracted metadata and gives them a more "semantic" dimension. Managing this heterogeneity is an essential step for implementing a holistic approach.
The automation of « in vivo » capturing and observation using non-intrusive devices without predefined scenarios introduces various issues that are related to data (i) privacy and security; (ii) heterogeneity; and (iii) volume. Hence, within the holistic approach we propose (1) a privacy-preserving comprehensive data model that grants decoupling between metadata extraction and social interaction analysis methods; (2) geometric non-intrusive eye contact detection method; and (3) French food classification deep model to extract information from the video content. The proposed approach manages heterogeneous cues coming from different modalities as multi-layer sources (visual signals, voice signals, contextual information) at different time scales and different combinations between layers (representation of the cues like time series). The approach has been designed to

operate without intrusive devices, in order to ensure the capture of real behaviors and achieve the naturalistic observation. We have deployed the proposed approach on OVALIE platform which aims to study eating behaviors in different real-life contexts and it is located in University Toulouse-Jean Jaurès, France.
Keywords: Observational studies, Social interaction analysis, Heterogeneous Social cues, Eating behavior analysis

RÉSUMÉ

Une interaction sociale désigne toute action réciproque entre deux ou plusieurs individus, au cours de laquelle des informations sont partagées sans « médiation technologique ». Cette interaction, importante dans la socialisation de l'individu et les compétences qu'il acquiert au cours de sa vie, constitue un objet d'étude pour différentes disciplines (sociologie, psychologie, médecine, etc.). Dans le contexte de tests et d'études observationnelles, de multiples mécanismes sont utilisés pour étudier ces interactions tels que les questionnaires, l'observation directe des événements et leur analyse par des opérateurs humains, ou l'observation et l'analyse à posteriori des événements enregistrés par des spécialistes (psychologues, sociologues, médecins, etc.). Cependant, de tels mécanismes sont coûteux en termes de temps de traitement, ils nécessitent un niveau élevé d'attention pour analyser simultanément plusieurs descripteurs, ils sont dépendants de l'opérateur (subjectivité de l'analyse) et ne peuvent viser qu'une facette de l'interaction. Pour faire face aux problèmes susmentionnés, il peut donc s'avérer utile d'automatiser le processus d'analyse de l'interaction sociale. Il s'agit donc de combler le fossé entre les processus d'analyse des interactions sociales basés sur l'homme et ceux basés sur la machine.

Nous proposons donc une approche holistique qui intègre des signaux hétérogènes multimodaux et des informations contextuelles (données "exogènes" complémentaires) de manière dynamique et optionnelle en fonction de leur disponibilité ou non. Une telle approche permet l'analyse de plusieurs "signaux" en parallèle (où les humains ne peuvent se concentrer que sur un seul). Cette analyse peut être encore enrichie à partir de données liées au contexte de la scène (lieu, date, type de musique, description de l'événement, etc.) ou liées aux individus (nom, âge, sexe, données extraites de leurs réseaux sociaux, etc.) Les informations contextuelles enrichissent la modélisation des métadonnées extraites et leur donnent une dimension plus "sémantique". La gestion de cette hétérogénéité est une étape essentielle pour la mise en œuvre d'une approche holistique.

L'automatisation de la capture et de l'observation « in vivo » sans scénarios prédéfinis lève des verrous liés à i) la protection de la vie privée et à la sécurité ; ii) l'hétérogénéité des données ; et iii) leur volume. Par conséquent, dans le cadre de l'approche holistique, nous proposons (1) un modèle de données complet préservant la vie privée qui garantit le découplage entre les méthodes d'extraction des métadonnées et d'analyse des interactions sociales ; (2) une méthode géométrique non intrusive de détection par contact visuel ; et (3) un modèle profond de classification des repas français pour extraire les informations du contenu vidéo. L'approche proposée gère des signaux hétérogènes provenant de différentes modalités en tant que sources multicouches (signaux visuels,

signaux vocaux, informations contextuelles) à différentes échelles de temps et différentes combinaisons entre les couches (représentation des signaux sous forme de séries temporelles). L'approche a été conçue pour fonctionner sans dispositifs intrusifs, afin d'assurer la capture de comportements réels et de réaliser l'observation naturaliste. Nous avons déployé l'approche proposée sur la plateforme OVALIE qui vise à étudier les comportements alimentaires dans différents contextes de la vie réelle et qui est située à l'Université Toulouse-Jean Jaurès, en France.
Mots clés : Études observationnelles, Analyse des interactions sociales, Descripteurs sociaux hétérogènes, Analyse du comportement alimentaire

PUBLICATIONS

Our ideas and contributions have already been published in the following scientific publications:

International conference papers

1. Mahmoud Qodseya, Franck Jeveme Panta, and Florence Sèdes. Visual-based eye contact detection in multi-person interactions. In *2019 International Conference on Content-Based Multimedia Indexing, CBMI 2019, Dublin, Ireland, September 4-6, 2019*, pages 1–6, 2019

2. Mahmoud Qodseya. Visual non-verbal social cues data modeling. In *Advances in Conceptual Modeling - ER 2018 Workshops Emp-ER, MoBiD, MREBA, QMMQ, SCME, Xi'an, China, October 22-25, 2018, Proceedings*, pages 82–87, 2018

3. Mahmoud Qodseya, Mahdi Washha, and Florence Sèdes. Dievent: Towards an automated framework for analyzing dining events. In *34th IEEE International Conference on Data Engineering Workshops, ICDE Workshops 2018, Paris, France, April 16-20, 2018*, pages 163–168, 2018

4. Franck Jeveme Panta, Mahmoud Qodseya, André Péninou, and Florence Sèdes. Management of mobile objects location for video content filtering. In *Proceedings of the 16th International Conference on Advances in Mobile Computing and Multimedia, MoMM 2018, Yogyakarta, Indonesia, November 19-21, 2018*, pages 44–52, 2018

5. Franck Jeveme Panta, Mahmoud Qodseya, Geoffrey Roman-Jimenez, André Péninou, and Florence Sèdes. Spatio-temporal metadata querying for cctv video retrieval: Application in forensic. In *Proceedings of the 9th ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pages 7–14. ACM, 2018

SUMMARY

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

GDPR General Data Protection Regulation

IoT Internet of Things

ML Machine Learning

PWS Prader–Willi syndrome

IP Internet Protocol

WDR Wide Dynamic Range

HDTV High-definition television

RF Random Forest

SVM Support Vector Machine

TP True Positives

FP False Positives

TN True Negatives

FN False Negatives

TPR True Positive Rate

FPR False Positive Rate

ANN Artificial Neural Network

RNN Recurrent Neural Network

CNN Convolutional Neural Network

DCNN Deep Convolutional Neural Network

ResNet Residual Neural Network

DenseNet Densely Connected Convolutional Network

CCTV Closed-Circuit TeleVision

VNAM Analysis Method based on Visual Nonverbal cues

FER Facial expression Recognition

GPS Global Positioning System

IMU Inertial Measurement Unit

SER Speech Emotion Recognition

LPCs Linear Prediction Coefficients

MFCCs Mel-Frequency Cepstrum Coefficients

NAS Network-attached storage

PoE Power over Ethernet

1 INTRODUCTION

Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human.

— Aristotle, Politics ca. 328 BC

1.1 Face-to-Face interaction

In sociology, face-to-face interaction is a concept describing social interaction carried out without any mediating technology [32]. Georg Simmel, one of the earliest social science scholars to analyze this type of interaction, observed that sensory organs play an essential role in interaction, discussing examples of human behavior such as eye contact [121].
Social interaction is a dynamic relationship of social cues/signals exchange between two or more individuals within a group. It has a vital role to play in the evolution of learners, which is not a straightforward generalization of complex environments[123]. Authors of [95] found a statistical correlation between the amount of social interaction and individual mental health. Thus, the study of social interaction provides a better understanding of human behavior in different contexts and scenarios.
Social interaction analysis can be useful in many domains like industry (e.g., restaurants) in which stakeholders can get feedback illustrating the satisfaction level of the clients regarding the provided service; medical services by which health issues such as eating disorders can be detected in a person's eating behavior; Internet of Things (IoT) applications where we can provide a fast and reliable way to measure user's experience with new devices testing; and observational studies in which individual's behaviors are systematically observed and recorded in order to describe the relationship the observed behaviors with a variable or set of variables.

1.2 Social cues

Social cues can be categorized into verbal (word) and nonverbal (wordless/visual) information [7], as shown in Figure 1.1. The verbal behavioral cues take into account the spoken information among persons, such as 'yes/no' responses in answering question context. Nonverbal behavioral cues represent a set of temporal changes in neuromuscular and physiological activities, which send a message about emotions, mental state, and other characteristics [135]. Nonverbal

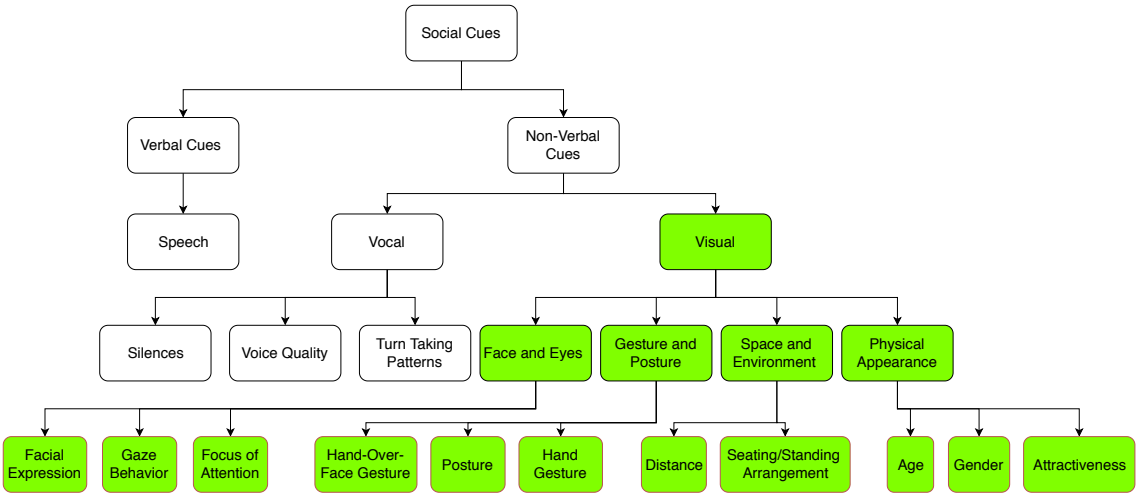


Figure 1.1: Taxonomy of social cues [7].

cues are accessible to our senses by sight and hearing, as shown in Figure 1.2. This fact means they are detectable through microphones, cameras, or other suitable sensors (e.g., accelerometer). Nonverbal cues can be taxonomized into vocal and visual cues, where: (i) vocal cues include voice quality, silences, turn-taking patterns, nonlinguistic vocalizations, and linguistic vocalizations; and (ii) visual cues include physical appearance (e.g., gender, height, ethnicity, age), face and eyes cues (e.g., facial expression, gaze direction, focus of attention), gesture and posture, and space and environment [136].

1.3 Social interaction analysis

In the context of testing and observational studies, several mechanisms are used to analyze social interactions from various perspectives. For example, if a teaching institute want to evaluate a new teaching technique (focus group, team working, etc.) in a class, one of the following method may be used: (i) directly observing students' behavior and recording notes during the class; (ii) asking students to fill pre-defined form (questionnaire), then analyze it; or (iii) recording the lecture and perform video analysis by human operators (observer). However, such mechanisms are expensive in terms of processing time, requiring a high level of concentration and attention to analyze several cues in parallel and dependent on the observer' personal beliefs or feelings (subjectivity of the analysis).

1.4 Towards «in vivo» automatic social interaction analysis

To face the observers' subjectivity and their limited ability to track multiple cues in the observational studies, the need for automatization of the analysis procedure

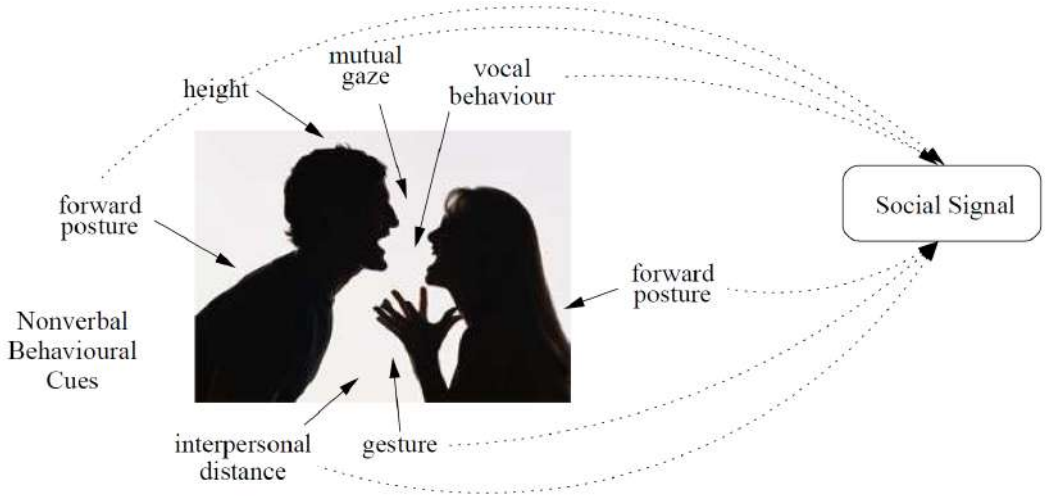


Figure 1.2: A group of nonverbal behavioral cues is recognized as a social signal [136].

is highlighted. So it is a question of bridging the gap between the human-based and machine-based social interaction analysis processes. Such an approach will eliminate the observers' subjectivity and will allow the analysis of multi "signals" in parallel (where humans are able only to focus on one). This analysis can be further enriched by data related to the context of the scene (location, date, type of music, event description, etc.), individuals personal data (name, age, gender, education levels), or data extracted from individuals' social networks (subscribers, friends, groups, shared contents, etc.).
Various studies have been performed to detect, analyze, and assess social interactions using automatic machine-learning methods, including automatic extraction of nonverbal social signals corresponding to multimodal (e.g., eye contact, touching, etc.) social cues [7]. These studies have been applied in a tremendous range of applications and domains, including role recognition [40], social interaction detection in a smart meeting [102], and work environments [69], detecting deceptive behavior [4], detecting dominant people in conversations [63], and studying parent-infant interaction [12].
Social signals that have been investigated during social interactions are primitive and context-independent because they are not semantic and often occur unconsciously. These signals include frequency and duration of social behavioral cues occurrences, such as the number of eye contact actions that happened between two persons. Different from the previous studies, our proposal take the"context" into account as an independent variable. We focus on visual social cues (highlighted in green color in Figure 1.1) and precisely the following:

- Facial expression since it is one of the most effective, natural, and universal signals for human beings to convey their emotional states [35].
- Eye and gaze behavior, as it is an important social cue for performing a wide range of analysis and studies such as a dominant person detection [51]. It provides multiple functions in the two-person contacts such as information

seeking, establishment and recognition of social relationships, and signaling that the "channel is open for communication" [10].

1.5 Social interaction analysis automization: the challenges

To face the observer's subjectivity and limited observation ability, we need to automate social interaction analysis process. This automation introduce three main challenges that must be considered in the proposed approach:

- the gap that appears from the transforming human-based social interaction analysis observation tasks into machine-based computational tasks;
- keeping the observation naturalistic, which means anything can affect the data subject's behavior is not allowed, like using intrusive device (head-mounted camera), having a camera in front of data subject, attaching microphone to the data subject;
- the different time scales that social cues can be changed within. For example, hand-over-face gesture can change one time every twenty seconds, but the gaze direction can be changed up to one hundred times per second.

In addition to the three main challenges, there are multiple technical challenges related to multiple technical challenges related to:

- Data heterogeneity, multiple social cues in face-to-face interaction need to be tracked. These cues are related to head, hands, face, body, and voice, which means each cue has a different representation, and this causes heterogeneity in data. For example, facial expression can be represented as discrete values, whereas the gaze direction is represented as a vector.
- Data volume, the minimum time scale that can be used in the analysis of observational studies records is the frame-level. Thus, the amount of the extracted data (social cues) proportionally increases with respect to the frame rate and video length and number of tracked cues. For example, performing analysis on the SALSA [9] dataset needs to extract social signals from 216,000 $(60 \times 60 \times 15 \times 4)$ frames belonging to four videos with a length of one hour and frame rate equals to fifteen frames per second, related to only one experiment. Therefore, with this example, analyzing hours of experiments will produce a massive data.
- Personal data privacy and security, with the progress in the computer vision field is raising concerns about individuals' privacy since visual information can be misused to profile/track them against their will. From another perspective, the General Data Protection Regulation (GDPR), which is a regulation in EU law on data protection and privacy for all individuals within the European Union [137], has to be considered.

1.6 Contributions

The main contribution of this thesis consist of the automation of social interactions analysis in the context of observational studies. In particular, we propose a generic architecture for a holistic approach for social interactions analysis. The proposed approach integrates different components: (i) data acquisition methods (cameras, microphones, etc.); (ii) context-aware feature extraction methods; (iii) (meta)data model; and (iv) social behavior analysis methods.
In (meta)data model component, we propose a comprehensive (meta)data model within which is the heart of the proposed approach due to many reasons: (1) it decouples feature extraction methods from the analysis methods; (2) it facilitates heterogeneous data fusion from different modalities; and (3) it encapsulates the recorded video, so we will (if it is needed) share (meta)data and analysis results only instead of sharing the recorded videos.
In the data acquisition methods, we propose a novel geometric-based method to detect eye contact in natural multi-person interactions without the need for eye tracking devices or any intrusive in order to do naturalistic observation studies. Furthermore, we propose a find-tuned deep model for food classification that will be used to extract contextual information from the video content.
In the social behavior analysis methods, we analyze the heterogeneous social cues coming from different modalities as multi-layer sources (visual cues, voice cues, contextual information (complementary "exogenous" data)) at different time scales and different combinations between layers. In our context, the layer is the representation of the cue as time series. So, we represent the cues as time series with a common time scale (sampling time such as millisecond, second, minute, etc.). This enables the aggregation between the heterogeneous social cues and contextual information dynamically at different time scales, and optionally according to their availability or not.
Finally, we have deployed the proposed holistic approach on the OVALIE [1] platform which is located in University Toulouse II-Jean Jaurès (UT2J - France) and it aims to study eating behaviors in different contexts .

1.7 Thesis organization

The rest of the thesis is structured into 5 chapters.
Chapter 2 gives an introduction to basic concepts related to social cues, machine learning, transfer learning, and Prader–Willi syndrome (PWS).

Chapter 3 reviews related work of common social interaction detection methods with a focus on the visual nonverbal cues based approach. Beside that, we reviews the existed methods of food classification based on the deep learning,

1 https://certop.cnrs.fr/plateforme-experimentale-ovalie-shs-alimentation/

since the food is commonly exist in several contexts (restaurant, hospital, etc.) and it consumption has an effect on the social interaction.

Chapter 4 Presents a novel geometrical method to detect eye contact in natural multi-person interactions without the need of any intrusive eye tracking device. We have experimented our method on 10 social videos, each 20 minutes long. Experiments demonstrate highly competitive efficiency with regards to classification performance, compared to the classical existing supervised eye contact detection methods. Also, this chapter evaluates the effectiveness of deep convolutional neural network (DCNN) in classifying French food images task.

Chapter 5 presents our holistic approach for social interaction analysis architecture that combines various methods together using a comprehensive (meta)data model that able to store heterogeneous (meta)data.
Chapter 6 presents OVALIE platform floor plan, hardware, software. After presenting the platform, we introduce the dataset that collected to perform our experiments. Finally, we introduce a PWS observation study which will be performed in OVALIE platform.

Chapter 7 concludes this thesis, and discusses findings and perspectives.

2

BACKGROUND: OBSERVATIONAL STUDIES, SOCIAL CUES, AND MACHINE LEARNING

A good stance and posture reflect a proper state of mind.

— Morihei Ueshiba

In this chapter, we introduce the basic concepts and definitions that we are going to use in the thesis. First, we introduce the observational study in social sciences, then we present «in vivo» experimental platform for eating behavior analysis (OVALIE), which is in the context of observational studies. Second, we introduce the main social signals that we are interested in within this thesis. After that, we provide a general overview of machine learning and transfer learning since we developed machine learning-based methods in this thesis to detect social cues. Finally, we present Prader–Willi syndrome (PWS) because we are going to study the eating behavior of children born with it (work in progress).

2.1 Observational study

In social sciences, observational study is a non-experimental social research method in which a researcher records and observes ongoing behavior in a natural setting. It aims to draw inferences from a sample to a population where the independent variables are not under the control of the researcher. The collected data in observational research studies are often qualitative. Based on the involvement of the observer, we can classify the observational methods into participant observation and non-participant observation.
Observational studies have many advantages: it is one of the main bases of formulating hypotheses; it has higher accuracy compared with other methods like interviews, questionnaires, etc. However, it is a time-consuming process, and it could affected by the observer subjectivity (Personal Bias of the Observer).

2.2 OVALIE platform

In the context of observational studies, OVALIE [3] is an experimental platform in human and social sciences, is located in University Toulouse-Jean Jaurès (UT2J - France) , aims to observe, analyse and study the influence of physical and social context on eating behaviors. In addition to the use of behavioral research equipment and software that can perform a wide variety of tasks, including facial expression analysis, audio analysis, spatial behavior tracking, and eye tracking.

This platform will take the "context" into account as an independent variable, differing from previous studies which neutralized the context. Additionally, through a partnership with Taylor's University (TU - Malaysia), an identical platform will be established, facilitating cross-cultural studies in eating behaviors.

2.3 Social cues

Social cues are critically an essential aspect of communication. In the following, we identify social cues and explore some examples of how they influence social interaction and engagement. More technical details related to the automatic detection of thesis social cues will be presented in Section 3.2. Imagine that you are talking to a friend at a party and suddenly he turns and walks away. Would you follow him and keep talking or would you simply go talk to someone else? Likely you would go find someone else because you recognize that walking away from you in the middle of a conversation is an indication that your friend is not interested in what you were talking about. You understood the fact that your friend was not interested in your talk because he gave you a social cue.

Figure 2.1: A smile gives an indication that the person is pleased or amused [126].

A typical example of a social cue is a **smile** (see Figure 2.1). Although smiles are not always genuine, people generally smile (consciously and unconsciously) when they are happy or amused. For instance, if you are in a group and tell a joke that makes people smile, then you can assume that the joke go well because people have provided a social cue (smile). Conversely, if your joke does not elicit a smile, then you can assume that the joke did not went well since people did not provide a smile (lack of smile is a social cue indicating that the group did not think what you said was funny). In both cases, social cues give you an indication of what to do next. The smile indicates that they like your humor and a lack of smile suggests that you should try a different style.
Having a well-developed understanding of social cues and the strong ability to interpret them can greatly increase your skills as a communicator because you will be able to read a person's behavior and appropriately respond. On the other

hand, under-developed awareness of social cues makes it challenging to create relationships with others and interact in social situations.
Social cues are symbols expressed through facial expressions, body posture, gestures, eye movement, pitch and tone of voice, or words that are intended to send a message from one person to another.

2.3.1 *Facial expressions*

Facial expressions are social signals that we make by moving our facial muscles. Facial expressions generally signify an emotional state. Authors of [43] defined six basic emotions based on cross-culture study, each emotional state of mind has a specific facial expression[42]. These facial expressions are anger, disgust, fear, happiness, sadness, and surprise as shown in Figure 2.2.

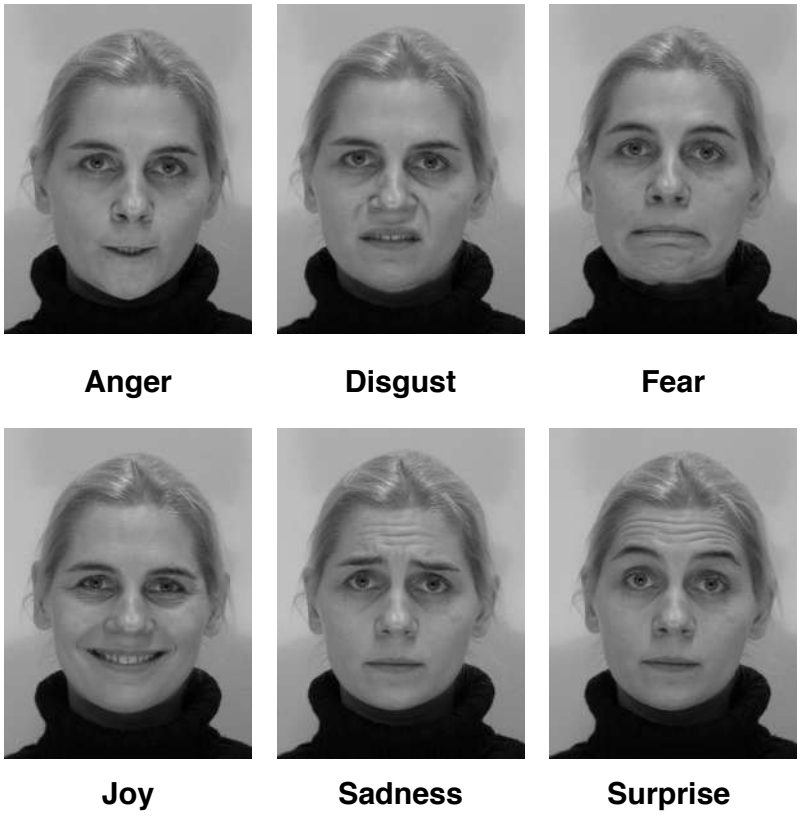


Figure 2.2: Samples of basic six emotions displayed by facial expression from MMI dataset[133].

2.3.2 *Body posture*

Body posture is the position in which someone holds their body while standing, sitting, or lying down. Posture can reveal significant information, such as a person's current state of mind, emotions, and attitudes [30]. Figure 2.3 shows

slumped and erect postures. Slumped posture is an example of a depressed attitude, whereas erect posture is an example of a more energized attitude.

Figure 2.3: Two examples of body postures. On the left is a slumped posture, on the right is a erect posture

2.3.3 *Gestures*

People often use gestures during their communication. Gestures are hands, face, or another part of the body movement to send a message in place of the speech or in parallel with it [68].

Figure 2.4: On the left "OK" and "cross figures" gestures, on the right examples of hand-over-face gestures taken from [80].

They may be conscious like the sign language or unconscious like the hand-over-face gestures; Figure 2.4 shows some examples of gestures. Gestures are culture-specific and may carry different meanings in different cultures. For example, the "OK" gesture in the USA signifies OKay; in Japan, it symbolizes money; in Brazile, is a rude gesture.

2.3.4 *Eye contact*

Eye contact detection is defined as a task of detecting whether two people look at each other's eyes or face simultaneously as shown in Figure 2.5. It is

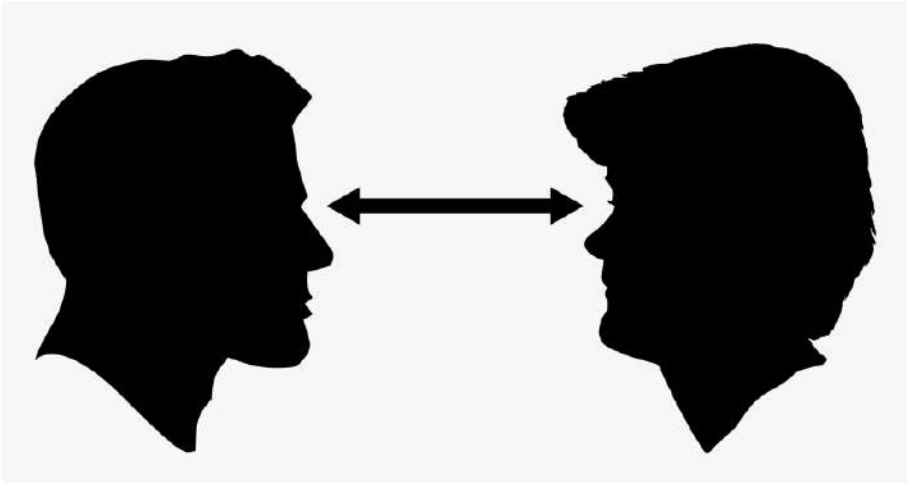


Figure 2.5: Eye contact clipart.

an important feature for better understanding human social behavior. It has numerous applications. For example, it is a key component in attentive user interfaces and it is used to analyze turn-taking, social roles, and engagement during multi-person interactions. Even more, we can deduce many things based on the eye contact [10]: (i) the topic nature, in which, there is more eye contact in case of the topic being discussed is straightforward and less personal, whereas, there is less eye contact during the hesitating passages; (ii) the relation between two persons, in which, there is more eye contact if the two persons are positively interested in each other.

2.3.5 *Pitch and tone of voice*

When communicating with others, the pitch and tone of your voice (intonation) can be a good indicator of how do you feel at that moment. For example, if someone was speaking very quickly in a shaky tone. His rapid speech is a sign of anxiety and urgency, which means you should assume something is wrong and requires an immediate attention.

2.4 Machine learning

In this section, we will briefly introduce the machine learning field. Based on it, we will be able to detect the introduced social cues in Section 2.3.
Machine learning is a field in computer science that shows the abilities of machines in learning to solve problems from given experimental data instead of explicitly programmed. The behavior of most machine learning algorithms is controlled by a set of parameters that define a model. The main purpose of machine learning is to estimate the parameters of the model to learn regular patterns from data observations, with avoiding learning the training samples "by heart". In practice, given a dataset of training examples, an algorithm is expected to learn a model to solve a specific task. **Learning from Examples** is one of the most commonly adopted learning strategies as well as it provides the most flexibility with enabling computer programs to completely develop unknown skills or find unknown structures and patterns in a given data [21]. Learning from examples is a technique that is often leveraged in classification tasks to predict the class label of new, properly unseen, data entries based on a dynamic set of known examples.

2.4.1 *Machine learning styles*

There are so many machine algorithms available that follow the learning from examples strategy. However, these algorithms can be categorized based on learning style (supervised, unsupervised, and semi-supervised) [67] as shown in Figure 2.6. **Supervised Learning** includes every task in which the algorithm has access

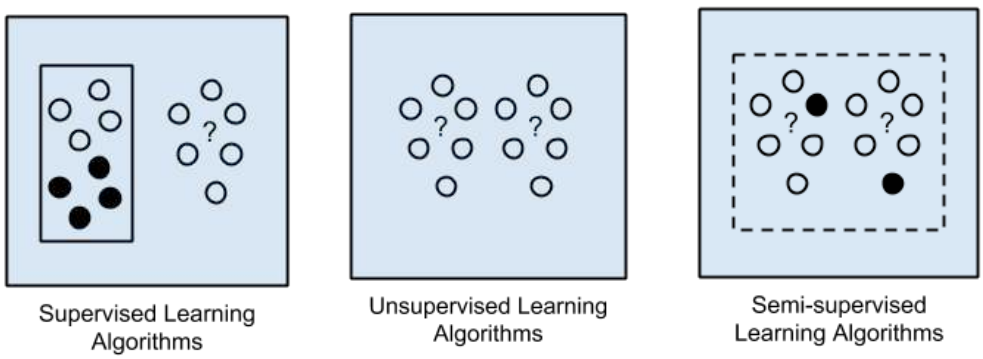


Figure 2.6: Three common learning styles adopted in machine learning field [17].

to input and output values. Herein, input values can be defined as the external information that the algorithm can use, such as attribute values, while output values, are the specific target labels of the class attribute. It means that the structure of the data is known and the purpose of these programs is to predict the correct class for a given new data. **Unsupervised Learning** is used for the tasks that have no access to output values and thus try to find structures within the data through

creating classes on their own. **Semi-Supervised Learning** is adopted for solving tasks that have a mixture input data between labeled and unlabeled examples. Indeed, those tasks can be viewed as a prediction problem, but the model has to learn the structures to organize the data.

2.4.2 *Clustering*

Clustering is one of the common data analysis techniques used to get knowledge about the structure of the data. It is the task of identifying subgroups (clusters) in the data such that data points in the same cluster are very similar based on similarity measures such as euclidean-based distance or correlation-based distance. Clustering is an unsupervised learning method since it does not need the ground truth. **K-means** is one of the most common used clustering algorithms due to its simplicity. We used the K-means algorithm in person tracking in our proposed geometrical eye contact detection algorithm in Section 4.1.1.
K-means algorithm is an iterative algorithm aims to partition the data into K clusters where each data point belongs to only one cluster. It works as following:

1. determine number of clusters K,
2. initialize centroids by shuffling the input data and then randomly selecting K data points for the centroids,
3. keep iterating until there is no change to the centroids:
    - compute the sum of the squared distance between data points and all centroids,
    - assign each data point to the closest cluster (centroid),
    - compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

2.4.3 *Classification*

Various supervised learning algorithms have been designed and well-implemented to build data-driven models using a given training set at hand. In this work, we focus on commonly classification learning methods, including decision tree, random forests, artificial neural networks, and conventional neural networks and we used these algorithms to make comparison with our proposed geometrical eye contact detection algorithm in Section 4.1.1.

2.4.3.1 *Decision trees and forests*

Decision tree-based learning algorithms are one of the most common methods used for non-linear classification problems. Their popularity increased because

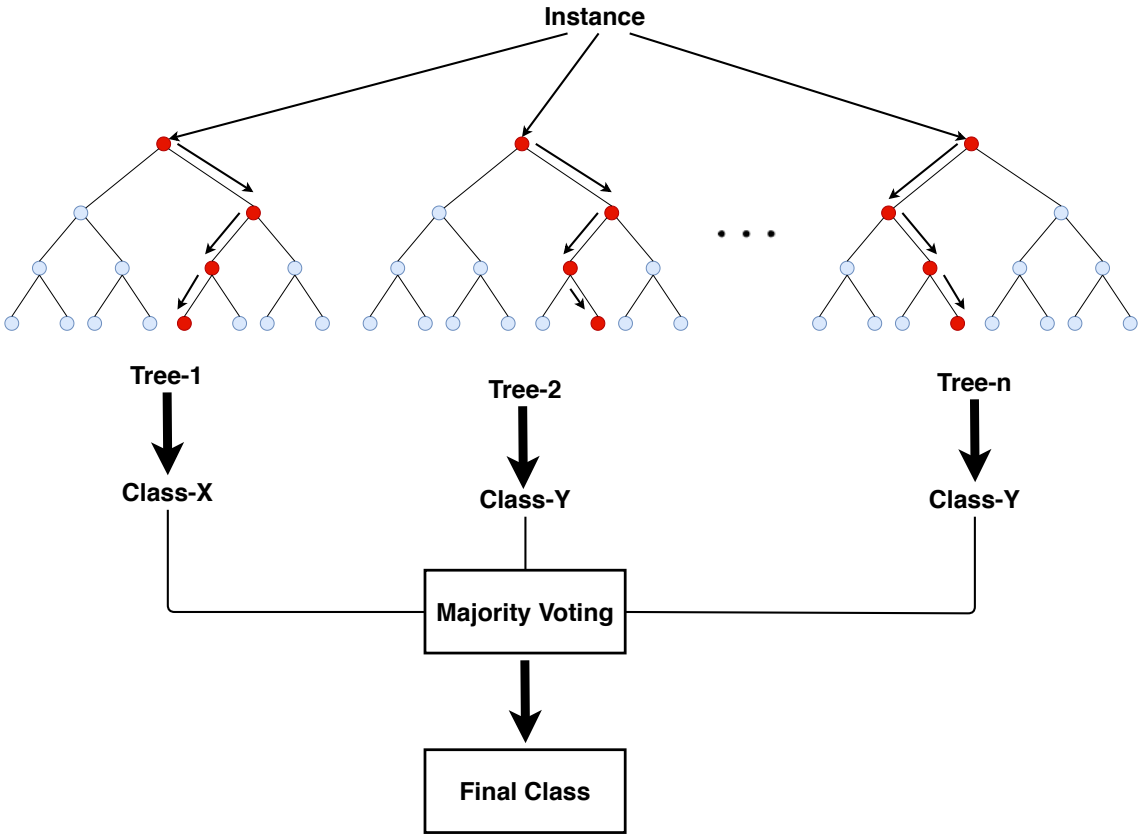


Figure 2.7: The random forest algorithm relies on multiple decision trees that are all trained slightly differently; all of them are taken into consideration for the final classification.

of their intuitive representation in the decision-making process. Decision trees are composed of a set of internal nodes where each one is labeled with an input feature. The output of each node represents a test resulting of the branch based on a certain threshold value. Both the actual branch feature and the threshold value are computed using an optimization procedure. Individual branches represent the outcome and lead to child nodes with subsequent tests, a target class label in the case of a leaf node. While single trees are useful for demonstration purposes, an ensemble of trees (i.e., tree forests) is widely used for classification problems than relying on single decision trees. Forests of trees [23] have many benefits over only adopting single trees where lower error margin and better generalization are the most important aspects. Random Forest (RF) is an ensemble of unpruned trees that can be used for both classification and regression problems. The key point of random forest is to build as a set of decision trees as shown in Figure 2.7. This method involves random feature selection for building individual and different trees. The final classification result is computed using an aggregating (voting) scheme in case of classification and averaging for regression problems over the members. RF has shown a massively improved performance compared to the traditional single decision trees [23, 20] such as C4.5 and J48. The generalization

of forests decreases as the number of trees increases, because of the randomness in the sampling process adopted for building the individual trees.

2.4.3.2 *Artificial Neural Networks (ANN)s*

ANNs are computing systems designed to mimic the human brain information processing mechanism. Such systems "learn" to execute tasks by considering examples without being programmed with any task rules and they have self-learning capabilities that make them produce better results when more data become available. For example, if someone wants to identify images that contain salad dishes, he can use example images that have labeled as "salad" or "no salad" to train an ANN and use the trained network to identify salad in other new images. ANNs perform the tasks without any prior knowledge about salad dishes. Instead, they automatically generate identifying characteristics from the learning material (e.g., labeled salad images).
An ANN consists of a collection of connected nodes called artificial neurons, which model the neurons in a human brain. Each connection can transmit a signal from one artificial neuron to another. The artificial neuron (signal receiver) can process it and then transmit to the connected artificial neurons.
In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called 'edges'. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after crossing multiple layers.

2.4.4 *Deep learning*

In this Section, we introduce deep learning since we used the deep model in the food classification in Section 4.2. Deep learning is a subset of machine learning methods based on artificial neural networks[115]; it uses multiple layers to extract higher-level features from the raw input [38]. Deep learning architectures such as convolutional neural networks (CNN)s and recurrent neural networks (RNN)s have been applied to fields including natural language processing, computer vision, speech recognition, and audio recognition, where they have produced results comparable to human experts [71, 26].
In deep learning, each layer learns to transform its input data into a more abstract representation. For example, in face recognition application, the raw input is a matrix of pixels; the first representational layer will abstract the pixels and

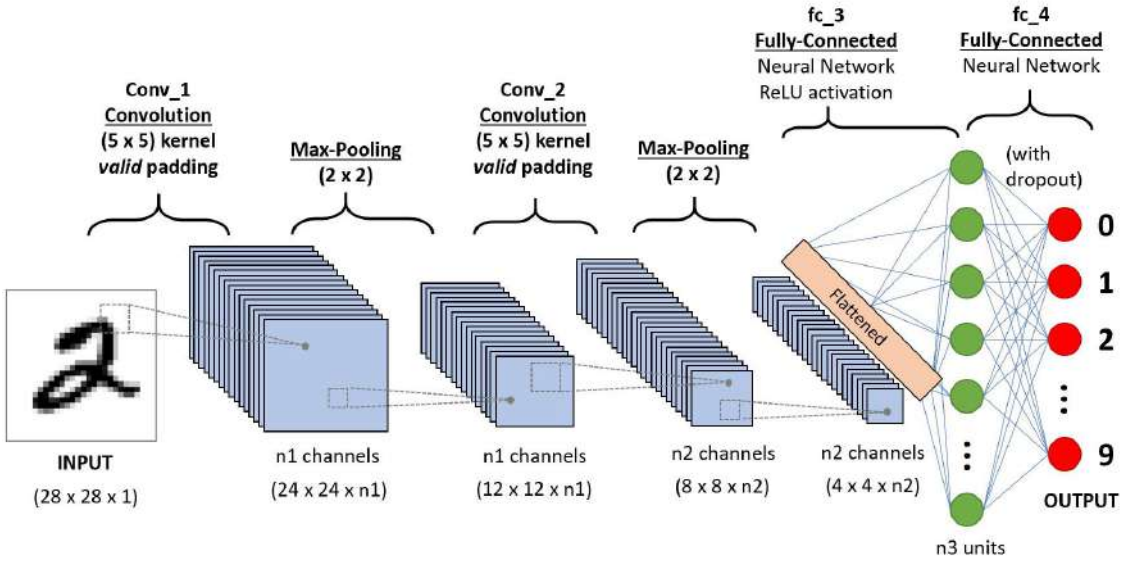


Figure 2.8: A CNN sequence to classify handwritten digits [2].

encode edges; the second layer may encode arrangements of edges; the third layer may encode a nose and eyes; and the fourth layer may recognize that the image contains a face [38].
The name of CNN drove from the employment of a mathematical operation called convolution in the network. In other words, CNN is a neural network that uses convolution operation instead of classical matrix multiplication within at least one layer [53]. A CNN consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of a series of convolutional layers that convolve with multiplication or other dot product. The activation function is commonly a RELU layer and is subsequently followed by additional convolutions such as pooling layers, fully connected layers, and normalization layers referred to as hidden layers because their inputs and outputs are masked by the activation function and final convolution as shown in Figure 2.8.

2.4.5 *Transfer learning*

Many machine learning methods work well under a common assumption: the training and test data are drawn from the same feature space and the same distribution. When the distribution changes, most statistical models need to be rebuilt from scratch using newly collected training data [96]. In many real world applications, it is expensive or impossible to re-collect the needed training data and rebuild the models. It would be nice to reduce the need and effort to re-collect the training data. In such cases, knowledge transfer or transfer learning between task domains would be desirable.

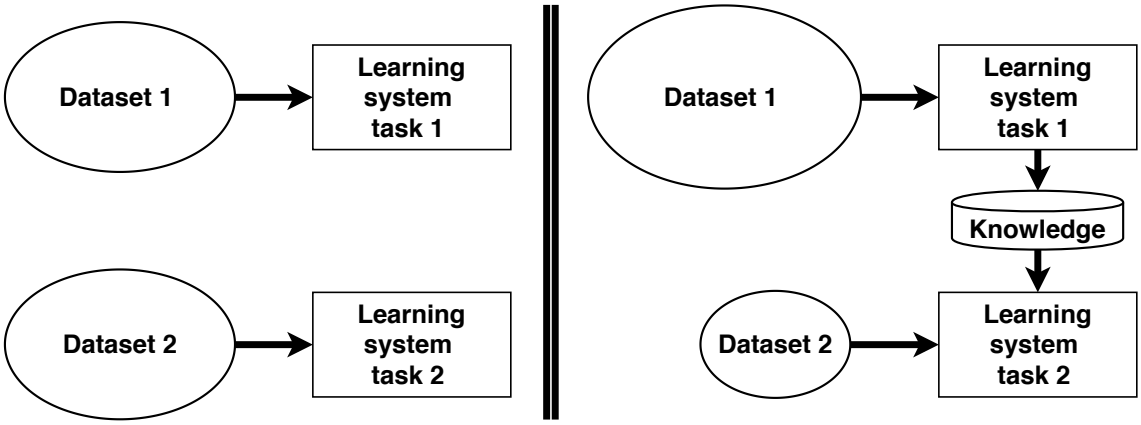


Figure 2.9: On the left learning process of traditional machine learning; On the right learning process of transfer learning.

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task as shown in Figure 2.9. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems [53].
As shown in Figure 2.9, traditional learning is isolated and occurs purely based on specific tasks, datasets and training separate isolated models on them. No knowledge is retained which can be transferred from one model to another. In transfer learning, you can leverage knowledge (features, weights etc) from previously trained models for training newer models and even tackle problems like having less data for the newer task.

2.4.5.1 *Transfer learning strategies*

In transfer learning, we need to determine which part of knowledge can be transferred across domains or tasks. After discovering which knowledge can be transferred, we need to develop learning algorithms to transfer the knowledge. Based on different situations between the source and target tasks and domains we can categorize the transfer learning into *inductive transfer learning*, *transductive transfer learning* and *unsupervised transfer learning*.
In the *inductive transfer learning* setting, the target task is different from the source task but they are related, no matter if the source and target domains are the same or not. In the *transductive transfer learning* setting, the source and target tasks are the same, while the source and target domains are different. Finally, for the *unsupervised transfer learning* setting the target task is different from the source task but they are related, similar to *inductive transfer learning* setting. However, the *unsupervised transfer learning* focus on solving unsupervised learning tasks in

the target domain, such as clustering [34], dimensionality reduction and density estimation [140].

2.4.5.2 *Deep transfer learning strategies*

Deep learning has made remarkable progress in recent years. This progress has enabled researcher to undertake complicated problems and yields amazing results. However, the required amount of data and the training time for such deep learning systems are much more than comparing with the traditional ML systems. There are various deep learning networks with the state-of-the-art performance that have been developed and tested across fields such as computer vision and natural language processing. In most cases, people share the details of these networks for others to use. These pre-trained models form the basis of *inductive transfer learning* in the context of deep learning (deep transfer learning). The two most commonly used deep transfer learning strategies are

- **Off-the-shelf Pre-trained Models as Feature Extractors.** Deep learning models are layered architectures that learn different features at different layers (hierarchical representations of layered features). These layers are then finally connected to the last layer to get the final output. This layered architecture allows utilizing a pre-trained network by removing the final layer and use the rest of the network as a feature extractor for other classification tasks, as shown in Figure 2.10.

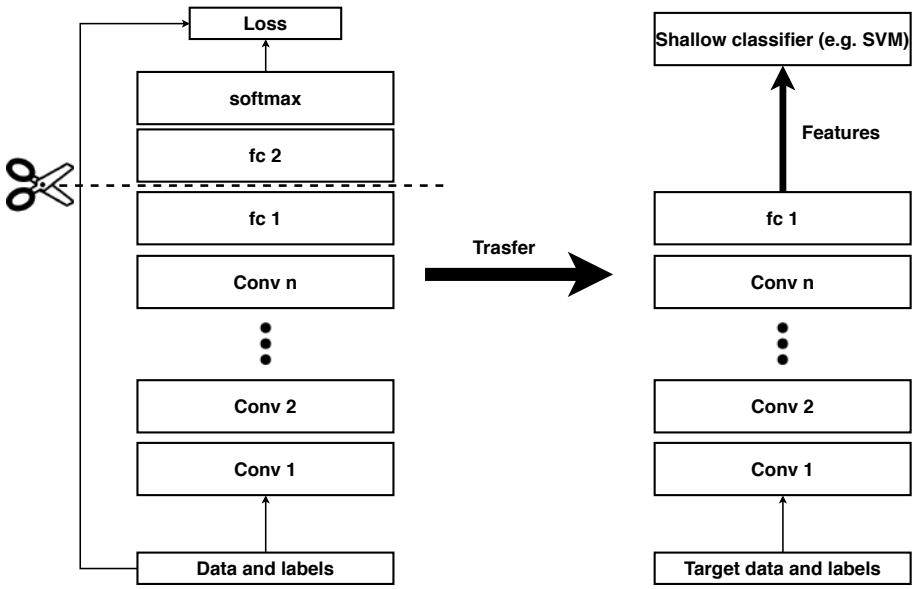


Figure 2.10: Transfer Learning with Pre-trained Deep Learning Models as Feature Extractors.

The principal idea is utilizing the pre-trained model's weighted layers as features extractor without updating the weights of the model's layers during the new task's training phase.

- **Fine Tuning Off-the-shelf Pre-trained Models.** Deep neural networks are highly configurable architectures with various hyper-parameters. Using this fact, particular layers can be frozen (weights are fixed) while the rest layers are retrained or fine-tuned to satisfy the new task, as shown in Figure 2.11. In this way, the network's architecture is utilized and used as a starting point for the retraining step.

Figure 2.11: In fine-tuning process, all convolutional layers (blue layers) in the network are fixed and gradient is backpropagated through the fully connected (FC) layer only.

2.4.6 *Image augmentation for deep learning*

Image augmentation is another solution to overcome the limited number of available annotated images. It increases the size of the available data by applying some image transformation operations to the existing images from a training dataset to produce new versions of existing images. Image transformation operations include rotation, shearing, translation, zooming, etc. these random transformations will produce different images each time.

2.4.7 *Model evaluation and metrics*

Model evaluation is a crucial part when developing data-driven models. The purpose of any predictive model is to correctly predict the target class value for unseen data instances with the highest possible accuracy. Thus, it is required to have a way of evaluating model performance, typically by quantifying it using some measure of model error. This same measure must be used to train the model to obtain high accuracy performance. One of the significant pitfalls when creating a predictive model is evaluating the trained model on the same or almost similar data to the training ones [49]. Adopting incorrect measures and evaluation methods may lead to generate overfitted and over-optimistic models.

2.4.7.1 *Model evaluation*

There are two main methods of model evaluation in machine learning: (i) **hold-out validation**; (ii) and **cross-validation**. Both methods use a separated test set of unseen data in model performance evaluation process. While in model training, the objective is to minimize the training error based on the chosen metric.

- **Hold-Out Validation.** It is called a train/test-split method which requires a part of the original data to be held-out from the training process. The final

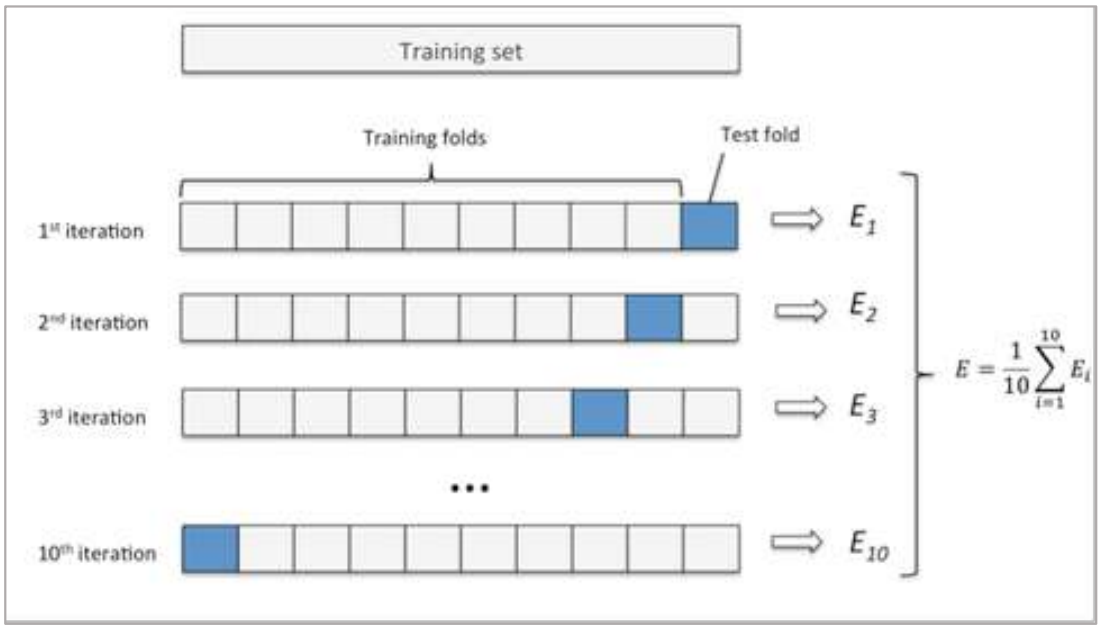


Figure 2.12: 10-fold cross validation. The designated training set is further divided up into K folds (K=10), each of these will now function as a hold-out test set in K iterations. Finally, the scores obtained from the model on individual iterations are summed and averaged into the final score.

evaluation score is only computed through experimenting the test set on the produced model. This method is simple, relatively fast, and it ensures that the model is tested on unseen data. The main disadvantage in this method is that a part of the original data is removed from the training set of the model. Moreover, there is a risk to have high variance in the predictions.

- **Cross-Validation [58, 88].** This method is an extension for the hold-out validation method. Cross-validation is a widely accepted and used in the state-of-the-art predictive methods for ensuring model reliability and generalization ability. In $k-fold$ cross validation 2.12, k represents the number of partitions that the training set will be divided into. Also, k represents the number of iterations that the trained model will execute, resulting an evaluation score at each iteration. The evaluation scores of k iterations are averaged to obtain the final score. For each iteration, (k-1)/k ratio of the original data is used for training a model, while the remaining partition with size of $1/k$ is used for model validation. With each iteration, the unused partition is set aside for the validation test, and a new model is trained from the remaining partitions as described above. The final score is computed through averaging over the number of iterations, k. The benefits of $k-cross$ validation are in reducing variance due to the averaging effect. However, this process is quite slow and requires high resource requirement.

2.4.7.2 *Metrics*

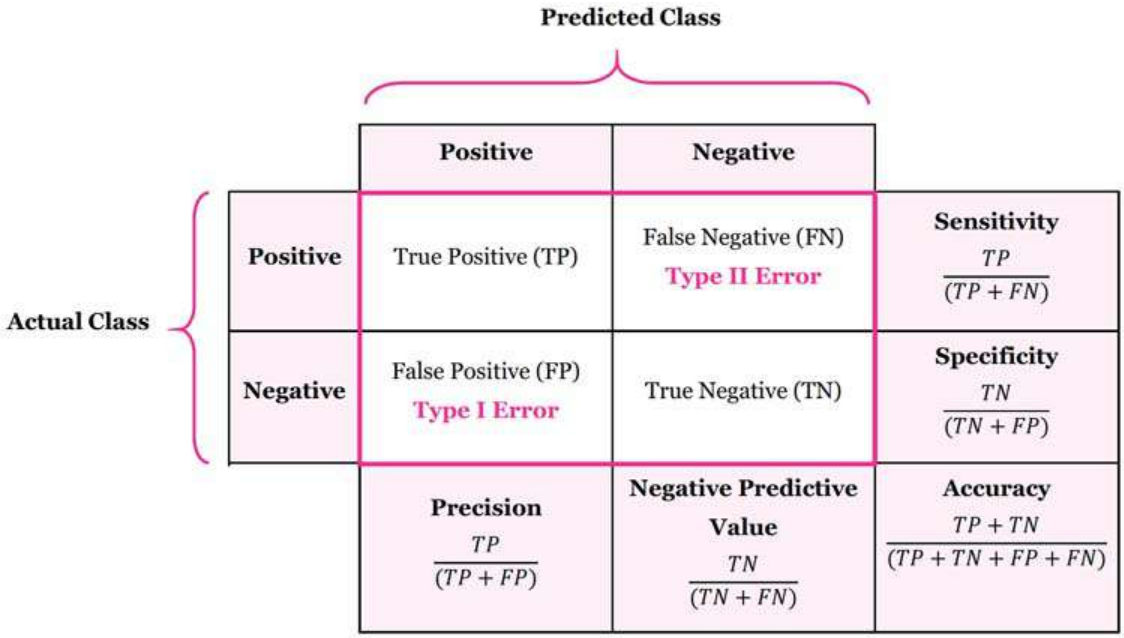


Figure 2.13: An illustrative depiction of the (binary) confusion matrix and a selection of the measures that may be derived directly from it.

To evaluate our purposed methods in Section 4.1.1 and compare with other machine learning methods, we used the most common classification models evaluation metrics, which include:

- **Confusion Matrix.** A confusion matrix provides a very intuitive and complete overview about classification models performance. The matrix has dimension of N × N, where N is the number of target class labels of the considered problem. The ground-truth (true label of instances) is matched with the predictions resulted by the trained model, showing information how much the model is accurate in the predictions for each class with providing the distribution of misclassified instances in each class. The confusion matrix is the basis for computing most evaluation metrics in machine learning. In a binary (two class) example 2.13, four basic counts are obtained from the matrix: True Positives (TP), False Positives (FP), True Negatives(TN) and False Negatives (FN). Based on these terms, we can derive most of the metrics described below.

- **Accuracy.** The most adopted and cited performance metric, defined as the percentage of correctly classified examples (instances) out of the total number of examples. Accuracy (ACC) is a good metric when the class distribution is balanced. The problem of imbalance class distribution becomes apparent when one class dominates other class(es). For example, in a dataset of 900:100 (class 0:class 1) binary class distribution, classifying blindly all instances as negative will result 90% of accuracy.

- **Precision and Recall.** Precision is defined as the ratio of correctly predicted labels out of the total predicted positive labels. Recall is defined as the ratio of correctly predicted positives to the total number of positive labels (ground-truth) in the data.

- **F-measure.** It is known as the balanced F-measure. It is a single scalar value metric summary that combines both Precision and Recall metrics together. It is used in performance evaluation of binary classification problems. F1 is defined as the harmonic mean of precision and recall, equally weighs precision and recall. As it is based on recall and precision, the F-score only considers the positive predictions. F1-score is generally considered in the evaluation when class imbalance is an issue [8]. A high F-measure score is a good indicator of a good performing classifiers *w*.r.t. minority classes.

2.5 Prader–Willi syndrome (PWS)

Prader–Willi syndrome (PWS) is a genetic disorder caused by an error in one or more genes located in a particular region of chromosome 15. It was first described in detail in 1956 by Andrea Prader, Heinrich Willi, and Alexis Labhart [103]. PWS affects approximately one out of every 15,000 births (males and females with equal frequency) [94].
Signs and Symptoms of PWS can vary among individuals and may slowly change over time from childhood to adulthood. In newborns, symptoms include poor muscle tone, distinct facial features (e.g., almond-shaped eyes, turned-down mouth), a poor sucking reflex which causes poor feeding, and slow development (mental or physical) [27]. Other sings and features appear at the beginning of childhood [61]. These sings may include:

- **Sleep disorders** could be caused by breathing pauses during sleep. These disorders can result in excessive daytime sleepiness and worsen behavior problems.

- **Cognitive impairment**, such as issues with thinking, reasoning, and problem-solving.

- **Infertility** since sex organs (testes in men and ovaries in women) of PWS individuals produce little or no sex hormones.

- **Behavioral problems.** PWS Children and adults are extremely stubborn and prone to anger. They may throw temper tantrums, especially when denied food.

- **Food craving and weight gain.** A classic sign of PWS is a constant craving for food (constantly hungry), resulting in rapid weight gain, starting around age two years. Constant hunger leads to eating often and consuming large portions.

Figure 2.14: Eight-year-old with PWS: Note presence of morbid obesity [29].

A key feature of PWS is a constant sense of hunger. They never feel full (hyperphagia), and they usually have trouble controlling their weight which cause morbid **obesity** as shown in Figure 2.14. **Obesity** is the main reason of many complications of PWS. Unfortunately, PWS has no cure [93], but treatment may improve outcomes, especially if carried out early for example, obesity can be

controlled externally by diet restrictions and behavior modification [93]starting around the age of three, in combination with an exercise program.

2.6 Conclusion

This chapter describes the basic concepts and definitions that will be used in the thesis. Since our work is in the context of observational studies, we present a brief introduction to these studies. Then, we present social cues that are commonly observed in observational studies. After that, we illustrate some concepts about machine learning methods, deep learning, transfer learning, and metrics that are commonly used to evaluate trained models. The presented machine learning methods are used to detect social cues as illustrated in Chapter 3 and Chapter 4. Finally, we introduce the PWS since one of the data subjects of OVALIE platform is children who born with this syndrome.

3

RELATED WORK: EXPERIMENTAL PLATFORMS, SOCIAL INTERACTION DETECTION AND ANALYSIS, AND FOOD CLASSIFICATION

Social interaction is based on interpretative analysis rather than statistical or empirical observation.

— Erving Goffman

In this chapter, we present the related work to observational studies analysis automation. First, we present some existed experimental platforms that aim to study individual's eating behaviors and they commonly ignored the effect of the context on the eating habit. Second, then we introduce most common existed methods that are used for automatic social cues capturing and that deployed or could be deployed in the experimental platforms. Finally, we present the most common methods and applications for food recognition task since the food will be part of the physical context in case of food habit studies.

3.1 Experimental platforms for eating behavior observation

Experimental observation platforms have appeared as a way to study social interactions effects linking to eating habits. These platforms allow for the solid improvement of observational strategies through the use of technical devices for automatic capture and processing thanks to recent technological advancements in the computer sciences. Several experimental platforms are tackling several food-related studies such as food choice and preference, nutrition, sensory analysis, and consumer behavior. These platforms are designed to record meals and then analyze the behavior of eaters. The most well-known platforms include the Living Lab [18], the Restaurant of the Future [114], and the Centre for Taste and Feeding Behavior (CSGA) [33].
These platforms have flexible spaces suitable for the construction of various consumption settings, data collection instruments (such as cameras and microphones), data processing software from a centralized location in order to account for various eating contexts, such as different catering systems (fast food, casual restaurant, family meals, etc.). However, they neutralized effect of the context and considered it as controllable variable and ignored it.

3.2 Social interaction detection and analysis

Several methods are proposed for detecting and analyzing social interactions using machine learning-based methods. These methods have been adopted in a wide range of applications and domains such as robotic, medical, economics, sociology, and Internet of Things (IoT) applications. The developed approaches use one or more primitive social signals for performing social detection and analysis. Social cues are classified into *verbal* and *nonverbal* cues as shown in Figure 1.1.

3.2.1 *Verbal cues detection*

Verbal communication is the use of sounds and words (verbal cues) to express yourself [91]. Verbal cues consist of words and linguistic units of sounds and speech organs take a prominent position among the production and transmission of signals. Machines can detect the verbal cues through speech recognition methods (engine) [84, 119, 48, 55]. The principal function of the speech recognition engine is to process spoken input and translate it into a text, as shown in Figure 3.1. In addition to audio input speech recognition engines require two inputs (models) to recognize speeches. First, a language model contains a list of words and their occurrence probability in a given sequence, language models are used in dictation application. Second, an acoustic model contains a statistical representation of the distinct sounds that make up each word in the language model.

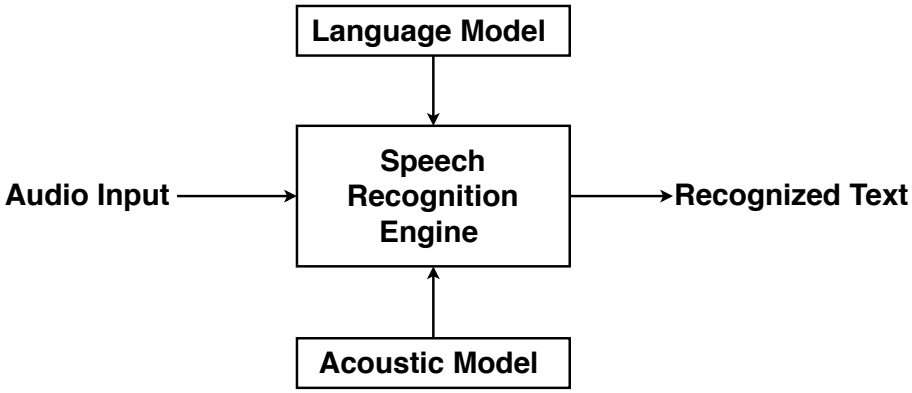


Figure 3.1: The general pipeline of speech recognition engines.

However, it is still a challenging task to have a high performance in multiple speakers free setup (e.g., restaurant) since it is more challenging to partitioning an input audio stream according to the speaker identity (speaker diarization) in such a noisy environment.

3.2.2 *Vocal nonverbal cues detection*

Nonverbal cues consists of *vocal* and *visual* cues. *Vocal* cues can contain information regarding events like the existence of music, speech, emotions, etc. Extracting such information can enhance the recorded video content analysis. Like the verbal cues machines can detect them using speech analysis methods.

3.2.2.1 *Speaker diarization*

Speaker diarization is the process of partitioning an audio stream into speaker related segments. On other words, it is a speaker segmentation followed by speaker clustering. It answers the question "who spoke when?" in a multi-speaker environment. Authors of [6] proposed speaker diarization method that detect the active speaker through a pre-trained audio visual synchronization model. The model achieve a close result comparing with other complex speaker diarization state-of-the-art. Speaker diarization is essential step in social cues analysis in a meeting as it allows the mapping between the detected vocal cues and the speaker.

3.2.2.2 *Speech emotion recognition (SER)*

Automatic speech emotion recognition is the task of predicting the speaker's emotional state (anger, sadness, etc.) using speech analysis techniques. Figure 3.2 shows the general SER pipeline. From the input speech segment feature vector is created by extracting the acoustic features. Then, the most relevant features are selected in the next step in order to achieve higher accuracy and reduce the training and processing time. Finally, a trained model is used to obtain the final classification result.

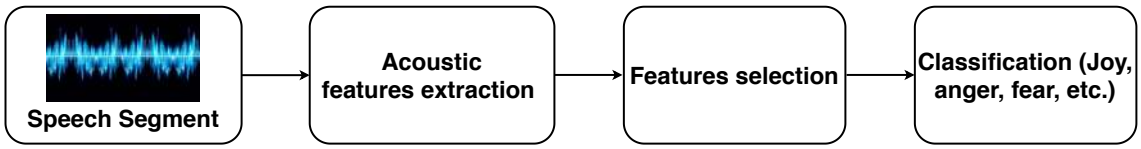


Figure 3.2: The general pipeline of speech emotion recognition.

All SER methods are following this pipeline and they differences in the type of features that are used or the classification model. The most common extracted fetures in SER methods are energy, pitch, formant, linear prediction coefficients (LPCs), and mel-frequency cepstrum coefficients (MFCCs) [44]. Many SER systems used the SVM for emotion classification [120, 116, 97], RNN is suitable for learning time series data, and it has shown improved performance for SER task [73, 87].

3.2.3 *Visual nonverbal cues detection*

The second category under the nonverbal cues is the *visual* cues. These cues include physical appearance (e.g., gender, height, ethnicity, age), face and eyes cues (e.g., facial expression, gaze direction, focus of attention), gesture and posture, and space and environment [136].

3.2.3.1 *Facial expression recognition (FER)*

Various studies have been performed to detect and analyze facial expressions using automatic machine-learning methods [14, 65, 89, 118, 85]. These studies have been applied in a tremendous range of applications and domains, including sociable robotics, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems.
Automatic deep facial expression recognition starts with pre-processing step, then deep feature learning step, followed by deep feature classification steps as shown in figure 3.3 [75].
Having sufficient labeled training data that include as many variations of the populations and environments as possible is important for the design of a deep expression recognition system. In the following we introduce the most known databases for FER. The Extended CohnKanade (CK+) [78] database is the most extensively used laboratory-controlled database for evaluating FER systems, it contains 593 video sequences from 123 subjects. MMI [100, 133] database also is a laboratory-controlled and it includes 326 sequences from 32 subjects. JAFFE [79] the Japanese Female Facial Expression database, which is a laboratory-controlled and it contains 213 samples of posed expressions from 10 Japanese females. EmotioNet [46] is a large-scale database with one million facial expression images collected from the Internet. However, all the proposed datasets are not sufficient to have a wild facial expression recognition model that ables to perform well when we have non-frontal face photos.

3.2.3.2 *Gaze and eye tracking*

Eye tracking has been developed in the context of studying human visual selection mechanisms and attention (see [64, 57, 50] for a review on eye detection and gaze tracking in video-oculography). A lot of information could be acquired through the study of the eye movement, where we can know about people's thinking based on where (who or what) they are looking. Also, it is well known that the points toward which humans direct the gaze are crucial for studying human perception and his ability to select the regions of interest out of a massive amount of visual information [131]. However, eye tracking and detection remains a very challenging task due to several spacial problems, including illumination, viewing angle, occlusion of the eye, head *pose*, etc.

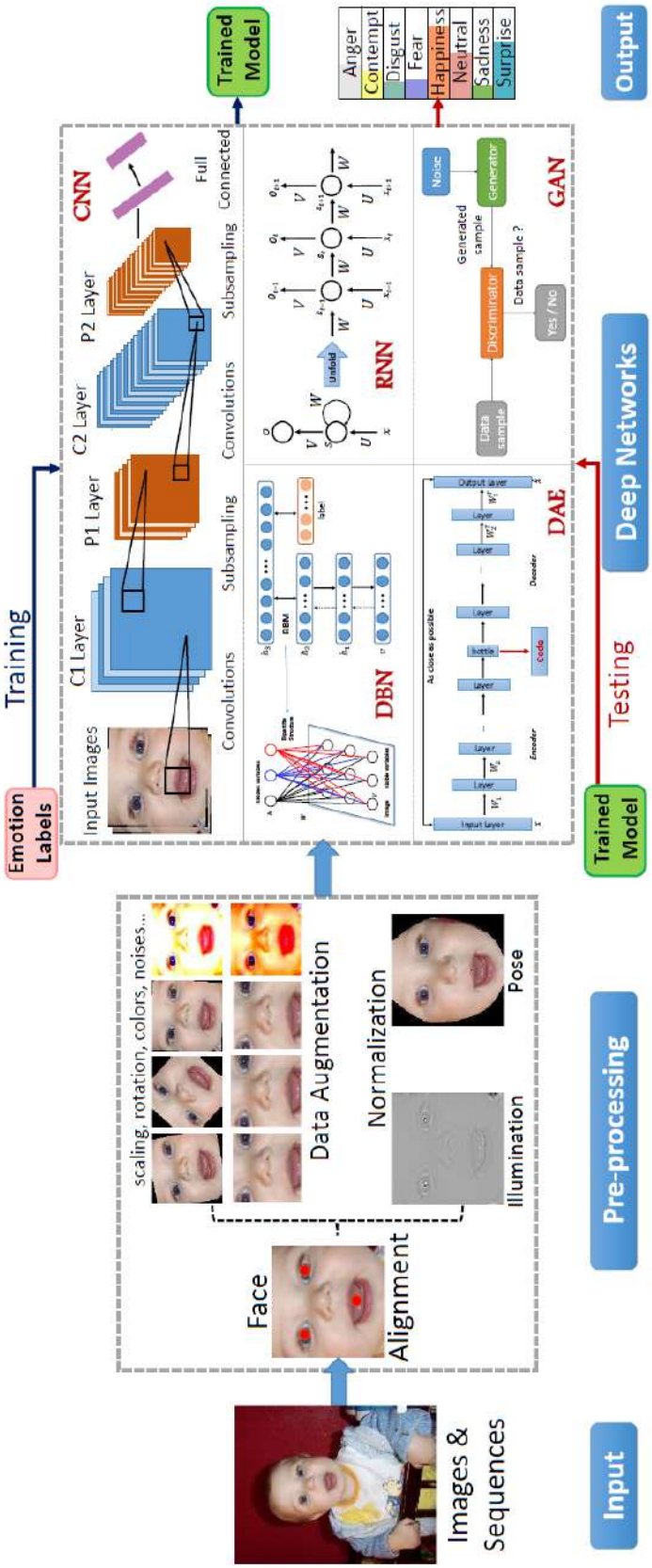


Figure 3.3: The general pipeline of deep facial expression recognition systems [75].

3.2.3.3 *Eye contact detection*

Eye contact detection is a binary decision on whether someone's gaze falls onto a target (e.g., face, screen) or not. Many methods have been developed to handle this issue by either using a head-mounted device [25, 5, 108] or requiring LEDs attached to the target [125]. More works focus on developing non-intrusive e.g., the authors of [124] trained a classification model to determine whether there is eye contact with a camera or not. However, their method requires prior knowledge about the size and location of the target. Authors of [144] have presented a method for eye contact detection during dyadic (two-person) interactions; however, their method works only for a single eye contact target that must be the closest object to the camera. This assumption does not hold for multi-person interactions in which multiple targets are available.

3.2.4 *Analysis methods based on visual nonverbal cues (VNAM)*

Many methods are proposed for detecting and analyzing social interactions based on the visual nonverbal cues. Generally, the existing methods rely on using fixed or mobile camera(s) for data acquisition. Furthermore, these methods could varying based on the type of social cues that are used as features, the way of fusing features together, and the type predicted social behavior that the system may provide. For instance, in [5], the authors have introduced an approach that detects the social interactions through using a user wearable low frame rate camera (mobile) to capture images, and then two primitive social signals (distance and orientation) are extracted to predict whether a social interaction existing among the desired persons. On the other side, the system introduced in [24] detects the social interactions in working environments through capturing the data from a fixed camera mounted at a particular height, and then the social interaction is predicted using a trained support vector machine (SVM) classification model, applied on the relative head orientation and distance between the pair of people features. The existing visual nonverbal-based methods can be grouped in three different levels as shown in Table 3.2.4:

1. **One or multi Social Signal.** VNAMs may leverage one or more of primitive social signals for performing social detection and analysis. Thus, highlight 12 common used social signals.

2. **Detected Social Behavior Type.** VNAMs are not common in their internal purpose so that some of them are dedicated for detecting different social behaviors such as emotion, mental state, rapport (people are "in sync" with each other) and dominance speakers.

3. **Time Variant v.s. Time Invariant signals.** Social signals are further categorized based on their relation with time. Hence, the social signals that evolve over time are classified as *Time Variant* signals, while the signals

don't change *w.r.t.* time are categorized under *Time Invariant* signals. The height of persons is a simple example on the time invariant signals, while the changing of head pose is time variant one since the persons may change their head orientation and positions.

Table 3.1: Visual nonverbal signals associated to the most common social behaviors.

Social Signals	Social Behavior						Methods for Automatic Detection
	Emotion	Dominance	Mental State	Deception	Rapport	Status	
Time Variant Signals							
Posture	✓	✓		✓	✓	✓	[139]
Head Pose		✓	✓	✓	✓		[15]
Facial Expression	✓	✓	✓	✓	✓	✓	[31, 82, 74]
Hand Over Face Gesture	✓	✓	✓	✓	✓	✓	[80]
Hand Gesture	✓		✓	✓	✓		[142]
Gaze Behavior	✓	✓	✓	✓	✓	✓	[72, 101, 15]
Visual Focus of Attention	✓	✓		✓	✓	✓	[11, 13]
Time Invariant Signals							
Height		✓				✓	
Age		✓			✓	✓	
Gender		✓			✓	✓	
Ethnicity		✓					
Attractiveness		✓			✓	✓	[92]

As shown in Table 3.2.4, the time variant signals have a high correlation with the applicable automated detection methods. The time invariant signals are not commonly used in the automated social signals detection and analysis because they are subjective. For instance, no standard attractiveness parameters, each person has his parameters. We can notice that there are possibilities to use multiple social cues from different modalities to detect social behavior. To determine the dominance speakers in a meeting, we could use facial expression, hand gestures, and even we can utilize the vocal cues that can be detected using the speech analysis methods. Face and eye cues are the most common targeted signals in the automated social signals detection and analysis methods.

3.3 Food recognition and classification

To study and analysis eating behavior in different contexts (one of the OVALIE platform goals), we need to recognize the consumed food during the observational study (in the content of the recorded video(s)). One of the earliest works in the field of food classification and recognition appeared in [143], where au-

thors studied the spatial relationships between different food ingredients. They deployed Semantic Texton Forest to segment the input image into eight different types of ingredients; then, a multi-dimensional histogram was computed using pairwise statistics, later it classified with a support vector machine (SVM) classifier. Authors of [47] introduce the UNICT-FD889 dataset to study the representation of food images; they have benchmarked their dataset with PRICoLBP [104], SIFT [76], and Bag of Textons (BoT) [134] descriptors. BoT descriptors achieved the best result and demonstrated that nearest neighbor and color descriptors are relevant for food classification task. The authors of [60] proposed a general framework for food analysis based on the integration of multimodal content, context, and external knowledge, including recipe analysis, food recommendation, restaurant oriented applications.

3.3.1 *Food datasets*

There are several open-access food image datasets with different categories such as Food-101 [19], UECFood-256[66], UECFood-100 [83], which are used to train a classifier and evaluate the trained model. Food-101 database is the most popular dataset in food domain, it includes 101 food classes with 1000 image of each class as shown in Figure 3.4.

Figure 3.4: This Figure shows one example for 100 out of the 101 classes Food-101 dataset. [19].

3.3.2 *Deep learning and food recognition*

CNNs have been widely used in food/nonfood classification, food category discrimination, and ingredients identification. Authors of [122] created a database named Food-5K consisting of 2500 food images and 2500 images of other objects. Then they fined-tuned GoogLeNet [127] model to classify the images. Authors of [110] do food/nonfood classification by coupling fine-tuned AlexNet with a binary SVM classifier.
The overall process of learning methods proposed for image-based food recognition in surveyed papers was basically the same. The first step is dataset preparation. Next, image preprocessing like normalization, resizing, is followed to

weaken the interference caused by nonuniform illumination, resolution inconsistency, and so on. If the dataset is not large enough, data augmentation should be performed to enlarge the dataset by random clipping, rotation, and flipping, to simulate shooting from different perspectives as illustrated in Section 2.4.6. Then the prepared dataset is always divided into training set for training the network, validation set for fitting the hyperparameters, and evaluation (or testing) set for confirmation of the predictive ability of the model. The generalization ability of the trained CNN-based model should be examined on different datasets.

3.3.3 *Food recognition and classification applications*

Many applications have been designed based on the food recognition and classification task.

- Food Calorie Estimation, authors of [86] designed a mobile application named *Im2Calories* for food calorie estimation from images. For desktop application, authors of [41] proposed a new network called multitask CNN for estimating food calorie from a food image, the in Figure 3.5.

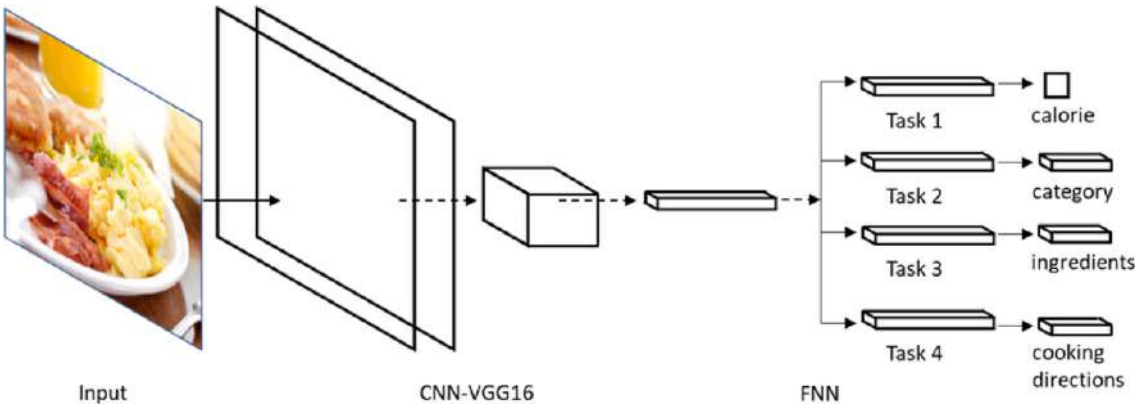


Figure 3.5: The multitask CNN used a VGG16 architecture for feature mining and the learned features were fed into four parallel subnetworks to predict the calorie and other attribute of food [41].

- Quality Detection of Fruits and Vegetables, authors of [111] developed a classification approach using the stacked sparse auto encoder combined with CNN for cucumbers defect detection based on hyperspectral imaging. Authors of [112] focused on the discrimination of plum varieties (Black Splendor, OwentT, and Angelino) at early maturity stages using deep learning technology. Authors of [129] aimed at realizing artificial intelligence (AI)-based alerting system for pests and diseases of apple. CNN was applied for recognition of apple skin lesion image collected via an infrared video sensor network.

3.4 Conclusion

This chapter presents the related works and tasks that will help to build a holistic framework for social interaction analysis in the context of observational studies automation. First, we introduce some existing experimental platform for eating behavior observation and we notice that commonly they ignored the context effect on the eating behavior. Second, we present a general overview of the existing methods that aim to detect and analyze social cues using machine learning methods and highlight the importance role of visual nonverbal cues. Then, we focus on the face and eye cues that are used as main cues for the social interaction analysis methods. Finally, we introduce the existing methods datasets of food recognition and classification task since it can be used in the context of eating behavior observation and analysis.

4

CONTEXT-AWARE FEATURE EXTRACTION METHODS

Eyes are the windows to the soul.

— William Shakespeare

In this chapter, we present our proposed context-aware methods that we will use to build our holistic approach for social interaction analysis. First, we present our geometrical eye contact detection methods then we compare the geometrical based development with machine learning based one. Second, we present a new dataset for evaluating find-tuned deep models for food classification. Then we discuss how the use of contextual information will add semantic dimension to the deep model and enhance the accuracy.

4.1 Eye contact detection in Face-to-Face interactions

Eye contact detection is defined as a task of automatically detecting whether two people look at each other's eyes or face simultaneously. It is an important feature for better understanding human social behavior. Eye contact detection has numerous applications. For example, it is a key component in attentive user interfaces and it is used to analyze turn-taking, social roles, and engagement during multi-person interactions. Even more, we can deduce many things based on the eye contact [10], e. g., the topic nature, in which, there is more eye contact in case of the topic being discussed is straightforward and less personal, whereas, there is less eye contact during the hesitating passages.
At the social cues extraction level, eye contact is an important social cue than can be used to perform a wide range of analysis and studies such as a dominant person detection [51]. It provides multiple functions in the two-person contacts such as information seeking, establishment and recognition of social relationships, and signaling that the "channel is open for communication" [10]. Indeed, extraction of this social cue must be fully automated, accurate at detection level, and compatible with simple capturing devices such as closed-circuit television (CCTV) cameras. However, existing state-of-the-art methods require expensive special devices for detecting any contacts at the eye-level. Such methods are based on supervised machine learning techniques to produce eye contact classification models, raising the need for ground truth datasets as a difficult and time consuming task.

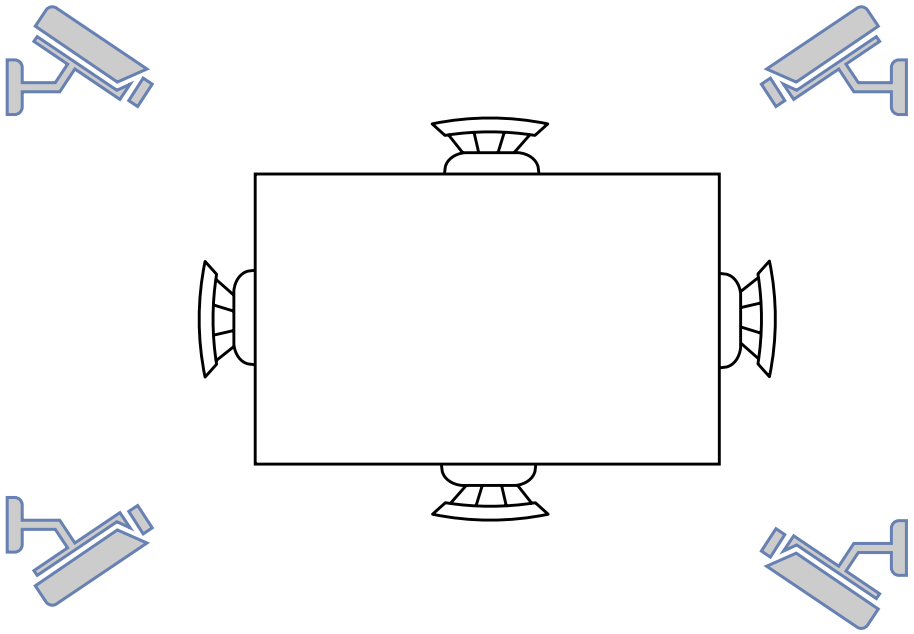


Figure 4.1: Camera setup used for the dataset recording.

4.1.1 *Geometrical eye contact detection*

We propose a novel geometrical method to detect eye contact in small group interactions using multiple cameras. Our method first extracts all participants' head pose from several ambient cameras and then map them to a common reference frame. After that, a check is performed for each detected person if there is an intersection between his/her gaze direction with other detected persons. Then, a temporal *square* matrix is built by which we can check whether an eye contact between two participants holds or not. Our proposed method does not need for eye tracking or any intrusive devices, which allows recording natural social behavior during a face-to-face social event.

4.1.1.1 *Cameras setup*

Our eye contact detection approach uses CCTV cameras mounted at a particular height in the place where the participants set around a table. The number of cameras is conditioned by arrangement of participants around the table, for example, a single camera is enough if the participants set in a horizontal way and the camera covers the participants' frontal face, otherwise we need more cameras to have a frontal and semi-frontal face photos. To evaluate our method, we equipped a room with a table and four cameras around the table as shown in Figure 4.1. The camera setup ensure that we have high-quality data (e.g., frontal face photo to have more accurate facial expression prediction). However, to record the natural behaviors of the participants, we need to keep the sensors far as much as possible from them.

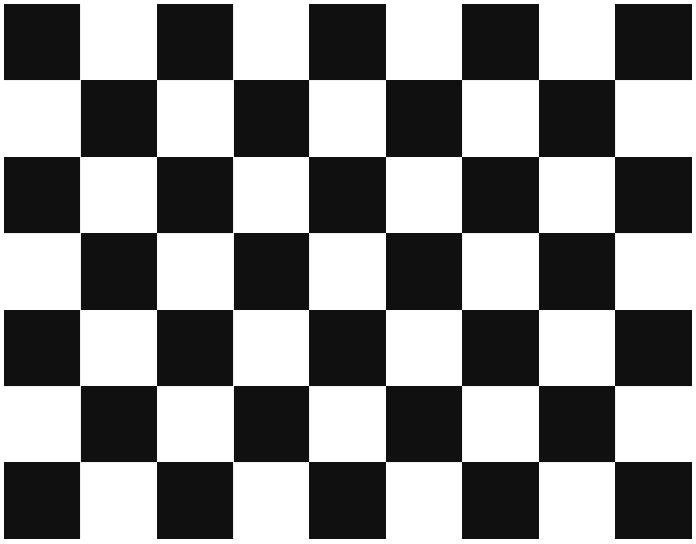


Figure 4.2: Calibration checkerboard contains 8×6 internal corners, 9×7 squares, square size = 4cm.

After installing the cameras, we need to estimate the cameras *intrinsic* parameters. First, we prepared a checkerboard with 8×6 internal corners as shown in Figure 4.2. Second, start the camera calibration procedure, which is implemented in *opencv* library [1], to estimate the *intrinsic* camera parameters.

4.1.1.2 *Person detection and tracking*

To detect the eye contact, we deploy OpenFace toolkit [14] to detect persons in videos frames, their head *pose*, and their gaze direction. However, having multiple views setup means that we will have a redundancy in the detected persons since the same person will appear in more than one view. To solve the redundancy issue, we map all detected persons into a common space then we track them during the video as following:

- **Mapping to common space (Perspective-n-Point).** In order to have a common reference space for all of our cameras we need to calibrate our cameras with the real-world (e. g., 3D location of the room corner). First, we need estimate a transformation matrix that map the 3D camera coordinate to a world coordinates for each cameras that we used. We estimate the transformation matrix that brings points from the world coordinate system to the camera coordinate system using [28], then we compute the inverse of the obtained matrix to map the estimated head *pose* for each detected person to a common reference.

1 https://docs.opencv.org/3.4/dc/dbb/tutorial_py_calibration.html

- **Geometrical grouping.** To remove the redundancy that caused by the multiple cameras usage by keeping only one detected person in the common reference within *30 cm* radius, since we can assume that we will have only one person head within that radius. So, for each detected person in the common if there are two or more persons fall within a radius less than 30 cm we will keep the one with the higher detection confidence value.
- **Spacial Location Tracking.** To track the detected persons within the video we use the K-means. First, we compute the location centers of the participants based on the first minute of the videos, where the methods compute n location centers based on the filtered head *pose* in the common space. Second, we classify the head *poses* based on the distance between the head pose and the centers.

4.1.1.3 *LookAt()*

LookAt() function determines if person x is looking to person y or not. Formally, we can assume that $\mathrm{LookAt}(x, y) \in \{0, 1\}$ at time t is a binary value that determines whether the participant x looks towards the participant y.

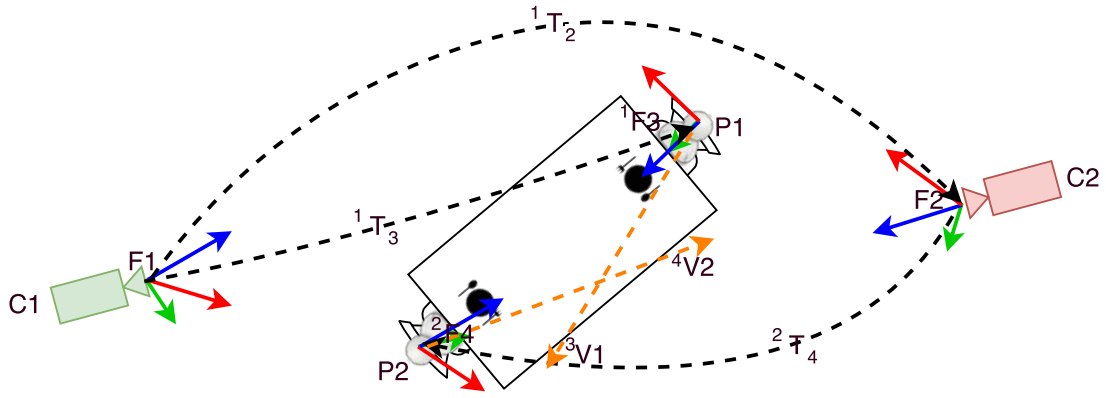


Figure 4.3: LookAt() evaluation between two persons. C1, C2 are first and second cameras; P1, P2 are first and second persons; F1 is the reference frame of C1, F2 is the reference frame of C2; ${}^{1}F3$ is P1 head pose *w.r.t.* F1, ${}^{2}F4$ is P2 head pose *w.r.t.* F2; ${}^{i}T_{j}$ is the pose of Fj *w.r.t.* Fi; ${}^{3}V1$ is the gaze direction of P1 *w.r.t.* ${}^{1}F3$, ${}^{4}V2$ is the gaze direction of P2 *w.r.t.* ${}^{2}F4$.

We can calculate the values in the LookAt() function using two different approaches. The first approach is based on supervised machine learning methods by training a classifier that takes the head pose of two participants and return whether the first one is looking at the other one. The second approach is a geometrical one that does not require any training dataset. We illustrate the second approach through an example of two participants and two cameras as follows:

1. Assign reference frames as illustrated in Figure 4.3, where F1 is the reference frame of first camera (C1), F2 is the reference frame of second camera (C2),

${}^{1}F3$ is the first person (P1) head pose *w*.r.t. F1, and ${}^{2}F4$ is the second person (P2) head pose *w*.r.t. F2.

2. Compute the transformation between frames, where ${}^{1}T_2$ is equal to the pose of C2 *w*.r.t. F1, ${}^{1}T_3$ is equal to the pose of P1 head *w*.r.t. F1, and ${}^{2}T_4$ is equal to the pose of P2 head *w*.r.t. F2. The transformation ${}^{i}T_j$ is used to transform a vector ${}^{j}V$ from Fj to Fi as

$$ {}^{i}V = {}^{i}T_j \times {}^{j}V \tag{4.1} $$

3. Check whether Pk stares at Pl. In particular, we have to check if the Pk gaze vector intersects with a sphere centered at Pl head position. Hence, both the line and the head position must be in the same reference frame. Assuming that F1 is the reference frame, and Pk is seen by C1 (Pk = P1) and Pl seen by C2 (Pl = P2), we transform ${}^{2}Vl$ to F1 based on equation 4.1 as follows:

$$ {}^{1}Vl = {}^{1}T_2 \times {}^{2}T_4 \times {}^{4}Vl \tag{4.2} $$

Next, we model Pk head as a sphere:

$$ \|\mathbf{x} - \mathbf{c}\|^2 = r^2 \tag{4.3} $$

where $\mathbf{c}$ is the sphere center, r is the sphere radius, and $\mathbf{x}$ is a point on the sphere. Geometrically, any line can be defined as:

$$ \mathbf{x} = \mathbf{o} + d\mathbf{l} \tag{4.4} $$

where $\mathbf{o}$ is the origin of line, $\mathbf{l}$ is the direction of the line, d is the distance along the line from the line starting point, and $\mathbf{x}$ is a point on the line.

Finally, we check the intersection through searching for points that are on the line and on the sphere. Thus, we combine equations 4.3 and 4.4, solve them for d, and substitute: (i) Pk head position (${}^{1}F3$) as the sphere center; (ii) the head position of Pl *w*.r.t F1 (${}^{1}F4 = {}^{1}T_2 \times {}^{2}F4$) as starting point of the line, and ${}^{1}Vl$ as the line direction:

$$ d = \frac{-(\mathbf{{}^{1}Vl}.(\mathbf{{}^{1}F4} - \mathbf{{}^{1}F3})) \pm \sqrt{w}}{\|\mathbf{{}^{1}Vl}\|^2} \tag{4.5} $$

$$ w = (\mathbf{{}^{1}Vl}.(\mathbf{{}^{1}F4} - \mathbf{{}^{1}F3}))^2 - \|\mathbf{{}^{1}Vl}\|^2 \left(\|\mathbf{{}^{1}F4} - \mathbf{{}^{1}F3}\|^2 - r^2\right) $$

If the value of $w \in \mathcal{R}^+$, then there are two intersection points crossing the sphere and Pl is looking at Pk; otherwise the line is either tangent to the sphere or not passing through the sphere at all and Pl is not looking to Pk.

4.1.1.4 *Time variant LookAt squared matrix*

After evaluation the lookAt function for all possible combinations among the participants will be able to build a square matrix ($n \times n$, where n is the number of the participants) named *LookAt square matrix* (LAM). We need to call the lookAt function $n(n-1)$ times to fill the time variant Look_At square matrix as shown in Figure 4.4. If we sum the matrix over the video time we can determine the dominant speaker in the face-to-face as he will have the maximum value among the participants.

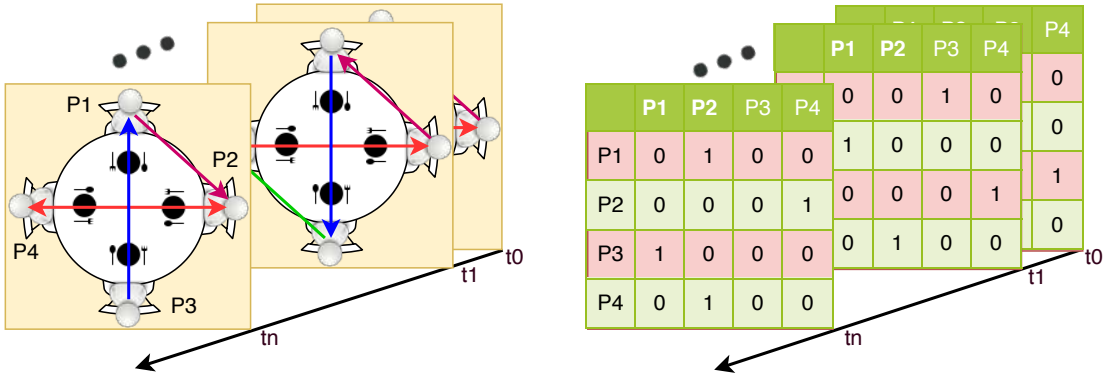


Figure 4.4: *Look At square matrix* example. Pi is the i^{th} person; on the table, the value of (x, y) is 1 if Px is looking at Py else it is 0.

4.1.1.5 *Eye contact detection*

An eye contact holds between two persons if $\mathrm{LookAt}(x, y) = \mathrm{LookAt}(y, x) = 1$. For example, in Figure 4.4, eye contact holds between P2 and P4.

4.1.1.6 *Experimental setup*

Dataset and Ground-truth. The adopted dataset in performing experiments has been recorded to study multi-person social interactions. It consists of 10 videos (average recording time is 20 minutes), and four participants in each video instructed to discuss a general conversational topic. The recording has been performed in a quiet office room equipped with four cameras as shown in Figure 4.1. Cameras have been slightly placed above the participants to provide a near frontal view of faces of all participants taking into account turning their heads during the conversation. To obtain the participants' gaze behaviour, we have asked five annotators to label the dataset with looking_At ground-truth. The annotators have identified for each participants whose face is being looked or not looked at a particular moment.

Performance Metrics. We treat the eye contact detection as a binary classification problem. Thus, we adopt various metrics to evaluate a classification model accuracy, precision, recall and F-Measure that has been introduced in Section 2.4.7.

Baseline. We define a baseline to compare our method with. The baseline reflects the results obtained when applying supervised machine learning algorithms on 18 features, divided as follows: (i) head pose of person P_i; (ii) head pose of person P_j; and (iii) world frame pose *w*.r.t. to camera frame reference. Weka tool [56] provides many learning algorithms. From these algorithms we exploit Naive Bayes, Random Forest, J48, and Artificial Neural Network (NN) to evaluate the performance of mentioned state-of-the-art features.
Parameters Setting. Our proposed method doesn't have parameters to be configured or may affect the results. Furthermore, the selected supervised learning methods in Weka tool are controlled by important parameters that may have impact on the classification performance. Thus, for the Naive Bayes method, we set the "useKernelEstimator" and "useSupervisedDiscretization" options to false value as default values set by Weka. For Random Forest, we set the option max depth to 0 (unlimited), with studying the effect of changing number of $trees \in \{20, 30, 100\}$. For J48 method, we set the minimum number of instances per leaf to 2, number of folds to 3, and confidence factor to 0.2. For neural network learning algorithm, we study the impact of having different numbers of hidden layers (from 1 to 4) each layer has 18 neurons.

Table 4.1: LookAt performance results of our proposed method (Geometrical approach) compared with multiple supervised approaches: Random Forest (RF), Random Tree (RT), J48, Naïve Bayes, and Neural Network (NN), in terms of Accuracy, Precision, Recall, and F-Measure for NotLooking class (0) and Looking class (1). Results averaged over 10 videos when performing 10-fold validation on each video.

	Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F-Measure 0	F-Measure 1
Proposed Method	78 %	85 %	59 %	85 %	59 %	85 %	59 %
RF (# Trees =10)	91 %	92 %	86 %	96 %	78 %	94 %	82 %
RF (# Trees =20)	92 %	92 %	87 %	96 %	80 %	94 %	84 %
RF (# Trees =30)	92 %	93 %	87 %	96 %	81 %	95 %	84 %
RF (# Trees =100)	92 %	94 %	87 %	95 %	83 %	95 %	85 %
RT	87 %	91 %	75 %	91 %	76 %	91 %	75 %
J48	89 %	93 %	79 %	93 %	79 %	93 %	79 %
Naïve Bayes	72 %	77 %	43 %	89 %	42 %	82 %	30 %
NN (#HL=1)	85 %	88 %	73 %	91 %	66 %	90 %	69 %
NN (#HL=2)	85 %	90 %	73 %	91 %	70 %	90 %	72 %
NN (#HL=3)	81 %	86 %	64 %	88 %	61 %	87 %	63 %
NN (#HL=4)	84 %	88 %	72 %	91 %	65 %	89 %	68 %

4.1.1.7 *Experimental results*

We have performed two types of experiments: (i) 10-folds cross validation at video frame level; (ii) and 10-folds cross validation at video level. The main purpose of the first type is to study the impact of performing training a set of frames and testing on other set of frames where both sets are related to same video. At higher

level, the second type of experiments give a strong indication about any possible dependency among same video frame level and different video levels. Table 4.1 reports the results of performing 10-fold cross validation at single video frame level, while Table 4.2 reports 10-folds cross validation at video level. The 10-fold cross validation is performed for each video with producing performance results in terms of the mentioned metrics. The ultimate performance result value for the first type is averaged over the entire video data-set. The results of first type of experiments show that the supervised learning-based methods have generally high classification performance compared to our geometrical proposed method in terms of accuracy metric.

Table 4.2: 10-fold cross validation (video level) Looking_At performance results of our proposed method (Geometrical approach) compared with multiple supervised approaches: Random Forest (RF), Random Tree (RT), J48, Naïve Bayes, and Neural Network (NN), in terms of Accuracy, Precision, Recall, and F-Measure for NotLooking class (0) and Looking class (1).

	Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F-Measure 0	F-Measure 1
Proposed Method	**78** %	**85** %	59 %	85 %	**59** %	85 %	**59** %
RF (# Trees =10)	76 %	78 %	60 %	95 %	22 %	85 %	32 %
RF (# Trees =20)	77 %	78 %	65 %	94 %	21 %	**86** %	32 %
RF (# Trees =30)	77 %	78 %	66 %	**96** %	23 %	**86** %	33 %
RF (# Trees =100)	77 %	78 %	**68** %	**96** %	22 %	**86** %	33 %
RT	69 %	79 %	41 %	79 %	42 %	79 %	41 %
J48	71 %	81 %	45 %	76 %	45 %	81 %	45 %
Naïve Bayes	65 %	74 %	25 %	83 %	15 %	78 %	17 %
NN (#HL=1)	76 %	80 %	56 %	90 %	36 %	85 %	43 %
NN (#HL=2)	78 %	83 %	59 %	88 %	47 %	85 %	52 %
NN (#HL=3)	77 %	83 %	58 %	87 %	51 %	85 %	53 %
NN (#HL=4)	72 %	80 %	46 %	82 %	44 %	81 %	44 %

The Random Forest learning method with different numbers of trees provides almost high classification performance in terms of accuracy and other class-based metrics. These results are expected since the nature of Random Forest is building many random trees (acting as uncorrelated experts), then voting among the trees to provide the ultimate prediction value. The high variation in the precision values of NotLooking class, compared to Looking one, shows that the dataset adopted in training is unbalanced at the class level, making the classification model biased towards a particular class, which is NotLooking class in our case. The results of the first type of experiments show that supervised learning-based methods are the winner in providing accurate and precise classification LookingAt model. However, the introduced results in Table 4.2 of the second type of experiments provide different conclusions: (i) training a classification LookAt model on a video is not necessary to perform very well on other video (social experiment), raising concerns about the degree of sensitivity when participants change their sitting/arrangement around the table; (ii) from the machine learning perspective,

the accuracy decreasing in the supervised-based learning methods results shows an over-fitting problem occurred, meaning that the classification models of first type experiments are not generalized enough to cover all patterns of looking among participants.
According to the results of second type experiments, our method outperforms most of the supervised classification models in terms of accuracy and other class-based metrics. Indeed, the key features of our proposed method are: (i) no prior training dataset required and thus avoiding the annotation step since it is time consuming; (ii) it has classification performance almost at the same level with supervised-based ones; and (iii) it does not require any intrusive devices.

4.2 Deep model for French food classification

As we presented in Section 2.2, the main goal of OVALIE platform is to study individuals' eating behavior in different contexts. Thus, having a model for recognizing what they are consuming (eating) is crucial for eating habits analysis and other health-care applications as introduced in Section 3.3. Also, such a model is useful for content-based retrieval for businesses based on the food industry. For instance, we can create an automated dietary planar application based on the requirements of the user and retrieve relevant images and recipes for the appropriate food items.
The effectiveness of Deep Convolutional Neural Network (DCNN) have been proved for large-scale object recognition at ImageNet [113]. However, food classification is a challenging problem due to the large number of categories, high visual similarity between different foods (inter-class similarity), food photos within the same class may have significant variability (intra-class diversity) as shown in Figure 4.5, as well as the lack of datasets for training state-of-the-art deep models.

Figure 4.5: Examples of "crepe" that shows intra-class diversity.

4.2.1 *Dataset collection*

Since OVALIE platform is located in France, we utilize this fact by selecting the relevant data. And so, we collected and cleaned a dataset for **French** food. This dataset contains 16373 food images divided into 37 classes as listed in Table 4.2.1, samples from our dataset are shown in Figure 4.6.

Table 4.3: List of the 37 classes in our French food dataset.

Aligot	Blanquette de veau	Bouillabaisse
Caneles	Coq au vin	Endives au jambon
Escargots de Bourgogne	Huitres	La Choucroute Garnie
La crepe	La galette bretonne	La piperade
La pissaladiere	La potee	La poule au pot
La quiche Lorraine	La salade nicoise	La soupe a l'oignon
La tarte Normande	La tarte aux Maroilles	La tartiflette
Le Paris-Brest	Le boeuf Bourguignon	Le cassoulet
Le clafoutis	Le couserans croustade	Le gratin dauphinois
Le hachis parmentier	Le lonzu	Le magrets de canard
Le pot au feu	Le rougail saucisses	Le steak tartare
Les boles de picolat	Moules	Paupiettes
Petits pates de Pezenas		

4.2.2 *Methodology*

As we discussed earlier about transfer learning in Section 2.4.5, we can perform fine-tuning on a pre-trained DCNN and fine-tune it to perform image classification and recognize classes it was never trained on (our French food dataset).
As shown if Figure 2.11, fine-tuning requires that we not only update the CNN architecture but also re-train it to learn new object classes. It is include the following steps:

- Eliminate the last layer of the DCNN (The fully connected layer, where the class label predictions are made).
- Replace the fully connected layer with new one.
- Freeze earlier convolutional layers in the network.
- Start training for the new fully connected layer.

Figure 4.6: Typical examples of our French food dataset.

Hence, we change the size of the last output layer to the same number as the number of the food classes (37 classes), and randomly initialize the weights of the new layer. In the following, we provide an overview about the ResNet50 and ResNet18 [59], DenseNet201 [62], and InceptionV3 [128], we find-tuned these models using our French food dataset.

4.2.2.1 *Residual Neural Network (ResNet)*

When deeper networks starts converging, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated and then degrades rapidly. Instead of optimizing the direct mapping of $x \rightarrow y$ with a function $H(x)$, we can optimize the residual mapping function $F(x)$. Let us define the residual function using $F(x) = H(x) - x$, which can be re-framed into $H(x) = F(x) + x$, where $F(x)$ and x represents the stacked non-linear layers and the identity function(input=output) respectively as shown in Figure 4.7. It is easy to optimize the residual mapping function $F(x)$ than to optimize the original, unreferenced mapping $H(x)$ [59].

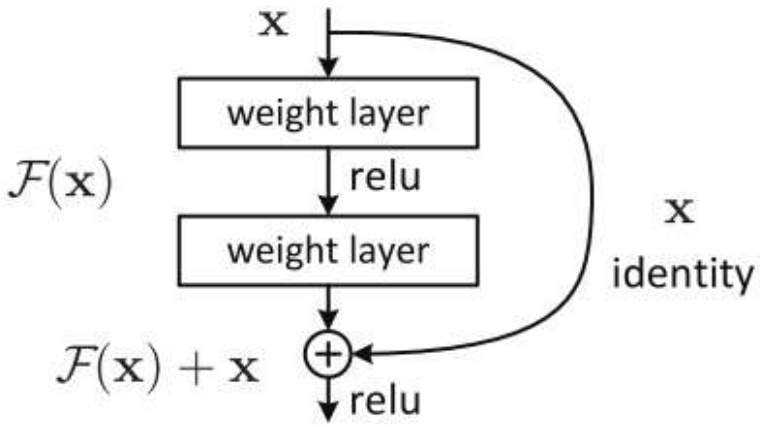


Figure 4.7: Residual learning: a building block [59].

4.2.2.2 *Densely Connected Convolutional Networks (DenseNet)*

DenseNet is a new CNN architecture that reached State-Of-The-Art (SOTA) results on classification datasets (CIFAR [70], SVHN [90], ImageNet [37]) using less parameters since it uses residual, which allows to go deeper than the usual networks. It is composed of **Dense blocks**, within each block the layers are densely connected with each other, where each layer take all outputs of the previous layer as its inputs as shown in Figure4.8.

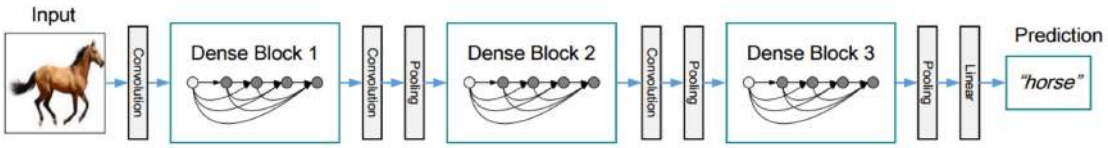


Figure 4.8: A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling [62].

4.2.2.3 *Inception-V3*

Inception v3 is a common used model for image recognition,it achieved greater than 78.1% accuracy on the ImageNet dataset.
The model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers as shown in Figure 4.9.

4.2.3 *Experimental results*

In our experiments, we used our French food dataset, 70% were used for training, 15% for validation, and for evaluation 15%, also we use the data augmentation

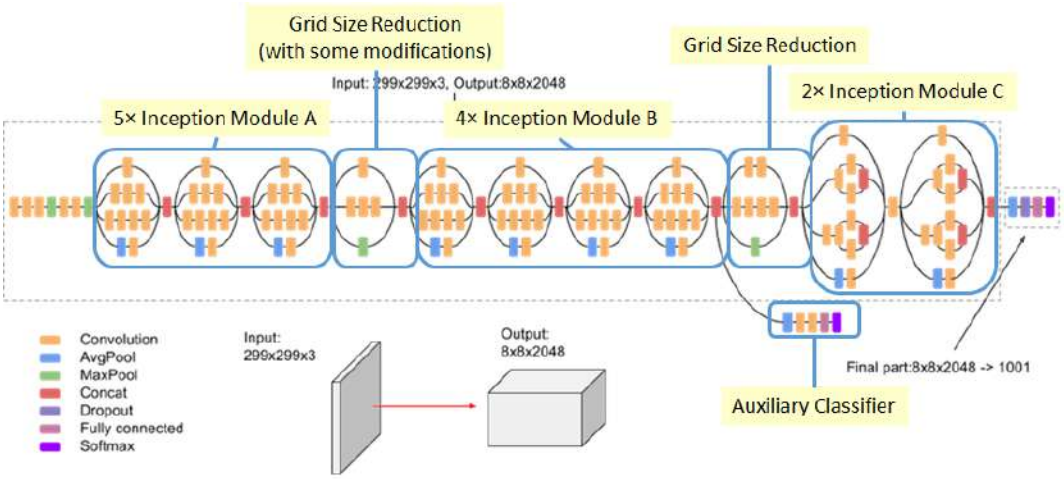
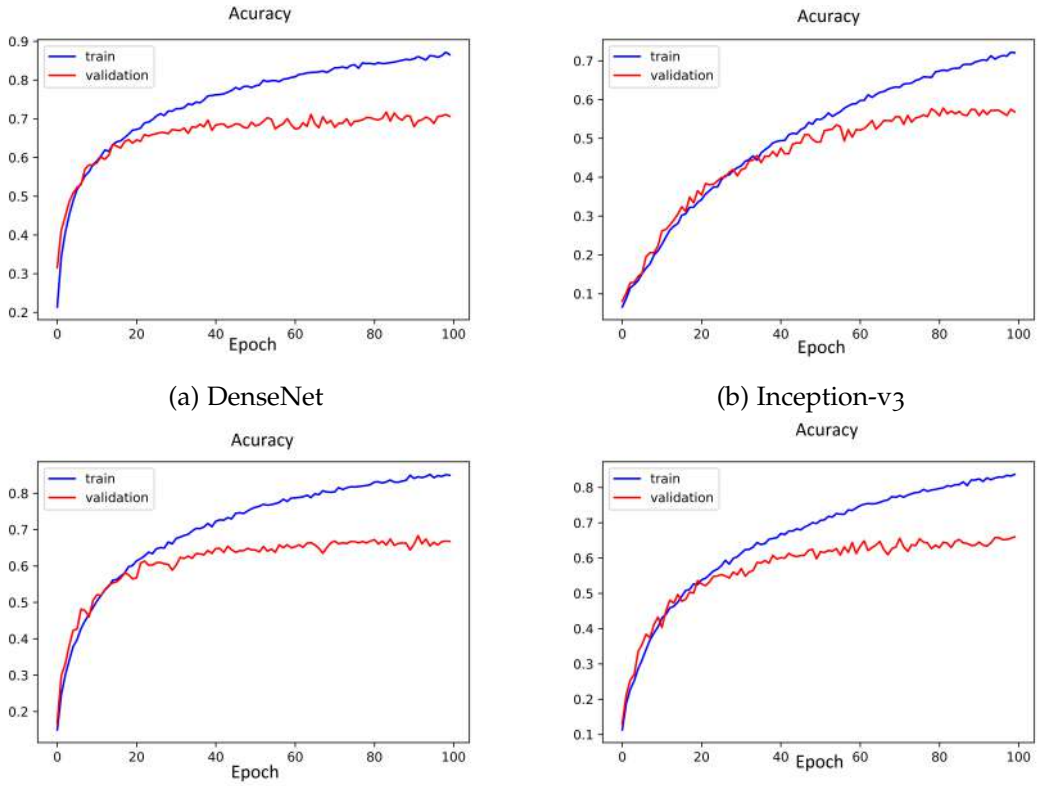


Figure 4.9: Inception-v3 Architecture [130].

techniques to increase the size of our data, as explained in Section 2.4.6. As shown in Figure 4.2.3, the models accuracy increased over each epoch, overfitting started at around 20 epochs for denseNet and resNet models, and around 35 epochs for Inception-v3 model.



(a) DenseNet

(b) Inception-v3

(c) ResNet18

(d) ResNet50

Figure 4.10: Training and validation accuracy of the fine-tuned models.

Table 4.4: Top–1 and Top–5 performance test on our French food dataset achieved by fine-tuned DCNN models. Best results are highlighted in boldface font

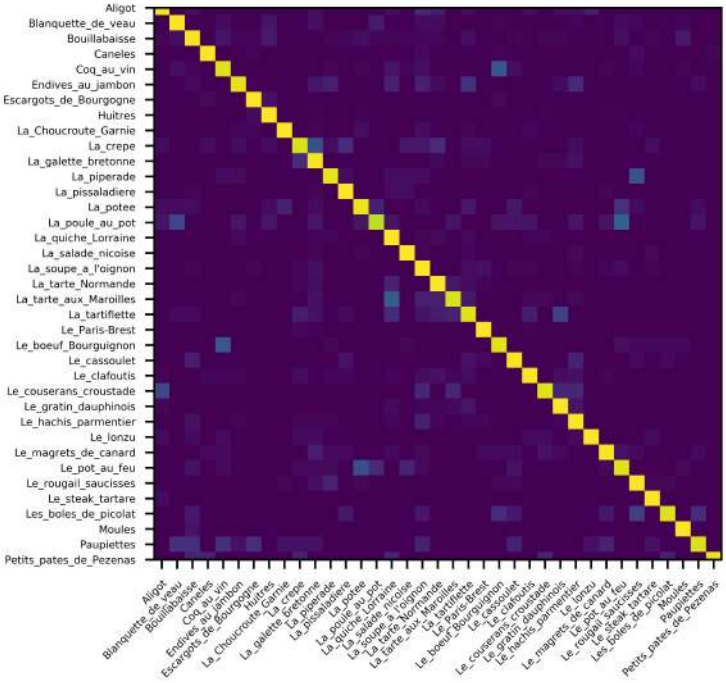
Model	Top-1 (%)	Top-5 (%)
DenseNet201	**75.23**	**92.40**
InceptionV3	63.91	88.97
ResNet18	71.35	92.24
ResNet50	71.71	91.15

Table 4.2.3 shows the results achieved by fine-tuned DCNN models. We got the best results using DenseNet201 model with 75.53% top-1 accuracy and 92.4% top-5 accuracy.
Figure 4.11 shows some wrong predictions using the tuned model. We can notice that there is a visual similarity between the classes (inter-class similarity). Also, we can notice that the models have missed classification on average between 5 classes, as shown in Figure 4.13. However, saying wrong prediction is subjective. For example, if we take the first row in Figure 4.11 the target class is "la crepe" and the first wrong prediction is "la galette bretonne", it was a wrong prediction since in the collected dataset "la crepe" and "la galette bretonne" are considered as two different classes, but some people consider "la galette bretonne" as synonym of "la crepe".

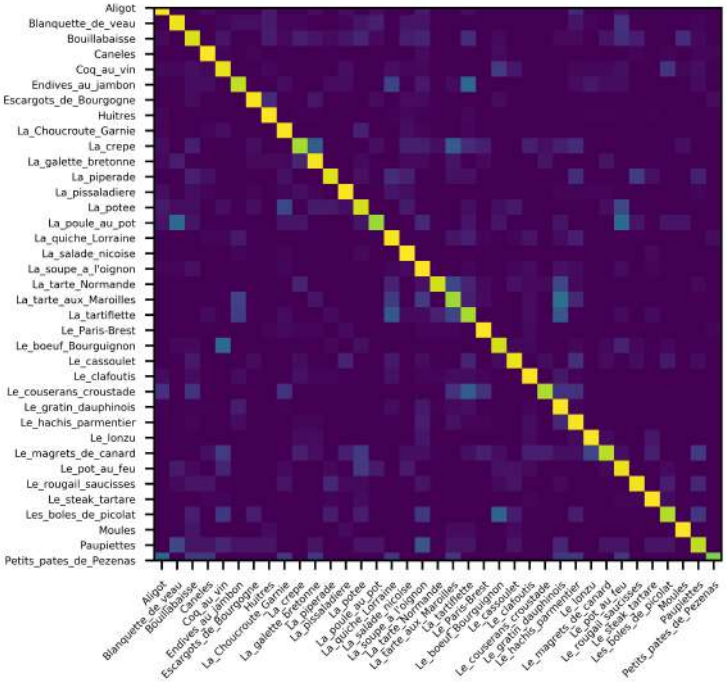
Figure 4.11: Some examples of wrong prediction. First column show the target class, the second one show correct prediction, and the rest shows wrong prediction.

As shown in Figure 4.12, the models have missed classification on average between 5 classes and it is compatible with the higher top-5 accuracy. This might be explained by that fact that foods have similar texture and color. On other words, food photos within the same class may have significant

variability (intra-class diversity), and photos from different classes have visual similarities(inter-class similarity).

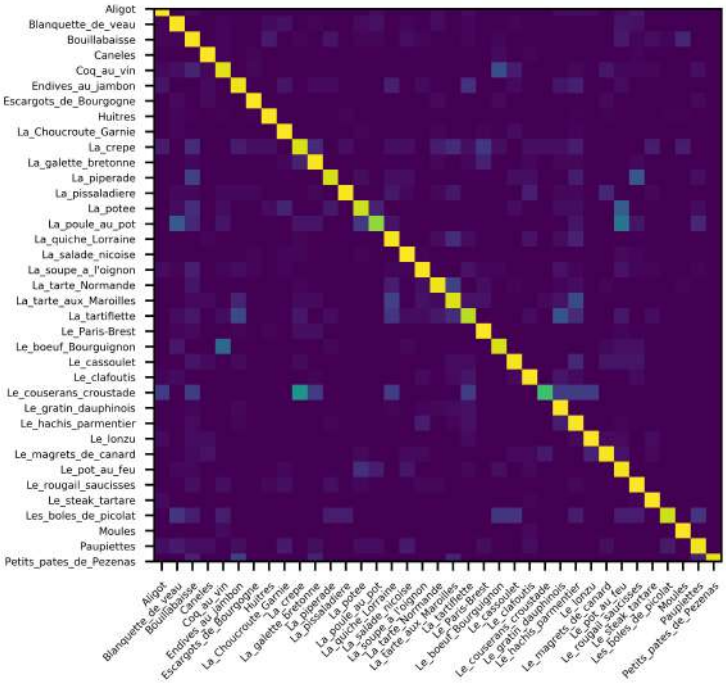


(a) DenseNet

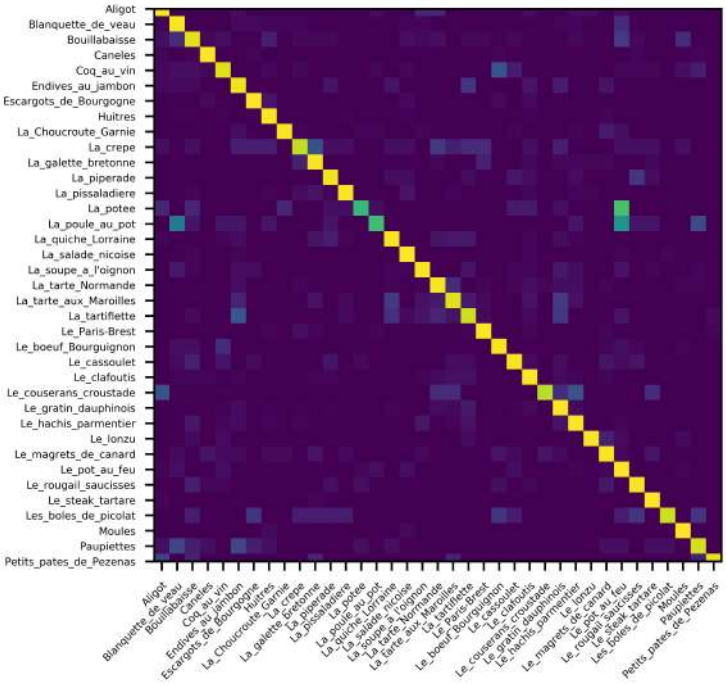


(b) Inception-v3

Figure 4.12: Confusion matrix of the fine-tuned models.



(c) ResNet18



(d) ResNet50

Figure 4.12: Confusion matrix of the fine-tuned models (cont.).

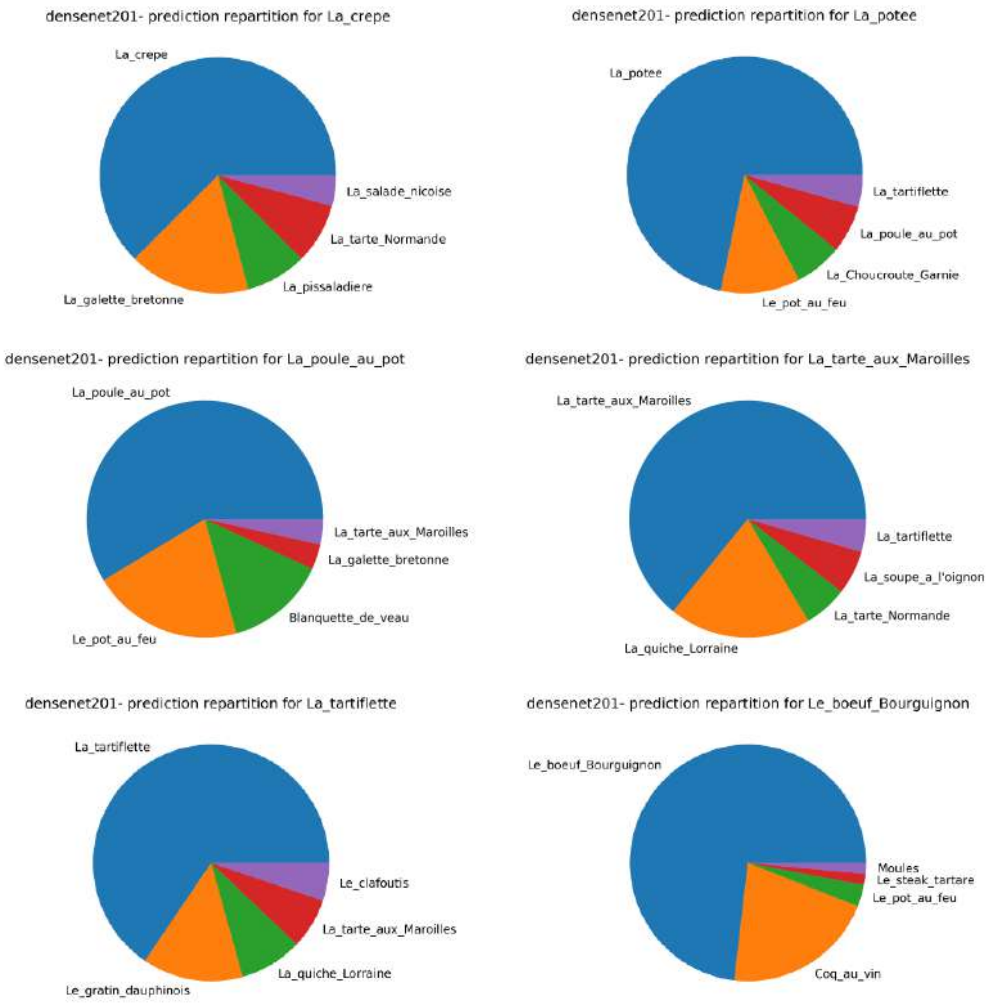


Figure 4.13: Some examples of prediction repartition per class of the denseNet.

We can notice that some of the miss classified plates are distinguishable by the dining event time-line (multicourse meal sequence). For example, three-course meal consists of "Soup/Salad", "Main Course", and "Dessert". If we associate the meal time line with the detected food images then we will be able to reject a prediction of "chocolate mousse" during the "Main Course" time slot. Another example, assuming that we have a full menu from the location where we collect the food images from, we can use this menu to filter out all the wrong predictions (predictions that are not exist in the menu). These examples open a question, can we use the contextual information to add a semantic side to our deep model?

4.3 Conclusion

This chapter presents two methods to extract features ((meta)data) from the context. First, we propose a novel geometric-based method to detect eye contact in natural multi-person interactions without the need for eye tracking devices or any intrusive, which allows recording natural social behavior. We evaluate our method on a recent dataset (10 social videos, where each video is 20 minutes long) of natural group interactions, which we annotated with LookAt() ground truth, and showed that it is highly efficient with regards to classification performance, and comparing to the classical supervised eye contact detection methods. Eye contact detection could be used to analyze turn-taking, social roles, and engagement during multi-person interactions.
Second, we propose find-tuned deep models (DenseNet, ResNet, and Inception-v3) for food classification that will be used to extract contextual information from the records (e.g., type of food in the social event). We evaluate the tuned models using a new dataset that we collected for 37 types of French food. Results show that the DenseNet achieved the best top-5 accuracy, and it opens a question related to the possibility of considering contextual information to enhance the accuracy of the *non-semantic* deep models.

5

TOWARDS A HOLISTIC APPROACH (FRAMEWORK) FOR SOCIAL INTERACTION ANALYSIS

A system is a network of interdependent components that work together to try to accomplish the aim of the system.

— W. Edwards Demings

We propose a generic framework architecture that integrates various components and methods together in order to automate the social interaction analysis in the context of observational study.

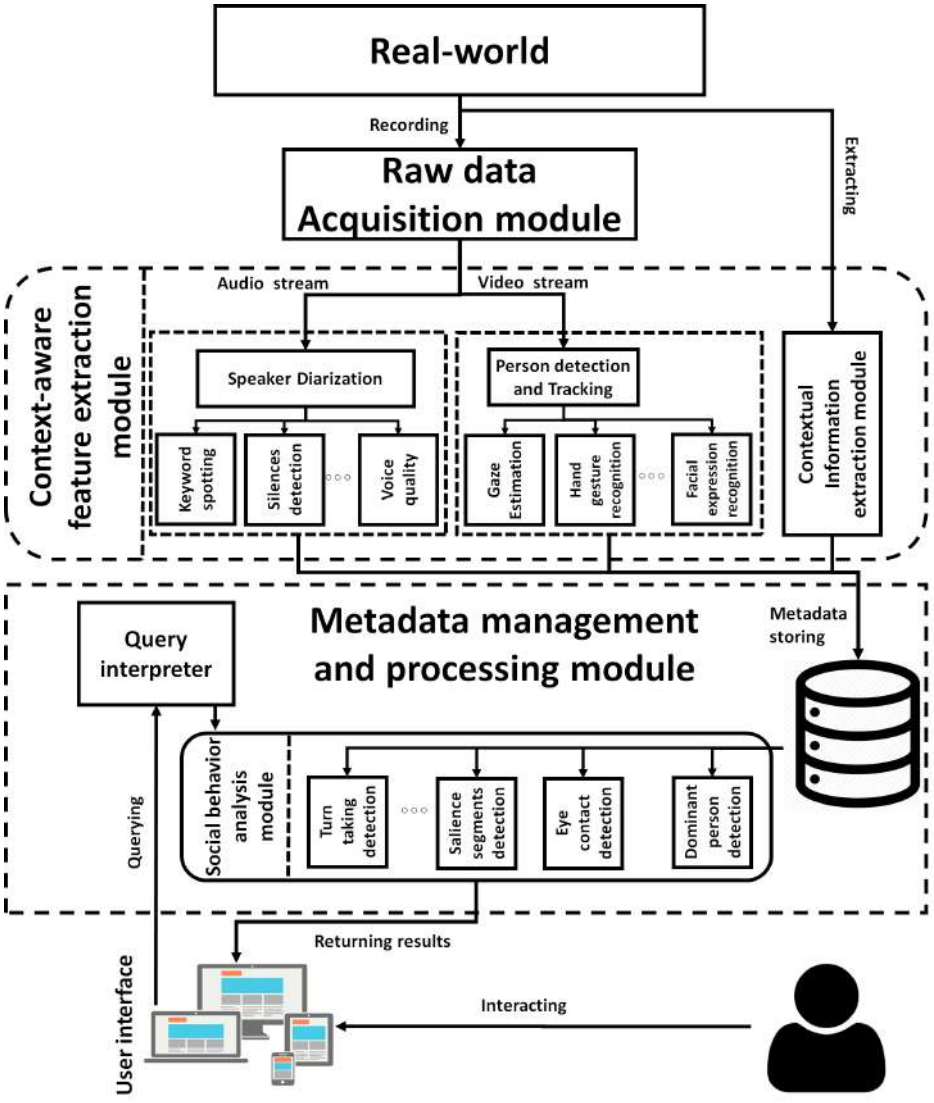


Figure 5.1: Social interaction analysis framework architecture.

The proposed framework architecture consists of: (i) raw data acquisition module to digitize the real-world environment as raw data (audio, video, temperature, etc.); (ii) context-aware feature extraction module to extract the (meta)data from the real-world and recorded data; (iii) (meta)data management and processing module; and (iv) user interface for results and (meta)data visualization, as shown in Figure 5.1.

5.1 Raw data acquisition module

We used internet of thing (cameras, microphones, temperature sensors, etc.) to digitize the real-world environment as raw data (audio, video, temperature, etc.) as shown in Chapter 6. Before start recording or collecting the data, we need to setup the sensors to ensure that we have high-quality data (e.g., frontal face photo to have more accurate facial expression prediction). However, to record the natural behaviors of the data subjects, we need to keep the sensors as far as possible from them.

5.2 Context-aware feature extraction module

Context-aware feature extraction methods include contextual information collection methods and context-based features extraction methods. There are two types of contextual information. First, physical context such as date, location, type of event, dishes, etc. Second, social context such as the data subjects' social media which can be summarized by content (videos, audio, texts, etc.), relationships (friends, followers, groups, etc.), and events (birthdays, parties, etc.). The relationship between all this information is represented in the (meta)data model shown in figure 5.2. For example, we purposed a French food classification deep model (see Section 4.2) to extract (meta)data from the environment.
As shown in figure 5.1 the context-based features extraction methods are deployed in parallel, which will reduce the processing time in case we use more hardware. For the video stream, first, we detect the experiment participants using openFace toolkit [15]. Second, we track them within the videos and apply geometrical filtering over the detect persons to minimize the detection redundancy. Third we used Affectiva Software Development Kit (SDK) [85] to extract participants' facial expressions, openFace toolkit for gaze direction estimation, and we can use more visual-based methods to detect more visual cues. On the other side, we need to process the audio stream in order to determine the speakers by applying speaker diarization [138], then we can apply different methods in order to obtain the turn taking [141], or detect laughing [77], etc.

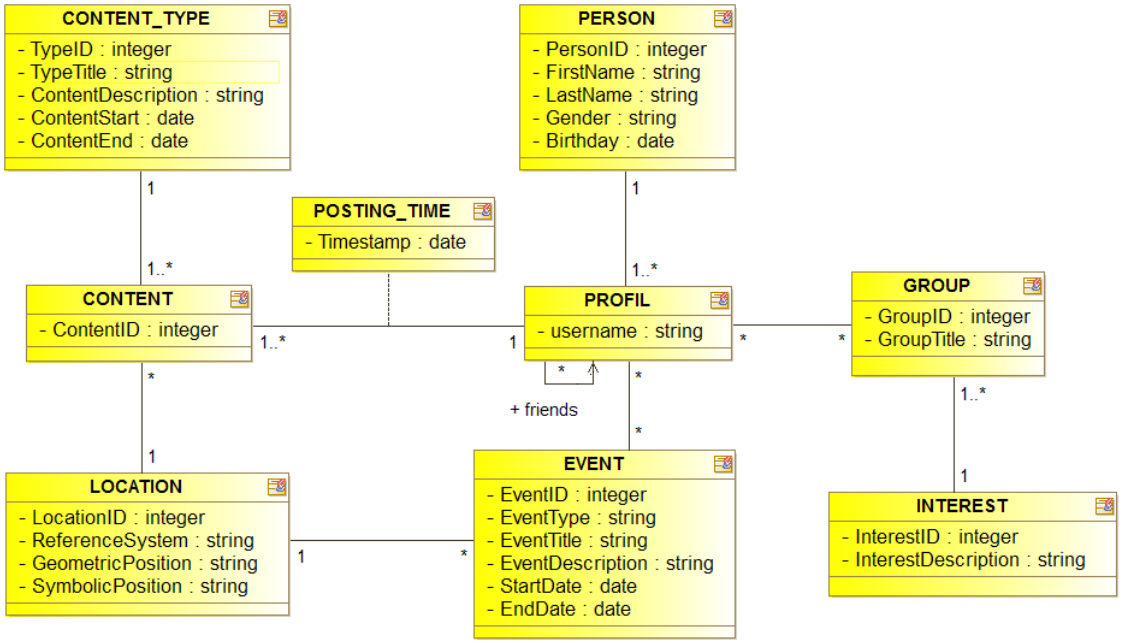


Figure 5.2: Social media (meta)data model.

5.3 (Meta)data management

To handle the high variety of the social cues, we propose a comprehensive (meta)data model for the visual nonverbal cues [106]. This model consists of four groups of entities: (i) acquisition group to store the used sensors' metadata (e.g., owner details, model number, transmission mode, data format, etc.); (ii) experiment group used to store the experiment's description including title, data, responsible person, and location, also the list of algorithms that are used to extract the social cues; (iii) video group used to store metadata related to the recorded video such as segments start/end timestamps, and frames information; and (iv) features group to store the extracted social cues for each detected person in a given *conceptual* frame (*conceptual* frame is multiple frames that have a common timestamp and have to be analyzed together) as shown in figure 5.3.

5.3.1 *Experiment group*

Researchers are interested in performing experiments using different configurations of algorithm types and parameters. Thus, the "Experiment" class is dedicated to hold simple information describing the experiment, including title, data, responsible person, location, and description metadata. An experiment contains a set of videos along with the list of algorithms that used to extract the social cues.

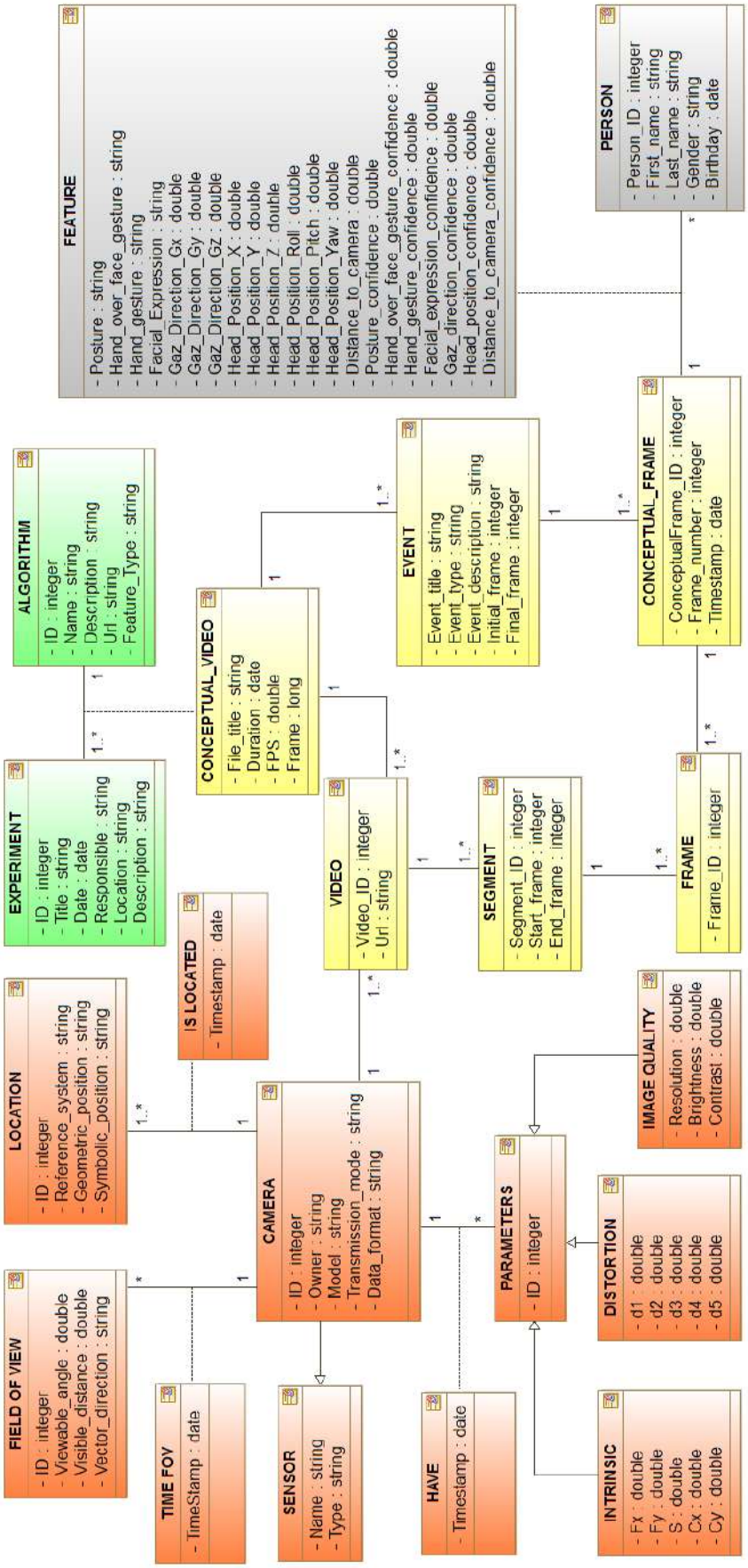


Figure 5.3: Generic (meta)data model.This generic data model for visual nonverbal social cues shows the relationships that exist between experiment, acquisition, video, and feature groups of entities, which are color-coded as green, orange, yellow, and gray respectively.

5.3.2 *Acquisition group*

Generally, different types of sensors (e.g., camera, GPS, IMU, and microphone) are used for social cues (verbal/nonverbal) acquiring. In the context of social interaction analysis cameras are widely adopted, so we cover the camera relevant information within this group in which "CAMERA" class contains attributes for holding information about the adopted camera(s) in conducting experiments. These attributes include the identity number (e.g., 58395FX), owner (e.g., IRIT), model (e.g., Axis F44 Dual Audio), transmission mode (wired/wireless), and data format (e.g., .mp4) of the camera. Cameras are controlled by *time invariant* parameters at different frequencies, while these parameters include camera intrinsic parameters, location, field of view, distortion, and image quality. Thus, we propose separated classes for each one of them as follows: (i) INTRINSIC class attributes include camera focal length (Fx,Fy), image sensor format(S), and principal point (Cx,Cy), (ii) LOCATION class contains a system reference as well as the symbolic and geometric (extrinsic camera parameters) position. In computer vision methods the intrinsic and extrinsic camera parameters are used in the computation of the camera projection matrix, (iii) FIELD OF VIEW class contains the attributes (viewable angle, visible distance, and FOV direction) that used to determine how wide an area of a camera field of view, (iv) DISTORTION class has five attributes that are used for lens distortion correction, (v) and the IMAGE QUALITY class include common image features as resolution, brightness, and contrast.

5.3.3 *Video group*

The classes VIDEO, SEGMENT, and FRAME represent a decomposition relationship as a video clip decomposing into segments which represents sequence of frames. An event is an action involving content items at a particular place and over a particular time interval (e.g., type of the played music, intensity of illumination). So, a video clip could be decomposed into events that contains a sequence of frames. Although the event is similar to the segment, but the event time interval can be longer or shorter than the video segment. So the EVENT class is directly related to the VIDEO in our model. A conceptual frame is representing one to N frames (N is the number of the adopted cameras within the experiment) that have a common time stamp and have to be analyzed together. A conceptual video is pointing to one to N videos. Thus, we introduce them to handle metadata fusion at the *frame-level* within the multiple cameras views scenarios.

5.3.4 *Features group*

The "FEATURE" class is designed as an association class containing the extracted social cues. The attributes of feature class are extracted for every detected person

in a given conceptual frame. "PERSON" class contains information (name, age, and birthday) about the experiments' participants.

5.4 Social behavior analysis module (Multi-layer aggregation)

To go deeper in the social behavior analysis we need to be able to analyse the stored social cues (e.g., gaze direction, head pose, facial expression) in the proposed (meta)data model in Section 5.3. So, we need to define a common representation for the retrieved social cues to aggregate them together and derive a new conclusions at the behavioral level of the detected persons.
Thus, we represent the retrieved social cues as layers at different time scales. In our context, the layer is the representation of the social cue as time series. Figure 5.4 shows facial expression, gaze direction, or speaking not speaking cue of a person during a dinner, eye contact between two person during the dinner, the dinner time line (e.g., French dinner include "entry course", "main course", "desert course") layers.

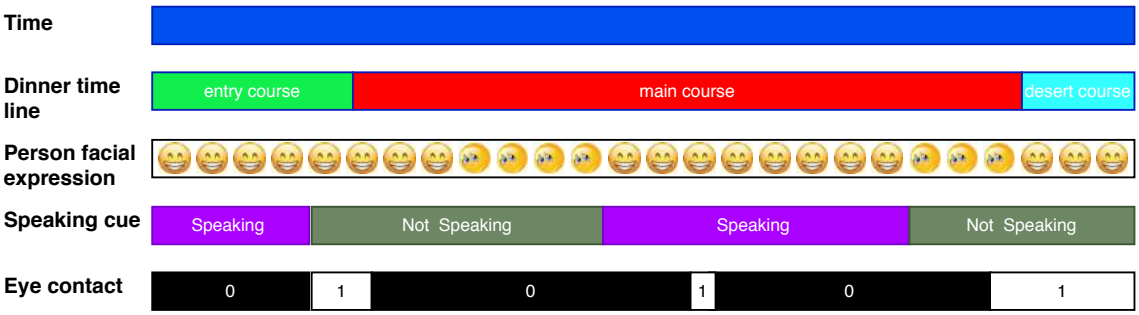


Figure 5.4: Social cues representation as multi layer.

Before the multi-layer aggregation, we need to have a common time scale (minimum time period) between the layers by scaling down or scaling up the minimum time period as shown in Figure 5.5. In the scaling up example, we can notice how the minimum time period is increased from one scale to another and the cues values are assigned based on the majority.

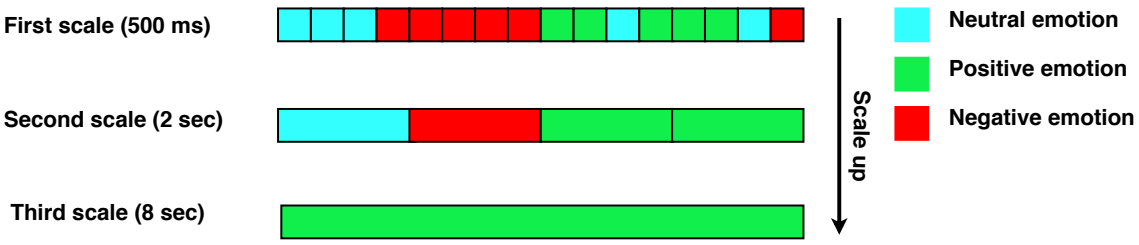


Figure 5.5: Emotion layer scaling up example.

The multi-layer aggregation analysis is used to provide more useful information and elicit conclusions at the behavioral level of the detected persons. For example, if we observe a focus group meeting, and we would like to detect the dominant speaker, we can aggregate multiple layers which may include participants' facial expression layer, participants' speaking cue layer, received gaze cue (received gaze by others participants) layer, social relationship between the participants (we can represent it as a layer with constant value) layer, and more layers based on the deployed methods for cues detection. Furthermore, we can consider the event time line (discussed topic time line) as an additional layer, this will allow determining the dominant speaker per topic not only for the whole event.

5.5 Visualization tools

Our approach displays the analysis result using a customized visualization tool as shown if Figure 5.6. It shows multiple view records aggregated with detected persons' facial expression, gaze direction, and eye contact.

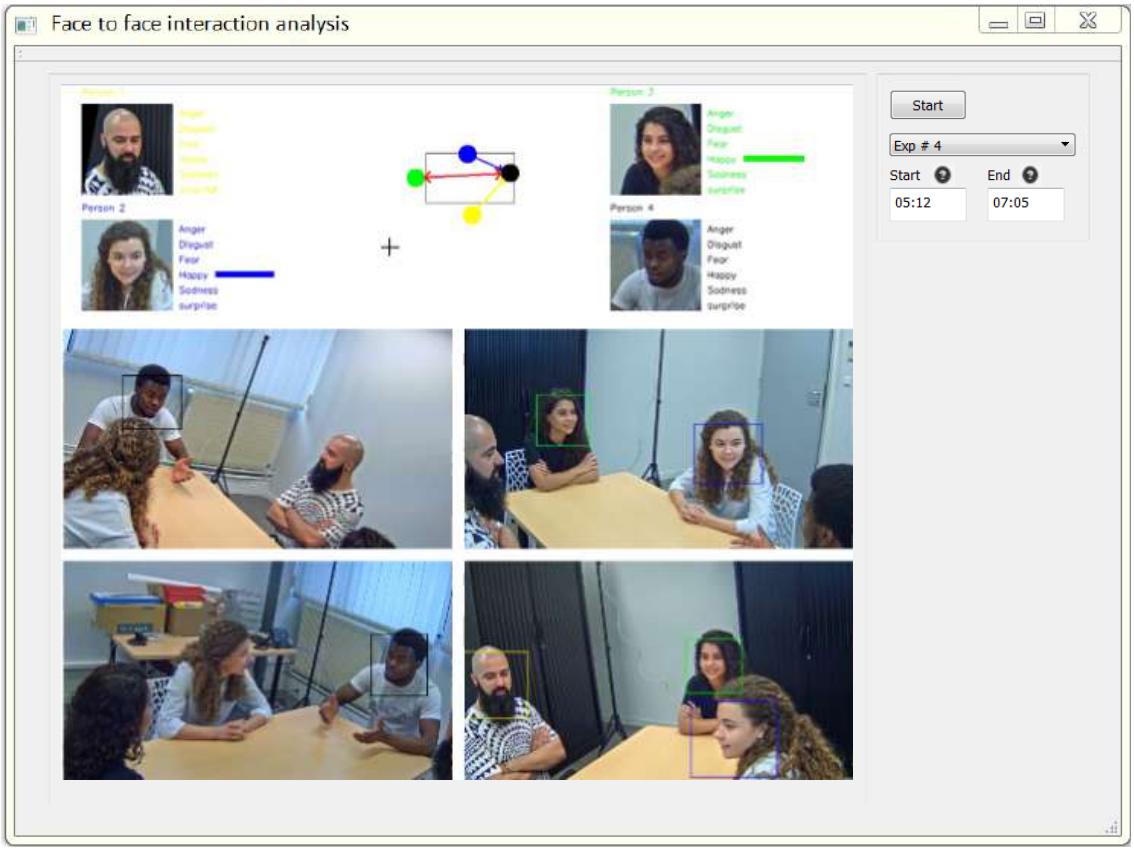


Figure 5.6: Face-to-Face visualization tool.

Furthermore, we visualize the extracted (meta)data statistics and (meta)data aggregation using Kibana (the visualization plugin of the search engine ElasticSearch [54]), as shown in Figure 5.7, the Mongo database was synchronized

with an ElasticSearch cluster using Monstache [1] which performs real-time synchronization of MongoDB replica sets to Elasticsearch clusters.

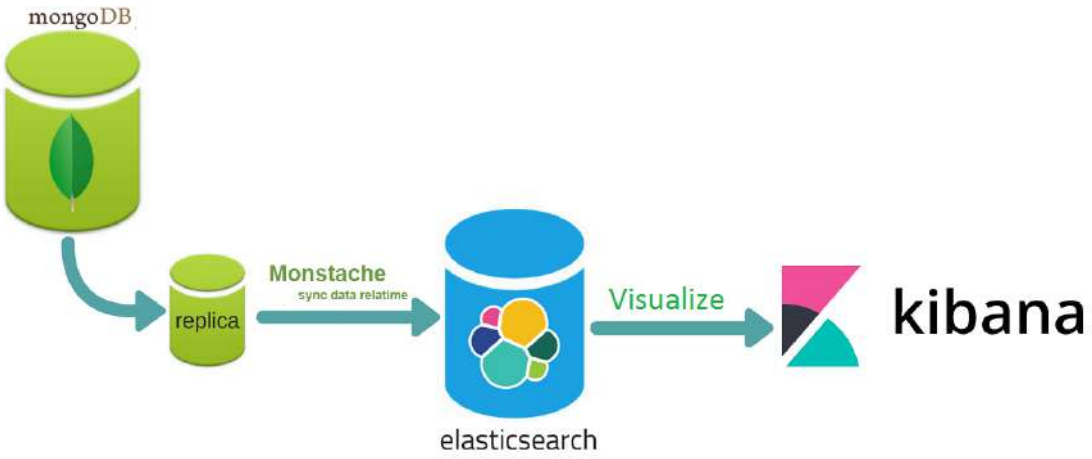


Figure 5.7: (Meta)data visualization pipeline using kibana.

Then we use Kibana as a graphical user interface to construct graphical visualizations to aggregate and visualize the social interaction (meta)data (charts, graphics, metrics, time-series, etc.), interact with these visualizations to select, filter and navigate through (meta)data, and perform social interpretations of the observed scene, as shown in Figure 5.8.

Figure 5.8: Statistics visualization using Kibana.

5.6 Conclusion

In this chapter, we propose a holistic framework architecture for social interaction analysis. The architecture allows the integration of various components and

methods. The integrated components can be grouped into (meta)data sources group, (meta)data processing and management group, the visualization tools group.
The (meta)data sources group includes raw data acquisition components for recording the experiments, contextual-based methods for extracting (meta)data from the real environment contextual information, and features extraction methods to extract (meta)data from the raw data.
The (meta)data processing and management group consists of two main parts first, a (meta)data repository with a comprehensive (meta)data model that handles the heterogeneity of collected, extracted, and processed (meta)data. Second, social behavior analysis module (multi-layer aggregation) that allow for dynamic analysis of the social cues and elicit conclusions at different time scales.
The visualization and aggregation tools group reduces the required time for eliciting conclusions from social events by human operators. Also, the aggregation of contextual information (extracted from the scene or collected from social networks) is used to bridge the gap between human-based and machine-based social interaction analysis processes.

6

EXPERIMENTAL ENVIRONMENT (OVALIE PLATFORM)

To experiment our approach and to perform "in vivo" social events capture and record without predefined scenarios, we use OVALIE platform to collect the related raw data from the target scene (audio, video, temperature, etc.). In the following sections, we describe the OVALIE platform plan and design, data acquisition module, and study of eating behavior of children, adolescents and adults with PWS.

6.1 OVALIE platform floor plan

In the context of naturalistic observation studies, we focus on modularity of the physical context in order to facilitate the creation of different observational environments: restaurant, dining room, hospital room, focus group room. Figure 6.1 shows a floor plan of OVALIE platform. In the floor plan, calibration area, secure area, restaurant area, kitchen, focus group area and adjustable area are respectively highlighted using red, pink, yellow and green rectangles.

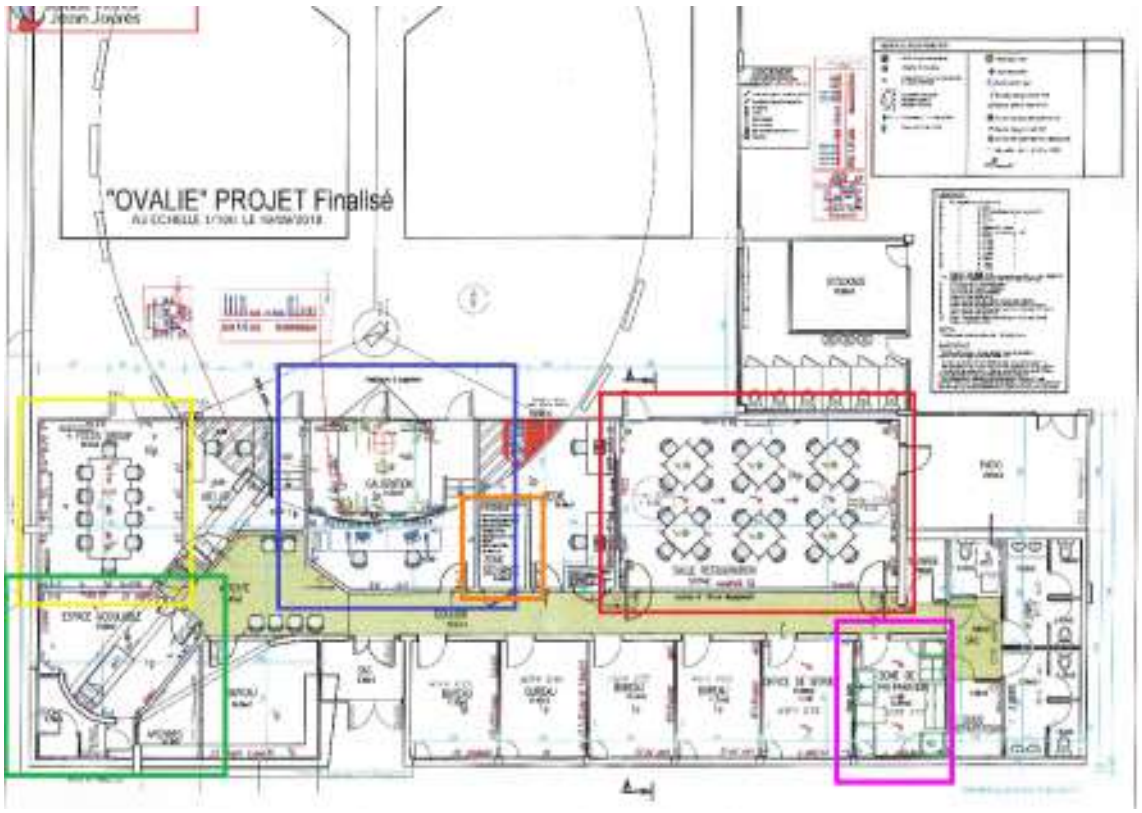


Figure 6.1: OVALIE floor plan.

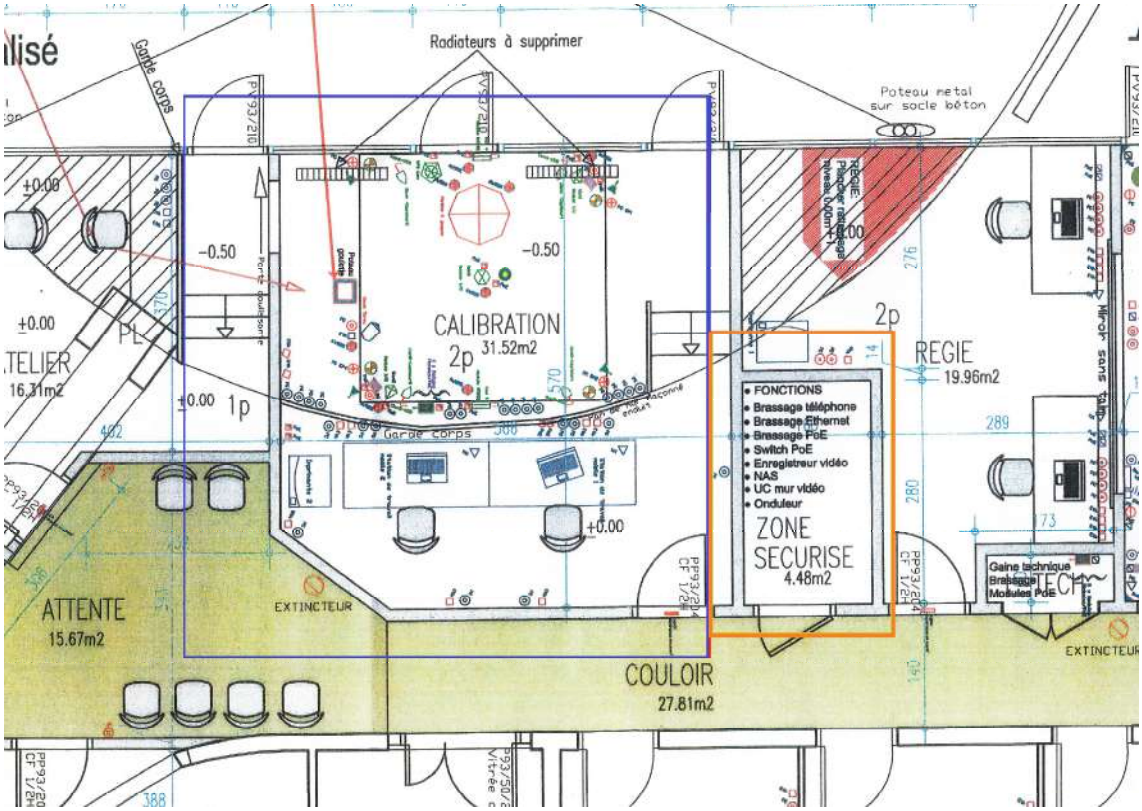


Figure 6.2: Calibration and secure areas floor plan.

Figure 6.2 shows the calibration and secure areas floor plans, the calibration area contains a mobile structure (tent) with five cameras are used for prototyping, validating and testing purposes before deploying in the other areas. Network-attached storage (NAS), which contains the recorded data, is located in the secure area. For security reasons, OVALIE platform has alarm system with personal key for disactivating the alarm in addition to electronic badge to ensure that only the authorized people can enter OVALIE.

Figure 6.3: Sample from one ceiling camera fixed inside the kitchen.

The kitchen is equipped with multiple ceiling cameras in order to observe the food preparation process (cooking), as shown in Figure 6.3

As shown in Figure 6.4, the restaurant area contains six tables, each table is surrounded by four ceiling cameras plus one camera over the center of the table.

Figure 6.4: Restaurant area floor plan.

Currently, the adjustable place is redesigned to be like a hospital room, as shown in Figure 6.5 there are multiple ceiling camera too in order to observe the scene from multiple views. Focus group area is the last area in OVALIE, a focus group

Figure 6.5: Hospital room setup in the adjustable area.

is a demographically diverse group of people whose reactions are studied in

guided or open discussions about a new product or something else to determine the expected reactions from a larger population [45]. In the social sciences and urban planning, a focus group allows members to interact and influence each other during the discussion and consideration of ideas [81]. It allows more natural conversation pattern than typical one-to-one interview. Figure 6.6 shows a sample from the multiple views for the focus group observation area in OVALIE.

Figure 6.6: A sample from the multiple views for the focus group observation area in OVALIE.

6.2 Raw data acquisition module

Raw data acquisition module is an internet of things we use to digitize the real-world environment, so we used different types of cameras, microphones, NAS, PoE switches, and AXIS Camera Station software for recording different areas in OVALIE. In the following, we will introduce the cameras, the microphones, and the AXIS Camera Station software that have been used within OVALIE platform.

6.2.1 *Axis cameras*

An Internet Protocol (IP) camera, it is a digital video camera that receives control data and sends image data via the Internet. They are commonly used for surveillance. Unlike analog closed-circuit television (CCTV) cameras, they require no local recording device, but only a local area network. Most IP cameras are webcams, but the term IP camera or netcam usually applies only to those used for surveillance that can be directly accessed over a network connection.
Axis Communications AB is a Swedish manufacturer of network cameras for the physical security and video surveillance industries. It provides wide range of network video surveillance cameras. First, we present Axis F-series that we used in the prototyping phase of our framework. Second, AXIS P23 series which have been used in OVALIE platform.

6.2.1.1 *Axis F Series*

AXIS F Series offers flexible, high-performance HDTV cameras for extremely discreet indoor, outdoor and in-vehicle surveillance applications. It is based on a divided network camera concept, where the camera is split into a sensor unit—made up of a lens and image sensor with a cable and a main unit, which is the body of a camera. The divided concept enables flexibility in the choice of hardware, as well as in the installation. The small sensor unit can be installed discreetly in tight places, while the long cable from the sensor unit to the main unit provides the flexibility to place the main unit where there is space. The main or sensor unit can be easily relocated or changed after the initial installation, giving users additional flexibility.

(a) F44 Main Unit. (b) AXIS F1015 Sensor Unit.

Figure 6.7: Axis F44 parts

For prototyping purpose, we used Axis F44. AXIS F44 Main Unit supports up to four AXIS F Sensor Units as shown in Figure 6.7a. Also, it streams 1080p/HDTV 720p videos from four sensor units simultaneously, it includes Axis Forensic Wide Dynamic Range (WDR), and it supports two-way audio.
We used four Axis F1015 Sensor Units with the F44 main unit. The sensor unit can be installed in tight places and flush-mounted in a wall or ceiling as it small size as shown in Figure 6.7b, while the main unit can be placed further away where there is space. AXIS F1015 comprises an image sensor with 1080p resolution (1920×1080 pixels) and a *varifocal* lens that provides between a 52^{o} and 97^{o} horizontal field of view. A *varifocal* lens gives users the flexibility to adjust the field of view to suit the application.

6.2.1.2 *AXIS P33 Series*

AXIS P33 Network Camera Series offers versatile fixed dome cameras for cost-efficient and flexible installation. They are suitable for a wide range of surveillance applications, such as in retail stores and education and healthcare facilities.

Figure 6.8: AXIS P3367 network camera.

AXIS P33 offers a *varifocal* lens. While streamlined in design, these cameras are robust and vandal-resistant. It includes Axis Forensic WDR for high-quality images even when there is both dark and light areas in the scene. It offers two-way audio and I/O connectivity so it's easy to complement your surveillance installation. It also includes variants with Axis Zipstream with support for both H.264 and H.265 and enhanced security features such as signed firmware and secure boot. Figure 6.8 shows a P3367 dome camera. It includes remote zoom and focus capabilities to eliminate the need for hands-on fine tuning. It is designed for effortless installation, they can be mounted flush to a wall or ceiling.

6.2.2 *Microphones*

We deployed U843R three directional boundary microphone that is shown in Figure 6.9. It is offering customizable coverage in mono or stereo for a variety of audio and video conferencing applications. Its three cardioid condenser elements can be utilized separately or together to realize cardioid, omnidirectional, or figure-8 polar patterns. It delivers clear, intelligible speech tonality, thanks to its 70 Hz to 15 kHz frequency response and 80 Hz low-cut filter.

Figure 6.9: U843R three directional boundary microphone.

6.2.3 *AXIS Camera Station software*

AXIS Camera Station is a video management software for surveillance specially developed for small and mid-sized installations. Retail stores, hotels, schools and manufacturing industries are just some of the companies that enjoy full control and protection of their premises and can quickly take care of incidents. AXIS Camera Station is powerful and easy to use with an intuitive interface so

Figure 6.10: AXIS Camera Station multiple views example.

anyone can manage the system. It is easy to add features like network speakers to communicate with staff and deter intruders, network video door stations for

audiovisual identification and remote entry control, video analytics to improve operator efficiency and radar to follow intruders. Figure 6.10 shows a the multiple views that can provided by the software.

6.3 Multi-person social interactions analysis

To perform and experiment "in vivo" multi-person social interactions analysis without pre-defined scenarios, we record a new dataset. It consists of ten videos (average recording time is twenty minutes), and four participants in each video instructed to discuss a general conversational topic. The recording performed in a quiet office room equipped with four cameras (four views), as shown in Figure 5.6. Cameras have been slightly placed above the participants to provide a near frontal view and to avoid occlusion. In addition to the automatic extracted social cues in section 4.1 we annotated manually hand position (hand_over_table,hand_under_table,hand_over_face,hand_other) and speaking/not_speaking cues for each participant.

	P1	P2	P3	P4
P1	0	12	5	1914
P2	29	0	7	984
P3	8	677	0	1310
P4	393	325	1118	0
Sum	430	1014	1130	4208

Figure 6.11: Participants' gaze direction summation during the analyzed segment. Pi is a person with index i.

6.3.1 *Qualitative analysis based on the eye gaze and the sum eye gaze over time*

Figure 5.6 shows the detected persons in the multiple view recorded scene, their facial expression, and their gaze direction. In the middle of the upper part of the figure, each person is coded with color and his/her gaze direction is coded with an arrow that has the same color (e.g., at the frame's timestamp person 2 was looking at person 4 while she was showing happy face expression). Whenever there is eye contact, it is visualized as a red double arrow. Furthermore, if we

make the summation of participants' gaze direction over the video time, we will get a square matrix ($n \times n$, where n is the number of the participants) named Look_At_square_matrix (LAM). Figure 6.11 shows LAM for the same experiment of figure 5.6. We can notice that P4 (person 4) looked 1118 times towards P3 and P3 looked 1310 times towards P4. Also, we can conclude that P4 is the dominant person as he received the maximum sum of gaze, and it is 4208; the time unit is equal to one frame = 1/15 second.

6.3.2 *Social media aggregation for better interpretation*

Figure 6.12 shows some collected information from the social networks of the participants of the shown experiment in figure 5.6. Intuitively we can say that person 2 and person 3 are close friends and they know person 1 very well, and they never met person 4 before the recording day. Based on this contextual information (collected from social media), we can say that person 4 was the dominant person since he was the new person and they would like to have more information about him. This example shows that aggregation of the contextual information will enhance the analysis process.

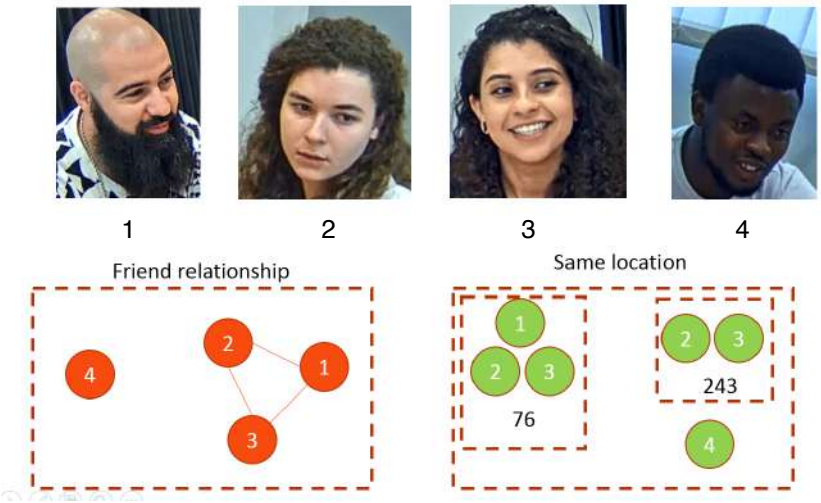


Figure 6.12: Example of the collected contextual information from social networks.

6.3.3 *(Meta)data aggregation and statistics visualization using Kibana*

Figure 6.13 shows a dashboard that contains multiple charts that visualize statics about the (meta)data. In addition to that it allows to apply multiple filter, e.g., you can keep the meta data that extracted from the experiment number 4 by applying the filter "experiment_id=4".

Figure 6.13: Face-to-Face social interaction analysis dashboard .

The pie chart in Figure 6.14 shows that person 4 from the face-to-face experiment received the max sum percentage of the gaze during the fourth segment of the experiment.

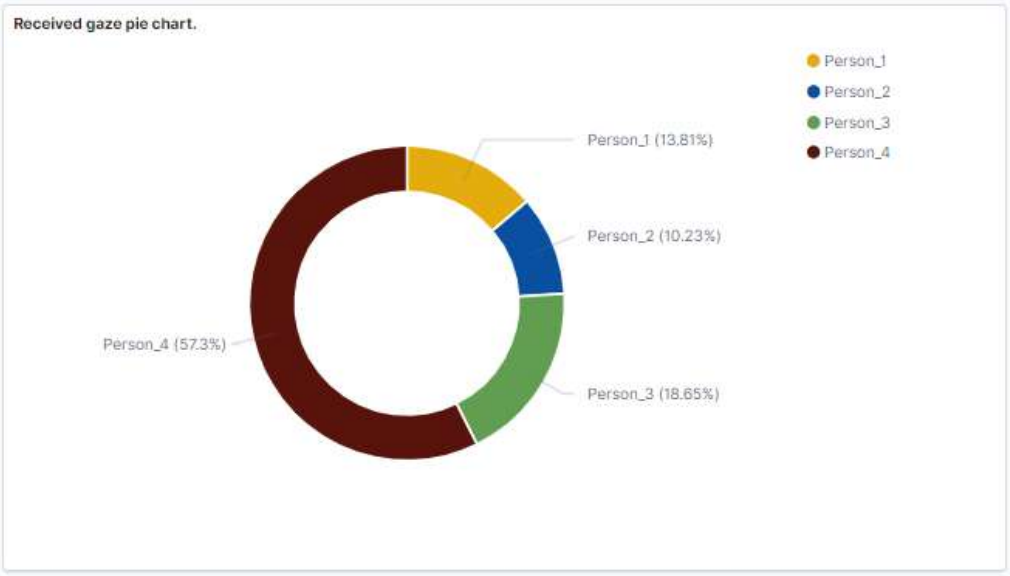


Figure 6.14: Face-to-face social interaction experiment sample received gaze pie chart.

We can elicit the same conclusion from the gaze direction heat-map as Figure 6.15, however we can notice that person 4 was receiving more gaze from person 3.

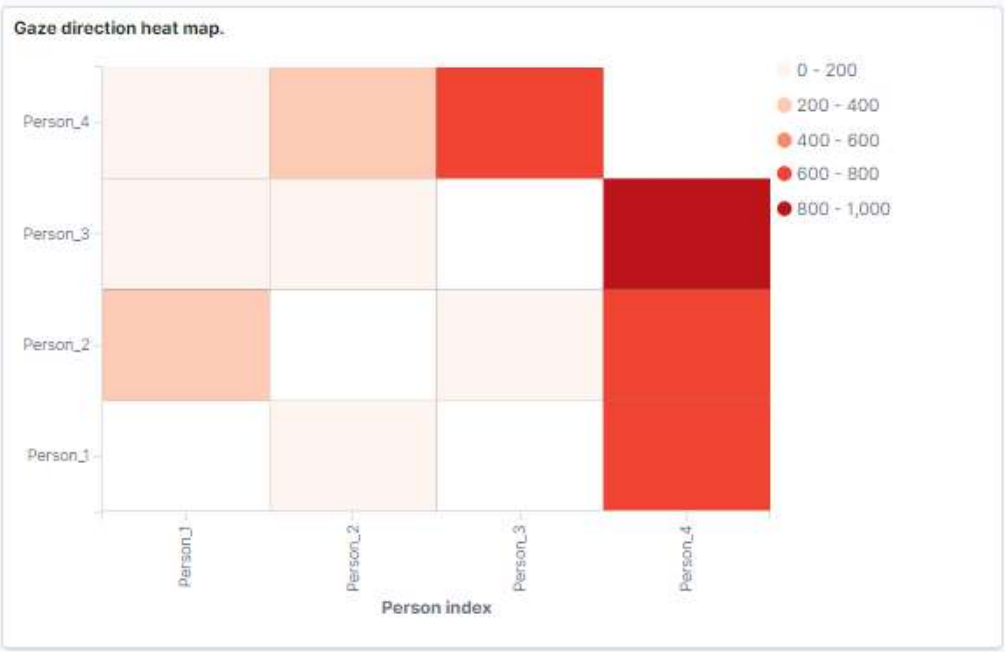


Figure 6.15: Face-to-face social interaction experiment sample gaze direction heat-map.

Again person 4 has the maximum speaking percent (65%) among the other participants, as shown in Figure 6.16. Based on the visualization of the (meta)data statics, we can elicit that person 4 was the dominant speaker during the 4th segment of the experiment.

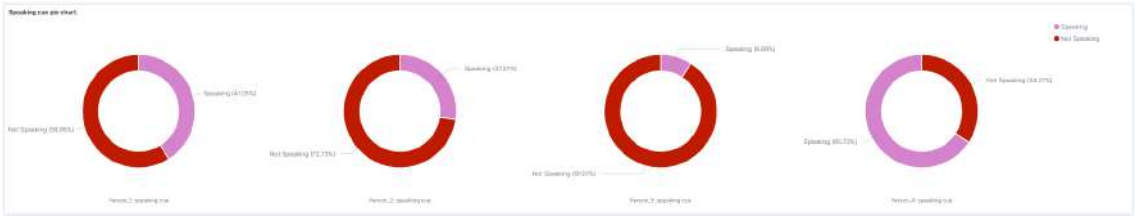


Figure 6.16: Face-to-face social interaction experiment sample speaking cues pie chart.

6.4 Study of eating behavior of the children with Prader–Willi syndrome (work in progress)

In PWS, behavioral problems and food craving remain a major difficulty through the development of children with pws, as introduced in Section 2.5. At the table, this food addiction and behavioral problems have effects on the health of these people and consequences on their social life as well as their family circle. This study aims to understand the social dimensions of eating disorders in children, adolescents and adults with PWS. We are going to observe 15 families, with a child or adolescent with PWS aged 7 to 18 years eating, having a meal within OVALIE platform. As shown in Figure 6.17, the observation flow starts with introducing the platform and the objectives of the observation, then the lunch in the restaurant area, after the break we start a reflexive interview in the focus group area. During the reflexive interview, different segments of the observation video will be present to the children and their families in order to get a better understanding and unambiguous meaning of the behaviors that appear within these segments.

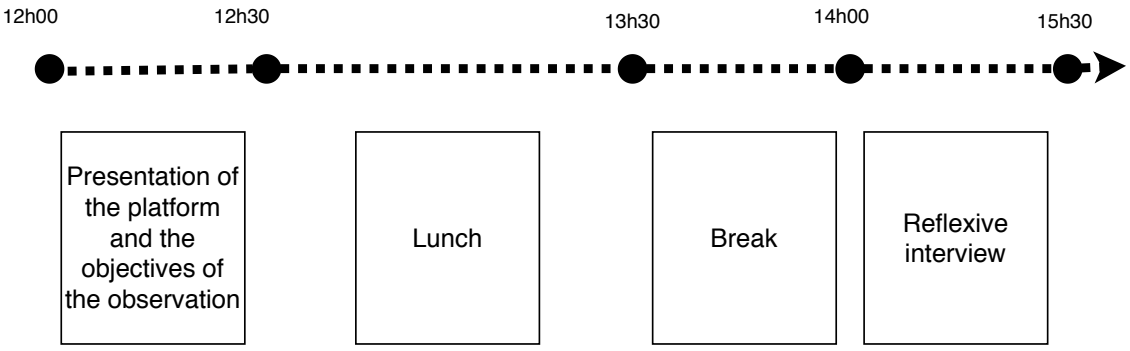


Figure 6.17: PWS observation flow.

6.5 Conclusion

In this chapter, we introduce different parts of OVALIE platform. The modularity of the physical context of OVALIE will enable the study of behaviors in real-life dining environments: restaurant, dining room, hospital room, focus group room. Then, we introduce the deployed sensors (cameras, microphones) for the scene recording along with Axis camera station software, which provides flexibility in multi-view recording and retrieving. After that, we present experiments for "in vivo" multi-person social interactions analysis without predefined scenarios. The results show that we can detect the dominant speaker based on the gaze direction, speaking cue, and facial expression. Also, social media aggregation enhances the result interpretation. Finally, we present a PWS observational study (work in progress) that aims to study the social dimension of such syndrome on the people who born with it and their family.

7

CONCLUSION

A conclusion is simply the place where you got tired of thinking.

— Dan Chaon

Social interaction analysis applied in many domains such as industry, medical services, and observational studies. In the context of observational studies, several mechanisms (for example, questionnaire, online observation and analysis, and recording then analysing) are used to analyze social interactions from various perspectives, but these methods suffer from many limitations related to subjectivity of the observer and his limited ability to track multiple social cues at the same time. Hence, the best way to handle such limitations is to automate the analysis process. However, this automation introduces several challenges. First, we have to keep the observation naturalistic (no predefined scenarios, non-intrusive devices). Second, we need to handle the gap between the human-based tasks and the machine-based computational tasks. Third, there are multiple technical challenges related to data privacy and security, data heterogeneity, and data volume.

So, we propose a holistic approach for social interaction detection and analysis that includes: (i) raw data acquisition module for recording the experiments; (ii) context-based features (social cues) extraction module to extract the metadata from the raw data; (iii) contextual information module for extracting the metadata from the real environment contextual information; (iv) comprehensive (meta)data model for storing the metadata (v) social behavior analysis module that includes multiple methods for aggregating the extracted metadata to analyze the primitive social cues; and (vi) user interface for visualizing both the analysis results and the extracted metadata.

In the holistic approach, the proposed (meta)data model [105] is: (1) privacy-preserving since it facilitates the data anonymization; (2) extendable to cover the vocal and verbal cues; (3) smoothing the data fusion among multiple modalities; and (4) decoupling the social cues extraction from the social interaction analysis. The usage of the (meta)data model helps us to track the collected and extracted personal data, which enables the data update possibility. Also, the (meta)data model helps us to encapsulate the recorded videos by preventing direct access to them. This encapsulation with the possibility of updating the personal data are essential requirements for making the proposed holistic approach GDPR compliant and incorporated with privacy and security.

In the features extraction methods, first, we propose a novel geometric-based method for eye contact detection in natural multi-person interactions that (a) does not require any prior training dataset, (b) has classification performance almost at

the same level of the well-known supervised-based ones, and (c) does not require any intrusive device, which allows to perform naturalistic observation. Second, we propose find-tuned deep models to extract contextual information from the recorded video content (type of food in the social event).

In the analysis part of the proposed holistic approach, the multi-layer analysis provides a flexibility for the framework as we can aggregate different combinations of the cues at different time scales, also it allows to aggregate the contextual information with the social cues. The aggregation of contextual information is helping to bridge the semantic gap between the description generated by the machine and the annotation made by human operators. Finally, the analysis result and metadata visualizing reduces the required time for eliciting conclusions from social events by human operators.

As future work, we are going to complete the PWS observation study in order to understand the desocialization eating habits of the children with PWS. Second, we will collect and annotate a benchmark dataset from the recorded experiments, which will help researchers to establish standardization to evaluate, configure, and compare future methods of analyzing social behavior. Moreover, we will record and annotate a new dataset for cooking activities, since the kitchen in OVALIE platform is equipped with many ceiling cameras,. Furthermore, we intend to implement additional social cues extraction methods and include vocal cues that satisfy constraints of naturalistic observation.

Finally, through a partnership with Taylor's University in Malaysia, an identical OVALIE platform will be established there, facilitating cross-cultural studies in eating behaviors. The multicultural society of Malaysia will allow exploring the different cultures within Malaysia, as well as comparative studies between Europe and Asia.

BIBLIOGRAPHY

[1] "MONSTACHE" 2019. Monstache: Sync mongodb to elasticsearch in real-time. https://rwynn.github.io/monstache-site/, 2019. Accessed: 2019-09-30.

[2] "TowardsDataScience " 2019. A comprehensive guide to convolutional neural networks. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, 2019. Accessed: 2019-10-15.

[3] "CERTOP UMR 5044". Ovalie platform. https://certop.cnrs.fr/plateforme-experimentale-ovalie-shs-alimentation/, 2019. Accessed: 2019-11-30.

[4] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security*, 12(5):1042–1055, May 2017.

[5] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva. With whom do i interact? detecting social interactions in egocentric photo-streams. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2959–2964. IEEE, 2016.

[6] Rehan Ahmad, Syed Zubair, Hani Alquhayz, and Allah Ditta. Multimodal speaker diarization using a pre-trained audio-visual synchronization model. *Sensors*, 19(23):5163, 2019.

[7] Z. Akhtar and T. H. Falk. Visual nonverbal behavior analysis: The path forward. *IEEE MultiMedia*, 25(2):47–60, Apr 2018.

[8] Josephine Akosa. Predictive accuracy: a misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum*, pages 2–5, 2017.

[9] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1707–1720, 2015.

[10] Michael Argyle and Janet Dean. Eye-contact, distance and affiliation. *Sociometry*, pages 289–304, 1965.

[11] Stylianos Asteriadis, Kostas Karpouzis, and Stefanos Kollias. Visual focus of attention in non-calibrated environments using gaze estimation. *International Journal of Computer Vision*, 107(3):293–316, May 2014.

[12] Marie Avril, Chloë Leclère, Sylvie Viaux, Stéphane Michelet, Catherine Achard, Sylvain Missonnier, Miri Keren, David Cohen, and Mohamed Chetouani. Social signal processing for studying parent–infant interaction. *Frontiers in Psychology*, 5:1437, 2014.

[13] S. O. Ba and J. M. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, Jan 2011.

[14] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[15] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[16] Ernest Barker and RF Stalley. *The politics*. Oxford University Press, 2009.

[17] Bhanu Prakash Battula and R Satya Prasad. An overview of recent machine learning strategies in data mining. *Proceeding of International Journal of Advanced Computer Science and Applications*, 4(3), 2013.

[18] Institut Paul Bocuse. Living lab. Accessed January 25, 2020.

[19] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014.

[20] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[21] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. Machine learning: a historical and methodological analysis. *AI Magazine*, 4(3):69–69, 1983.

[22] Dan Chaon. Conclusion quotes. https://www.azquotes.com/quote/473841, 2019. Accessed: 2019-11-30.

[23] Chao Chen, Andy Liaw, Leo Breiman, et al. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110(1-12):24, 2004.

[24] Chih-Wei Chen, Asier Aztiria, Somaya Ben Allouch, and Hamid Aghajan. Understanding the influence of social interactions on individual's behavior pattern in a work environment. In *International Workshop on Human Behavior Understanding*, pages 146–157. Springer, 2011.

[25] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):43:1–43:20, September 2017.

[26] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[27] Mayo Clinic. Prader-willi syndrome. Accessed January 05, 2020.

[28] Pengyu Cong, Zhiwei Xiong, Yueyi Zhang, Shenghui Zhao, and Feng Wu. Accurate dynamic 3d sensing with fourier-assisted phase shifting. *IEEE Journal of Selected Topics in Signal Processing*, 9(3):396–408, 2014.

[29] FM Cortés, MA R Alliende, AR Barrios, BL Curotto, L María V Santa, XO Barraza, LA Troncoso, CS Mellado, and RV Pardo. Clinical, genetic and molecular features in 45 patients with prader-willi syndrome. *Revista medica de Chile*, 133(1):33–41, 2005.

[30] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28(2):117–139, 2004.

[31] Marco Cristani, R. Raghavendra, Alessio Del Bue, and Vittorio Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86 – 97, 2013. Special issue: Behaviours in video.

[32] David J Crowley and David Mitchell. *Communication theory today*. Stanford University Press, 1994.

[33] CSGA. Centre for taste and feeding behavior (csga). Accessed January 25, 2020.

[34] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Self-taught clustering. In *Proceedings of the 25th international conference on Machine learning*, pages 200–207. ACM, 2008.

[35] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[36] W Edwards Deming. *The new economics for industry, government, education*. MIT press, 2018.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[38] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[39] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[40] W. Dong, B. Lepri, F. Pianesi, and A. Pentland. Modeling functional roles dynamics in small group interactions. *IEEE Transactions on Multimedia*, 15(1):83–95, Jan 2013.

[41] Takumi Ege and Keiji Yanai. Image-based food calorie estimation using recipe information. *IEICE TRANSACTIONS on Information and Systems*, 101(5):1333–1341, 2018.

[42] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[43] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[44] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[45] Oxford Dictionaries English. Focus group. Accessed January 02, 2020.

[46] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[47] Giovanni Maria Farinella, Dario Allegra, and Filippo Stanco. A benchmark dataset to study the representation of food images. In *European Conference on Computer Vision*, pages 584–599. Springer, 2014.

[48] James L Flanagan. *Speech analysis synthesis and perception*, volume 3. Springer Science & Business Media, 2013.

[49] Scott Fortmann-Roe. Accurately measuring model prediction error. *Online: http://scott. fortmann-roe. com/docs/MeasuringError. html*, 2012.

[50] Wolfgang Fuhl, Marc Tonsen, Andreas Bulling, and Enkelejda Kasneci. Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art. *Machine Vision and Applications*, 27(8):1275–1288, 2016.

[51] Yuki Fukuhara and Yukiko Nakano. Gaze and conversation dominance in multiparty interaction. In *2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction*, volume 9, pages 9–16, 2011.

[52] Erving Goffman. *Relations in public*. Transaction Publishers, 2009.

[53] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[54] Clinton Gormley and Zachary Tong. *Elasticsearch: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2015.

[55] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[56] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[57] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009.

[58] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[60] Luis Herranz, Weiqing Min, and Shuqiang Jiang. Food recognition and recipe analysis: integrating visual content, context and external knowledge. *arXiv preprint arXiv:1801.07239*, 2018.

[61] Vanja A Holm, Suzanne B Cassidy, Merlin G Butler, Jeanne M Hanchett, Louise R Greenswag, Barbara Y Whitman, and Frank Greenberg. Prader-willi syndrome: consensus diagnostic criteria. *Pediatrics*, 91(2):398, 1993.

[62] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[63] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.

[64] Qiang Ji, Harry Wechsler, Andrew Duchowski, and Myron Flickner. Special issue: eye detection and tracking. *Computer Vision and Image Understanding*, 98(1):1–3, 2005.

[65] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.

[66] Yoshiyuki Kawano and Keiji Yanai. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *European Conference on Computer Vision*, pages 3–17. Springer, 2014.

[67] John D Kelleher, Brian Mac Namee, and Aoife D'arcy. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press, 2015.

[68] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[69] Yuki Kizumi, Koh Kakusho, Takeshi Okadome, Takuya Funatomi, and Masaaki Iiyama. Detection of social interaction from observation of daily living environments. In *Future Generation Communication Technology (FGCT), 2012 International Conference on*, pages 162–167. IEEE, 2012.

[70] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[71] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[72] Otto Lappi. Eye movements in the wild: Oculomotor control, gaze behavior and frames of reference. *Neuroscience and Biobehavioral Reviews*, 69:49 – 68, 2016.

[73] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth annual conference of the international speech communication association*, 2015.

[74] K. Lekdioui, Y. Ruichek, R. Messoussi, Y. Chaabi, and R. Touahni. Facial expression recognition using face-regions. In *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6, May 2017.

[75] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[76] Tony Lindeberg. Scale invariant feature transform. 2012.

[77] Hirofumi Lnaguma, Masato Mimura, Koji Inoue, Kazuyoshi Yoshii, and Tatsuya Kawahara. An end-to-end approach to joint social signal detection and automatic speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6214–6218. IEEE, 2018.

[78] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[79] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[80] Marwa Mahmoud and Peter Robinson. Interpreting hand-over-face gestures. In Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction*, pages 248–255, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[81] Catherine Marshall and Gretchen B Rossman. *Designing qualitative research*. Sage publications, 2014.

[82] Aleix M Martinez. Visual perception of facial expressions of emotion. *Current Opinion in Psychology*, 17:27 – 33, 2017. Emotion.

[83] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. Recognition of multiple-food images by detecting candidate regions. In *2012 IEEE International Conference on Multimedia and Expo*, pages 25–30. IEEE, 2012.

[84] Robert McAulay and Thomas Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

[85] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 3723–3726. ACM, 2016.

[86] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. Im2calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1233–1241, 2015.

[87] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231. IEEE, 2017.

[88] Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

[89] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[90] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[91] Sik Hung Ng and James J Bradac. *Power in language: Verbal communication and social influence.* Sage Publications, Inc, 1993.

[92] Tam V. Nguyen, Si Liu, Bingbing Ni, Jun Tan, Yong Rui, and Shuicheng Yan. Towards decrypting attractiveness via multi-modality cues. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(4):28:1–28:20, August 2013.

[93] National Institutes of Helth. Is there a cure for prader-willi syndrome (pws)? Accessed January 05, 2020.

[94] U.S. National Library of Medicine. Prader-willi syndrome. Accessed January 05, 2020.

[95] Eisuke Ono, Takayuki Nozawa, Taiki Ogata, Masanari Motohashi, Naoki Higo, Tetsuro Kobayashi, Kunihiro Ishikawa, Koji Ara, Kazuo Yano, and Yoshihiro Miyake. Relationship between social interaction and mental health. In *2011 IEEE/SICE International Symposium on System Integration (SII)*, pages 246–249. IEEE, 2011.

[96] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[97] Yixiong Pan, Peipei Shen, and Liping Shen. Speech emotion recognition using support vector machine. *International Journal of Smart Home*, 6(2):101–108, 2012.

[98] Franck Jeveme Panta, Mahmoud Qodseya, André Péninou, and Florence Sèdes. Management of mobile objects location for video content filtering. In *Proceedings of the 16th International Conference on Advances in Mobile Computing and Multimedia, MoMM 2018, Yogyakarta, Indonesia, November 19-21, 2018*, pages 44–52, 2018.

[99] Franck Jeveme Panta, Mahmoud Qodseya, Geoffrey Roman-Jimenez, André Péninou, and Florence Sèdes. Spatio-temporal metadata querying for cctv video retrieval: Application in forensic. In *Proceedings of the 9th ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pages 7–14. ACM, 2018.

[100] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005.

[101] Tomislav Pejsa, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. Authoring directed gaze for full-body motion capture. *ACM Trans. Graph.*, 35(6):161:1–161:11, November 2016.

[102] Mannes Poel, Ronald Poppe, and Anton Nijholt. Meeting behavior detection in smart environments: Nonverbal cues that help to obtain natural interaction. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[103] Andrea Prader. Ein syndrom von adipositas, kleinwuchs, kryptorchismus und oligophrenie nach myatonieartigem zustand im neugeborenenalter. *Schweiz Med Wochenschr*, 86:1260–1261, 1956.

[104] Xianbiao Qi, Rong Xiao, Chun-Guang Li, Yu Qiao, Jun Guo, and Xiaoou Tang. Pairwise rotation invariant co-occurrence local binary pattern. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2199–2213, 2014.

[105] Mahmoud Qodseya. Visual non-verbal social cues data modeling. In *Advances in Conceptual Modeling - ER 2018 Workshops Emp-ER, MoBiD, MREBA, QMMQ, SCME, Xi'an, China, October 22-25, 2018, Proceedings*, pages 82–87, 2018.

[106] Mahmoud Qodseya. Visual non-verbal social cues data modeling. In *Advances in Conceptual Modeling - ER 2018 Workshops Emp-ER, MoBiD, MREBA, QMMQ, SCME, Xi'an, China, October 22-25, 2018, Proceedings*, pages 82–87, 2018.

[107] Mahmoud Qodseya, Franck Jeveme Panta, and Florence Sèdes. Visual-based eye contact detection in multi-person interactions. In *2019 International Conference on Content-Based Multimedia Indexing, CBMI 2019, Dublin, Ireland, September 4-6, 2019*, pages 1–6, 2019.

[108] Mahmoud Qodseya, Marta Sanzari, Valsamis Ntouskos, and Fiora Pirri. A3d: A device for studying gaze in 3d. In *European Conference on Computer Vision*, pages 572–588. Springer, 2016.

[109] Mahmoud Qodseya, Mahdi Washha, and Florence Sèdes. Dievent: Towards an automated framework for analyzing dining events. In *34th IEEE International Conference on Data Engineering Workshops, ICDE Workshops 2018, Paris, France, April 16-20, 2018*, pages 163–168, 2018.

[110] Francesco Ragusa, Valeria Tomaselli, Antonino Furnari, Sebastiano Battiato, and Giovanni M Farinella. Food vs non-food classification. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 77–81. ACM, 2016.

[111] Lankapalli Ravikanth, Digvir S Jayas, Noel DG White, Paul G Fields, and Da-Wen Sun. Extraction of spectral information from hyperspectral data and application of hyperspectral imaging for food and agricultural products. *Food and bioprocess technology*, 10(1):1–33, 2017.

[112] Francisco J Rodríguez, Antonio García, Pedro J Pardo, Francisco Chávez, and Rafael M Luque-Baena. Study and classification of plum varieties using image analysis and deep learning techniques. *Progress in Artificial Intelligence*, 7(2):119–127, 2018.

[113] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[114] HE Schepers, R De Wijk, J Mojet, and AC Koster. Innovative consumer studies at the restaurant of the future. *Measuring Behavior 2008*, page 366, 2008.

[115] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[116] Thapanee Seehapoch and Sartra Wongthanavasu. Speech emotion recognition using support vector machines. In *2013 5th international conference on Knowledge and smart technology (KST)*, pages 86–91. IEEE, 2013.

[117] William Shakespeare. *The complete works of William Shakespeare*, volume 19. Ginn, 1900.

[118] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.

[119] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.

[120] Peipei Shen, Zhou Changjun, and Xiong Chen. Automatic speech emotion recognition using support vector machine. In *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, volume 2, pages 621–625. IEEE, 2011.

[121] Georg Simmel. *The sociology of georg simmel*, volume 92892. Simon and Schuster, 1950.

[122] Ashutosh Singla, Lin Yuan, and Touradj Ebrahimi. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 3–11. ACM, 2016.

[123] Rory Smead. The role of social interaction in the evolution of learning. *The British Journal for the Philosophy of Science*, 66(1):161–180, 2014.

[124] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 271–280, New York, NY, USA, 2013. ACM.

[125] John D. Smith, Roel Vertegaal, and Changuk Sohn. Viewpointer: Lightweight calibration-free eye tracking for ubiquitous handsfree deixis. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, UIST '05, pages 53–61, New York, NY, USA, 2005. ACM.

[126] Study.com. What are social cues? - definition and examples. Accessed September 25, 2019.

[127] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[128] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[129] Wenxue Tan, Chunjiang Zhao, and Huarui Wu. Intelligent alerting for fruit-melon lesion image based on momentum deep learning. *Multimedia Tools and Applications*, 75(24):16741–16761, 2016.

[130] Sik-Ho Tsang. Review: Inception-v3. Accessed November 25, 2019.

[131] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1-2):507–545, 1995.

[132] Morihei Ueshiba. *The art of peace*. Shambhala Publications, 2002.

[133] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France, 2010.

[134] Manik Varma and Andrew Zisserman. A statistical approach to texture classification from single images. *International journal of computer vision*, 62(1-2):61–81, 2005.

[135] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2011.

[136] Alessandro Vinciarelli, Hugues Salamin, and Maja Pantic. Social signal processing: Understanding social interactions through nonverbal behavior analysis. In *2009 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 42–49. IEEE, 2009.

[137] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[138] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. Speaker diarization with lstm. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5239–5243. IEEE, 2018.

[139] Wen-June Wang, Jun-Wei Chang, Shih-Fu Haung, and Rong-Jyue Wang. Human posture recognition based on images captured by the kinect sensor. *International Journal of Advanced Robotic Systems*, 13(2):54, 2016.

[140] Zheng Wang, Yangqiu Song, and Changshui Zhang. Transferred dimensionality reduction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 550–565. Springer, 2008.

[141] Hendrik Wesselmeier, Stefanie Jansen, and Horst M Müller. Influences of semantic and syntactic incongruence on readiness potential in turn-end anticipation. *Frontiers in human neuroscience*, 8:296, 2014.

[142] Pei Xu. A real-time hand gesture recognition and human-computer interaction system. *CoRR*, abs/1704.07296, 2017.

[143] Shulin Yang, Mei Chen, Dean Pomerleau, and Rahul Sukthankar. Food recognition using statistics of pairwise local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2249–2256. IEEE, 2010.

[144] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Everyday eye contact detection using unsupervised gaze target discovery. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, UIST 2017, Quebec City, QC, Canada, October 22 - 25, 2017*, pages 193–203, 2017.