Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13816-w

1226: DEEP-PATTERNS EMOTION RECOGNITION IN THE WILD

Interpersonal relation recognition: a survey

Hajer Guerdelli[1,3] · Claudio Ferrari[2] · Stefano Berretti[3]

Received: 2 December 2021 / Revised: 8 August 2022 / Accepted: 6 September 2022

Abstract

People spend a considerable amount of their time in social activities, where person-to-person relations are of main relevance. Recently, there has been an increasing research interest in automatically analyzing interpersonal relations, for the social and behavioral implications, and the many practical applications it may have. However, to the best of our knowledge, there is not a systematic study providing a harmonized view of the literature in the field. On this ground, we summarize in our work interpersonal relation recognition datasets and methods aiming to help researchers to have a better understanding of the characteristics of the state-of-the-art. In the proposed study, we distinguish between methods that address *objective* relations that do not depend on behavior or emotional state, and methods that consider *subjective* ones that depend on emotions. It turns out quite evidently that aiming at the latter recognition task is more challenging, with the existing methods that provide convincing results only on limited and very specific cases. For both the broad categories, we discuss datasets and methods according to the different behavioural and psychological models used to annotate and classify the data. We conclude our review work, by providing a comprehensive discussion pointing out current limitations and future research perspectives.

Keywords Interpersonal relation recognition · Facial expression · Emotions

✉ Hajer Guerdelli
hajer.guerdelli@unifi.it

Claudio Ferrari
claudio.ferrari2@unipr.it

Stefano Berretti
stefano.berretti@unifi.it

[1] Université de Tunis El Manar, Institut Supérieur d'Informatique d'El Manar, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l'Information et de la Connaissance (LIMTIC), 2 Rue Abou Rayhane Bayrouni, Ariana-Tunis, 2080, Tunisia

[2] Department of Engineering and Architecture, University of Parma, Parco Area delle Scienze 181/A, Parma, 43124, Italy

[3] Department of Information Engineering, University of Florence, via Santa Marta 3, Florence, 50139, Italy

1 Introduction

Over the last few years, automatic understanding of the human behavior has attracted an increasing interest [39, 41, 46]. Several factors contributed to push the research in this direction. On the one hand, a variety of digital devices make it available an unprecedented amount of data, either images, videos or 3D, captured in laboratory settings or in the wild conditions, that have humans as main target. This has encountered the deep neural network revolution that owes a large part of its success to the abundance of data, thus giving rise to effective and efficient solutions largely surpassing previous approaches based on hand-crafted features. On the other, the analysis of the human behavior opens the way to many concrete applications, such as in video surveillance, where anomaly or unusual behavioral patterns can be searched for, in the analysis and recognition of human actions and emotions for social and medical investigations, in identity recognition based on face deformations (e.g., expressions, action units) and body actions, etc. [2, 9, 15, 21, 23, 36, 44].

Despite this rapid development in methods that target human behavior understanding, there are some research directions in this area that received less attention and are still at their initial steps. A clear example is represented by the recognition of *interpersonal relations*. This is evidently a topic with a great potential for the many social implications and possible application scenarios. As a matter of fact, people organize their social live in terms of their relations with other people [16]. It is estimated that most people spend from 80% to 90% of their time in some form of interpersonal communication [25], whether at home, at work, or with a friend. We communicate interpersonally face-to-face, by phone, text, social media, etc. Recent research in computer vision attempted to understand and recognize human relationships from face-to-face nonverbal communication, where people constantly interact through different attributes such as emotions, head position, gender, age and facial expressions [35, 49, 52, 53].

We can think of the vast literature on facial expression and emotion recognition from images and videos as a necessary and preliminary step to target the task of interpersonal relation recognition. In particular, we note the clear direction of passing from the analysis of posed expressions, often acted by trained subjects in constrained laboratory conditions, to more realistic scenarios where spontaneous emotions are shown in the wild [4, 14, 55].

The used models for expression and emotion categorization have also evolved from the simple six expression model to more continuous representations, like that represented by the valence-arousal space [1, 7, 50]. This has advanced the knowledge and the potential for concrete application of such methods, but we think that posing the investigation of human expression and emotion in the context of an interpersonal relation provides an additional value. Indeed, in many cases expressions and emotions come as reaction to an external stimuli in a person-to-person interaction (though other forms of interaction can be considered, such as person-to-object). In such a context, it becomes relevant the mutual analysis of the interacting subjects, and the analysis of facial expressions is complemented by other visual features, like gestures, body posture, head pose, gaze attention. Other soft-biometric features can become relevant, like the gender, the age, etc. The short-term prediction of the interpersonal relation also becomes of interest as it can happen in a teacher-to-student interaction, where early signs of loss of attention or boredom can be noted in advance. In the following of our survey, we will refer to these relations as *subjective* because they can vary dynamically and depend from the subjective emotional status of the interacting subjects.

A different typology of relations that can be automatically investigated has instead an *objective* nature, in that they do not change and not depend on the emotional state. An example of this is given by the kinship relation, or other affective or work relations. In

general, the recognition of these relations can rely on different features and models than those used for the subjective category. For example, investigating the kinship relation can require the use of facial or body traits to discover similarity, thus making it of potential interest the use of features extracted for face recognition.

In this survey, we provide a comprehensive overview of the recent literature related to the problem of estimating humans interpersonal relations in images or videos. In particular, we analyze and discuss recent methods and datasets that have been proposed to address this problem. To this end, we identified a broad categorization in *objective* and *subjective* interpersonal relations. The former category includes methods and datasets designed to classify family or work relations, while the latter category comprises social and emotional relations. For each category, we identified the different behavioural models that have been proposed in the literature, and relate them with datasets and methods. The survey is concluded by a comprehensive discussion on the existing literature, its current limitations and open questions, together with the perspectives for future investigations. Figure 1 illustrates the evolution of the literature of interpersonal relations, and how it is grown in the recent years. Prior to this developing, tools to analyze complex relationships were lacking; now that sufficient datasets and methods do exist, we are able to analyze many types of relationships between two or more subjects. Our proposed survey includes recent studies published between 2015 and 2021.

Looking to the literature on interpersonal relation recognition, we were not able to find a survey paper summarizing the work done and the developments in this recent area of research. We think that providing an overview of the datasets and methods in this domain, while also organizing the existing literature in a comprehensive way can be useful for researchers working in this field and for new enthusiasts interested to approach it. We also think that presenting what has been done, the current limitations and future perspectives can contribute to form a common and shared background.

The rest of the manuscript is organized as follows: in Section 2, we provide a general introduction to the main characteristics of social relations and motivate the separation into *objective* and *subjective* relations; The models, databases and methods referring to the former category are presented and discussed in Section 3; The discussion on the literature for the latter category in instead deepened in Section 4; Section 5 concludes the paper

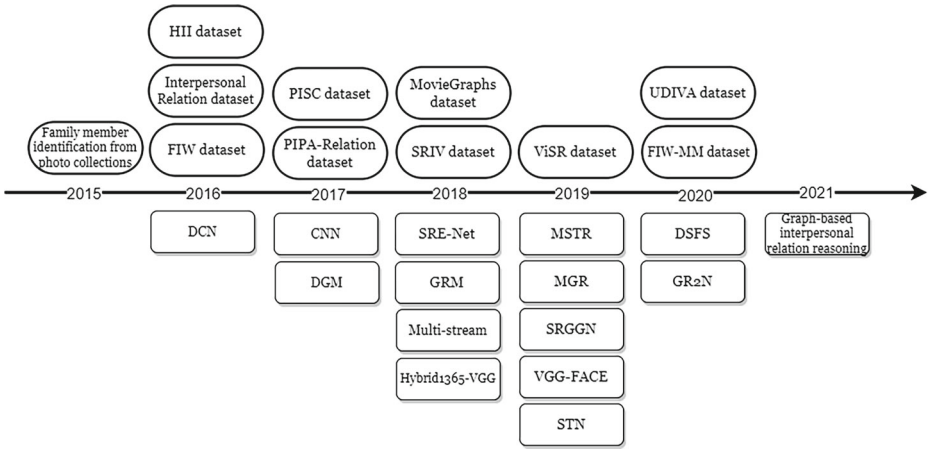

Fig. 1 Evolution of the datasets and methods presented in the proposed study

by discussing positive aspects and limitations of the existing methods and datasets, also prospecting future research directions.

2 Social relation traits

Social relations can be described under different points of view, and involve various aspects. Social relations can be defined according to *objective* terms or based on *subjective* behavior. Among others, this classification was also followed in [31], where authors classified their dataset according to objective and subjective relations [22, 24]. For the *objective* relations case, we can think of relationships such as kinship or work roles that do not depend on the particular behavior or emotional state. Differently, subjective relations depend on the perceived emotion, and can be defined using categories such as friendly/hostile, cooperative/submissive to mention some. These relations are directly originated from the particular interaction between people.

The large categorization introduced above can be further specialized by considering the different behavioral and social psychology models that were followed by the research in this field. In our survey, the objective relations comprise Kinship recognition [10, 37, 38], Asian social psychology and the Bugental's social psychology theory [28, 35, 54]; the subjective relations, instead, include the interpersonal circle theory as defined by Kiesler [31, 52], and the Fieske's theory of relational models [16].

In order to automatically detect and analyse such different relations using a computer vision/machine learning algorithm, a major problem is that of identifying some distinguishing features that can provide useful clues to the specific task. Clearly, they depend on the different type of relation of interest. For example, to determine a mother/son relation, it could be necessary to rely on some facial traits such as hair or eyes color, skin color, shape of mouth, eyes and nose, age difference. In this case, face recognition features can be useful. If, instead, we wanted to determine whether two people are in a friendly relation, analyzing expressions or actions is surely more informative.

Subjective relations can change for the same class of objective relations though. For example, a mother and a son can either have a good relationships, namely, warm, friendly, involved and demonstrative, or a bad one, namely, strict, harmful, distant and silent. Same as work relations, i.e., boss and employee, it can be dominant, trusting, friendly and assured or equal, mistrusting, hostile and unassured. This makes the estimation of subjective relations significantly more challenging than the objective ones.

Several social psychology theories have been proposed to formalize the above [6, 8, 16, 17, 22, 24, 32, 34]. Among them, we choose those that are used in the interpersonal relation datasets collected in our survey. Figure 2 illustrates our proposed organization of the literature works: we first categorized the datasets and methods according to the fact they address objective or subjective relations. For each category, several theories and relational models were proposed and formalized in the literature, each of which focuses on different yet related aspects. We first divided the datasets according to the five theories that we analyzed. Then, methods were ranked according to the used dataset and connected to it in the graph (we used a color-coding for methods that appear more than once). In this way, it is immediate to identify the topic of interest of each approach, and which datasets can be used to address a specific task. In what follows, we separately describe datasets and methods based on the type of relation they refer to, either objective or subjective.

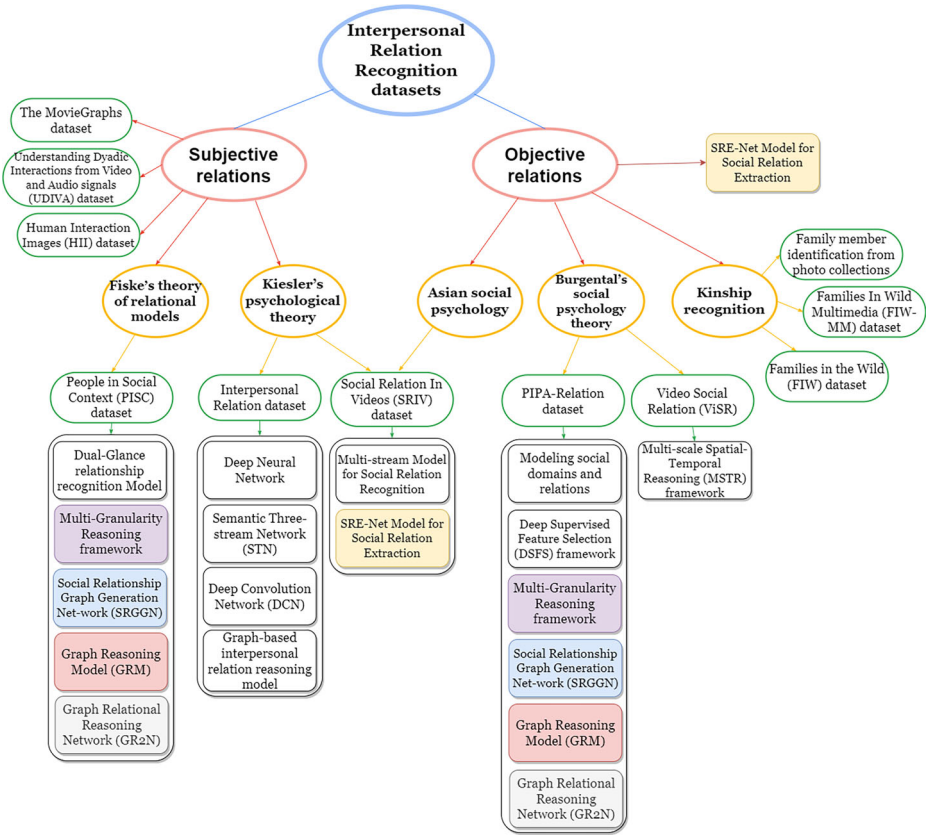

Fig. 2 Our proposed organization of the literature on interpersonal relations. Yellow circles illustrate the theories we analyzed; green circles indicate the datasets, while the rounded rectangles refer to frameworks and methods. A color coding is used for the frameworks and methods to evidence methods that appear more than once in the figure and are applied to different datasets

3 Objective relations

Objective relations are defined in terms of social relationships that people have with respect to others, such as teacher/student or mother/child. Given the large and different number of such relations that exist in the real life, some effort was put in defining meaningful categories. In the literature, the two models most largely employed are the *Bugental's Social Psychology Theory* [6], and the *Asian Social Psychology Theory* [22]. Both develop on the idea of categorizing people relationships as they likely influence their behavior. In fact, in social contexts, people are expected to behave differently depending on the type of relation they have. It is easy to imagine that one can change his/her behavior significantly whether the interaction is with a friend, a brother/sister or one's boss/supervisor. While the above theories thoroughly consider both kinship as well as other general types of relationships, some literature works instead focus only on the former, others focus on the different levels of family relationship [11]. In the following, we will first briefly present datasets and method related to kinship recognition in Section 3.1, and then describe the two theories in

Sections 3.2 and 3.3, respectively. Tables 1 and 2 summarize the existing datasets for objective relations with their context/size and main annotations, and methods in the literature that were designed for objective relation recognition. For each method, we also reported information about the network architecture and the data used in the training.

3.1 Kinship recognition

The objective of kinship recognition is to determine the exact type of kinship (e.g., parent-child) rather than if the subjects are related or not. Furthermore, methods described here only deal with kinship verification and family recognition. Datasets for kinship recognition are differentiated by the type of relation, e.g., Parent-Child [13, 29, 48], Twin Pairs [43], Siblings [5]. In the following, we present the datasets that are labeled with multiple-types of kin relations from pair of people or more images and videos.

Families in the Wild (FIW) The Families in the Wild (FIW) dataset [38] was collected using various search engines (e.g., Google, Bing, Yahoo) and social media outlets (e.g., Pinterest). It contains 10,000 family photos of 1,000 families, each with at least 3 family members. It divided the relationships into three categories with a total of 11 types: the *Parent-child* relation includes Father-Daughter (F-D), Father-Son (F-S), Mother-Daughter (M-S), Mother-Son (M-S); the *Grandparent-grandchild* relation comprises: Grandfather-Granddaughter (GF-GD), Grandfather-Grandson (GF-GS), Grandmother-Granddaughter (GM-GD), Grandmother-Grandson (GM-GS); Finally, the *Siblings* relation includes Sister-Brother (SIBS), Brother-Brother (B-B), Sister-Sister (S-S).

Families In the Wild Multimedia (FIW-MM) The Families In the Wild Multimedia (FIW-MM) dataset [37] is an extended dataset from the FIW dataset with multimedia (MM) data (video, audio, and text captions). It contains 550 subjects in 660 videos, subset of 200 FIW families. In FIW-MM the data have three types: non-speaking face tracks (visual only), speech segments (audio only), and face tracks of speakers (visual-audio). This dataset is annotated in the same way as the FIW dataset, with families IDs, Member IDs, gender information and relationship types.

Family member identification from photo collections This dataset [10] contains photo collections of 16 different families taken at amusement parks, annotated with multiple labels such as social role, for example: “child1”, “father”, etc., while non family members are given identities such as “femaleAdult1”, “maleAdult1”, a bounding box around the face and the body skeleton.

3.2 Asian social psychology theory

The Asian social psychology theory [22] asserts the need to include the context of relationships in any study of social behavior because of the profound impact that relationships have on individual behavior. Interestingly, it classifies relationships into fourteen relations according to the basis of their formation: Kinship (by blood–consanguinity–, marriage, adoption, or godparenthood), Connection by birth, Political authority, Subjugation (by military conquest, slavery, or colonialism), Social class, Office or employment, Residential location, Institutional affiliation, Social connections based on ascription (e.g., inheritance) or achievement, Tutelage apprenticeship or guardianship, Professional consultation, Companionship affection or sexual attraction, Situational temporary or chance encounters. To

Table 1 Objective relations datasets: Publicly available audiovisual human-human (face-to-face) non-acted interaction datasets

Dataset	Year	Annotations	Content/Size	Observations
Families In Wild Multimedia (FIW-MM) [37]	2020	families IDs, Member IDs, gender information, relationship types	Videos: 550 subjects in 660 videos	Kinship dataset with image, audio and video
Video Social Relation (ViSR) [28]	2019	Eight common social relations	Videos: over 8,000 videos	Limited types of social relations
Social Relation In Videos (SRIV) [31][a]	2018	Eight subjective relations and eight objective relations	Videos: 3,124 videos	The multi-label strategy makes the relation ambiguous
PIPA-Relation [35]	2017	Head bounding box, identity number and social relations	Images: extend PIPA with 26,915 person	Semantic attributes are collected from both body and head images according to Bugental's theory
Families in the Wild (FIW) [38]	2016	families IDs, Member IDs, gender information, relationship types	Images: 10,000 family photos	The largest kinship verification dataset
Family member identification from photo collections dataset [10][b]	2015	Identity label shows social role, bounding box indicating location of the face and the body skeleton	Images: photos of 16 families	The volume of the dataset is small

[a,b]Not-available

Table 2 Objective relations: Methods and architectures

Method	Year	Architecture	Training Set (Dataset)	Observations
Deep Supervised Feature Selection (DSFS) framework [45]	2020	Architecture of the proposed deep supervised feature selection (DSFS) framework	PIPA-relation dataset	The outcome is that body attributes contribute more than face attributes to recognize social relationships
Graph Relational Reasoning Network (GR2N) [26]	2020	Graph Relational Reasoning Network	PIPA-Relation and PISC dataset	Generate a reasonable and consistent social relation graph
Multi-scale Spatial-Temporal Reasoning (MSTR) framework [28]	2019	- Triple Graphs model - Pyramid Graph Convolutional Network (PGCN)	ViSR dataset	A graph network to capture social relations according to long-term and short-term temporal cues in the video
Multi-Granularity Reasoning framework [53]	2019	deep CNN: - Person-Object Graph (POG) - Person-Pose Graph (PPG)	PIPA-Relation and PISC dataset	Predict social relations by fusing scene information as surrounding objects, but malfunction is observed for images with noisy features
Social Relationship Graph Generation Network (SRG-GN) [3]	2019	ConvNet architecture	PIPA-Relation and PISC dataset	Relationships are treated independently on the same image, which neglects the logical constraints between social relationships on an image
Graph Reasoning Model [47]	2018	Gated Graph Neural Network (GGNN)	PIPA-Relation and PISC dataset	Building a graph where persons and objects are fully connected, then using GGNN to predict the social relation
SRE-Net Model for Social Relation Extraction [51]	2018	MoCNR algorithm for Character Nodes Recognition	SRIV and LFW dataset	Extracting social relations between characters based on scene to uncover the relations within the same scene
CNN models trained end-to-end and trained for semantic attributes [35]	2017	double-stream CaffeNet	PIPA and PIPA-Relation dataset	Recognize social relations from a group of semantic attributes

the best of our knowledge, only the following dataset exists that is collected and annotated according to this categorization.

Social Relation In Videos (SRIV) The Social Relation In Videos (SRIV) dataset [31] was proposed to recognize social relationships from a video, where different behaviors are classified along two-dimensions: Subjective Relations (Sub-Relation) and Objective Relations (Obj-Relation). It contains 3,124 videos, collected from movies and TV dramas. The social relations were extracted from movies and television domains, and classified based on two classification approaches: the Kiesler theory (this theory is described in Section 4.1), and the Asian Social Psychology theory. According to this, the dataset was divided into 16 subclasses: eight subclasses were derived based on the Kiesler theory, and eight subclasses were arranged based on the Asian Social Psychology theory, using from the last one only three types of relationships (work, kinship and other) divided into eight subclasses (supervisor-subordinate, peer, service, parent-daughter, mating, sibling, friendly and hostile).

In the same work [31], authors proposed a Multi-stream Model for social relation recognition, using a ConvNet architecture. The network used RGB images of videos to learn clues of social relations, such as video scenes and people representations. Then, a multi-stream ConvNet including spatial, temporal and audio features was proposed, where the action features were fused by a temporal segment network, which is used to learn spatial and temporal features. In doing so, Inception with Batch Normalization (BN-Inception) was chosen to achieve a balance between accuracy and efficiency, and GoogleNet was adopted to learn audio features using audio spectra.

In [51], the authors used the objective relation classes from the SRIV dataset with their SRE-Net Model for social relation extraction. To predict the social relation (leader-member, peer, service, parents-offspring, lover, sibling, friend and enemy) between characters, they proposed a MoCNR method to identify the number of people and characters that appear in the video by clustering people's facial features (keyframe extraction, face detection, face alignment, face feature extraction and clustering). Then, the social relation recognition method was introduced based on scene segmentation: the scene changes in videos are detected and used to split the video into separate clips. The GoogleNet model was used to extract audio features with a C3D model to capture the time sequence information of video. Finally, AdaBoost (Adaptive Boosting), XGBoost, GBDT and LightGBM classifiers were used to accomplish social relation prediction.

3.3 Bugental's social psychology theory

The Bugental's domain-based theory [6] partitions social life relations into 5 domains, namely, Attachment, Reciprocity, Mating, Hierarchical power, and Coalitional groups. Them, from this partition 16 fine-grained categories are derived: father-child, mother-child, grandpa-grandchild, grandma-grandchild, friends, siblings, classmates, lovers/spouses, presenter-audience, teacher-student, trainer-trainee, leader-subordinate, band members, dance team members, sport team members and colleagues. We observe that both kinship and other types of relations are defined in this theory, which makes it widely exploited with several methods addressing the problem of recognizing these classes.

PIPA-Relation The PIPA-Relation dataset [35] is an extended version of the People In Photo Albums (PIPA) dataset [54] with 26,915 person pair annotations for social relations. The dataset is labeled with five social domain according to Bugental's theory. The PIPA dataset

is collected from Flickr photo albums for the task of person recognition and contains 37,107 photos with 63,188 instances of 2,356 identities, labeled by head bounding box and identity number, where the same person can appear in multiple albums.

In [35], authors experimented with two models: a CNN model trained end-to-end, and a CNN model for semantic attribute recognition derived from the social domain theory. This latter network is capable of identifying attributes such as age and age difference, the latter being motivated by its importance to distinguish relations, gender, head appearance such as straight hair, wavy hair, wearing earring, wearing hat and so on, head pose and facial expressions. Other attributes that are classified are clothing such as hat, tShirt, jeans, actions such as holding hands, high five, hug, and activities such as adjusting, ailing, applauding, arranging, attacking, ballooning, baptizing and so on. Then, the method uses the concatenated feature to learn a linear SVM and categorize the interaction. The results of the experiments are quite preliminary in terms of accuracy, since authors reported a relation recognition accuracy for all the attributes of 57.2%, which is the same for body attributes, and only 44.8% for head attributes.

In [45], the authors adopted 12 face and body attributes including age, gender, appearance, emotion, pose, location & scale, face appearance, face pose, face emotion, body clothing, body proximity and body activity. A Deep Supervised Feature Selection (DSFS) framework is proposed, where the input photo can include more than two subjects that are segmented and used as input to the DSFS framework as pairs (the DSFS framework is illustrated in Fig. 3). Then, the feature extraction module, included in the DSFS, extracts the 12 attribute's features using a pre-trained Double-Stream CaffeNet, a pre-trained CNN-CRF, or multi-task RNN. To deal with the problem of noises and redundancies, authors used a feature selection module with two feature selection policies: group feature selection to extract the optimal feature subset based on contributions of attributes, and dimensional feature selection to learn the optimal feature subset at a fine-grained level and so remove most of the redundancy. Then, the final classification was performed by a Softmax classifier used to compute the probability distribution of the input pair on social relationship categories. The accuracy of social relationship recognition obtained in this work was 61.51%. Authors also tried to eliminate some attributes but that effected the accuracy; an interesting outcome

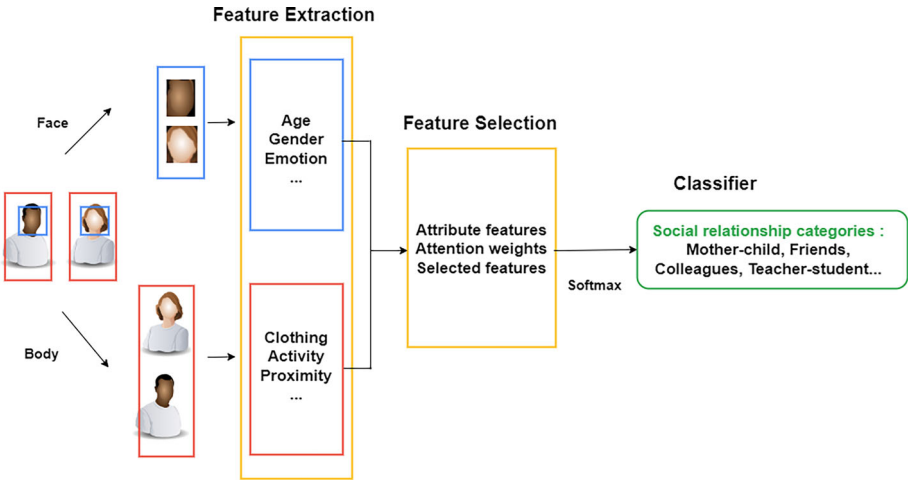

Fig. 3 The deep supervised feature selection (DSFS) framework proposed in [45]

of this work is that body attributes contribute more than face attributes to recognize social relationships, which is quite in line with the observation that objective relations are independent from the emotional status of the person, and behavioral features such as actions or body proximity are more informative to recognize the social interaction.

The works in [3, 26, 47, 53] also used the PIPA-Relation and PISC datasets [27] (this latter dataset is introduced in Section 4.2). Given that the goal is to understand the interaction among people, the common intuition of these methods is that of modeling the interaction with graphs. The Multi-Granularity Reasoning framework as proposed in [53] has two branches. The first one used a CNN to learn knowledge about the scenes from the whole image. The other branch focused on regional cues and fine interactions among persons and contextual objects, with three main procedures: a CNN object detection model to crop persons and objects in an image, a human pose estimation method, and a Person-Object Graph (POG) to connect each person with the other person and the objects. They built a Person-Pose Graph (PPG) to model the interaction between two persons. Social relation reasoning was performed on the two graphs by GCNs. The social relation was predicted by the global feature (deep convolutional neural network ResNet-101) from CNN and the reasoning feature from the GCNs. They obtained the accuracy of 64.6% for Friends, 67.8% for Family, 60.5% for Couple, 76.8% for Professional, 34.7% for Commercial, and 70.4% for No Relation on the PISC dataset. An accuracy of 64.4% was reported on the PIPA-relation dataset.

In [3], authors proposed a Social Relationship Graph Generation Network (SRGGN) with two modules: A Multi-Network Convolutional Neural Network (MN-CNN) module for Attribute and Relationship representations followed by a Social Relationship Graph Inference Network (SRG-IN) module for generating a structured graph representation. The MN-CNN module has two sub-modules, SN1 and SN2, with an input image and a set of bounding box annotations for the people in image. The input to SN1, the Attribute ConvNet architecture, is the cropping for a single-body image of a person, with the network having layers for each attributes: age, gender and clothing. The sub-module SN2 is a network of pairwise-relationship ConvNet architectures, with two VGG-16 architectures to compute activity and scene features from the context images of people. The SRG-IN module where the relationships is predicted in an image according to (person1, relation, person2), and the social relationships between people are classified in the form of a social graph. Gated Recurrent Units (GRUs) are used to improve the prediction of relationships between persons, where each relationship in an image gets information from its nearby nodes (person attributes) and also its nearby edges (relationships).

A Graph Reasoning Model (GRM) was finally proposed in [47] to detect the correlations between social relationships and semantic objects in the scene. To this end, a social relationship and an object nodes are included. The GRM takes an image and a person pair of interest and extracts features from the regions of the person pair to initialize the relationship node. For the object node, it uses a Faster-RCNN detector to extract their features. Then, to explore the interaction of the persons with the contextual objects it uses the Gated Graph Neural Network (GGNN), and finally the graph attention mechanism is used to select the most informative nodes.

In [26], authors created a social relation graph called the Graph Relational Reasoning Network (GR2N). It jointly gathers all relations within a single image by constructing several virtual relation graphs to explicitly model the logical constraints among different types of social relations. They presented each person by a node in a graph, and the edge between

each node is the relation between the people. GR2N is created to predict the existence of edges and the type of edges, in order to solve the problem of unknowing topology of the graph at the beginning, then generate a reasonable and consistent social relation graph.

Video Social Relation (ViSR) The Video Social Relation (ViSR) dataset [28] contains 8,000 video clips with more than 200 movies. The ViSR dataset is labeled with eight social relations such as Parent-offspring, Couple, Leader-subordinate, Service, Sibling, Friend, Colleague and Opponent. A Multi-scale Spatial-Temporal Reasoning (MSTR) framework was also proposed in that work. The approach started by cropping the persons and objects from frames with Mask R-CNN. Then, a triple Graphs model was designed: with an Intra-Person Graph (IntraG) for the same person, an Intra-Person Graph (IntraG) to model the spatial and temporal representation of persons and objects, and a Person-Object Graph (POG) to capture the co-existence of persons and contextual objects. Then, a relation reasoning by Pyramid Graph Convolutional Network (PGCN) was used to learn dynamics in varied temporal range.

3.4 Discussion

All the datasets and methods described so far are related to the problem of recognizing and classifying objective social relationships, so identifying the type of connection among individuals. Overall, we can draw some conclusions; first, most of the methods rely more on body, e.g., age difference, gender, clothing, or action features, e.g., proximity, interaction, rather than emotional features, e.g., facial expressions, to discriminate the relationships. This somehow confirms that behavioral patterns are more informative in this scenario. We also observe another common way of addressing the problem is that of using graphs, which makes sense inasmuch as connections among people indeed form sort of graphs, where edges can be useful to verify the consistency of relations and perform reasoning on the graph. As an example, if subject A is the father of subject B, and subject C is the brother of subject B, then subject A must be also the father of subject C. Observing the overall performance of the reported methods though, we can conclude that the problem is rather challenging, and that there is still room for large improvements.

4 Subjective relations

We defined subjective social relations as those involving emotional states and behaviors, independently from the objective relationship that exists among subjects. Differently from objective relations, these are much more complex, variegated and challenging both to define and detect. In fact, they depend on both the particular emotional status of the individuals, the person whom he/she is interacting with, and other possibly varying aspects. We identified two major theories related to subjective relations: the *Kiesler's Circle Theory* [24], and the *Fiske's Theory of Relational Models* [16]. Both analyze and categorize social relations in terms of perceived emotional status, even though the latter include also a partial objective categorization. In the following, we separately describe the two theories and the related datasets and methods that are most close to them. Tables 3 and 4 summarize the datasets, architectures and methods belonging to this category.

Table 3 Subjective relation datasets: Publicly available audiovisual human-human (face-to-face) non-acted interaction datasets

Dataset	Year	Annotations	Content/Size	Observations
Understanding Dyadic Interactions from Video and Audio signals (UDIVA) [33]	2020	Personality scores (self- and peer-reported), sociodemographics, mood, fatigue, relationship type	Videos: 147 subjects in 188 sessions	The self-report annotation is difficult to understand
Social Relation In Videos (SRIV) [31][a]	2018	Eight subjective relations and eight objective relations	3,124 videos	The multi-label strategy makes the relation ambiguous
MovieGraphs [42]	2018	Interactions between characters, relationships, reasons behind certain interactions, situation, scene, and natural language	7,637 videos	An exhaustively annotated dataset. Some of the information are incomplete due to the discarded parts between scenes
People in Social Context (PISC) [27]	2017	social relationship	22,670 images	Scene and global contextual cues to predict social relationships
Interpersonal Relation [52]	2016	Faces' bounding boxes and interpersonal relation traits	8,016 images	Interpersonal relationship traits are divided into 16 segments, thus describing a wide range of relations
Human Interaction Images (HII) [40]	2016	face locations and orientations	1,971 images	Recognizing interactions between people by face information

[a]Not-available

Table 4 Objective relations: Methods and architectures

Method	Year	Architecture	Training Set (Dataset)	Observations
Graph-based interpersonal relation reasoning model [19]	2021	multi-scale features with the GNNs	Interpersonal relation dataset	Exploiting the interactions between two faces for interpersonal relationship recognition
VGG-FACE model [20]	2019	Siamese-like architecture	Interpersonal relation dataset	Extracting information from whole images to predict group-level emotions
Semantic three-stream network (STN) [49]	2019	Siamese network	Interpersonal relation dataset	Using the semantic information to understand the social relation from the entire image
Multi-stream Model for Social Relation Recognition [31]	2018	Network Architecture (deep learning)	SRIV dataset	Using visual and audio features that represent social relations of people to ameliorate the recognition
Hybrid1365-VGG model [42]	2018	-	MovieGraphs dataset	Multi-labels to summarize and localize each scene, generating explanations for human interactions
Dual-Glance Model [27]	2017	Attentive RCNN	PISC dataset	Bringing attention on regions of persons and objects to predict social relations
Deep convolutional network (DCN) [52]	2016	Siamese-like architecture	Interpersonal relation dataset	Purchasing the inherent correspondences between facial expressions and other heterogeneous attributes

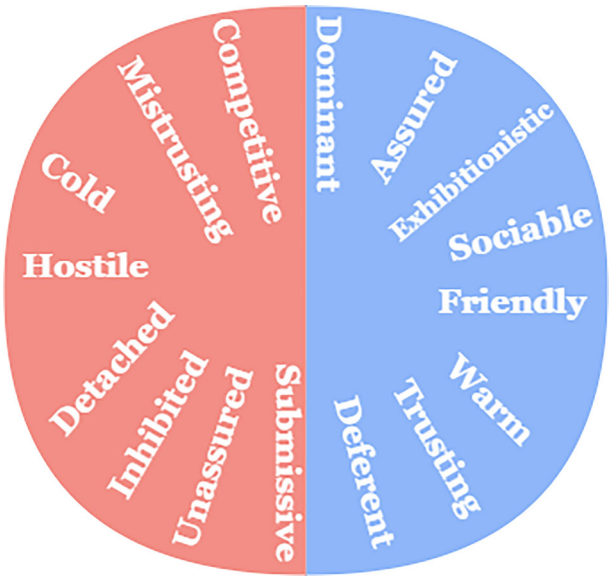

Fig. 4 Visualization of the Kiesler's model using the Interpersonal Circle

4.1 Kiesler's circle

The Kiesler's psychological theory [24] proposed to organize the interpersonal relations into a circle called interpersonal circle or circumplex, where the relations are divided into 16 segments, with 8 pairs: Dominant, Competitive, Trusting, Warm, Friendly, Involved, Demonstrative and Assured/Submissive, Deferent, Mistrusting, Cold, Hostile, Detached, Inhibited and Unassured. Figure 4 shows the Kiesler circle, with the 16 segments and 8 pairs.

In the following, we present the interpersonal relation datasets and methods that defined the social relation traits based on the interpersonal circle proposed by Kiesler.

Interpersonal Relation In [52], authors built two datasets: the Expression in-the-Wild (ExpW) for facial expression recognition and the Interpersonal Relation dataset. This latter contains 8,016 videos and movies collected from the web which are annotated with interpersonal relation traits. The authors proposed a Deep Convolution Network (DCN) to capture a rich face representation and a novel attribute propagation method that, despite the different datasets, takes advantage of the inherent correspondences between facial expressions and other heterogeneous attributes. Next, they jointly considered pairwise faces for interpersonal relation prediction by arranging two identical DCNs obtained in a Siamese-like architecture (the interpersonal relation model is illustrated in Fig. 5). Using the interpersonal relation dataset, they trained the new Siamese network end-to-end to map raw pixels of a pair of face images to relation traits. Then, they performed relation traits reasoning using face representation and additional spatial cues. In this work, authors proposed a method that uses a pre-trained model, which already learned face attributes (facial expression, gender, age, and head pose), and therefore, to predict interpersonal relations. They obtained 71% average accuracy across all the relation traits prediction performance (dominant, competitive, trusting, warm, friendly, involved, demonstrative and assured).

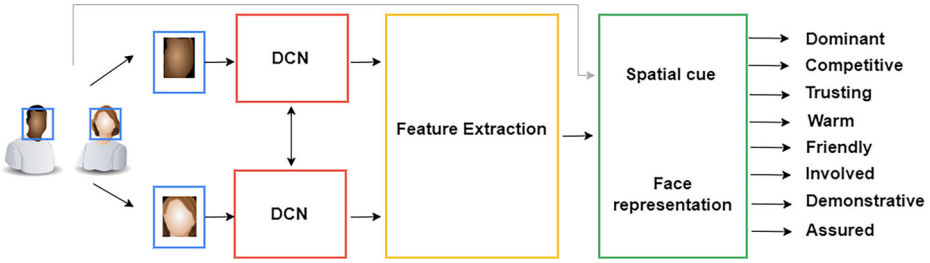

Fig. 5 The model for interpersonal relation prediction as proposed in [52]

In [20], authors proposed a Deep Neural Network combining a Siamese-like Network with two VGG-Face model branches to learn the social relation between two people and a Deep Model Based to extract information (such as layout of objects, posture of people and background of a scene) from whole images to predict group-level emotions. The authors used the same evaluation metric and the same spatial cues as [52]. In particular, the used measures are the balanced accuracy to account for the imbalanced positive and negative attribute samples as evaluation metric, and the two faces' positions, the relative faces' positions and the ratio between the faces' scales as spatial cues. They selected VGG, Inception-v2 and ResNet to be fine-tuned on the social relation dataset. The reason behind this choice, is because of the performance of these networks in the ImageNet Large Scale, Visual Recognition Challenge (ILSVRC) and the information learned from it. This resulted useful to infer high level social relation, such as layout of objects, posture of people and background of a scene.

In [49], the authors proposed a semantic three-stream network (STN) for social relation recognition. They used a Siamese network with a semantic augmentation structure to extract features from a pair of face images to reduce parameters, since they tried to understanding social relationships from the entire original image. They mapped each face separately, using a Siamese network to extract features and using a ResNet where convolutional networks share weights. They also proposed a semantic augmentation structure to help the network to better sense objects in the environment, which contains five parts: a size adjustment module, a channel adjustment module, a multi-view module, a feature transition module, and a semantic fusion module.

For interpersonal relation recognition, authors in [19] proposed a graph-based interpersonal relation reasoning model with multi-scale features. Their proposed architecture is divided into a feature extraction, the GNN module, and the classifiers. In the feature extraction stage, a pair of faces is cropped, represented in a bounding box and fed to two VGGFace models. Then, the joint area of every face pair is cropped as an input to another ResNet. The extracted features are represented as five nodes to construct the GNN graph structure. The GNN module based graph reasoning extract hierarchical features from the connected nodes. Finally, they designed eight binary classifiers using cross-entropy for making a final multinomial prediction.

Social Relation in Videos (SRIV) As mentioned in Section 3.2, the Social Relation In Videos (SRIV) dataset [31] uses subjective relations and objective relations, from the Asian social psychology theory along with the Kiesler theory, using the eight subclasses. Other than the method in [31], we are not aware of other approaches reporting results on subjective relations on this dataset. Still, we included it in this discussion as it provides annotation

of subjective interactions, and can be of interest for developing approaches addressing this specific task.

4.2 Fiske's theory of relational models

The Fiske's theory of relational models [16] targets the reflection of personal history onto the cognitive individual experiences and differentiates social relations into four parts: Communal sharing, Authority ranking, Equality matching, and Market pricing. This allows dividing the relationships into intimate, non-intimate, no relation, friends, family, couple, professional, commercial. Even though some of these classes are actually objective rather than subjective, e.g., family, couple, other classes are defined based on the personal experience and how this is reflected onto interpersonal relations. Given this, we included this model within subjective relation models.

People in Social Context (PISC) The People in Social Context (PISC) dataset [27] consists of 22,670 images and 76,568 manually annotated labels from 9 types of social relationship (No Relation, Has Relation, Intimate Relation, Non-Intimate Relation, Friends, Family Members, Couple, Professional, Commercial) and consists of 66 annotated occupation categories. It collects around 40k images containing people from a variety of sources, including Visual Genome, MS-COCO, YFCC100M, Flickr, Instagram, Twitter and commercial search engines (i.e., Google and Bing), using a combination of key words search (i.e., co-worker, people, friends, etc.) and people detector (Faster RCNN). The proposed Dual-Glance relationship recognition Model [27] has two glances. The first one takes in three inputs, two bounding boxes covering each person and one for the union region insert into three CNNs. In the second glance, they adapted Faster RCNN to process the input image with a Region Proposal Network (RPN) to generate a set of region proposals, the process with a CNN. Next, they allocated attention to each region, and aggregated their outputs to refine the score. Their experiments show 63.2% accuracy on three-relationships (Intimate, Non-Intimate and No relation), and 79.7% on six-relationship (Friends, Family, Couple, Professional, Commercial and No Relation).

4.3 Others

Here, we separately report datasets for subjective interpersonal relation analysis that do not follow any of the previous models/theories.

Understanding Dyadic Interactions from Video and Audio signals (UDIVA) The UDIVA dataset [33] includes face-to-face dyadic interaction videos based on free and structured tasks, recorded from different camera positions. It is composed of 90.5h of recordings of dyadic interactions between 147 voluntary participants. It includes socio-demographic (age, gender, ethnicity, occupation, maximum level of education and country of origin), self- and peer-reported personality, internal state (pre- and post-session mood and fatigue), and relationship profiling. Authors propose a novel method for self-reported personality (characterized by the basic Big Five traits (Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), or OCEAN, based on self-reported assessments) using R(2+1)D network for the visual chunks and VGGish for audio to extract feature, with a visual input (face, local context, and extended context chunks), audio input (raw chunks), and metadata input (both interlocutors' characteristics, and session and dyadic features), to get an OCEAN scores as an output.

The MovieGraphs The MovieGraphs dataset [42] presents 20 relationships such as family (e.g., parent, spouse), friendship/romance (e.g., friend, lover), or work (e.g., boss, coworker). It contains 51 movies and 7,637 video clips annotated with scene label, situation label (topic), description, graph with 8 nodes: character, attributes (age, gender, ethnicity, profession, appearance, mental and emotional states), relationship, interaction (verbal or non-verbal), topic, reason, time stamp.

In this work, authors proposed three tasks starting with Graph-Based Situation Retrieval to retrieve relevant clips using as nodes character, attribute, relationship, interaction, topic, and reason. Then Interaction Ordering where they used an attention-based decoder RNN. Finally, Reason Prediction using information about the scene in the form of attributes of each character, their relationship, and an interaction to predict the reason of the interaction.

Human Interaction Images (HII) The Human Interaction Images (HII) dataset [40] includes 10 human interaction classes: boxing-punching, dining, handshaking, highfive, hugging, kicking, kissing, partying, speech and talking. HII contains 1,972 images with 150 images for each class. There can be two, three or more subject sin an images, with a visible facial region for at least one of them.

Authors proposed a novel descriptor based on facial regions to recognize the human interactions. The proposed approach started with face detection, then a set of descriptors were used: Histogram of Face Orientations (HFO) for face orientations distribution, Histogram of Face Directions (HFD) for direction frequencies in the images distribution, Distances of Faces (DF) for the location of each face and the distances between them, Circular Histogram of Face Locations (CHFL) for the relative layout of people using a histogram of their locations by fitting a circle to the center of the extracted faces, and Grid Histogram of Face Locations (GHFL) same as CHFL for the spatial layout of the multiple people. Finally, Scene features with GIST descriptors and Bag-of-Words (BoW) were used to detect the scene and deep features with deep learning and Deep feature Convolutional Neural Network (CNN).

4.4 Discussion

Differently from the objective relations category, the different types of subjective interpersonal relations are way more variegated and complex, making it hard to define clearly separated categories. Despite referring to two main common behavioral models, all the reported datasets slightly differ from each other in terms of data annotation and focus. This makes it complex to gather comprehensive conclusions that are shared among the described approaches. In any case, a very clear difference that can be noted with respect to objective relations is that much more attention and emphasis is put on analyzing faces. Indeed, many of the described approaches use networks that are pre-trained either for face or expression recognition. Still, the less developed literature represents a piece of evidence of the challenging nature of the problem, which is yet far from being solved.

5 Conclusions, limitations and future perspectives

Humans understand images by looking at the whole scene rather than only to the objects under consideration [3]. This evidences the importance of the context in our visual perception, also suggesting the need for incorporating such trait when designing automatic methods to analyse image and video content. This urgency for taking into consideration such

aspect is reinforced when the objective is the analysis of human facial expressions and emotions that, in most of the cases, happen as reaction to external stimuli (e.g., visual, emotional, tactile). In this case, focusing on just the face of a subject could be misleading and not fully representative. For these reasons, understanding the interpersonal relations promises a clear advancement in the way images and videos are processed to derive automatic annotations of humans status and behavior.

As summarized in Fig. 2, in this survey we proposed an at large categorization of the datasets and methods for interpersonal relation recognition into two classes given by, respectively, *objective* and *subjective* relations. Indeed, we observed that most of the existing datasets and methods fall into the first category. This can be motivated by the fact that objective relations are somewhat easier to detect also thanks to the availability of data that are accurately annotated. Datasets and methods in the second category are still less consolidated and there is room for substantial improvement.

In several of the works, facial expression features were extracted and used to characterize, together with other features, the interpersonal relations. Facial expressions are the activity of facial muscle according to the emotional states. Emotions are beyond the six basic emotions happiness, sadness, disgust, fear, surprise, and anger identified by Ekman [12], they are contagious [30] and facial expressions are one of the most important channels for them. Emotions can be an imitation of another person, a reaction or a reflex response to others [18]. Since the aim is to identify relation traits such as dominance, warm, and friendliness, social relation recognition goes beyond facial expression recognition, that is why most of the works has used facial expressions as one of several attributes to recognize interpersonal relationship. For example, in [45], authors focused on both face and body, the face attributes include age, gender, appearance, pose, location & scale and emotion. In [35], to predict the social relation, authors used several attributes: age, gender, location & scale, head appearance, head pose and face emotion (anger, happiness, sadness, surprise, fear, disgust and neutral), clothing, proximity and activity. In [52], the facial expression is one of the used attributes (together with gender, age, and head pose). Wanting to go beyond facial expressions, for the task of understanding human-centric situations in videos, authors in [42] used different attributes such as age, gender, ethnicity, profession, appearance, mental and emotional states such as happy, worried, calm, excited, quiet, amused.

Limitations In summary, though some interesting and useful methods and datasets have been proposed for understanding the interpersonal relations, from the current literature it appears evident that the existing solutions are still preliminary, and can provide effective results only in constrained scenarios. We identified some factors that can have an impact on this:

- **Which social relation model?** – There are several theories for describing the human social interactions; the choice of the best model to use also depends on the final application. However, in the design, acquisition and annotation of the existing datasets it is often missing a clear relation from the data and the existing theoretical models. This makes the related evaluations only partial;
- **Which datasets?** – Related to the above is the challenge of capturing large annotated datasets. Interpersonal relations happen in social person-to-person interaction, which is, on the one hand, difficult to capture in the wild and, on the other, complicated to be simulated in laboratory conditions. In addition, using laboratory settings with enrolled participants or actors reduces the variability of cases as well as the spontaneity in manifesting spontaneous behavior. Despite of this, there are some specific contexts, like

the teacher/student one, where real interpersonal relation data are produced in large quantity;

- **Which annotations?** – Annotating video data is probably the most complicated task in this domain. While online, distributed methods, like Mechanical Turk, can work properly when the objective of the annotation requires low experience and can be afforded by non trained people, the scenario totally changes for the interpersonal relation case. The annotations required for most of the behavioral and social models imply a quite large degree of variability and subjectivity that needs for experienced personnel.
- **Which techniques?** – The limitations and difficulties listed above are reflected in the methods developed so far. It is difficult to compare methods because they are often strictly related to specific contexts and data, so that it becomes also complicated to test the methods on different datasets. One trend that can be observed, as in most of the literature in computer vision and multimedia, is the shift to deep learning solutions, with methods that also exploit the temporal dimension in the data.
- Other limitations are related to the number of participants and their variability, the number of persons in the scene and their frontal/non-frontal pose, the number of recordings, the number and position of camera views, the quality of images/videos, the presence of audio track, and the different duration of a video clip that can lead to some ambiguity in the detection of interpersonal relations.

Perspectives The limitations evidenced by the current literature are clear directions for investigation and future development. In addition to this, we also prospect some additional line of interest.

We think predicting the short term evolution of the interpersonal relation can result in concrete applications. Indeed, having automatic methods capable of understanding the specific interpersonal relation is an important goal, but anticipating the future behavior based on the mood of the past and current interaction can be decisive. For example, a teacher/student scenario is an evident case, where detecting early signs of boring, loss of attention, or of learning difficulty, can be important to change and tune the teaching modality by adding interaction or using other ways to attract attention.

Acknowledgements Hajer Guerdelli was partially supported by the MOBIDOC scheme, funded by the Ministry of Higher Education and Scientific Research through the PromEssE project and managed by the ANPR.

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement. Hajer Guerdelli was partially supported by the MOBIDOC scheme, funded by the Ministry of Higher Education and Scientific Research through the PromEssE project and managed by the ANPR.

Declarations

Conflict of Interests Authors declare no conflicts of interest.

References

1. Alhagry S, Fahmy AA, El-Khoribi RA (2017) Emotion recognition based on EEG using LSTM recurrent neural network. Emotion 8.10:355–358
2. Ariano L, Ferrari C, Berretti S, Del Bimbo A (2021) Action unit detection by learning the deformation coefficients of a 3D morphable model. Sensors 21(2):589
3. Arushi G, Ma KT, Cheston T (2019) An End-To-End network for generating social relationship graphs. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
4. Bargal SA et al (2016) Emotion recognition in the wild from videos using images. Proceedings of the 18th ACM International Conference on Multimodal Interaction
5. Bottino AG, De Simone M, Laurentini A, Vieira T (2012) A new problem in face image analysis: finding kinship clues for siblings pairs. Conf Pattern Recognit Appl Methods
6. Bugental DB (2000) Acquisition of the algorithms of social life: a domain-based approach. Psychol Bull 126(2):187–219
7. Chanel G, Ansari-Asl K, Pun T (2007) Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: IEEE international conference on systems, man and cybernetics. IEEE, p 2007
8. Clark MS, Mills J (1979) Interpersonal attraction in exchange and communal relationships. J Person Soc Psychol 37(1):12
9. Cristani M, Raghavendra R, Bue AD, Murino V (2013) Human behavior analysis in video surveillance: A social signal processing perspective. Neurocomputing 100:86–97
10. Dai Q, Carr P, Sigal L, Hoiem D (2015) Family member identification from photo collections. In: Applications of computer vision, pp 982–989
11. Dehshibi MM, Bastanfard A (2010) Unsupervised feature based facial family similarity recognition. In: Proc International conference on image and video processing and computer vision (IVPCV-10), ISRST: 132–138
12. Ekman P (1992) An argument for basic emotions. Cogn Emot 6:169–200
13. Fang R, Tang KD, Snavely N, Chen T (2010) Towards computational models of kinship verification. In: International conference on image processing (ICIP) IEEE
14. Ferrari C, Berretti S, Pala P, Del Bimbo A (2018) Rendering realistic subject-dependent expression images by learning 3DMM deformation coefficients. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp 0–0
15. Ferrari C, Lisanti G, Berretti S, Del Bimbo A (2015) Dictionary learning based 3D morphable model construction for face recognition with varying expression and pose. In: 2015 international conference on 3d vision. IEEE, pp 509–517
16. Fiske AP (1992) The four elementary forms of sociality: framework for a unified theory of social relations. Psychol Rev 99(4):689
17. Foa EB, Foa UG (1980) Resource theory., Social exchange. Springer, US, pp 77–94
18. Frith C (2009) Role of facial expressions in social interactions. Philosophical Transactions of the Royal Society B: Biological Sciences 364.1535:3453–3458
19. Gao J, Qing L, Li L, Cheng Y, et Peng Y (2021) Multi-scale features based interpersonal relation recognition using higher-order graph neural network. Neurocomputing 456:243–252
20. Guo X, Polanía LF, Garcia-Frias J, Barner KE (2019) Social relationship recognition based on a hybrid deep neural network. In: 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019), Lille, France, pp 1–5. https://doi.org/10.1109/FG.2019.8756602
21. Hajer G, Claudio F, Walid B, Haythem G, Berretti S (2022) Macro-and micro-expressions facial datasets: A survey. Sensors 4(1524):22
22. Ho DY, relationships I, dominance relationship (1998) An analysis based on methodological relationism. Asian J Soc Psychol 1(1):1–16
23. Ingo S, Kim H, David P-H, Andreas W (2013) Human behaviour in HCI: Complex emotion detection through sparse speech features, 8212. https://doi.org/10.1007/978-3-319-02714-2-21
24. Kiesler DJ (1983) The 1982 interpersonal circle: a taxonomy for complementarity in human transactions. Psychol Rev 90(3):185
25. Klemmer ET, Snyder FW (1972) Measurement of time spent communicating. J Commun 22.2:142–158
26. Li W, Duan Y, Lu J, Feng J, Zhou J (2020) Graph-based social relation reasoning. European Conference on Computer Vision (ECCv). Springer, Cham
27. Li J, Wong Y, Zhao Q, Kankanhalli MS (2017) Dual-glance model for deciphering social relationships. In: 2017 IEEE international conference on computer vision (ICCV). Venice, Italy, pp 2669–2678. https://doi.org/10.1109/ICCV.2017.289

28. Liu X et al (2019) Social relation recognition from videos via multi-scale spatial-temporal reasoning. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). Long Beach, USA, pp 3561–3569. https://doi.org/10.1109/CVPR.2019.00368
29. Lu J, Hu J, Liong VE, Zhou X, Bottino A, Islam IU, Vieira TF, Qin X, Tan X, Chen S et al (2015) The fg 2015 kinship verification in the wild evaluation. In: Conference on automatic face and gesture recognition (FG), vol 1. IEEE, pp 1–7
30. Lundqvist L-O, Dimberg U (1995) Facial expressions are contagious. J Psychophysiol 9:203–203
31. Lv J, Liu W, Zhou L, Wu B, Ma H (2018) Multi-stream fusion model for social relation recognition from videos. In International Conference on Multimedia Modeling, Springer, Cham, pp 355–368.
32. MacCrimmon KR, Messick DM (1976) A framework for social motives. Behav Sci 21(2):86–100
33. Palmero C, Selva J, Smeureanu S, Junior JCS, Clapés A, Mosegui A, Zhang Z, Gallardo D, Guilera G, Leiva D, Escalera S (2020) Jacques context-aware personality inference in dyadic scenarios: Introducing the UDIVA Dataset, IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW). arXiv:2012.14259v1 [cs.CV]
34. Parsons T, Shils EA, Smelser NJ (eds) (1965) Toward a general theory of action. Theoretical foundations for the social sciences. Transaction Publishers, Routledge
35. Qianru S, Schiele B, Fritz M (2017) A domain based approach to social relation recognition. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 435–444
36. Rázuri G, Francisco J et al (2015) Recognition of emotions by the emotional feedback through behavioral human poses. Int J Comput Sci Issues 12.1:7–17
37. Robinson JP et al (2020) Families in wild multimedia (FIW-MM): A multi-modal database for recognizing kinship. arXiv:2007.14509: n. pag, 2020
38. Robinson JP, Shao M, Wu Y, Fu Y (2016) Families in the wild (FIW): Large-scale kinship image database and benchmarks. In: MM '16: Proceedings of the 24th ACM international conference on Multimedia, pp 242–246
39. Siyang S, Shen L, Valstar M (2018) Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In: 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018). IEEE, pp 158–165
40. Tanisik G, Zalluhoglu C, Ikizler-Cinbis N (2016) Facial descriptors for human interaction recognition in still image. Pattern Recognit Lett, ISSN 0167–8655
41. Tyshchuk Y, Wallace WA (2018) Modeling human behavior on social media in response to significant events. IEEE Trans Comput Soc Syst 5(2):444-457. https://doi.org/101109/TCSS.2018.2815786
42. Vicol P, Tapaswi M, Castrejon L, Fidler S (2018) MovieGraphs: Towards understanding Human-Centric situations from videos. 2018 IEEE conference on computer vision and pattern recognition (CVPR)
43. Vijayan V, Bowyer KW, Flynn PJ, Huang D, Chen L, Hansen M, Ocegueda O, Shah SK, Kakadiaris IA (2011) Twins 3d face recognition challenge. In: 2011 international joint conference on biometrics (IJCB). IEEE, pp 1–7
44. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, Kyoto, Japan, pp 3551–3558
45. Wang M, Du X, Shu X, Wang X, Tang J (2020) Deep supervised feature selection for social relationship recognition. Pattern Recognit Lett 138:410–416
46. Wang S, Gao JZ, Lin H, Shitole M, Reza L, Zhou S (2019) Dynamic human behavior pattern detection and classification. In: 2019 IEEE Fifth international conference on big data computing service and applications (BigDataService). IEEE, pp 159–166
47. Wang Z, Chen T, Ren J, Yu W, Cheng H, Lin L (2018) Deep reasoning with knowledge graph for social relationship understanding. In: 27th int. joint conference on artificial intelligence (IJCAI'18). AAAI Press, pp 1021–1028
48. Yan H, Hu J (2018) Video-based kinship verification using distance metric learning. Pattern Recognit
49. Yan H, Song C (2019) Semantic three-stream network for social relation recognition. Pattern Recognit Lett 128(2019):78–84
50. Yisi L, Sourina O, Nguyen MK (2011) Real-time EEG-based emotion recognition and its applications. In: Transactions on computational science XII. Springer, Berlin, pp 256–277
51. Zhou L, Wu B, Lv J (2018) SRE-net model for automatic social relation extraction from video. In CCF Conference on Big Data, Springer, Singapore, pp 442–460.
52. Zhang Z, Luo P, Loy CC, Tang X (2016) From facial expression recognition to interpersonal relation prediction. arXiv:1609.06426v2
53. Zhang M, Liu X, Liu W, Zhou A, Ma H, Mei T (2019) Multi-granularity reasoning for social relation recognition from images. In: IEEE international conference on multimedia and expo (ICME), pp 1618–1623. https://doi.org/10.1109/ICME.2019.00279

54. Zhang N, Paluri M, Taigman Y, Fergus R, Bourdev L (2015) Beyond frontal faces: Improving person recognition using multiple cues. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 4804–4813
55. Zeng Z et al (2007) Audio-visual spontaneous emotion recognition. Artifical intelligence for human computing. Springer, Berlin, pp 72–90