Visual Non-verbal Social Cues Data Modeling

Mahmoud Qodseya

Visual Non-verbal Social Cues Data Modeling

Mahmoud Qodseya(✉)

Institut de Recherche en Informatique de Toulouse (IRIT),
Université de Toulouse, Toulouse, France
Mahmoud.Qodseya@irit.fr

Abstract. Although many methods have been developed in social signal processing (SSP) field during the last decade, several issues related to data management and scalability are still emerging. As the existing visual non-verbal behavior analysis (VNBA) systems are task-oriented, they do not have comprehensive data models, and they are biased towards particular data acquisition procedures, social cues and analysis methods. In this paper, we propose a data model for the visual non-verbal cues. The proposed model is privacy-preserving in the sense that it grants decoupling social cues extraction phase from analysis one. Furthermore, this decoupling allows to evaluate and perform different combinations of extraction and analysis methods. Apart from the decoupling, our model can facilitate heterogeneous data fusion from different modalities since it facilitates the retrieval of any combination of different modalities and provides deep insight into the relationships among the VNBA systems components.

Keywords: Social signal processing · VNBA systems
Metadata modeling · Visual non-verbal cues

1 Introduction

Observing and understanding human reactions and interactions with other human beings are challenging and interesting research directions for a wide variety of applications from customer satisfaction estimation to social robots modeling. Thus, the main purpose of SSP field is to automatically recognize and understand the human social interactions by analyzing their sensed social cues. These cues are mainly classified into verbal (word, e.g., semantic linguistic content of speech) and non-verbal (wordless and visual) social cues. The verbal cues take into account the spoken information among persons like 'yes/no' response in the answering question context. The non-verbal social cues represent a set of temporal changes in neuromuscular and physiological activities, which send a message about feelings, mental state, personality, and other characteristics of human beings [1].

https://doi.org/10.1007/978-3-030-01391-2_16

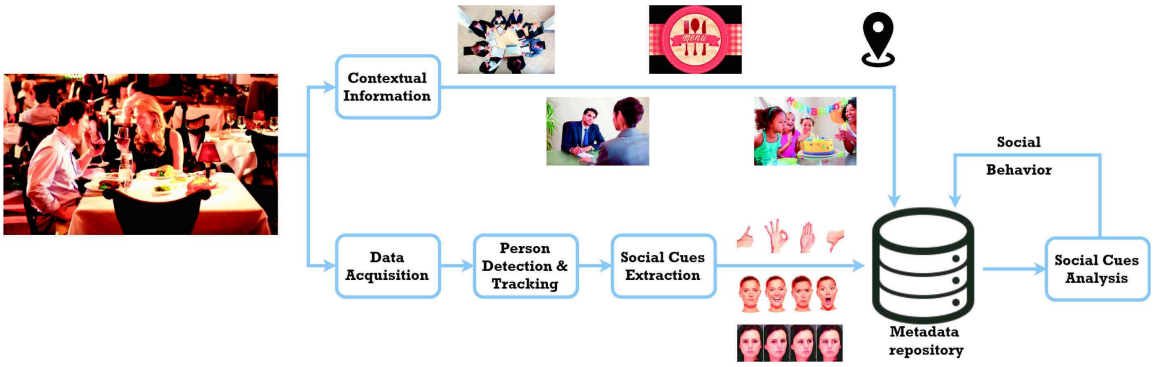

Fig. 1. A general visual non-verbal behavioral analysis schema.

Since crowded places are often noisy, most of the existing methods have been directed toward small group interactions (i.e., meeting, dining). Therefore, various visual-based SSP methods have been introduced in the literature [1–4], which generally can be summarized in a common system schema consisting of five modules, depicted in Fig. 1: (i) data acquisition, (ii) person detection and tracking, (iii) social cues extraction, (iv) contextual information, and (v) social cues analysis.

Different types of sensors and devices like cameras and proximity detectors might be exploited in the data acquisition module to record social interactions. Thus, one or more dedicated computer vision and image processing based (e.g., face detection) techniques could be leveraged for treating the input data to detect and track person(s). The social cues extraction module takes as an input the detected person(s) to extract a feature vector (per person) describing the social cues such as head *pose* (position and orientation). The social cues understanding module deeply analyzes the primitive social cues through modeling temporal dynamics and combining signals extracted from various modalities (e.g., head *pose*, facial expression) at different time scales to provide more useful information and conclusion on the behavioral level of the detected persons. Indeed, this module might optionally leverage additional information describing the context in which the data is captured to provide a precise social behavior prediction and analysis. Finally, the existence of metadata repository decouples the analysis phase from the other components as Fig. 1 shows.

In this paper, we introduce a metadata model for the visual cues as they are the mostly perceived and unconsciously displayed by human among the non-verbal cues. Our data model has many benefits in: (i) fusing heterogeneous data from different sources and modalities via facilitating the retrieval process, (ii) decoupling the social signal extraction phase from the analysis phase, and (iii) preserving privacy by facilitating access control layer integration within the metadata repository [5] to confront the privacy and integrity threats. Therefore, the analysis methods are going to use the extracted metadata instead of the original input data. Moreover, it is easy to extend the designed model for supporting the vocal cues as our model considers *per-frame* metadata associated with the detected persons.

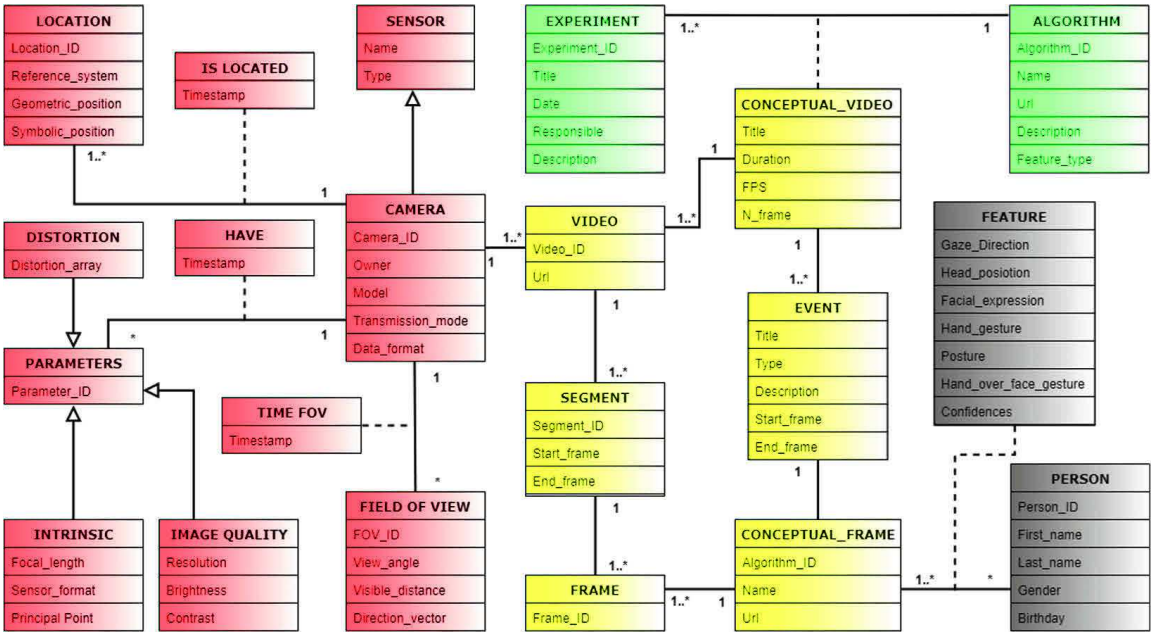

Fig. 2. Generic data model. This generic data model for visual non-verbal social cues shows the relationships that exist between experiment, acquisition, video, and feature groups of entities, which are color-coded as green, orange, yellow, and gray respectively. (Color figure online)

Consequently, this provides a capability to integrate the extracted vocal cues at the *frame-level* similar to *audio-video* synchronization.

Paper Outline. Sect. 2 presents a detailed description of the designed data model.The current state of our research is presented in Sect. 3. The conclusion of our work is summarized in Sect. 4 with some insights for future works.

2 Visual Non-verbal Social Cues Data Model

Visual non-verbal social cues have been received more attention as they are not semantic in nature and often occur unconsciously. The explored visual non-verbal cues include body movements, hand movements, gaze behavior, and signals with higher level of annotation like smiling, facial expressions, hand gestures, hand-over-face gestures, head orientations, and mutual gaze. To handle the high variety of the social visual cues, we design a conceptual data model consisting of experiment, acquisition, video, and feature groups of entities as Fig. 2 shows.

2.1 Experiment Group

Researchers or experts in the VNBA systems are interesting in performing experiments using different configurations of algorithm types and parameters. Thus, the EXPERIMENT class is dedicated to hold such information that includes experiment title, date, responsible person, location, and description. On the

other side, researcher could be interested in additional information about the list of algorithms that are used to extract the social cues from a conceptual video. Therefore, the ALGORITHM class is devoted to carry the algorithm's name, URL, description, and feature type (e.g., facial expression, gaze direction) that can be extracted using this algorithm.

2.2 Acquisition Group

Generally, different types of sensors (e.g., camera, GPS, IMU, and microphone) are used for social cues (verbal/non-verbal) acquiring. In the context of VNBA systems, cameras are widely adopted, and thus 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.

To model these parameters, 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, where 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 are 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, and (v) the IMAGE QUALITY class includes common image features such as resolution, brightness, and contrast.

2.3 Video Group

VIDEO, SEGMENT, and FRAME represent a decomposition relationship since a video clip is decomposed into segments which represent sequence of frames. An event is a thing that happens or takes place over a particular time interval (e.g., type of the played music during a time interval). So, a video clip could be decomposed into event(s) that contain(s) a sequence of frames. Although both event and segment represent a part of video, but the event has semantic descriptions (contextual information).

In multiple cameras views scenarios, for each *time-stamp*, we need to process multiple frames (same number of used cameras) together, where same person could appear in more than one frame. Thus, we will have inconsistence feature values related to the same detected person. To overcome this challenge, we introduce both CONCEPTUAL-FRAME and CONCEPTUAL-VIDEO classes. A conceptual frame represents one to N frames (N is the number of the adopted cameras within the experiment) that have a common time stamp and must

be analyzed together. A conceptual video represents one to N videos that are recoded within an experiment and must be analyzed together. Therefore, we handle the latter mentioned challenge using the conceptual frame as we will fuse the extracted cues at the *frame-level*.

Since the event is related to particular time interval and location, which are common within the multiple cameras views scenarios, we present an association relationship between EVENT and CONCEPTUAL-VIDEO classes as Fig. 2 shows.

2.4 Features Group

The "FEATURE" class is designed as an association class containing the extracted social cues that are extracted for each detected person inside the conceptual frame. The attributes of feature class include the most common visual non-verbal cues such as gaze direction, head position, facial expression, hand gesture, hand-over-face gesture, and their detection confidences. "PERSON" class contains information (name, age, and birthday) about the experiments' participants.

3 Current Work

Due to the lack of data, we are not able to evaluate the proposed system, so we are implementing an acquisition platform named OVALIE for data recording. Our prototype consists of four cameras that surrounded a table with four chairs, where each person who is sitting on the table his face will appear in two cameras. Furthermore, we have been implemented a software that ables to read multiple videos, detect and track multiple persons based on an upgraded version of OpenFace toolkit [6], and detect the facial expression for each detected person using affectiva SDK[1].

Meanwhile, we implemented a physical model using Mongo database and test (using synthetic data) different quires that can be used for retrieving features for the analysis phase. Currently, we start collecting food images which will be used for YOLO [7] retraining. The customized YOLO deep learning based model could be used for detecting and localizing the different types of food on the table.

4 Conclusion and Future Work

In this paper, we propose a metadata model for the VNBA systems, which is: (i) privacy-preserving, (ii) easily expendable to cover the vocal and verbal cues, (iii) smoothing the data fusion among multiple modalities, (iv) decoupling the social cues extraction phase from the analysis phase, and (v) allowing to evaluate different combinations of the extraction and analysis methods.

As a future direction, we are going to extend our model to include the verbal and non-verbal cues. In addition to that, we intend to collect and annotate a dataset for experimenting and validating our data model.

[1] https://developer.affectiva.com/.

References

1. Akhtar, Z., Falk, T.: Visual nonverbal behavior analysis: the path forward. In: IEEE MultiMedia (2017)
2. Vinciarelli, A., et al.: Bridging the gap between social animal and unsocial machine: a survey of social signal processing. IEEE Trans. Affect. Comput. **3**, 69–87 (2012)
3. Cristani, M., Raghavendra, R., Del Bue, A., Murino, V.: Human behavior analysis in video surveillance: a social signal processing perspective. Neurocomputing **100**, 86–97 (2013)
4. Salah, A.A., Pantic, M., Vinciarelli, A.: Recent developments in social signal processing. In: 2011 IEEE International Conference on Systems, Man, and Cybernetics, pp. 380–385, October 2011
5. Kukhun, D.A., Codreanu, D., Manzat, A.-M., Sedes, F.: Towards a pervasive access control within video surveillance systems. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 289–303. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40511-2_20
6. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66 (2018)
7. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv (2018)