Measurement and Visual Analytics of Nonverbal Behavior and Interaction

Matus Gasparik, Carolin Bronowicz, Susanne Bleisch

In spite of the recent rapid advances in artificial intelligence, computer vision and a proliferation of off-the-shelf tools, studying and understanding the dynamics of nonverbal behavior (NVB) and interaction still proves challenging. New emerging research disciplines, such as Social Signal Processing (SSP), attempt to leverage the computational methods and machine-learning to automatically extract NVB from high-volume multi-modal video, audio and natural language data with moderate success. The success of these automatic approaches is predicated on the availability of high-quality extensive training datasets and is complicated by the issues concerning the theoretical and context-specific validity of the predicted constructs.

One underexplored approach to utilize sensor data with significant potential to improve our understanding of NVB is Visual Analytics (VA). In this work, we present a methodological approach using the VA framework to extract and analyze NVB and interaction from video data in context of collaborative-learning. Using state-of-the-art Computer Vision techniques, we extract high-resolution time series of face, hands and body landmarks from video recordings of small groups of students tasked with collaboratively solving a task on a computer. We describe a visualization pipeline that transforms the raw landmarks to relevant NVB signals and generates visual representations that facilitate an effective analysis of temporal patterns by a human observer. Additionally, we present some visual-mapping techniques to deal with the visualization of complex multi-dimensional data or with the loss of data due to aggregation. We showcase the potential of the VA for studying individual and dyadic NVB using the temporal patterns in head movement, posture and the mutual orientation (facing directions) of pairs of persons.

1 Introduction

The understanding of nonverbal behavior (NVB) and communication (NVC) in group processes is of great interest across many interdisciplinary domains and application contexts. A notable example is educational research on Computer-Supported Collaborative Learning (CSCL), where enhanced engagement and improved learning outcomes have been demonstrated (J. Chen et al., 2018). However, our understanding of the interpersonal and the group dynamics (e.g. the quality of engagement) and how they manifest through nonverbal behavior is still in its early stages (Paneth et al., 2023).

The progress in studying the dynamics of NVB is predicated on our ability to measure NVB with sufficient precision and temporal resolution. The measurement of NVB has traditionally relied on coding by trained coders to manually annotate video recordings. However, this approach has several drawbacks, most notably the prohibitive human effort and high cognitive load necessary to annotate extensive lengths of video data as well as the limited temporal resolution and granularity of the annotated data. For practical reasons, the coding of NVB is typically performed over small regular time segments with an arbitrary length (on the order of minutes) and the observed nonverbal cues (e.g. smiling, eye contact) are recorded as aggregated values. The kind of aggregation (e.g. count/frequency versus duration) needs to be decided upfront and must be informed by the research objective in question or multiple aggregation values are coded jointly at the cost of increased effort. Moreover, if the coding involves abstract, higher-order constructs (e.g. assessing a person's "cognitive engagement"), other issues arise regarding the theoretical validity of the construct and potential bias introduced by subjective perception of the coder.

The recent rapid advancements in AI, computer vision, natural language or audio processing, as well as the abundance of ubiquitous low-cost sensors have sparked optimism about the potential of leveraging these technologies to measure and study NVB and social interaction. The introduction of wearable sensors as "sociometric badges" into social sciences and group research made it possible to collect vast amount of data over large temporal and spatial extents and outside the controlled laboratory settings (Kim *et al.* (2008), blok.etal_2017, Zhang *et al.* (2018)). The sociometric badges are typically equipped with microphones, accelerometers as well as proximity and radio frequency sensors to capture high-resolution data of the wearer's speech, body movement, location as well as proximity to others (Parker *et al.*, 2020). While they only allow to capture some aspects of NVB and, by themselves, might not be sufficient to study NVB and its dynamics with high granularity (unless used in conjunction of other sensors), the sociometric sensors have helped user the social sciences into the "big data" era.

One promising approach that might help address the shortcomings of the traditional methods of collecting data is *Social Signal Processing*. Social Signal Processing (SSP) is emerging research domain encompassing behavioral and social sciences, computer science and engineering with the ambitious goal of providing machines with social intelligence, the ability to "sense and understand human social signals" (Vinciarelli et al., 2009). SSP posits that the nonverbal cues encoded in facial expressions, posture, hand gestures or non-linguistic aspects of speech (paralanguage) convey machine-detectable social signals amenable to automatic analysis by algorithms. The process in SSP involves 1) the extraction of the nonverbal cues from multi-modal sensor data and 2) the "social signal understanding". The main challenges in machine understanding of social behavior lie in understanding and representing the temporal dynamics of NVB and in fusion of multi-modal data with varying timescales (Vinciarelli et al., 2009). In practice this is usually addressed with various machine-learning and feature engineering approaches and requires carefully curated training datasets labelled by domain experts. It should be also noted that available research typically focuses only on specific aspects of nonverbal behavior with a goal to automatically predict, for example, turn-taking (Vinciarelli et al., 2009) or task and social cohesion (Lehmann-Willenbrock & Hung, 2023). Deep-learning approaches are superior to classical machine-learning in that they don't require feature engineering and typically achieve much better performance in many tasks. In the context of NVB and deep-learning, methods have been used to directly predict the theoretical construct of interest, e.g. the group learning engagement (Zheng et al., 2023).

However, the ability to collect large amounts of nearly continuous quantitative data in combination with data-driven algorithms is not a panacea when it comes to automatically detecting nonverbal behaviors and comes with its own drawbacks. First and foremost, it does not completely eliminate the manual coding as the training of models for automatic detection of NVB require substantial effort by domain experts to create and maintain high-quality labelled datasets (Lehmann-Willenbrock & Hung, 2023). The trained models will likely only work in narrow contexts depending on the specified objective of the model as well as the characteristics of the sensor data. Secondly, if the models are designed to predict abstract, higher-level constructs, the concerns surrounding their theoretical validity (Luciano et al., 2018; Müller et al., 2019), consideration of contextual and environmental factors and potential bias become even more pressing (Renier et al., 2021). More importantly, these methods are limited in terms of generating new insight into the NVB as they serve merely as a means to automate the existing coding routines.

Interactive visualizations offer an alternative and effective means of analyzing large and complex multi-dimensional data, especially in cases where the theoretical underpinnings of the observed phenomena are not yet fully established. We believe that adopting the visual analytics approach to study the NVB has the potential to improve our understanding of the complex interactions and temporal dynamics of interaction and group processes. Instead of reducing the raw input data into the model outputs in an opaque, black-box manner, visual analytics offers a semi-automatic approach, where a human observer engages and interacts with visual representations of the data in an iterative process, with the goal of generating insight from the data. In the visualization pipeline, the raw data is taken through a series of adjustable transformations that ultimately produce a low-dimensional (ideally 1D or 2D) graphical representation that facilitates rapid insight into the complex data thanks to the visual and cognitive capabilities of a human observer. These transformations operate on the data or on visualizations themselves and in the case of the former, should be limited to algebraic or well-established feature extraction approaches with little ambiguity in order to not negatively affect the theoretical validity. For example, instead of relying on a black-box model to predict some abstract construct (such as the quality of the group engagement), we seek to derive new signals using computer vision, face and human pose detections algorithms and perform some form of visual signal fusion to derive visualizations which are effective in revealing temporal patterns.

In this contribution we outline a methodological approach using visual analytics (VA) to study nonverbal behavior (NVB) and interaction-based video recordings of group learning activity. Concretely, we show a practical approach to:

- extract face, hands and body landmarks at arbitrary temporal resolution using simple monocular camera recordings and off-the-shelf computer vision tools,
- structure the data in a way that is optimized for working with large multi-dimensional time-series data,
- transform the landmark data into nonverbal signals corresponding to individual's head movements and changes in posture,
- derive an approximation of eye contact in dyadic interaction using the head's pose and simple vector algebra,
- use some non-standard visual mapping techniques to deal with high dimensionality of the data or the loss of detail due to data aggregation

We note that our goal in this work is to demonstrate some feasible methodological approaches and provide pointers for further research rather than trying to present a comprehensive and wholistic treatment of how to measure and analyze nonverbal behavior and group dynamics using the visual analytics framework. VA is agnostic to origin and type of the analyzed signals and lends itself perfectly to studying complex multi-modal such as social signals. The specifics of data collection, processing, and visualization procedures should be guided by careful consideration of the inherent trade-offs between the desired scope, quality, and granularity of the data and the required experimental effort.

2 Related work & state of the Art

Human visual perception and cognition are remarkably powerful systems capable of processing large amounts of visual information in parallel and in a non-linear fashion. Unlike language or auditory processing, which is inherently sequential and temporally constrained, the visual system enables rapid, simultaneous assessment of multiple spatial features such as shape, color, position, and motion (Ware, 2013). This parallelism allows humans to detect patterns, trends, and anomalies in complex visual scenes almost instantaneously, a capability that underlies the effectiveness of visualization as a cognitive aid. Visual representations have long served as external aids to human cognition, predating the advent of computers by centuries. From early cartographic representations and scientific diagrams to the use of charts in statistical analysis, visualizations have historically supported reasoning, discovery, and communication (Card et al., 1999; Shneiderman, 1996). They extend human cognitive capacity by externalizing abstract data into visual form, thereby leveraging the efficiency of the human perceptual system for analytical tasks. Visualization can be broadly defined as a mapping of data to graphical forms through a series of adjustable transformations. These transformations translate data attributes into visual variables, such as position, size, shape, color, or motion, that can be perceived and interpreted by human observers. The goal of this mapping process is to generate visual representations that facilitate the rapid discovery of patterns and relationships within data, supporting insight generation and hypothesis formation. Interaction plays a crucial role in this process. By allowing users to probe and explore data at arbitrary levels of detail, interaction transforms visualization from a static display into a dynamic exploratory environment. Common forms of interaction include modifying data transformations (e.g., filtering, aggregation), adjusting visual mappings (e.g., color encoding or scaling), and performing view transformations (e.g., panning, zooming, navigation). These principles are encapsulated in Shneiderman's Visual Information-Seeking Mantra: "Overview first, zoom and filter, then details on demand" (Shneiderman, 1996). The visual design of a visualization is equally critical. It involves developing visual metaphors that effectively represent abstract or complex data. Since many data types, particularly those without inherent spatial structure, lack intuitive graphical analogs, thoughtful design choices are required to ensure that visual forms convey the intended meaning and maintain perceptual clarity. Information Visualization is a subfield concerned with the visual representation of abstract, non-spatial data (C. Chen, 2010). Unlike scientific or geospatial visualization, which deals with physical or spatially grounded phenomena (e.g., molecular structures, terrain models), Information Visualization addresses data without natural spatial mappings, such as networks, text corpora, or interaction patterns. The core challenges of Information Visualization lie in the development of suitable visual encodings and interaction techniques for such abstract data, as well as the design of algorithms that can efficiently transform data into meaningful visual structures. Through iterative design and user-centered evaluation, Information Visualization seeks to create representations that make underlying data patterns and relationships perceptually salient and cognitively accessible.

Visual Analytics (VA) extends the principles of visualization and Information Visualization by integrating them with methods from data analysis, machine learning, and human–computer interaction to support analytical reasoning and decision-making (C. Chen, 2010). It is a multidisciplinary field encompassing (1) analytical reasoning techniques for deriving insight, (2) visual representations and interaction techniques for exploring data, and (3) data representations and transformations for managing complexity (Cook & Thomas, 2005). The central aim of VA is to combine the computational power of algorithms with the perceptual and cognitive strengths of humans, thereby creating a human–machine analytical loop. Through this integration, users can interactively guide computational processes, refine hypotheses, and iteratively explore data representations. Despite significant advances, the discipline of VA remains in an evolving state. Designing effective visual representations for complex, high-dimensional, or abstract data continues to be as much a craft as it is a science. A related concept is visual data mining, which leverages visualization to support the discovery of structures, correlations, and outliers in large datasets (Keim & Ward, 2003). Visual data mining and VA share the goal of enabling interactive exploration, but VA typically places stronger emphasis on reasoning and decision-making in complex, uncertain contexts.

In spite of the rapid advances in artificial intelligence, computer vision, and machine learning, understanding nonverbal behavior (NVB) and interaction dynamics remains a significant challenge. Automated methods in Social Signal Processing (SSP) aim to extract NVB from multimodal data using machine-learning techniques, yet

their success depends heavily on large, high-quality training datasets and raises questions about the theoretical validity of inferred constructs. In this context, Visual Analytics offers an underexplored but promising approach. By combining computational feature extraction with interactive visualization, VA enables the exploration of complex temporal and relational patterns in NVB data that are difficult to capture through purely algorithmic or statistical means. Through interactive visual mappings, researchers can observe synchrony, mimicry, and coordination patterns across individuals or groups, facilitating insights into the social and cognitive dynamics of interaction. Applying the VA framework to NVB analysis thus provides a complementary, human-centered approach to data interpretation, one that aligns with the strengths of our visual–cognitive system and allows analysts to discover emergent behavioral structures that might otherwise remain hidden in raw or aggregated data.

3 Methodological Framework

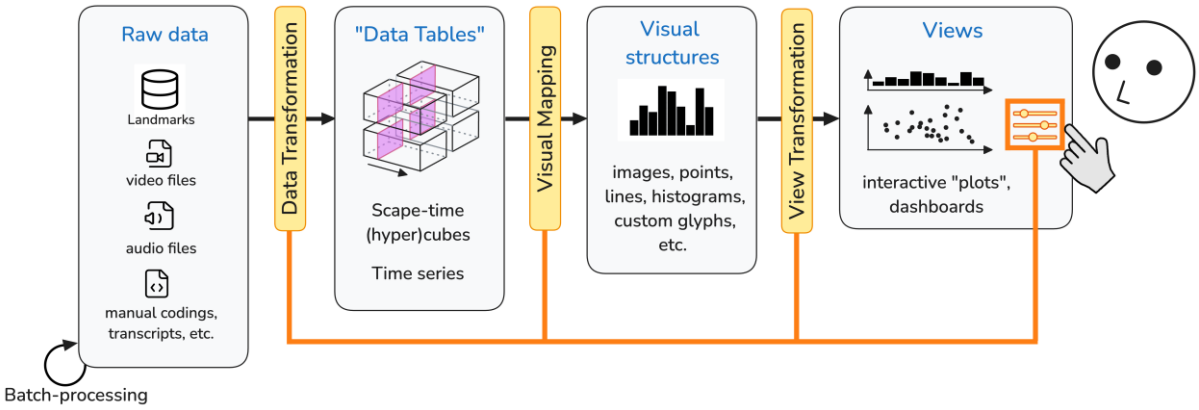


Figure 1: Schematic representation of a visualization pipeline contextualized to the analysis of nonverbal behavior (adapted from the original by Card *et al.* (1999, pg. 17)).

3.2 Data Acquisitions

Video recordings of groups of 3-4 students tasked with collaboratively solving an assignment on a computer were obtained for two types of settings: 1) a face-to-face (*f2f*) laboratory setting and 2) an online meeting for a study performed during the Covid-19 pandemics. The broader research context as well as the details on the study design and the compliance with ethical standards are reported in Paneth *et al.* (2024). In the face-to-face setting, the students were seated next to each other at a large table equipped with two monitors, a computer mouse and a keyboard. A video camera was installed at a distance of ca. 1.5 m from the table with a good view on all three participants. A typical consumer-grade digital camera (Panasonic HC-X909) with integrated microphone were used to record each learning session. For the online study, recordings of online meeting sessions were produced consisting of a mosaic of the individual participants' camera feeds. The lengths of the video recordings ranged from 75 minutes to 100 minutes, and from 25 minutes to 30 minutes, for the f2f and the online study, respectively.

3.3 Batch-processing

For the extraction of landmarks from video recordings we used the MediaPipe (Lugaresi *et al.*, 2019) library and specifically the "holistic" solution suite which combines the perception pipelines for detecting face, hand and body (pose) landmarks into a semantically consistent end-to-end solution. Our rationale for choosing MediaPipe was primarily based on the project's multi-platform and multi-language support, the ease of integration into our codebase and its real-time capabilities making the processing of large amounts of video data manageable. Table 1 shows the sizes of face, hand and pose landmarks in terms of the number of distinct points. A total of 543 landmarks are thus detected for a single person for each sampled video frame. In our case however, this didn't include the lower portions of the body (hips, legs, feet) as they were occluded from the camera view for the entirety of the video recording.

Table 1: Dimensions of landmarks extracted with the MediaPipe library (mediapipe-holistic, 2020)

Feature	Size	Coordinate system
Face	468 × 3	screen, metric
Pose	33 × 4	screen
Hands (lhand, rhand)	21 × 3	screen

A schematic of the batch-processing routine designed to extract the landmark data from videos is shown in Figure 2. Starting with the raw video recordings, first, the individual frames are sampled at a specified constant

sampling rate. We used an arbitrary sampling rate of 5 frames per second (5 Hz) in order to obtain data at sufficiently high temporal resolutions. This sampling rate is high enough that it enables capturing micro-expressions such as head nodding from the head movement data. In the online study, the video recordings consisted of a mosaic of individual participants' camera feeds. After the frame sampling step, each mosaic frame was divided into ROIs (regions of interest) corresponding to the individual camera feeds, and each ROI was then processed separately. Since MediaPipe's holistic solution lacked multi-person functionality at the time of the study, the pipeline incorporated a masking step. This step generated a sequence of frames, each masking all but one person, before passing them to the landmark extraction stage. We used the pre-trained YOLOv8 model to obtain the instance segmentation maps for the *person* class and applied simple heuristics (left-to-right sorting) to identify the individual persons in the frame. It should be noted that this simple approach is only applicable to static scenes (as in our case) where the participants don't change their positions relative to each other during the video recording. A more robust approach would need to involve some form of person tracking. Finally, the extracted landmarks for each individual person are stored together with corresponding timestamps and the metadata consisting of anonymous study, group and person identifiers in the HDF5 format which is optimized for large hierarchical and multi-dimensional array data. The HDF5 datasets contain the landmark data for all groups for all persons within each group and serve as inputs into the visualization pipeline.

An example of extracted landmarks for a single video frame is presented in Figure 3. It shows the face-mesh consisting of 468 discrete, the upper-body portion of the pose (body) landmarks and the left and the right hand landmarks each consisting of 21 discrete points and using different color-coding for right versus left hand.

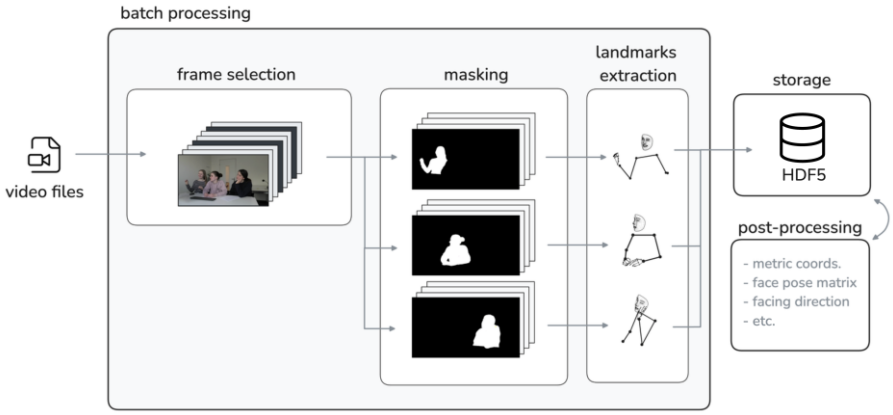


Figure 2: Schematics of the batch-processing pipeline for the extraction of landmarks from video recording.

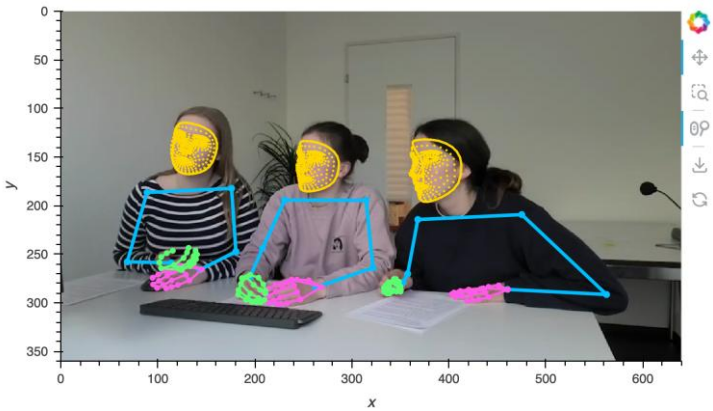


Figure 3: Example of a video frame with extracted landmarks shown as an overlay.

As the raw video recordings constitute personal, potentially sensitive data of the study participants, we opted for a strict separation of any person-related data in the processed datasets. The landmark data obtained from the batch-processing are stored in numeric coordinate values associated with the abstract person identifiers (designated as «P1», «P2», etc.) and without the corresponding video frame samples. In order to be able to quality-check the landmark detections or produce visualization outputs, we developed an efficient routine for combining the landmark data with the corresponding video frames on-the-fly, based on a performant random-access retrieval of arbitrary frames within a video file. This also had the positive side-effect of significantly reducing the file sizes required for storing the HDF5 datasets.

3.4 Coordinate spaces

The output of the landmark detection algorithm (MediaPipe) is given in 3D screen coordinates that represent the landmark position in the video frame, normalized to the frame's width and height, respectively (Figure 4, left). z-values are relative coordinates obtained under the *weak perspective camera model* and rescaled to match the scale of x (Kartynnik *et al.*, 2019). Thus, the z-coords are not metrically accurate but acceptable for the intended application domain of the MediaPipe suite of tools (AR). We denote as "pixel coordinates" the 2D screen coordinates scaled to the video frame's width and the height (Figure 4, middle). The pixel coordinates are thus dependent on the resolution of the video frame.

For the face landmarks, the MediaPipe library also includes a routine that estimates the metric coordinates based on a light-weight **Procrustes analysis** of a canonical face-mesh model which is specified in centimeter units (Figure 4, right). We utilized the *Python* version by Rasmus (2021) to obtain the 3D metric coordinates of the face-mesh as well as its position and orientation (pose) under the rigid-body transformation.

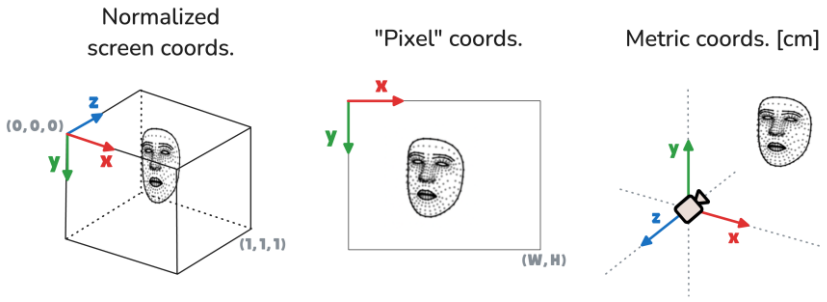


Figure 4: The different coordinate systems used for the landmarks at different stages of the visualization pipeline.

3.5 Data representation

Within the visualization pipeline, the processed data from the HDF5 store for an entire group (single video recording) is mapped to a specialized in-memory data structure which is optimized for handling multi-dimensional array data. Figure 5 shows the conceptual representation of this data structure which was implemented using the xarray Python library (Hoyer & Joseph, 2017). This structure uses a well-established API (Application Programming Interface) for handling complex numerical arrays in the Python ecosystem and facilitates slicing or aggregating the data along arbitrary dimensions. Moreover, performant mathematical operations thanks to the broadcasting and vectorization (Harris *et al.*, 2020) can be applied to entire multi-dimensional arrays which is crucial for handling large data withing the visualization pipeline.

This data representation aligns effectively with the conceptual approach of the space-time cubes for visualizing complex temporal data described by Bach *et al.* (2014). In fact, it can thought of as a generalization of the space-time cube into higher dimensions ("space-time hypercubes"). While it is beyond the scope of this work to elaborate on such a generalization of the conceptual framework, in practical terms, this data representation together with visualization APIs such as HoloViews (Stevens *et al.*, 2015) that support easy creation of interactive visualizations over dimensioned containers of data, make many of the operations described in Bach *et al.* (2014) trivial to implement.

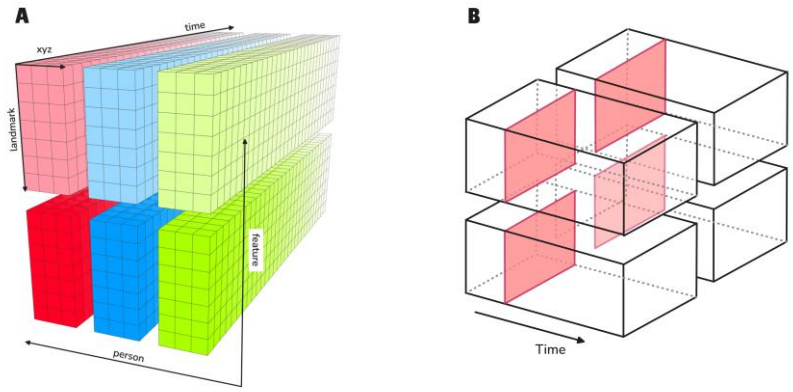


Figure 5: A: Conceptual representation of a data structure used for accessing the landmarks data of an entire group from the HDF5 store as a multi-dimensional array with different dimensions. The "feature" dimension denotes either a set of

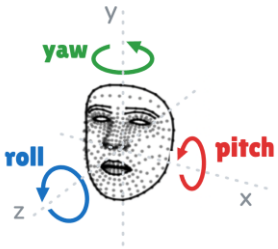
landmarks corresponding to distinct body parts (face, left hand, etc.) or a derived quantity calculated in the post-processing such as face-mesh position and direction, "landmark" refers to the individual landmark index within a set and "xyz" represent the 3D cartesian coordinates. B: Schematic representation of the "space-time hypercube" conceptual framework as an extension of the space-time cube by Bach *et al.* (2014) into higher dimensions.

3.6 Derived signals

The extracted landmarks represent dense geometric data from which it is possible to extract new derived signals relevant to NVB with suitable data transformations. A range of spatial queries can be made with respect to the instantaneous (e.g. position, orientation, distance) or dynamic values (e.g. velocity, acceleration). In this work, we explored only a few of such derived signals described below and much more could be done in future research. For example, the metric face-mesh data could be used to derive new signals related to the facial expressions within the permissible spatial and temporal resolution, or the hands and the pose landmarks could be used for fine-granular analysis of gestures. In the context of group processes research, even more interesting would be the exploration of the spatio-temporal patterns at the group level by considering signals derived from all persons simultaneously.

Through the Procrustes Analysis used by the MediaPipe library, we obtain the face-mesh in metric coordinates superimposed with the canonical face-mesh model. This procedure removes the scale, rotation, and translation components and additionally provides the 4x4 homogeneous transformation matrix representing the rigid-body transformation of the face-mesh. From this, we extract the translational components (t_x, t_y, t_z) representing the position of the face-mesh in the metric space and the axis-aligned elemental rotations corresponding to "pitch", "yaw" and "roll" (extrinsic Euler angles, Figure 6) using the scipy.spatial module (Virtanen *et al.*, 2020):

$$\mathbf{H} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{R} = \mathbf{R}(\text{pitch,yaw,roll}) \tag{1}$$



Angle name	Axis of rotation	Description
Yaw	y-axis	Left-right head rotation
Pitch	x-axis	Up-down head rotation
Roll	z-axis	Tilting head sideways

Figure 6: The rotation of the face-mesh around the x-axis (pitch), and its first two derivatives

Using the face-mesh pose, we estimate the position and the direction of individual's face (head). The "facing" direction ($\hat{d}$) is obtained by the matrix-vector multiplication of the face pose (H) with forward-facing vector of canonical face-mesh corresponding to [0,0,1]:

$$\hat{d} = H\hat{z} \qquad \hat{z} = [0,0,1] \tag{2}$$

Based on the positions and directions we can estimate the mutual facing orientations for each pair of individuals which serve as a proxy for eye-contact, or more generally, for evaluating how individuals are spatially oriented with respect to each other.

For a given pair (persons A and B) let p_A and p_B denote the positions, and $\widehat{d_A}$ and $\widehat{d_B}$ the facing-directions, respectively. The relative positional vector r_{BA} is defined as the vector from A to B:

$$r_{AB} = p_B - p_A \tag{3}$$

What we are looking for is a projection of $\widehat{d_A}$ onto r_{BA} :

$$u_{f1} = \text{proj}_{r_{AB}} \widehat{d_A} = \frac{\widehat{d_A} \cdot r_{BA}}{|r_{AB}|^2} r_{BA} \tag{4}$$

Using normalized vectors, this simplifies to a dot product:

$$\mathrm{dot}_{A\to B} = \widehat{d_A} \cdot \widehat{r_{AB}} \tag{5}$$

which yields a value between -1 (facing away) and 1 (facing directly towards the other person). Finally, to obtain the "facing orientation index" we map the value of the dot product to a more interpretable range of $[0,1]$:

$$\mathrm{orientation}_{A\to B} \approx \frac{\mathrm{dot}_{A\to B}+1}{2} \tag{6}$$

We refer to this value as "mutual orientation index" ("MOI") as it does not necessarily correspond to the eye contact. A representation of recorded scene from the bird's-eye perspective and the vector representation for the calculation of the MOI is shown in Figure 7.

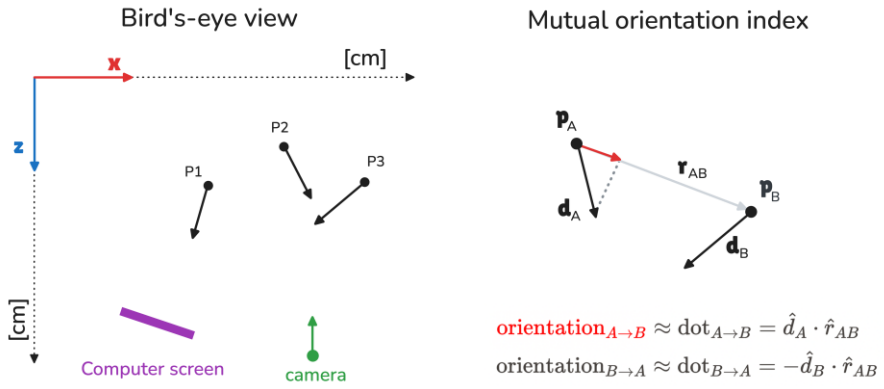


Figure 7: Left: Bird's-eye view of the recorded scene showing the positions and orientations of the face-meshes for persons P1 to P3 together with the placement of the camera a computer screen in metric coordinates. Right: Vector representation used for the calculation of the "*mutual orientation indices*" (MOI) between persons "A" and "B".

Kinematic Posture Model

To streamline the analysis of postural changes we implemented a kinematic model interface analogous to those used in robotics and human motion imitation (Kulić, 2019). The kinematic model is defined as a hierarchical graph of nodes representing anatomic landmarks (i.e. joints) or mechanical points (e.g. center of mass for the face mesh) starting with the midpoint between the shoulders as an arbitrarily defined root node. Due to occlusions, we limit the analysis to the visible portion of the upper body. Since the estimation of the metric coordinates in MediaPipe is only possible for the face landmarks (face-mesh), but not for hands and body landmark, all geometric queries are performed in pixel coordinates. While this is limiting and doesn't allow true spatial queries to be performed, it is nevertheless sufficient information to analyze temporal changes in the spatial configuration of the posture.

We note that the purpose of our kinematic model is not to provide the forward and inverse kinematics (Kulić, 2019) functionality as in robotics. Rather, its main purpose is to provide an intuitive API for defining geometric or motion-dynamic queries related to the posture (such as angles, accelerations or distances between distinct nodes, Figure 8) in the code. In a later section, we show how these data can be used as input features for a vector-embedding-based dimensionality reduction approach to explore temporal patterns in postural changes.

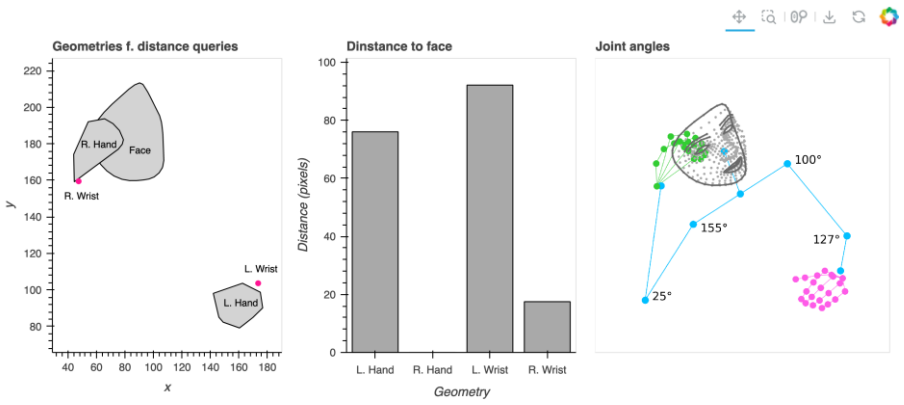


Figure 8: Posture analysis on an example of face-to-hand distance and joint angle geometric queries. (Left) Geometric objects used in face-to-hand distance queries defined as points (wrists) or polygons (face, hands). Middle: distances to face for wrist points and hand polygons, respectively, in pixel units. (Right) Joint angles in degrees.

3.7 Visual Mappings

Visual mapping refers to the mapping of data to visual (graphical) structures arranged spatially on the output medium (e.g. screen). Different data attributes or dimensions can be encoded along different visual channels such as position, length, area, color, shape, layout, etc. However, creating efficient visual representations for abstract or non-spatial data is a non-trivial task that must be guided by the principles of design and the Gestalt principles of visual perception and evaluated in terms of expressiveness and effectiveness of the visual design (Munzner, 2014). Visualizations of data with more than 3 dimensions present a similar challenge, whereby specially-designed visual mappings are used to map such data onto a 2D medium. For example, Beddow (1990) used a glyph-based visualization technique, which he dubbed "shape coding", to analyze the temporal patterns in 13-parameter magnetosphere and solar wind data. For the purposes of jointly visualizing higher-dimensional time-oriented data in the present work, we implemented a custom glyph-based visual mapping similar to the "shape coding" by Beddow (1990), which we refer to as "glyph arrays" and "glyph maps". Figure 9 depicts the general idea behind glyph arrays and glyph maps using random dummy data. The glyph arrays are made up of a 2D arrangement of elemental shapes (circle of square) where the values (or potentially two values) can be encoded using the size or the color (or both). Since the output medium for the visualization is limited to two dimensions, the higher dimensions (beyond 3 or 4) are handled by arranging the glyph arrays into hierarchies (glyph maps) using the proximity Gestalt principle in a recursive pattern (Figure 9, bottom). The dimensions need not necessarily represent distinct (spatial) dimensions and can instead represent a space-filling arrangement of an otherwise 1D data (similar to text-wrapping and arrangement into multiple columns). In this implementation the individual dimensions are strictly ordinal and in concrete visualizations (Figure 13) they are annotated with text marks or accompanied by an explanatory legend.

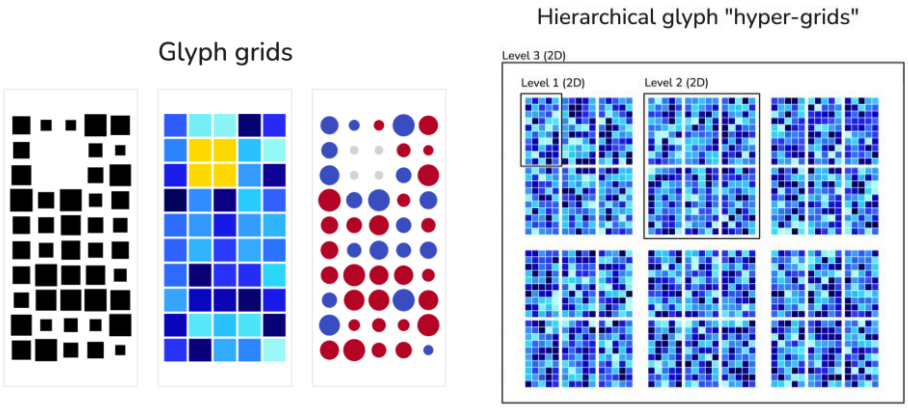


Figure 9: Visualization of higher-dimensional data using "glyph arrays" and "glyph maps". (left) example glyph arrays of random data with spatial dimensions of 10x5 encoding single or two values using different visual channels (color, size). The shape of the elemental glyph (square or circle) can be chosen deliberately but is not used as a visual encoding. Missing data (2. and 3. columns and rows) can be indicated with a distinct color or size (or both). (right) individual glyph arrays can be arranged hierarchically based on the proximity Gestalt principle to represent higher dimensions on a 2D medium.

4 Demonstrative Use Cases

Head movement patterns

The visualization of the raw motion data with very high temporal resolution is impractical for high-level tasks such as analyzing patterns in NVB as the data quantity is overwhelming and difficult to make sense of. Interactive visualizations alleviate this problem somewhat thanks to zooming and panning. For example, Figure 10 shows a segment of the x-rotation (pitch) data for a particular person with a notable peak corresponding to head-nod. Another way to reduce the information overload is to aggregate (resample) the raw data at lower time resolution at the cost of losing detail and the possibility to analyze small time-scale events (micro-expressions).

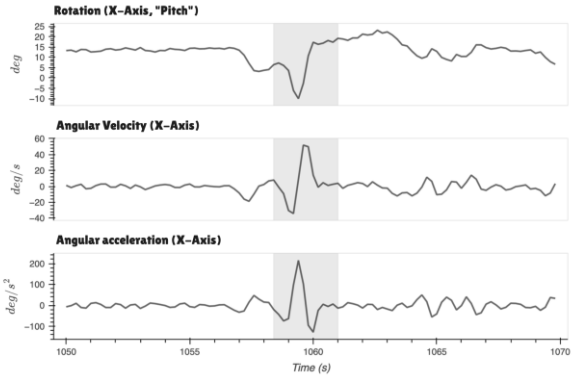


Figure 10: A zoomed-in time-series visualization of the face-mesh rotation around the x-axis (pitch) and its first two derivatives showing a segment corresponding to a head nod (highlighted area).

However, if one wants to understand the overall patterns of motion such as the head movement, the typical visualizations of temporal data (e.g. line plots), while providing accurate representations of variations in the signal, are not suitable for representing jointly multiple signals. For example, when studying the translational and rotational components of the head movement, we found that heatmaps were very effective in depicting the overall patterns of motion. Figure 11 shows an example of such a heatmap based on one group from the online study that has some interesting characteristics. In an interactive dashboard this representation could serve as an overview in the overview-detail linked plots making it possible to simultaneously observe general trends and small time-scale details.

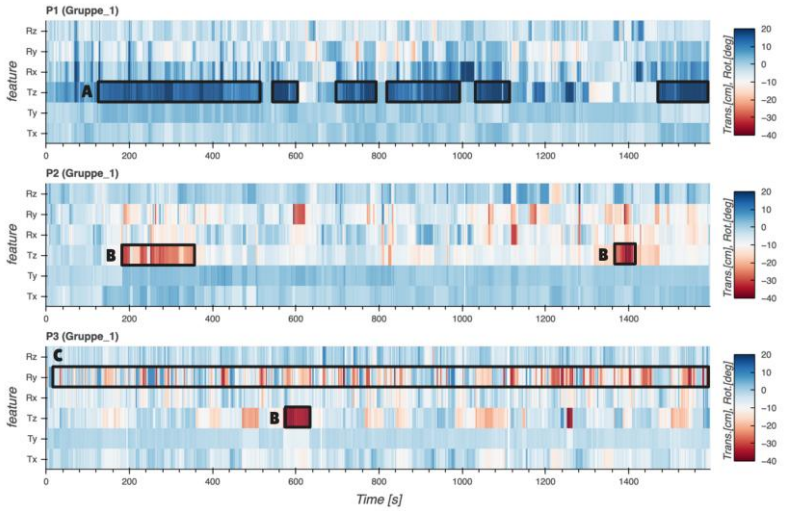


Figure 11: Heatmaps showing the temporal variations of face positions and orientations represented by the translational (Tx, Ty, Tz) and the rotational (Rx=yaw, Ry=pitch, Rz=roll) components for persons P1, P2 and P3. All values are represented as changes relative to the initial values at time = 0. x,y,z subscripts designate the axes of the metric (camera) coordinate space (Figure 4). Hence, for example a positive (or negative Tz) component represents positions closer to (or further from) the camera (areas A or B, respectively). The positive or negative values for the rotational components designate the sense (clockwise or counter-clockwise) of rotation. Notice the constant switching of the left-right rotation (Ry component) for the person P3 due to a presence of two screens (area C).

Dyadic interaction: MOI

Using the MOIs derived from face-mesh pose data we demonstrate the usefulness of custom visual mapping leveraging the glyph maps approach. For comparison, in Figure 12 we show the MOI data for each ordered pair of persons and for each person with respect to the computer screen. It is apparent that in this form the visualization isn't very helpful if one want to observe temporal patterns of MOIs in relation to one another.

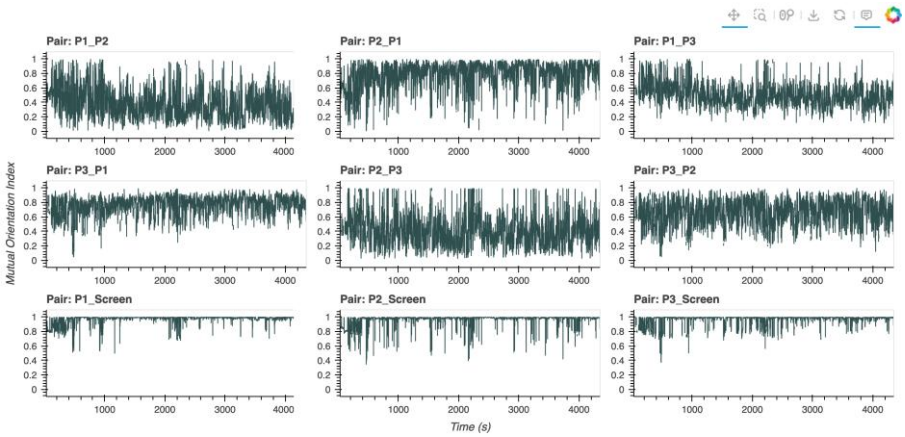


Figure 12: Time series of MOIs between individual persons within a group and between each person and the computer screen. While this kind of visual representation might be useful for revealing the temporal patterns of individual MOIs, it renders the comparison between the MOIs at each time instance impossible.

In contrast, Figure 13 shows a more efficient visual representation of the same data using the glyph maps. In this case, the data were aggregated in 5 seconds intervals based on maximum values to reduce the data quantity (though not practical, the glyph maps could be on used on full-resolution data). The individual glyph arrays are defined as a 3x3 grid where the diagonal elements represent the MOIs between each person and the computer screen and the off-diagonal elements represent the MOIs for individual ordered pairs of persons. The color-mapping is based on binarization of the original MOI values using an arbitrary user-controlled threshold value. The arrangement of the glyph arrays in the glyph map follows the space-filling pattern explained in the legend and in the first-level groups the glyphs arrays are based on 10-minute intervals. The second-level arrangement lays the 10-minute glyphs arrays into columns and rows based on available space. It should be noted, that in general, the hierarchies in this kinde of 2D glyph-maps can be based upon any dimension and can potentially include coordinate axes.

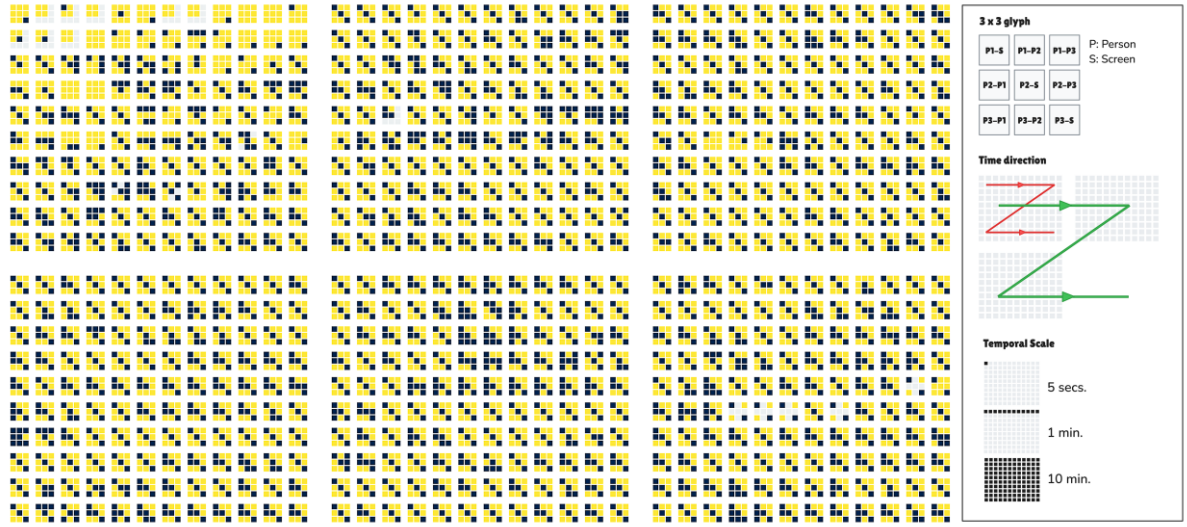


Figure 13: An alternative visual representation of the MOI data (resampled at 5 second intervals and using maximum as aggregation values) based on a custom-designed hierarchical 2D glyph-map. The elementary graphical marks (squares) are organized into hierarchies using the proximity gestalt principle. The color encodes values that are below (yellow) or above (dark blue) an arbitrary, user-controlled threshold MOI value (here, >= 0.94). Referring to the legend on the right-hand-side, the first hierarchy level is as a 3x3 grid of squares representing the MOIs for all pairs and between each person and the computer screen (diagonal elements). Next, the 3x3 square-glyphs are arranged in 12 columns and 10 rows representing increments of 5 seconds and 1 minute, respectively, in consecutive data and containing a total of 10 minutes worth of data. Finally, each of the 10-minute glyphs arrays are laid out in horizontal and vertical arrangement, so as to make efficient use of the available screen (medium) space. The time progression is analogous to the left-to-right text directionality.

The salient visual features in this representation result in readily observable patterns and after a short consultation of the legend, it is easy to identify specific MOI patterns in time (Figure 14). Additionally, more nuanced aggregation functions could be used to more faithfully summarize the instantaneous MOIs over a time interval and could be parametrized as user inputs. For instance, in our examples the maximum MOI values do not explicitly

consider the duration and capture transient head movements that only momentarily align with the other member of the MOI pair (in other words, this aggregation behaves similarly to peak-hold meters). This is noticeable in the prominent diagonal patterns of glyph arrays but is less of a problem for the MOIs between the persons.

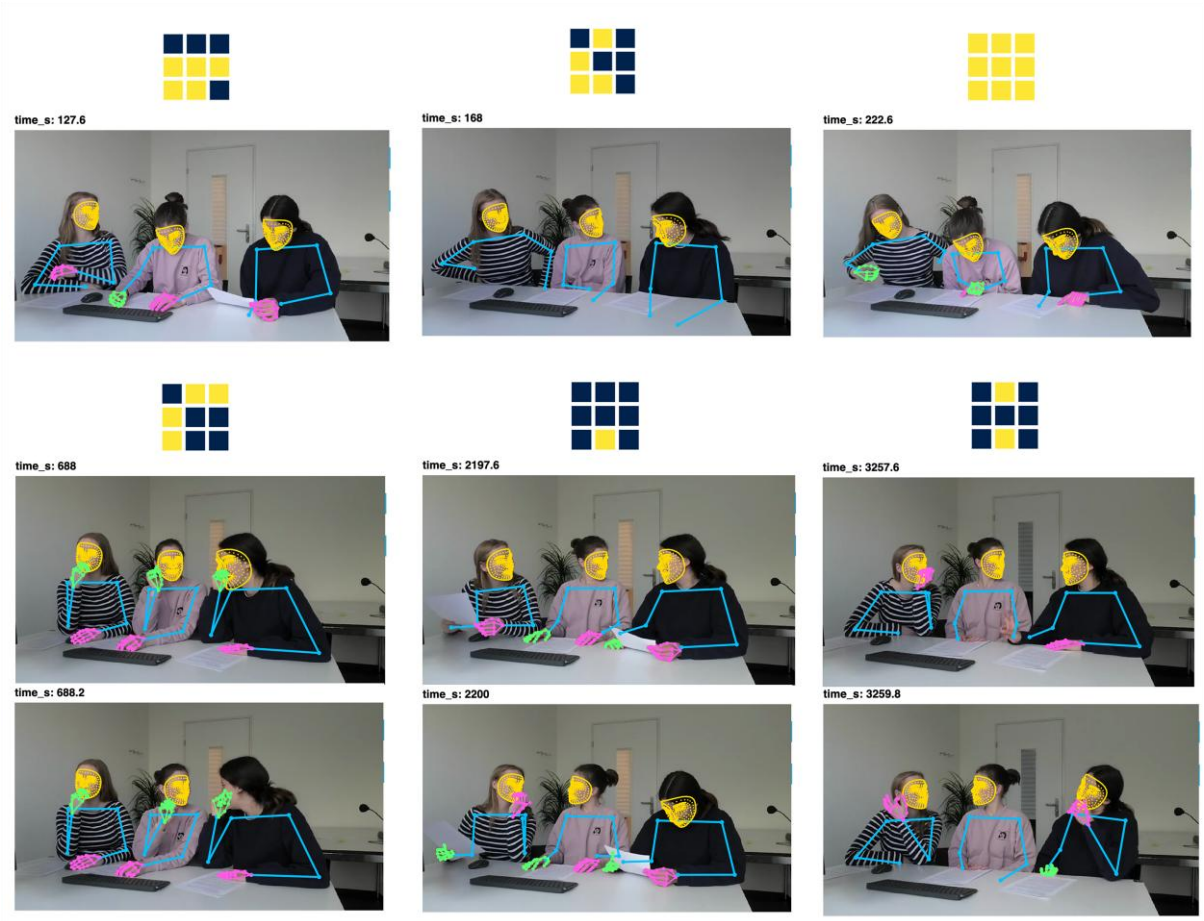


Figure 14: Examples of the 3x3 glyph arrays representing the binarized maximum MOIs during a 5-second segment and 1 or 2 example frames in this time segment that are characteristic for the pattern of the 3x3 glyph array. The example frames are overlaid with the extracted landmarks.

4.3 Dealing with data loss due to aggregation

As mentioned previously, resampling the data at lower temporal resolutions helps reduce its size but also leads to a data-loss that is more pronounces the larger the aggregation interval is. The aggregated values do not necessarily need to be scalars (such as mean, max., etc.) and could be represented by multiple values that could be mapped to distinct graphical marks (or encoded using distinct channels). In some cases, a less abstract but effective technique can be achieved with the combination of the "time-flattening" operation on the space-time (hyper-)cube (Bach *et al.*, 2014) and the "data-shading" technique originally developed for visualizing very large geometric data (Cottam *et al.*, 2014). This technique can be thought of as an image-based aggregation, where a (potentially large) sequence of time-based geometric data is mapped ("flattened") to a single image in which the pixels act as bins of a 2D-histogram. The colors values of the pixels can be rendered based on different aggregation functions.

An example of such image-based aggregation is presented in Figure 15, which shows a bird's-eye view of the recorded scene (a projection into the xz-plane). It shows rays emanating from points defined by the positions of the face-meshes and extending along the facing direction, or in other words, the projected facing directions. We used the Python Datashader library (https://datashader.org/) which handles issues related to over-plotting, offers functionality for flexible color-mapping and supports interactive visualizations (i.e. repainting the pixels in higher resolution upon zooming). A small multiple (time-juxtaposing) of image-based aggregations over smaller time intervals can be used to reveal changing patterns over time.

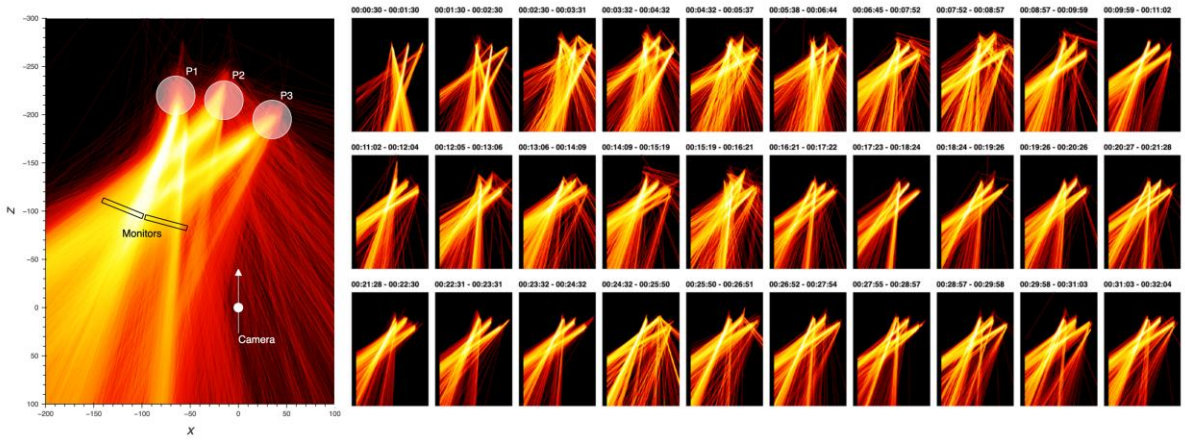


Figure 15: Image-based temporal aggregation of rays representing the facing directions projected to the xz-plane (bird's-eye view). Left: image obtained by aggregating 30 minutes' worth of data showing the general extent and distribution of rays. The approximate locations of two computer screens and the camera are indicated for reference. Right: small multiples of images representing consecutive 1-minute long aggregation intervals showing the changes with time. In the nomenclature of space-time (hyper)cube operations by Bach *et al.* (2014) this correspond to the "time-flattening" operation for the image on left and a compound operation of "time-chopping", "time-flattening" and "time juxtaposing" for the small multiples, respectively.

4.4 Posture Transitions Over Time

The kinematic model interface described in Kinematic Posture Model makes it easy to perform specific geometric queries with respect to the posture (e.g. is the person pointing with the hand?) or to generate and experiment with user-defined feature vectors as inputs to embedding or dimensionality-reduction algorithms. Vector embeddings into lower-dimensional (2D or 3D) spaces in conjunction with interactive visualizations areoften used as a means to visually explore high-dimensional data. For example, using the joint angles and the distances between the hands and between the hands and the face as input vectors, we created 2D vector embeddings using the Minimum-Distortion Embedding (MDE, Agrawal *et al.* (2021)) and the PyMDE Python library (https://pymde.org/citing/index.html) and applied k-means clustering with arbitrarily chosen number of clusters (Figure 16). Stressing that this is meant as a demonstrative example with the intent to highlight the Visual Analytics approach rather than present a rigorous embedding and clustering procedure, the visualization in Figure 16 can be used as an effective means to study distinct posture and their reoccurrence with time. For example, by clicking on the individual points in the embedding plot, the corresponding landmarks are presented in a linked plot and can be visually inspected. This approach can help to quickly identify distinct or reoccurring postures. Moreover, inspecting the high-distortion pairs of the MDE might reveal outliers such as samples with erroneous landmarks detections (Figure 17).

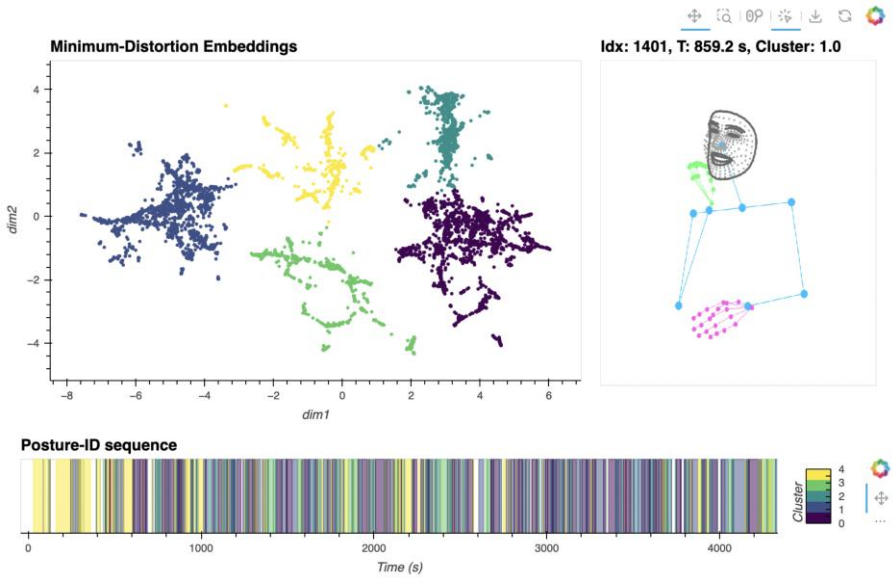


Figure 16: (Left) Visualization of the 2D vector embeddings of feature vectors derived from spatial queries (angles, distances) of the posture. The color-mapping is based on the results of the k-means clustering. Clicking on any point in the

embedding plot shows the landmarks for the corresponding sample (right). (Bottom) Temporal sequence of postures belonging to individual clusters.

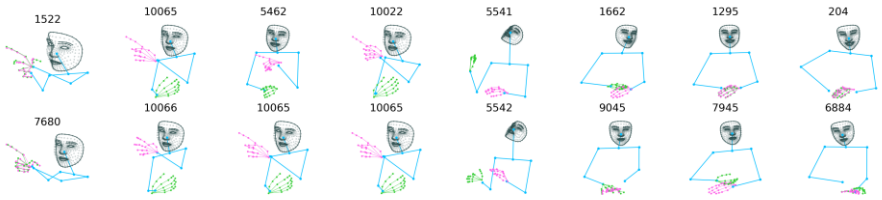


Figure 17: Examples of high-distortion pairs (columns) showing outliers due to erroneous landmarks detections.

5 Discussion

The results presented here showcase the potential of VA to provide rapid insights into complex temporal data of NVB using specially-crafted interactive visualizations. Unlike end-to-end, black-box machine-learning approaches, which require extensive training data and often lack interpretability, VA leverages a human-in-the-loop framework that combines data processing, visualization, and interactivity in an iterative process with the aim of generating insight from visual representations of data. As such it offers great flexibility to explore new avenues for extracting rich, high-resolution NVB signals from the sensor data and for designing visual representations and interfaces that facilitate the detection of patterns in the data. While we do not advocate the use of purely end-to-end machine-learning models, given the aforementioned challenges, we see a tremendous value in coupling Visual Analytics with robust ML and computer vision techniques with well-defined semantics when employed as feature extractors in the visualization pipeline. In the present work, we exemplify this with the extraction of landmarks to study body movements, but this approach could include other computer vision techniques, such as detection and tracking of objects in the environment, robust facial expression and gesture detection. By integrating these targeted methods, researchers can benefit from reliable automated feature generation without sacrificing the interpretability and flexibility that VA provides. The use of interactive visualizations offers a powerful means for researchers and analysts to explore complex data in a flexible and user-driven manner. By enabling users to determine whether, how, and to what extent the underlying data should be aggregated, such systems support dynamic hypothesis formation and allow multiple perspectives on the same phenomena. This flexibility, however, also places significant responsibility on the user: the choice of aggregation level and representation can substantially influence the patterns that emerge and, consequently, the interpretations drawn from them. Interaction thus not only facilitates data exploration but also introduces epistemological considerations regarding the validity and reproducibility of visual insights.

An important prerequisite for the effective use of visualizations is the user's visual literacy. When confronted with novel visual representations, particularly those that deviate from conventional plots or familiar metaphors, viewers must first learn how to interpret the graphical encoding. This learning curve can initially limit the accessibility of innovative visualization techniques, even if they provide deeper insights once understood. Therefore, visualization design should balance expressive power and interpretability, ensuring that visual metaphors remain both cognitively ergonomic and semantically transparent. Equally critical is the quality and contextual appropriateness of the underlying data. Without a suitable experimental or observational setting, the data available for visualization may be incomplete, biased, or contextually ambiguous, leading to misleading visual patterns and potentially erroneous conclusions. In the study of nonverbal behavior (NVB), for instance, subtle inaccuracies in sensor calibration, tracking, or temporal alignment can distort the apparent dynamics of interaction. Consequently, visualization-based analyses must be grounded in robust data collection methodologies that preserve the contextual integrity of behavioral phenomena. Our present work has primarily focused on NVB related to limb movements and gaze direction, as these are among the most readily quantifiable aspects of human interaction. However, nonverbal communication encompasses a much broader range of expressive behaviors, including impression management, affect display, and emotional regulation. These aspects are often still assessed through human interpretation, as they depend on nuanced contextual cues and subjective judgment. While machine learning (ML) approaches have made notable progress in detecting emotions or affective states from facial expressions and body posture, the interpretation of such signals within broader social and cultural contexts remains inherently subjective and situational. Moreover, NVB is deeply embedded in cultural frameworks that shape both how it is performed and how it is interpreted. Gestures, gaze norms, interpersonal distance, and expressive intensity vary across cultures, influencing the meaning of specific behavioral patterns. Consequently, both automated and visual-analytic approaches to NVB must account for cultural variability to avoid overgeneralization and misinterpretation. Integrating contextual knowledge, whether through metadata, annotation, or domain expertise, remains a key challenge for future visual analytics systems aiming to support nuanced interpretation of social interaction data.

In summary, while interactive visualizations provide a powerful lens for exploring nonverbal behavior, their effectiveness depends on several interrelated factors: the interpretability of the visual design, the quality and contextual validity of the data, and the user's ability to read and critically evaluate the visual encodings. Addressing these challenges requires interdisciplinary collaboration, combining advances in computer vision, visual analytics, and behavioral science with insights from cultural and cognitive psychology.

5.1 Limitations and future directions

Landmarks extraction

The extraction of landmarks from single-camera, monocular video data using human pose estimation algorithms has some limitations. These are related to general constraints of the computer vision problem of estimating 3D data from 2D images, as well as the intrinsic capabilities of the algorithms. In our landmarks extraction procedure, the landmarks detection may fail completely (no data) or produce erroneous results due to occlusions caused by other persons and objects in the environment. In the case of the face landmarks, the detection also fails when persons are oriented at obtuse angles relative to camera. Figure 19 shows a heatmap of missing data (dark-grey values) for the different body parts for an example dataset in the f2f study. The instances of missing landmarks data, most notably for hands landmarks, are primarily due to occlusions (e.g. hands under the table). Moreover, dynamic settings in which multiple persons are allowed to move around freely require a robust tracking algorithm in order to correctly attribute the extracted landmarks to corresponding persons. As the research in this domain is evolving rapidly, we expect that some of these issues can be addressed by a different choice of tools and algorithms or their future iterations. The mentioned limitations can also be addressed by adaptations to the experimental setup (e.g. use of multiple cameras and possibly depth sensors, using markers for tracking). However, we note that such improvements in the experimental procedure, especially by using a more elaborate equipment with multiple redundant sensors, have to be evaluated against the rapidly increasing complexity and the computational demand of data processing.

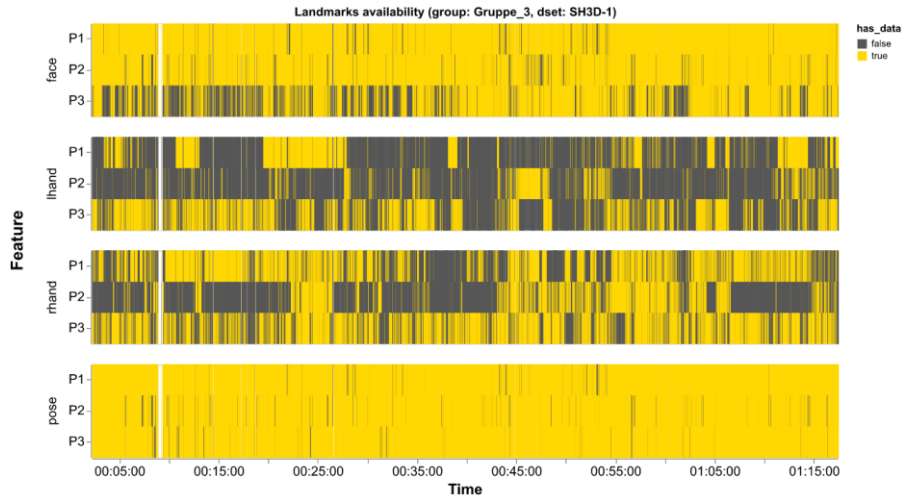


Figure 19: Heatmap showing missing data in the extracted landmarks (i.e. no detection) for each person (P1, P2 and P3) and grouped by the body part. Observe the many missing data for hands resulting mainly from occlusions (e.g. hands under the table).

Technological challenges

The implementation of VA interfaces (e.g. web-based dashboards) presents some additional technological hurdles. First of all, the necessity of scaling to large datasets can be challenging to realize with the existing visualization libraries. Resampling or binned aggregation can help alleviate the problem at the expense of data loss and inability to recover small time-scale variations (i.e. micro-expressions in NVB). Using techniques such as image-based aggregation (Figure 15) can be useful in certain scenarios involving geometric 2D data. Moreover, interactions such as zooming and filtering or linked views (brushing) involving large number of data points, if not responsive, might lead to decreased user experience with the visualization tool or render the tool unusable altogether. Another limitation concerns the design of novel visual representations, similar to the glyph-maps presented in Section 3.7. While existing libraries, such as HoloViews (https://holoviews.org/index.html), Vega (https://vega.github.io/vega/) and VegaLite (https://vega.github.io/vega-lite/), excel in providing expressive APIs for many types of common visualization types, they lack useful abstractions for easily defining such new glyph-based visual representations. Consequently, the researchers have to implement their designs using inconsistent abstractions and accept potential limitations in terms of interactivity.

6 Conclusion

In this contribution we demonstrated a methodological approach to analyse nonverbal behavior from video recordings using the Visual Analytics framework. We described a data-processing pipeline to extract rich high-temporal-resolution body movement data using off-the-shelf computer vision tools that detect face, hands and body landmarks. Based on these data, we showcased possible approaches to derive individual and dyadic nonverbal behavior, as well as techniques for designing effective visualizations of complex high-dimensional data. We believe that Visual Analytics unlocks new opportunities to utilize the high-volume multi-modal sensor data to study the complex dynamics of NVB and interaction and invite future research to further refine these methods and demonstrate their usefulness in real-world scenarios.

References

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds). (1999). Readings in information visualization: Using vision to think. Morgan Kaufmann Publishers Inc.

Chen, C. (2010). Information visualization. WIREs Computational Statistics, 2(4), 387–403. https://doi.org/10.1002/wics.89

Chen, J., Wang, M., Kirschner, P. A., & Tsai, C.-C. (2018). The Role of Collaboration, Computer Use, Learning Environments, and Supporting Strategies in CSCL: A Meta-Analysis. Review of Educational Research, 88(6), 799–843.

Cook, K. A., & Thomas, J. J. (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE Computer Society, Los Alamitos, CA, United States(US). https://www.osti.gov/biblio/912515

Keim, D., & Ward, M. (2003). Visualization. In M. Berthold & D. J. Hand (Eds), Intelligent Data Analysis: An Introduction (pp. 403–427). Springer. https://doi.org/10.1007/978-3-540-48625-1_11

Kulić, D. (2019). Human Motion Imitation. In A. Goswami & P. Vadakkepat (Eds), Humanoid Robotics: A Reference (pp. 1657–1677). Springer Netherlands. https://doi.org/10.1007/978-94-007-6046-2_34

Lehmann-Willenbrock, N., & Hung, H. (2023). A Multimodal Social Signal Processing Approach to Team Interactions. Organizational Research Methods, 10944281231202741. https://doi.org/10.1177/10944281231202741

Luciano, M. M., Mathieu, J. E., Park, S., & Tannenbaum, S. I. (2018). A Fitting Approach to Construct and Measurement Alignment: The Role of Big Data in Advancing Dynamic Theories. Organizational Research Methods, 21(3), 592–632. https://doi.org/10.1177/1094428117728372

Müller, J., Fàbregues, S., Guenther, E. A., & Romano, M. J. (2019). Using Sensors in Organizational Research—Clarifying Rationales and Validation Challenges for Mixed Methods. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.01188

Paneth, L., Jeitziner, L. T., Rack, O., & Zahn, C. (2023). A Multi-Method Approach to Capture Quality of Collaborative Group Engagement. 91–98. https://doi.org/10.22318/cscl2023.134087

Renier, L. A., Schmid Mast, M., Dael, N., & Kleinlogel, E. P. (2021). Nonverbal Social Sensing: What Social Sensing Can and Cannot Do for the Study of Nonverbal Behavior From Video. Frontiers in Psychology, 12, 606548. https://doi.org/10.3389/fpsyg.2021.606548

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. Proceedings 1996 IEEE Symposium on Visual Languages, 336–343. https://doi.org/10.1109/VL.1996.545307

Vinciarelli, A., Salamin, H., & Pantic, M. (2009). Social Signal Processing: Understanding social interactions through nonverbal behavior analysis. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 42–49. https://doi.org/10.1109/CVPRW.2009.5204290

Zheng, L., Long, M., Niu, J., & Zhong, L. (2023). An automated group learning engagement analysis and feedback approach to promoting collaborative knowledge building, group performance, and socially shared regulation in CSCL. International Journal of Computer-Supported Collaborative Learning, 18(1), 101–133. https://doi.org/10.1007/s11412-023-09386-0