SoftwareX 27 (2024) 101827

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

FT Xtraction: Feature extraction and visualization of conversational video data for social and emotional analysis

Tivan Varghese George, Hye Soo Park, Uichin Lee *

KAIST, Daejeon, South Korea

ARTICLE INFO

Keywords:
Human behavior analysis
Feature extraction from video data

ABSTRACT

Conversational video data are widely used to analyze social interaction patterns in various fields, such as social-emotional learning. However, current video processing technologies for social and emotional analysis lack unified feature extraction from video data and video overlays for feature visualization. This paper introduces FT Xtraction to address such limitations. For unified feature extraction, FT Xtraction extracts key base features, such as facial emotion, facial landmarks, and pose keypoints, which are then used to extract derived features such as emotion entropy/synchronicity, pose synchronicity, and physical activeness. FT Xtraction supports video overlay-based feature visualization that allows researchers to examine video data along with extracted features.

Code metadata

Current code version	0.1
Permanent link to code/repository used for this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-24-00035
Permanent link to Reproducible Capsule	–
Legal Code License	MIT
Code versioning system used	None
Software code languages, tools, and services used	Python
Compilation requirements, operating environments & dependencies	MacOS w/ the M1 Processor, Python 3.9, other requirements provided in requirements.txt
If available, link to developer documentation/manual	–
Support email for questions	tgeorge.engg@gmail.com

1. Motivation and significance

Video-based human behavior analysis has been used in a variety of applications, including sports coaching [1], security [2], gait detection [3,4], gesture detection [5] and education [6]. Furthermore, advances in machine learning and deep learning have enabled researchers to use video analysis for complex inference, such as detecting nuances that are known to be difficult to predict automatically. With the advancement of object detection technology [7] and context capturing [8] technology, video data analysis has been used in melancholia detection [9], biometric application by gait [10,11] emotion recognition of group [12], rapport detection [13] and conversation coaching [14]. Another important application is social signal processing, which recognizes human social signals and behaviors like turn-taking, politeness, and disagreement [15]. Social signal processing based on conversational data will enable a variety of applications, e.g., social-emotional learning (SEL) for children.

According to the framework first proposed by CASEL [16], SEL skills are composed of five core competencies: self-awareness, self-management, social awareness, relationship skills, and responsible decision-making [17], which include social interactions with others. There have been studies in which video feedback has been used to practice social-emotional learning and social skill training [18]. Mainly it is analyzed by an external coder [19] or the user watches the video and performs direct self-modeling [20] rather than automated video analysis. The literature around the video interaction guidance (VIG) framework provides evidence of how guided, post hoc reflection of

* Corresponding author.
E-mail addresses: tgeorge@kaist.ac.kr (Tivan Varghese George), hyehye@kaist.ac.kr (Hye Soo Park), uclee@kaist.ac.kr (Uichin Lee).

https://doi.org/10.1016/j.softx.2024.101827
Received 17 January 2024; Received in revised form 16 June 2024; Accepted 10 July 2024
Available online 24 July 2024

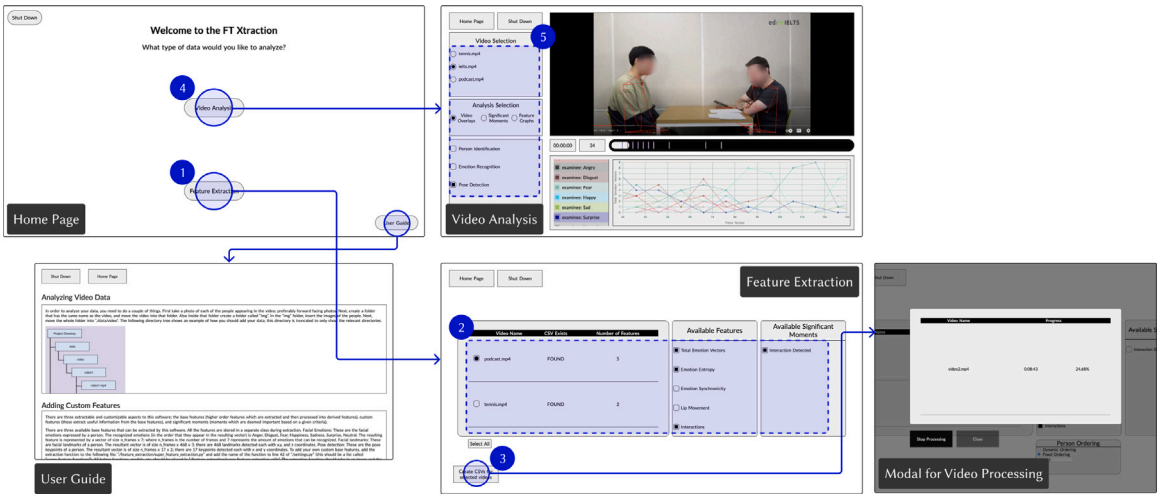

Fig. 1. User Flow - the users can process the video by the order of (1) Feature extraction (2) Select features (3) Process the Video (4) Start Visualization (5) Select Features to Visualize as marked in the figure. An additional guide for users is provided.

micro-moments, selected from video clips of everyday activities, can promote social skills learning [21]. VIG is an intervention where the clients are guided to reflect on video clips of their successful interaction in family activity [22]. This has traditionally been done by intervening human agents to detect situations where interaction is going well or not going well.

There is a growing trend to conduct research using automatic video analysis for simple metrics, such as coding social attention through gaze detection [23], interpreting body language [24], supporting facial emotion recognition [25] and detecting classroom engagement [26,27] among studies for child development. To conduct video analysis research for SEL, it is necessary to extract multiple features related to emotions and behaviors broadly from the video where multiple people appear. To support the conversational video data analysis of human behavior and emotion, researchers developed analytic tools, such as Noldus Observer XT [28], OpenSense [29] and ConAn [30]. Existing tools provide a GUI and visualization of extracted basic features such as facial emotion recognition and pose detection. However, the tools are specific purpose-oriented programs in which researchers can only use basic features embedded in the software. If the researcher should infer complex metrics for SEL research by composing social and emotional features, they should calculate it externally.

This leads us to propose FT Xtraction, a flexible software that supports conversational analysis for researchers in processing video data, analyzing the processed data, visualizing the data through charts and video overlays, and creating their custom features. Researchers will be able to use this software through a web browser after some initial setup.

FT Xtraction can be used for a multitude of behavioral and emotional analysis tasks. In addition to multi-people pose detection and facial emotion recognition, it calculates additional derived features such as emotion synchronicity and pose synchronicity within the people, lip distance, and gaze direction. These are important metrics that are used by social cues in social signal processing and social-emotional learning. Emotion synchronicity and pose synchronicity help to determine the degree of partner mimicry [31], which is widely believed to have a social function. The lip distance between people can be used to estimate interpersonal distance, which is influenced by the quality of the relationships between individuals. These derived features demonstrate its utility via SEL. We expect this tool that promotes the use of complex automatic video analysis, including pose and emotion synchronicity, can help develop new technology and conduct research with more quantitative and objective data in this research field.

2. Software description

2.1. FT Xtraction overview

FT Xtraction offers a web service that allows users to upload, process, and analyze social interaction videos. After setting up the system, the user must first process their video data on the extraction page, after which they can analyze their data on the video analysis page. The overall flow to utilize this software is described in Fig. 1. Detailed views of feature extraction and video analysis are in Figs. 2 and 3

The extraction page is where the users must first go in order to process their video data and extract features (see Fig. 2). The video extraction process is visualized in Fig. 4. After initialization (which involves selecting which features you want to process), the base features are extracted from every frame. Then, for every n frame (i.e., a window size of n), the derived features are extracted from the base features and recorded in a CSV file. When they click "Create CSVs for selected videos", it will show a progress message about the time left for completion. All the extracted data is saved in comma-separated value (CSV) files.

After processing the video data, the users can then analyze their videos on the video analysis page as shown in Fig. 3. The video overlays allow users to visualize certain features frame by frame. By selecting multiple overlays they can visualize multiple features at once. The inbuilt overlays are: (1) "Person Identification" - this draws bounding boxes around the faces of recognized people and displays their names as well; (2) "Emotion Recognition" - this displays the seven classes of emotions by a person; and (3) "Pose Detection" - this draws the pose keypoints and pose skeleton of each person.

The significant moments highlight moments of importance in the video along the video progress bar (in the image these are the green bars on the video progress bar). The only inbuilt significant moment is "Interactions Detected" - this displays moments where people interact with each other. Users can add their own significant moments by following the instructions in the software manual.

Lastly we have the feature graphs, which list the features extracted for the selected video; only one can be selected at a time. When a certain feature is selected, the appropriate sub-categories and data are displayed in the section below the video progress bar. Users can choose which sub-categories to see in case they want to focus on a certain person or selection of people.

Fig. 2. The feature extraction page; here users can process their video data.

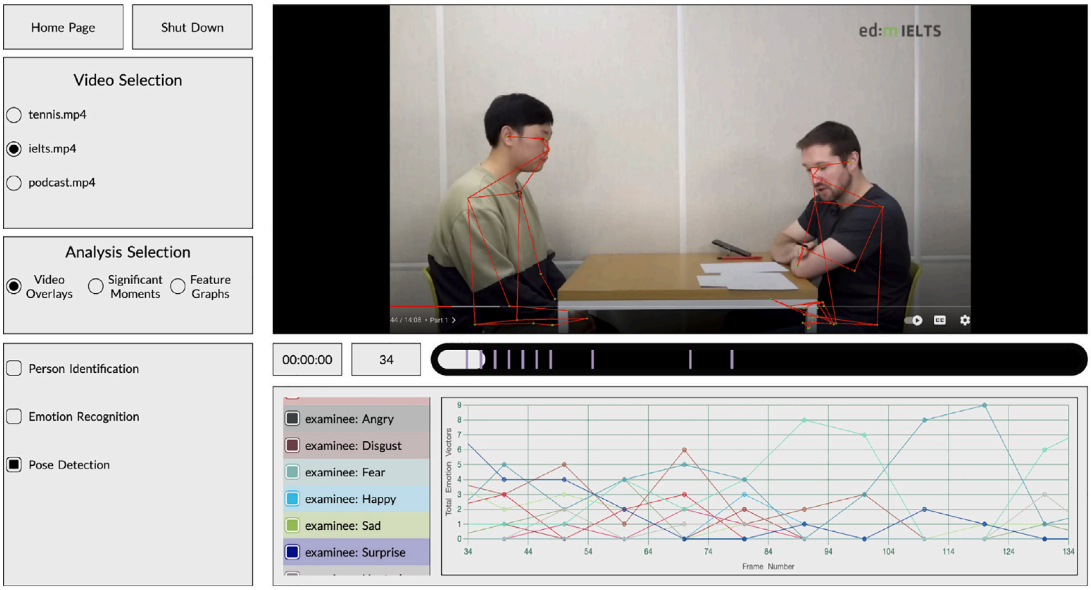

Fig. 3. The video analysis page: users can analyze their videos and the extracted features here.

2.2. FT Xtraction implementation architecture

FT Xtraction is based on client–server architecture. The server handles feature extraction and video overlay image generation, whereas the client offers a web-based user interface and data visualization. FT Xtraction was implemented using Python, HTML, JavaScript, and CSS. Python was chosen as the language used to handle the backend due to the plethora of available data science and machine learning libraries.

The server was implemented using FastAPI [32]. FastAPI is a modern web framework used to build APIs with Python. It is well known for both its high performance and ease of working with, allowing developers to create APIs extremely quickly. FastAPI was chosen over other frameworks, such as Django [33] and Flask [34], because of such benefits. The client-side code was implemented using HTML, JavaScript, and CSS. All requests made from the server to the client were handled using Starlette responses. All requests made from the client to the server were handled using the Fetch API for Javascript.

2.3. Feature extraction details

We consider base features: facial emotions, facial landmarks, pose keypoints, and person matching. These base features are then used to calculate the following derived features: total expressed emotions, emotion entropy, emotion synchronicity, lip distance, gaze direction, pose synchronicity, and physical activeness.

2.3.1. Base feature extraction

Facial Emotions: To extract facial emotions, we use a pretrained model [35], which claimed to have achieved a 73.11% performance on the FER2013 dataset [36] and a 94.64% on the CK+ dataset [37]. Though several other facial emotion recognition models exist, we chose this model since it performed well during preliminary evaluation while having a relatively small model size and the higher speed of processing, and could easily be downloaded with code (which minimizes the manual setup users have to do). This model is based on the VGG19 architecture and was trained on the FER2013 dataset. This model classifies the displayed emotion into one of seven unit emotions (i.e., anger,

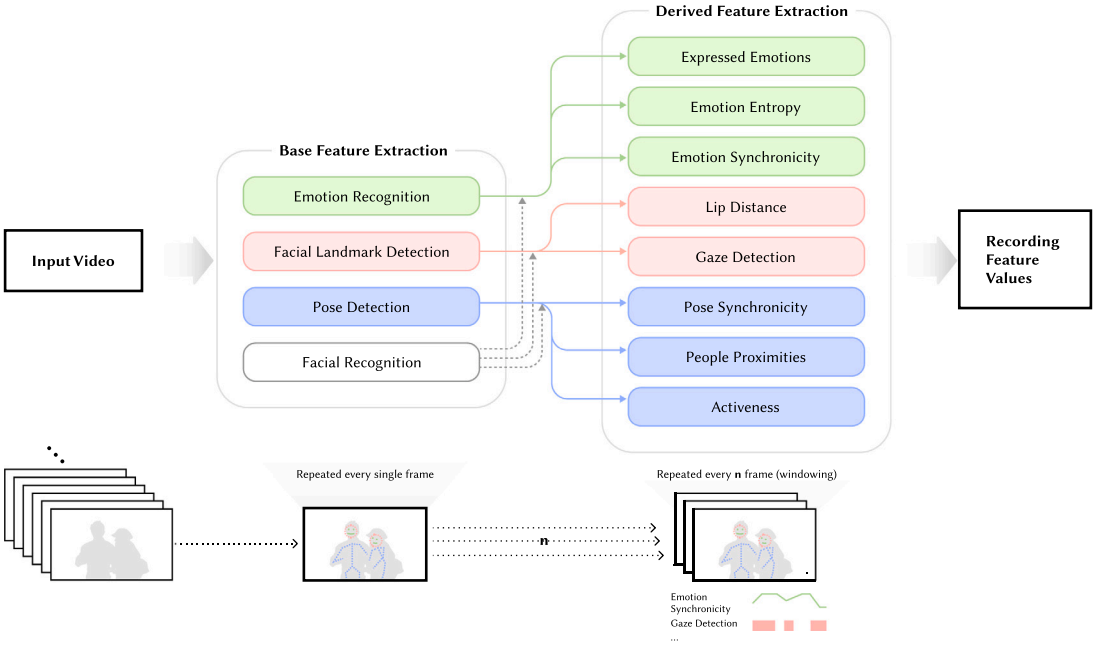

Fig. 4. An overview of the feature extraction process.

disgust, fear, happiness, sadness, surprise, and neutral) and returns an integer ranging from 0 to 6, signifying the index of the recognized emotion. In this software, the extracted emotions are stored in an $n \times 7$ array, where n is the number of frames and 7 refers to the number of emotions.

Facial Landmarks: To detect facial landmarks, the FaceMesh model from Google's MediaPipe library was used [38,39]. One of the key reasons we chose this model was that it outputs a comparatively larger amount of facial landmarks. This allows users more freedom when creating their own features without having to use a separate model. Furthermore, as stated before, this model has a relatively low IOD-MAD (Interocular Distance, Mean Absolute Distance) error rate of 3.96% [39] and is reasonably lightweight, which allows for reduced processing times. This model detects 468 facial landmarks, each with their own x, y, and z coordinates. As such, in this software, the facial landmarks of each person are stored in an $n \times 468 \times 3$ array, where n is the number of frames, 468 refers to the number of facial landmarks, and 3 refers to x, y, and z coordinates per landmark. This model is based on the Single Shot Detector architecture and was trained by Google on a custom dataset.

Pose Keypoints: To extract pose keypoints, Multipose Movenet from Tensorflow was used [40]. This model detects the pose keypoints of up to six people in a given image. We chose this model over other multi-pose models primarily due to its high OKS (Object Keypoint Similarity) score on the Active Person Image (0.840) as mentioned in model details of Multipose Movenet [40] which aligns with the intended use of our software such as social activity. Furthermore, this model is relatively fast and is easily accessible (this allows us to automate the model download process which minimizes the amount of setup the user does). For each person, this model outputs 17 keypoints, each with its own x and y coordinates, the bounding box of the detected person, and scores for each of the keypoints and the person as a whole. In this software, the keypoints of each person are stored in an $n \times 17 \times 2$ array, where n is the number of frames, 17 refers to the number of keypoints, and 2 refers to the coordinates per keypoint. This model uses a "MobileNetV2 image feature extractor with Feature Pyramid Network decoder followed by CenterNet prediction heads with custom post-processing logics" and is trained on both the COCO Keypoint Dataset Training Set 2017 and an active dataset training set. The second dataset consists of several images taken from YouTube fitness videos.

Person Matching: To match the extracted data to the correct person, we used face recognition; we would extract facial encodings from each person prior to feature extraction and then match these known encodings to the unknown encodings obtained during feature extraction. In order to do all of this, we opted to use the face-recognition library [41]. This library claims to have achieved an accuracy of 99.38% on the Labeled Faces in the Wild benchmark (this is a public benchmark used for face verification and recognition) [42], and several papers have used this library [43–45]. Also this library requires no separate models to be downloaded and has several utility functions related to facial recognition that users can be used in when making their own features; as such we have decided to use this library. The models used from this library use the dlib's shape predictor, which is based on "Ensemble of Regression Trees" and is trained on the dlib 5-point face landmark dataset.We compared it with other popular face recognition libraries such as "deepface", but found that the chosen face-recognition worked with the best accuracy for our uses through a manual video coding done by researchers.

2.3.2. Derived features

These derived features are only calculated after n frames.[1] This is because the Emotion Entropy, Lip Distance, and Activeness features require more than one frame to be calculated.

Total Expressed Emotions: The expressed emotions for person i, denoted as TEE_i, is the vector containing all the emotions (7 basic emotions) displayed by person i over n frames. This is calculated by simply adding all the values for one emotion, denoted as e_{ij} where j is between 0 and 6 (representing the 7 basic emotions), over all the frames in the base extracted emotions for person i, denoted as BEE_i, for each basic expressed emotion:

$$TEE_i = [e_{i0}, e_{i1}, \ldots, e_{i6}] \tag{1}$$

$$e_{ij} = \sum_{k=0}^{n-1} BEE_i[k][j] \tag{2}$$

[1] This variable will be used throughout this section. The default value $n = 10$ in the implementation.

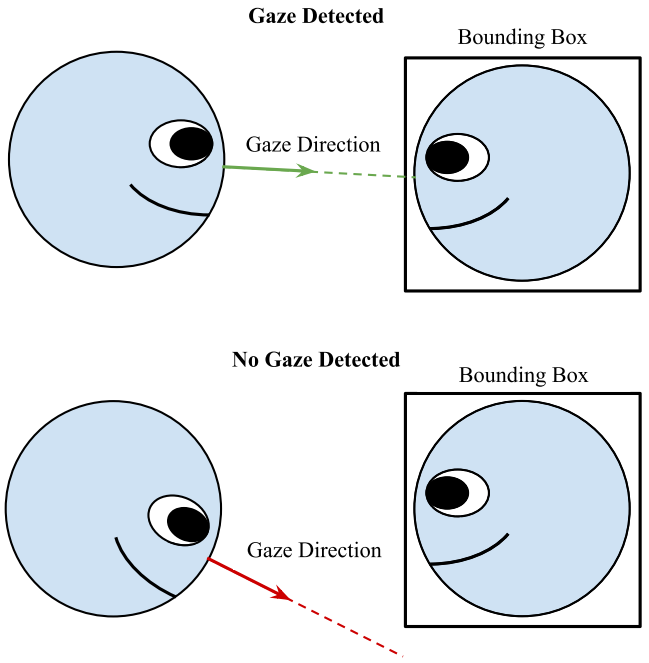

Fig. 5. This figure displays what is counted as a detected gaze.

Emotion Entropy: The emotion entropy of person i over n frames, denoted by $EEnt_i$, is calculated by applying Shannon's entropy [46] to a person's Total Expressed Emotions vector, TEE_i.

$$EEnt_i = -\sum_{j=0}^{6} p(TEE_i[j]) \cdot log_7(p(TEE_i[j]) + \varepsilon) \tag{3}$$

$$p(TEE_i[j]) = \frac{TEE_i[j]}{\sum_{k=0}^{6} TEE_i[k]} \tag{4}$$

where ε is an arbitrarily small value to avoid errors when encountering log(0).

Emotion Synchronicity: The emotion synchronicity between person i and person j over n frames, denoted by $ESync_{ij}$, is calculated by taking the Euclidean distance between the Total Expressed Emotions vectors of person i and person j, which are TEE_i and TEE_j respectively.

$$ESync_{ij} = \sqrt{\sum_{k=0}^{6} (TEE_i[k] - TEE_j[k])^2} \tag{5}$$

Lip Distance: The average change in lip distance of person i over n frames, denoted by LD_i, is calculated by first taking the average distance between the facial landmarks in the upper and lower lips, denoted by UL_i and LL_i respectively, for each of the n frames. For clarity, let the number of landmarks in the set LL_i be denoted as m_U, and the number of landmarks in the set UL_i be denoted as m_L. Let this average distance be denoted as $FrameLD_{im}$, where m is the frame number, ranging from 0 to $n-1$. Lastly, the average of the absolute value of the differences in the lip distances is taken.

$$LD_i = \frac{\sum_{j=0}^{n-2} |FrameLD_{ij} - FrameLD_{i(j+1)}|}{n-1} \tag{6}$$

$$FrameLD_{ij} = \frac{\sum_{k=0}^{m_U-1} LL_i[j][k]}{m_U} - \frac{\sum_{k=0}^{m_L-1} UL_i[j][k]}{m_L} \tag{7}$$

Social Gaze Detection: Social gaze is detected when the gaze of one person intersects the facial bounding box of another person, as detailed in Fig. 5. The gaze direction of a person is approximated as the line perpendicular to the line connecting the chin and forehead landmarks, as shown in Fig. 6. For clarity, the chin and forehead landmarks are denoted as $chin\_kp$ and $forehead\_kp$, respectively, in the figure. This figure also defines several variables used to determine if a gaze is detected or not. Social gaze detection is done in the following steps. First we define $gaze\_line = m_{gaze}x + b$ for some arbitrary b, where $m_{gaze} = \frac{x_{ch}-x_{fh}}{y_{fh}-y_{ch}+\varepsilon}$. ε is 0 unless $y_{fh} - y_{ch}$ is 0, in which case it is an arbitrarily small value. ε must be variable in order to avoid division by 0; if ε is static, then division by 0 is possible since $y_{fh} - y_{ch}$ can be any real number. Next, we project the gaze line of one person onto the closest vertical line originating from the other person's facial bounding box and calculate the y-coordinate of the projection; $y_{projected} = y_o + \Delta x \cdot m_{gaze}$. $\Delta x = x_{lower} - x_o$ if the subject is facing the right; else $\Delta x = x_{upper} - x_o$. If $x_o > \frac{x_{fh}+x_{ch}}{2}$, then the subject is considered to be facing left; otherwise, they are considered to be facing right. Finally, if $y_{lower} < y_{projected} < y_{upper}$, then a gaze is said to be detected; otherwise a gaze is not detected.

People Proximities: The proximity between person i and person j over n frames, denoted PP_ij, is calculated by taking the average distance between the 17 pose keypoints of person i and person j, denoted by PKP_i and PKP_j respectively, over the n frames.

$$PP_{ij} = \frac{\sum_{k_1=0}^{n-1} \sum_{k_2=0}^{16} \sqrt{dif(k_1,k_2,0)^2 + dif(k_1,k_2,1)^2}}{n} \tag{8}$$

$$dif(k_1,k_2,l) = PKP_i[k_1][k_2][l] - PKP_j[k_1][k_2][l] \tag{9}$$

The function dif calculates the difference between two coordinates of a given keypoint at a given frame of two people.

Pose Synchronicity: The pose synchronicity between person i and person j over n frames, denoted by PS_ij, is calculated by taking the average Euclidean distance between the 8 pose angles of person i and person j, denoted by PA_i and PA_j respectively. The pose angles are calculated by using the dot product of the 10 vectors connecting select keypoints of each person's pose skeleton, denoted by v_i and v_j for person i and person j respectively; the resultant angles are the only extractable angles using the given pose detection model. The pose angles and connecting vectors are detailed in Fig. 7; the pose angles are in blue, and the connecting vectors are in green.

$$PS_{ij} = \sqrt{\sum_{k=0}^{7} (PA_i[k] - PA_j[k])^2} \tag{10}$$

$$PA[j] = arccos(\frac{v_{j1} \cdot v_{j2}}{|v_{j1}||v_{j2}| + \varepsilon}) \tag{11}$$

Here $j1$ and $j2$ are integers ranging from 0 to 9, indicating the specific vector used for a given angle calculation. ε is an arbitrarily small value to avoid division by 0.

Physical Activeness: The activeness of person i over n frames, denoted by AC_i, is calculated by taking the average of the differences between the Euclidean distances of the pose keypoints of person i, denoted by PKP_i between each frame. Here, the function dif calculates the difference between two coordinates of a given keypoint at a given frame.

$$AC_i = \frac{\sum_{j=0}^{n-2} \sum_{k=0}^{16} \sqrt{dif(j,k,0)^2 + dif(j,k,1)^2}}{n-1} \tag{12}$$

$$dif(j,k,l) = PKP_i[j][k][l] - PKP_i[j][k+1][l] \tag{13}$$

2.4. Analysis of the software

2.4.1. Software speed

To employ this program in real-world scenarios, we recorded 12 videos and used FT-Xtraction to extract various features for a preliminary analysis. The videos used for this analysis featured three seated individuals, each comprising different people. The hardware configuration utilized for this analysis was an Apple Silicon M1 Max processor, supported by 32 GB of RAM. The average length of these videos was

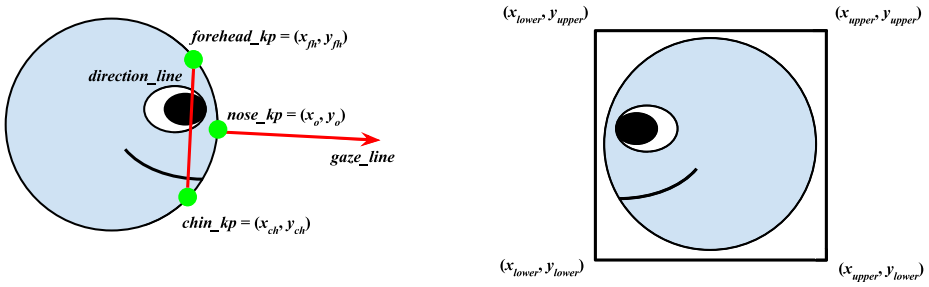

Fig. 6. This figure displays the values used to calculate whether a social gaze is detected or not.

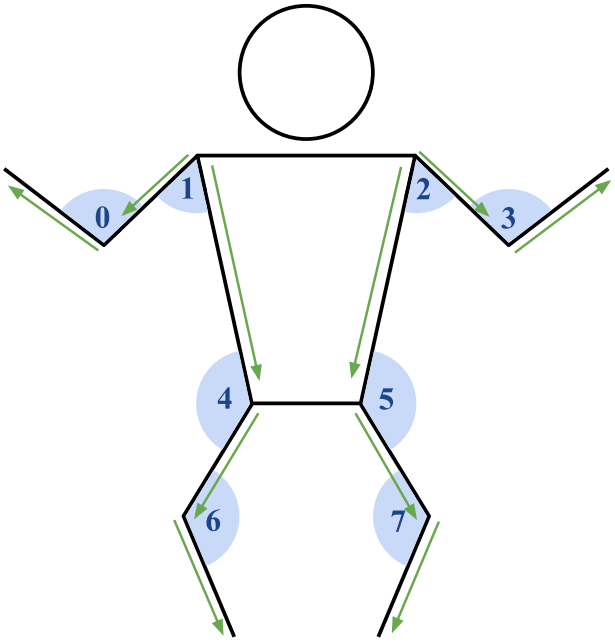

Fig. 7. This figure displays the pose angles, as well as the connecting vectors used to form them, that are used to calculate the pose synchronicity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

about 25 min, and when analyzed at a rate of one frame every 20 s, the total analysis time averaged 5 min. We confirmed that each frame took approximately 4 s to process, and through appropriate frame skipping and keyframe settings, we were able to achieve a sufficient speed for practical social and emotional video analysis.

2.4.2. Software performance

For the core models for base features which are facial emotion recognition and facial recognition, we ran the testing with the benchmark dataset. Also, the process of calculating the derived features is theoretically analyzed with the time complexity.

Facial Emotion Recognition: To test this model, we used the CK+ dataset as found in Kaggle [37]; the whole dataset was used for testing. As such we simply ran all the data through the model and stored all the results. After this, we used the data to calculate the above metrics using the scikit-learn library. As a result, the overall accuracy is 0.737.

Facial Recognition: To test this model, we use the Labeled Faces in the Wild test dataset [42]. We combined both the match and mismatch test datasets to create a larger test dataset. We simply ran all the data through the model and stored all the results. After this, we used the data to calculate the above metrics using the scikit-learn library. As a result, the overall accuracy is as high as 0.975.

Time Complexity: The time complexities of calculating the derived features with respect to the amount of frames n and amount of people m are as follows:

- Total Expressed Emotions: $O(nm)$
- Emotion Entropy: $O(m)$
- Emotion Synchronicity: $O(m^2)$
- Lip Distance: $O(nm)$
- Social Gaze Detection: $O(nm^2)$
- People Proximities: $O(nm^2)$
- Pose Synchronicity: $O(nm^2)$
- Physical Activeness: $O(nm)$

2.5. Visualization: video playing and overlays

Our HTML-based web page provides a fully functional inbuilt video player, accessible via the video tag, which can be used to easily play videos. To play the video by generating frames on the server side, the frames are encoded on the server side and decoded back into images on the client side. Since the client is now just receiving a stream of images instead of playing a video, it has no access to the current position of the video. This necessitates periodic communication between the server and the client in order to update the video progress. These updates were implemented using the Fetch API; a request from the client side is periodically sent to the server for the metadata of the current frame. This information is then sent back to the client and reflected on the progress bar. Whenever the video is paused, the update requests are also paused to avoid sending unnecessary requests.

3. Illustrative examples

As mentioned in Section 1, this software can be utilized to analyze the behavior and emotions of a child and their parents to support social-emotional learning. For example, by analyzing recorded videos of family interactions during meals, FT-Xtraction can quantify and visualize the dynamics of family interactions, providing children with opportunities for reflection on these conversations. A scenario for using this software would be as follows. After uploading the requisite files to the appropriate directories, we first move to the "Feature Extraction" page. In this example, we are interested in extracting the "Emotion Entropy", "Emotion Synchronicity", and "Interaction Features", as well as the significant moment "Interaction Detected"; as such, we select those and request the software to extract the data and make the necessary CSV files. After extraction is finished, we move to the "Video Analysis Page", where we use the video overlays to view the emotions expressed in each frame, the significant moments to see at which points the individuals are interacting, and the feature graphs to analyze the emotional entropy of the examinee. These derived features are specifically tailored to concentrate on the social-emotional dynamics observed within family activities. Moreover, users are empowered to develop both visualization and intervention tools aimed at fostering social-emotional learning with more interpretability than existing deep-learning model, by utilizing videos capturing family interactions. This software serves as a practical solution for visualizing derived features, debugging errors, and seamlessly implementing new model modifications.

4. Impact

FT Xtraction, as illustrated in the software description, enables users to process and analyze video data of human behavior and emotions using either inbuilt features or their own custom features, which is a feature that other softwares does not possess. Additionally, during video analysis, this software allows users to visualize certain features in real time while also viewing the graphical representations of the previously extracted features. Previously, users would either have to create their own software or use currently existing softwares which were specific to their task in order to process their video data. With FT Xtraction, however, users simply need to create their own features and add it to the software before using it. This means that FT Xtraction can be used to process video data for not only various behavioral analysis tasks, but also for any task in general that solely relies on video data. Furthermore, due to addition of the video overlays in the video analysis page, FT Xtraction can be used as a debugging tool for machine learning models that extract features from video data, as researchers can easily determine if the extracted features match up with the content of the video.

5. Conclusions

This paper introduced FT Xtraction which aids researchers in extracting and analyzing various behavioral features from video data. FT Xtraction excels in multifaceted behavioral and emotional analysis tasks, including multi-people pose detection and facial emotion recognition. It calculates additional derived features such as emotion synchronicity and poses synchronicity among individuals, lip distance, and gaze direction—key metrics for interpreting social cues in social signal processing and social-emotional learning (SEL). These advanced features underscore FT Xtraction's main contribution to advancing SEL by facilitating the detailed analysis of social interactions. We expect that supporting feature extraction and visualization will greatly help researchers facilitate video-based behavior analysis tasks in various fields. Furthermore, FT Xtraction will open up an avenue to the creation of a more general tool to process and analyze video data for various purposes. Looking ahead, future developments could focus on offering more features related to social and emotional analysis such as social gesture detection. This advancement will not only refine the accuracy of behavioral assessments but also expand the scope of applicable research areas in social sentiment analysis and beyond. FT Xtraction supports researchers in fluently utilizing various features to assess the social nuances of video content or to employ them in experiments. This flexibility enhances the tool's utility in a wide range of behavioral studies, providing researchers with robust, quantitative data that are crucial for developing deeper insights into social dynamics.

CRediT authorship contribution statement

Tivan Varghese George: Writing – original draft, Software, Conceptualization. **Hye Soo Park:** Writing – review & editing, Supervision. **Uichin Lee:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the KAIST Future Smart Home Research Center grant funded by The Taejae Research Foundation, South Korea and by the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. 2022M3J6A1063021).

References

[1] Wilson Barry D. Development in video technology for coaching. Sports Technol 2008;1(1):34–40.
[2] Xiang Tao, Gong Shaogang. Video behavior profiling for anomaly detection. IEEE Trans Pattern Anal Mach Intell 2008;30(5):893–908. http://dx.doi.org/10.1109/TPAMI.2007.70731.
[3] Parashar Anubha, Parashar Apoorva, Shabaz Mohammad, Gupta Deepak, Sahu Aditya Kumar, Khan Muhammad Attique. Advancements in artificial intelligence for biometrics: A deep dive into model-based gait recognition techniques. Eng Appl Artif Intell 2024;130:107712. http://dx.doi.org/10.1016/j.engappai.2023.107712, URL: https://www.sciencedirect.com/science/article/pii/S0952197623018961.
[4] Habib Zeeshan, Mughal Muhammad Ali, Khan Muhammad Attique, Shabaz Mohammad. Wifog: Integrating deep learning and hybrid feature selection for accurate freezing of gait detection. Alex Eng J 2024;86:481–93. http://dx.doi.org/10.1016/j.aej.2023.11.075, URL: https://www.sciencedirect.com/science/article/pii/S1110016823010803.
[5] Zhang Zhang, Tao Dacheng. Slow feature analysis for human action recognition. IEEE Trans Pattern Anal Mach Intell 2012;34:436–50. http://dx.doi.org/10.1109/TPAMI.2011.157.
[6] Fitzgerald Angela, Hackling Mark, Dawson Vaille. Through the viewfinder: Reflecting on the collection and analysis of classroom video data. Int J Qual Methods 2013;12(1):52–64. http://dx.doi.org/10.1177/160940691301200127.
[7] Zhang Yunzuo, Liu Ting, Yu Puze, Wang Shuangshuang, Tao Ran. SFSANet: Multi-scale object detection in remote sensing image based on semantic fusion and scale adaptability. IEEE Trans Geosci Remote Sens 2024.
[8] Zhang Yunzuo, Liu Yameng, Wu Cunyu. Attention-guided multi-granularity fusion model for video summarization. Expert Syst Appl 2024;249:123568.
[9] Bhatia Shalini, Hayat Munawar, Breakspear Michael, Parker Gordon, Goecke Roland. A video-based facial behaviour analysis approach to melancholia. In: 2017 12th IEEE international conference on automatic face & gesture recognition. 2017, p. 754–61. http://dx.doi.org/10.1109/FG.2017.94.
[10] Khan Muhammad Attique, Arshad Habiba, Khan Wazir Zada, Alhaisoni Majed, Tariq Usman, Hussein Hany S, et al. HGRBOL2: Human gait recognition for biometric application using Bayesian optimization and extreme learning machine. Future Gener Comput Syst 2023;143:337–48. http://dx.doi.org/10.1016/j.future.2023.02.005, URL: https://www.sciencedirect.com/science/article/pii/S0167739X23000468.
[11] Hamza Ameer, Khan Muhammad Attique, ur Rehman Shams, Al-Khalidi Mohammed, Alzahrani Ahmed Ibrahim, Alalwan Nasser, et al. A novel bottleneck residual and self-attention fusion-assisted architecture for land use recognition in remote sensing images. IEEE J Sel Top Appl Earth Obs Remote Sens 2024;17:2995–3009. http://dx.doi.org/10.1109/JSTARS.2023.3348874.
[12] Zeng Haipeng, Shu Xinhuan, Wang Yanbang, Wang Yong, Zhang Liguo, Pong Ting-Chuen, et al. EmotionCues: Emotion-oriented visual summarization of classroom videos. IEEE Trans Vis Comput Graphics 2021;27(7):3168–81. http://dx.doi.org/10.1109/TVCG.2019.2963659.
[13] Müller Philipp, Huang Michael Xuelin, Bulling Andreas. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In: 23rd international conference on intelligent user interfaces. New York, NY, USA: Association for Computing Machinery; 2018, p. 153–64. http://dx.doi.org/10.1145/3172944.3172969.
[14] Arakawa Riku, Yakura Hiromu. INWARD: A computer-supported tool for video-reflection improves efficiency and effectiveness in executive coaching. In: Proceedings of the 2020 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery; 2020, p. 1–13. http://dx.doi.org/10.1145/3313831.3376703.
[15] Vinciarelli Alessandro, Pantic Maja, Bourlard Hervé. Social signal processing: Survey of an emerging domain. Image Vis Comput 2009;27(12):1743–59. http://dx.doi.org/10.1016/j.imavis.2008.11.007, Visual and multimodal analysis of human spontaneous behaviour: URL: https://www.sciencedirect.com/science/article/pii/S0262885608002485.
[16] Advancing Social and Emotional Learning - CASEL. URL: https://casel.org/.
[17] Jones Stephanie M, Doolittle Emily J. Social and emotional learning: Introducing the issue. Future Child 2017;27(1):3–11, URL: http://www.jstor.org/stable/44219018..
[18] Melero Silvia, Morales Alexandra, Espada José Pedro, Orgilés Mireia. Improving social performance through video-feedback with cognitive preparation in children with emotional problems. Behav Modif 2022;46(4):755–81. http://dx.doi.org/10.1177/0145445521991098, PMID: 33511861.

[19] Rusconi-Serpa Sandra, Sancho Rossignol Ana, McDonough Susan C. Video feedback in parent-infant treatments. Child Adolesc Psychiatr Clin N Am 2009;18(3):735–51. http://dx.doi.org/10.1016/j.chc.2009.02.009, Infant and Early Childhood Mental Health. URL: https://www.sciencedirect.com/science/article/pii/S1056499309000236.

[20] Ulgado Rachel Rose, Nguyen Katherine, Custodio Van Erick, Waterhouse Aaron, Weiner Rachel, Hayes Gillian. VidCoach: A mobile video modeling system for youth with special needs. In: Proceedings of the 12th international conference on interaction design and children. New York, NY, USA: Association for Computing Machinery; 2013, p. 581–4. http://dx.doi.org/10.1145/2485760.2485870.

[21] Slovak Petr, Fitzpatrick Geraldine. Teaching and developing social and emotional skills with technology. ACM Trans Comput-Hum Interact 2015;in print. http://dx.doi.org/10.1145/2744195.

[22] Colley Bill. Video interaction guidance: a relationship-based intervention to promote attunement, empathy and wellbeing. Emot Behav Diffic 2013;18(3):347–9. http://dx.doi.org/10.1080/13632752.2012.683554.

[23] Higuchi Keita, Matsuda Soichiro, Kamikubo Rie, Enomoto Takuya, Sugano Yusuke, Yamamoto Junichi, et al. Visualizing gaze direction to support video coding of social attention for children with autism spectrum disorder. In: 23rd international conference on intelligent user interfaces. New York, NY, USA: Association for Computing Machinery; 2018, p. 571–82. http://dx.doi.org/10.1145/3172944.3172960.

[24] Balazia Michal, Müller Philipp, Tánczos Ákos Levente, Liechtenstein August von, Brémond François. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In: Proceedings of the 30th ACM international conference on multimedia. New York, NY, USA: Association for Computing Machinery; 2022, p. 70–9. http://dx.doi.org/10.1145/3503161.3548363.

[25] Garcia-Garcia Jose, Penichet Victor, Lozano María, Fernando Anil. Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions. Univ Access Inf Socs 2021;21. http://dx.doi.org/10.1007/s10209-021-00818-y.

[26] Raca Mirko, Kidzinski Lukasz, Dillenbourg Pierre. Translating head motion into attention - towards processing of student's body-language. In: Educational data mining. 2015, URL: https://api.semanticscholar.org/CorpusID:15798760.

[27] Sharma Prabin, Joshi Shubham, Gautam Subash, Maharjan Sneha, Khanal Salik Ram, Reis Manuel Cabral, et al. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In: Reis Arsénio, Barroso João, Martins Paulo, Jimoyiannis Athanassios, Huang Ray Yueh-Min, Henriques Roberto, editors. Technology and innovation in learning, teaching and education. Cham: Springer Nature Switzerland; 2022, p. 52–68.

[28] The Observer XT. URL: https://www.noldus.com/observer-xt.

[29] Stefanov Kalin, Huang Baiyu, Li Zongjian, Soleymani Mohammad. OpenSense: A platform for multimodal data acquisition and behavior perception. In: Proceedings of the 2020 international conference on multimodal interaction. New York, NY, USA: Association for Computing Machinery; 2020, p. 660–4. http://dx.doi.org/10.1145/3382507.3418832.

[30] Penzkofer Anna, Müller Philipp, Bühler Felix, Mayer Sven, Bulling Andreas. Conan: A usable tool for multimodal conversation analysis. In: Proceedings of the 2021 international conference on multimodal interaction. New York, NY, USA: Association for Computing Machinery; 2021, p. 341–51. http://dx.doi.org/10.1145/3462244.3479886.

[31] Lakin Jessica, Jefferis Valerie, Cheng Clara, Chartrand Tanya. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. J Nonverbal Behav 2003;27. http://dx.doi.org/10.1023/A:1025389814290.

[32] FastAPI. URL: https://fastapi.tiangolo.com/.

[33] Django. URL: https://www.djangoproject.com/.

[34] Flask. URL: https://flask.palletsprojects.com/en/3.0.x/.

[35] Facial-Expression-Recognition.Pytorch. URL: https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch.

[36] Goodfellow Ian J, Erhan Dumitru, Carrier Pierre Luc, Courville Aaron, Mirza Mehdi, Hamner Ben, et al. Challenges in representation learning: A report on three machine learning contests. In: Lee Minho, Hirose Akira, Hou Zeng-Guang, Kil Rhee Man, editors. Neural information processing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013, p. 117–24.

[37] Lucey Patrick, Cohn Jeffrey F, Kanade Takeo, Saragih Jason, Ambadar Zara, Matthews Iain. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition - workshops. 2010, p. 94–101. http://dx.doi.org/10.1109/CVPRW.2010.5543262.

[38] Face landmark detection guide | MediaPipe. URL: https://developers.google.com/mediapipe/solutions/vision/face_landmarker#get_started.

[39] Lugaresi Camillo, Tang Jiuqiang, Nash Hadon, McClanahan Chris, Uboweja Esha, Hays Michael, et al. MediaPipe: A framework for perceiving and processing reality. In: Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR) 2019. 2019, URL: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf.

[40] movenet | Kaggle. URL: https://www.kaggle.com/models/google/movenet/frameworks/tensorFlow2/variations/multipose-lightning/versions/1?tfhub-redirect=true.

[41] ageitgey/face_recognition. URL: https://github.com/ageitgey/face_recognition.

[42] Huang Gary B, Ramesh Manu, Berg Tamara, Learned-Miller Erik. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep. 07-49, Amherst: University of Massachusetts; 2007.

[43] Gandhi Vidhyotma, Singh Jaiteg. Intensified biological signature recognition in the wild: A case study. In: 2021 12th international conference on computing communication and networking technologies. 2021, p. 1–6. http://dx.doi.org/10.1109/ICCCNT51525.2021.9579578.

[44] He Li. Developing and refining a multifunctional facial recognition system for older adults with cognitive impairments: A journey towards enhanced quality of life. 2023, arXiv:2310.06107.

[45] Verma Rajesh, Bhardwaj Navdha, Singh Pushap Deep, Bhavsar Arnav, Sharma Vishal. Estimation of sex through morphometric landmark indices in facial images with strength of evidence in logistic regression analysis. Forensic Sci Int: Rep 2021;4. http://dx.doi.org/10.1016/j.fsir.2021.100226, URL: https://www.sciencedirect.com/science/article/pii/S2665910721000578.

[46] Entropy (information theory) - Wikipedia. URL: https://en.wikipedia.org/wiki/Entropy_(information_theory).