DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Artificial Intelligence for Team Productivity

Oleksandr "Alex" Pokras

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Artificial Intelligence for Team Productivity

Künstliche Intelligenz für Team-Produktivität

Author:	Oleksandr "Alex" Pokras
Supervisor:	Prof. Dr. Hans-Joachim Bungartz
Advisors:	Ivana Jovanovic Buha (TUM) Dr. Peter Alexander Gloor (MIT)
Submission Date:	15.05.2024

I confirm that this bachelor’s thesis in informatics is my own work and I have documented all sources and material used.

Munich, 15.05.2024

Oleksandr "Alex" Pokras

Abstract

This work explores the impact emotional entanglement between team members has on productivity on teams. Building on empirical findings, a novel, quantitative approach to analyzing team dynamics using artificial intelligence is introduced. The primary contributions include the development of Moody, an easy-to-use system designed to analyze team emotions during video calls and give actionable insights, as well as a comprehensive AI analysis pipeline encompassing face detection, emotional analysis, speaker diarization, and interruption dynamics. The proposed approach marks a significant departure from traditional, empirical methods, providing a robust framework for quantifiable assessments of the impact of emotions on team performance and productivity. The practical implications of this work are substantial, offering organizations easily employable tools to foster better emotional state of its members, enhance communication efficiency, and ultimately improve team effectiveness.

Contents

1 Introduction

1.1 What Influences Team Effectiveness?

The effectiveness of teams is crucial for the success of organizations, as it directly impacts productivity and performance. However, as teams grow larger, the impact of adding each new member on team productivity is less and lesss pronounced, leading to the phenomenon of diminishing returns. This effect is the root cause for the well-known challenge of maintaining optimal productivity as organizations expand. Additionally, communication between team members often results in loss of information, hindering effective collaboration and decision-making. Understanding the emotional connection between team members is essential, as it can significantly improve communication and team productivity.

The concept of diminishing returns in team productivity as the team or organization grows larger has been widely studied. Studies such as those by [1] have explored the impact of team size on productivity in various contexts, shedding light on the challenges associated with increasing team size. These studies provide valuable insights into the diminishing impact of each new member on overall productivity as the team size increases.

As an example, in software development, larger teams often face challenges in coordinating efforts, maintaining communication, and making decisions efficiently. As the team size increases, the complexity of managing dependencies between different pieces of code, coordinating schedules, and ensuring effective communication becomes increasingly difficult. This can lead to a decrease in productivity as more time is spent on coordination and communication rather than actual coding and problem-solving. [2]

Furthermore, larger teams can also lead to a diffusion of responsibility, where individual team members may feel less accountable for the overall outcome. This diffusion of responsibility can result in a decrease in the quality and speed of work, as team members may rely on others to pick up the slack. [3] [4]

The loss of information in communication between team members has been a subject of research as well, as highlighted in studies such as those by [5] and [6]. These studies emphasize the potential inefficiencies in coordination and communication that may arise from larger team sizes, contributing to the understanding of the challenges associated with communication within teams.

Diversity within a team has a significant influence on team productivity. When teams are composed of individuals with diverse backgrounds, experiences, and perspectives, they are more likely to generate a wide range of ideas and approaches to problem-solving. This diversity can lead to greater innovation and creativity, ultimately enhancing the team's overall productivity. Research has shown that diverse teams are better at decision-making and problem-solving, as they consider a broader range of options and perspectives. Additionally, diversity fosters a more inclusive and collaborative environment, as team members learn from one another and are exposed to different ways of thinking. Embracing diversity in teams can also improve communication and reduce the risk of groupthink, leading to more effective and efficient collaboration. Overall, diversity has been proven to be a driving force behind enhanced team productivity and performance. [7] [8] [9]

In order to address these challenges, organizations may need to reconsider their team structures and look for ways to foster more efficient collaboration and communication within larger teams. This may include implementing agile methodologies, breaking larger teams into smaller, more focused groups, or investing in tools and technologies that can streamline communication and coordination efforts. By addressing these challenges, organizations can work towards maintaining higher levels of productivity even as their teams grow in size. [10] [11] [12]

While the previous research has extensively explored the impact of team size, communication, and diversity on team productivity, a less researched topic that requires attention is the emotional entanglement within teams. Emotional entanglement refers to the complex interplay of emotions and relationships among team members, which can significantly influence how a team functions and performs.

1.2 Emotional Entanglement

Understanding and managing emotional entanglement within teams is crucial for maintaining a conducive work environment that fosters productivity. This less-explored aspect of team dynamics warrants further research and attention from organizations aiming to optimize their teams' effectiveness. Moreover, addressing emotional entanglement can complement existing strategies for improving team productivity and help organizations achieve better overall outcomes.

The emotional connection between team members has been recognized as a critical factor in improving communication and team productivity. Studies such as those by [13] and [14] have delved into the impact of emotional connections and team culture on team effectiveness, highlighting the significance of trust and cohesion in enhancing communication and productivity within teams.

Emotional entanglement in a team can impact productivity in both positive and negative ways. On the positive side, strong emotional bonds between team members can foster trust, facilitate open communication, enhance collaboration, and promote a supportive environment that is conducive to creativity and problem-solving, especially during challenging times.

When team members feel emotionally connected, they are more likely to be committed to the team's goals and work cooperatively towards achieving them. [15].

On the other hand, negative emotional entanglement can have detrimental effects on team productivity. Negative emotional entanglement can lead to conflicts if not managed properly. It may result in personal conflicts, misunderstandings, or an increased sensitivity to criticism, all of which can hinder team performance and productivity. Additionally, team members might become distracted by emotional dynamics and struggle to maintain focus on tasks efficiently. [16]

Effective conflict resolution and emotional intelligence within the team are key to harnessing the positive aspects of emotional entanglement while mitigating the negative effects. A team that can balance emotional connections with professionalism is likely to be more productive and successful [17] [18] [19].

In summary, optimizing team productivity and performance requires a deep understanding of the challenges associated with diminishing returns, communication inefficiencies, and the role of emotional connections. The efficiency of teams is crucial for organizational success, and emotional entanglement appears to be an important contributor to it.

By leveraging AI models that analyze the emotions of team members, we can gain valuable insights into their emotional entanglement and its impact on team productivity. This information can be used to identify areas of improvement, such as fostering stronger emotional bonds, promoting effective communication, and resolving conflicts.

1.3 AI and Emotion in Video Calls

With the advancement of artificial intelligence and machine learning, analyzing emotions during team interactions has become more feasible. AI models can be leveraged to interpret the emotional dynamics within a team, providing valuable insights for improving communication strategies and team engagement. By analyzing emotions of team members during their day-to-day interactions, organizations can gain a deeper understanding of their team's emotional entanglement and its impact on productivity.

The utilization of video call data as a source for emotional analysis is driven by its rich, spontaneous, and genuine nature. Video calls capture a wealth of non-verbal cues, such as facial expressions and voice tonality, which are paramount in understanding emotions. The real-time and spontaneous nature of video communication offers a unique window into team dynamics, extracting the emotional data efficiently and non-intrusively. The convenience and speed of accessing video data make it extremely useful for companies looking to improve teamwork and achieve success by understanding team emotions.

This approach aligns with the broader trend of leveraging AI for emotional analysis in various fields. In telehealth, for example, understanding a patient's emotions has long been considered crucial for providing quality care [20]. Analogously, in the context of team dynamics, analyzing emotions can significantly augment efforts to foster a more supportive and collaborative work environment.

As organizations aim to optimize team effectiveness, integrating AI-based emotional analysis into their communication strategies can offer a distinct advantage. The innovative approach introduced in this work not only addresses the less-explored aspect of emotional entanglement within teams but also aligns with the broader goal of improving productivity and collaboration in the modern workplace. By using AI to analyze team members' emotions and emotional entanglement, we can uncover valuable insights that contribute to team productivity and success.

In the course of this work, a system *Moody* has been developed to analyze team emotions during video calls and provide real-time feedback. Furthermore, state-of-the-art models have been utilized to build an analysis framework aimed at enabling the assessment of teamwork and deriving insights into how emotional intelligence contributes to enhancing team productivity. This represents a significant advancement as such insights were previously only based on empirical evidence.

Additionally, advanced SOTA models were employed in establishing an analytical framework designed to facilitate the evaluation of teamwork dynamics and offer valuable insights into how emotional intelligence can serve to improve overall team performance. It is worth noting that these advancements mark a departure from previous practices, as this area of research had so far relied mainly on empirical evidence for investigation. This framework builds a basis for quantifiable measurements and predictions, empowering researchers to quantify the effects of individual emotions, emotional entanglement and other observations on the productivity of teams. The discoveries made that way build the core for recommendations and assessments of *Moody*, providing them in a form that's fast and easy to digest for any team member, even without specialized knowledge on team productivity.

1.3.1 Moody: An Easy-To-Use Emotion Analysis Software

Moody (available at moody-v2.vercel.app) is a web app built with AngularJS to facilitate real-time emotion analysis in video calls. It is tailored towards the needs of individuals who are not experts in team dynamics. However, it also provides the opportunity to download a *.csv* file with detailed data that can be analyzed with the more advanced analysis toolkit.

The version 1 of *Moody* (available at https://www.moody.digital/) was developed at Massachusetts Institute of Technology Center for Collective Intelligence (MIT CCI) prior to this work. It was fully browser-run and utilized some small facial and emotion recognition models with unsatisfactory accuracy and speed, especially if not launched on a powerful PC. For this work, it was therefore decided to build the version 2 of *Moody* with vastly expanded functionality while still keeping it simple enough for a layman to use with a few minutes of introduction from a professional. It was also decided to deploy the AI backend on a separate compute cluster instead of launching everything in the user's browser.

Below, some functionality of *Moody v2* is showcased.

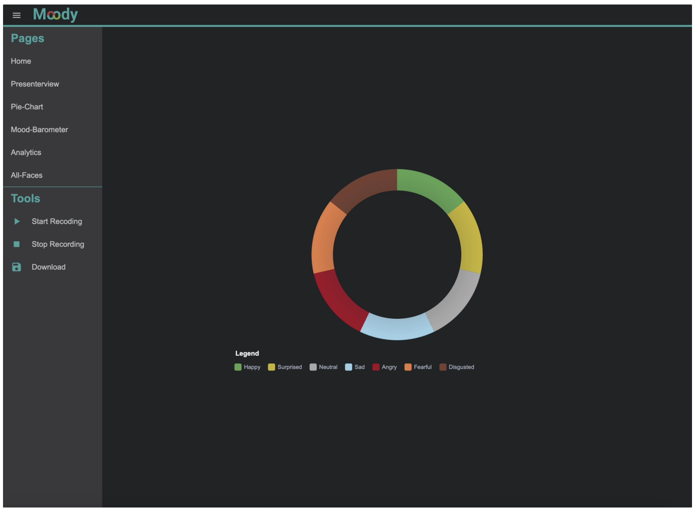

Figure 1.1: The *Pie chart* view shows the emotions averaged out among the team members.

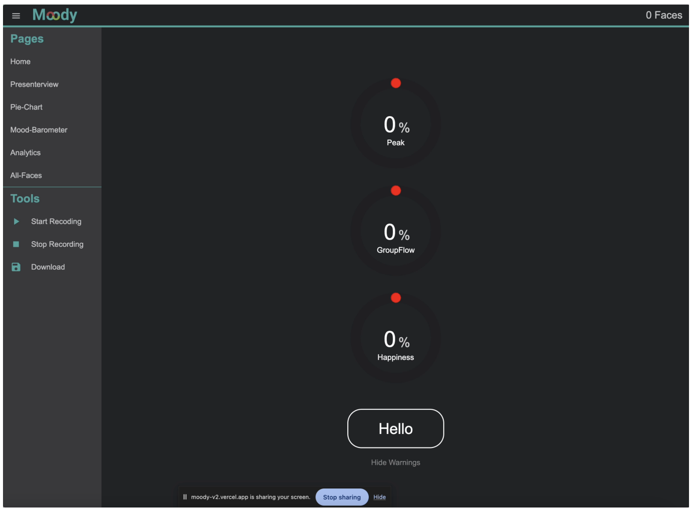

Figure 1.2: The *Presenter view* shows the main metrics providing rapid feedback to the speaker. At the bottom, an LLM-generated short actionable insight is shown.

2 Team Emotion Analysis with AI

2.1 Purpose of This Work

In the course of this work, a system *Moody* has been developed to analyze team emotions during video calls and provide real-time feedback. Furthermore, state-of-the-art models have been utilized to build an analysis framework aimed at enabling the assessment of teamwork and deriving insights into how emotional intelligence contributes to enhancing team productivity. This represents a significant advancement as such insights were previously only based on empirical evidence.

Additionally, a combination of AI models was employed in establishing an analytical framework for further use at MIT CCI. It was designed to facilitate the evaluation of teamwork dynamics and offer valuable insights into how emotional intelligence can serve to improve overall team performance. It is worth noting that these advancements mark a departure from previous practices, as this area of research had so far relied mainly on empirical evidence for investigation. The analytical framework builds a basis for quantifiable measurements and predictions, empowering researchers to quantify the effects of individual emotions, emotional entanglement and other observations on the productivity of teams. The discoveries made that way build the core for recommendations and assessments of *Moody*, providing them in a form that's fast and easy to digest for any team member, even without specialized knowledge on team productivity.

2.2 Challenges

The primary challenge in analyzing emotions in video calls lies in accurately detecting and interpreting the complex and subtle facial expressions of participants in real-time. This requires robust artificial intelligence (AI) models that can handle diverse lighting conditions, face orientations, and facial occlusions, which are common in video calls. Identity detection AI models also need to recognize team members robustly as they may join/quit calls without prior notice or suddenly get out of frame.

Another difficulty associated with analyzing emotions during video calls stem from the need to effectively detect and interpret intricate and nuanced facial expressions in real-time. Because video calls require a lot of data to be analyzed, a fine balance must be met between speed and accuracy when developing an appropriate setup of AI models. Another factor to be kept in mind that the cost per minute of running the AI models must be kept low enough (preferably on the order of magnitude of $0.01 USD) to keep the technology accessible for large-scale experiments and adoption in various organizations.

In addition to the technical aspects, the integration of user-friendly interfaces that facilitate the seamless collection and analysis of emotional data during video calls is pivotal. These interfaces should provide a clear and easy-to-understand visualization of emotional dynamics, allowing team members and leaders to gain insights and take necessary actions in real-time.

Ultimately, by overcoming the challenges through a comprehensive approach that encompasses advanced AI models, efficient computational resources, and user-friendly interfaces, *Moody* allows organizations to harness the power of emotional analysis in video calls to enhance team productivity and collaboration. Furthermore, *Moody* provides

2.3 AI Methods of Emotion Analysis

Several AI models have been developed to address these challenges, each with its strengths and specific applications. Here, we discuss some of the prominent categories of models used for emotion analysis. Further, we discuss how these models can be combined in a pipeline to tackle the objective of this work.

2.3.1 Face Detection Models

The first step in emotion analysis is detecting faces within a video frame. Models like RetinaFace are widely used for this purpose due to their high accuracy in detecting faces under various conditions.

Retinaface is a detection model designed for accurate and dense face localization [21] [22] [23] in complex and diverse real-world scenarios. Developed by Deng et al. in 2019, Retinaface operates as a single-stage model, making it highly effective and fast in detecting faces. The model is able to accurately detect faces in various poses, lighting conditions. This model utilizes a combination of convolutional neural networks (CNNs) to efficiently process image data and extract informative features for precise face localization. Additionally, the use of anchor mechanisms [24] [25] allows the model to accurately identify and localize faces across different scales and aspect ratios within the input images.

Retinaface is trained and tested on several large-scale face detection datasets, most prominently, WIDER FACE [26]. The WIDER FACE dataset consists of 32,203 images with 393,703 labeled faces in various poses and occlusions, making it well-suited for evaluating face detection models in real-world scenarios. This dataset enables Retinaface to achieve robust and accurate face localization in complex environments.

The loss function used in the training of Retinaface is the focal loss, which is particularly

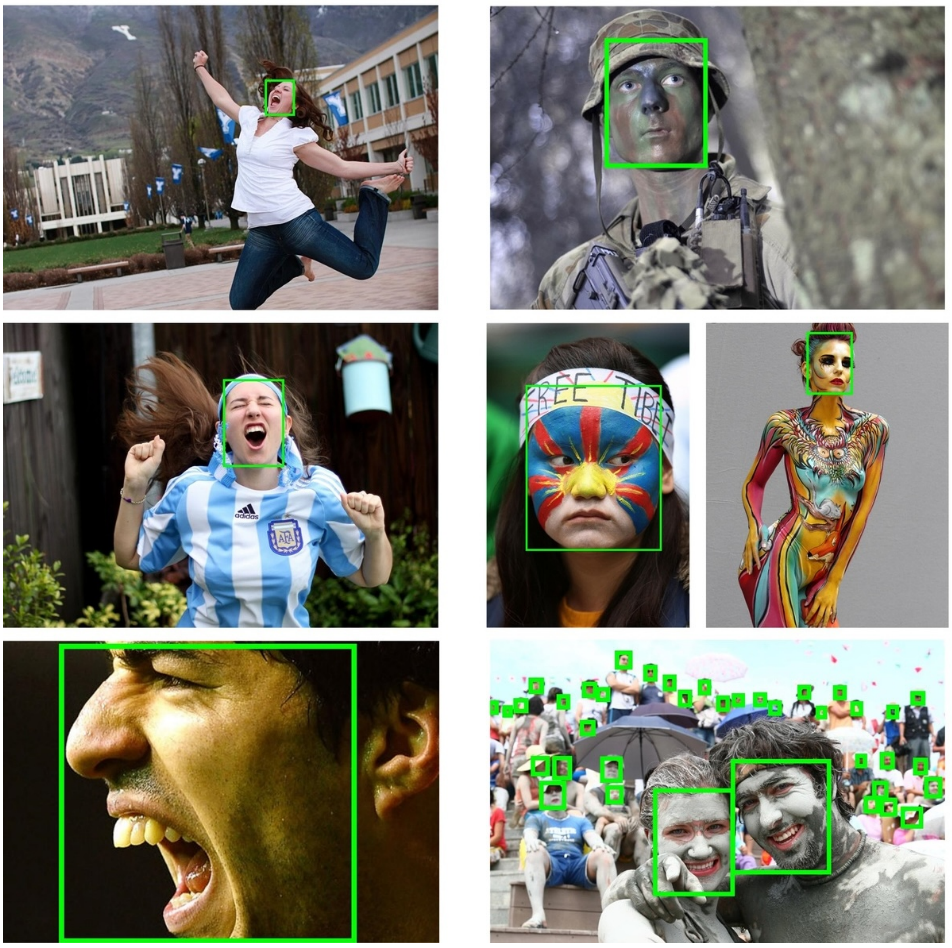

Figure 2.1: Sample images from the WIDER FACE dataset with ground truth bounding boxes. The WIDER FACE dataset consists of images with faces possessing a high degree of variability in scale, pose and occlusion. [26]

well-suited for dense object detection [27] tasks like face localization. The focal loss is effective at addressing the class imbalance that occurs in densely labeled datasets, such as the WIDER FACE dataset, by assigning less weight to well-classified examples and focusing more on the misclassified ones. This allows the model to effectively learn from examples and improve its accuracy in localizing faces, especially in scenarios with occlusions and varying lighting conditions.

In addition, Retinaface uses a regression function, such as smooth L1 loss [28], for bounding box regression. This regression function is beneficial because it is robust to outliers and prevents large errors from dominating the training process. By using smooth L1 loss, Retinaface can effectively learn to predict accurate bounding box coordinates for face localization while minimizing the impact of outliers and noisy annotations in the training data.

Overall, the choice of focal loss for classification and smooth L1 loss for bounding box regression in Retinaface contributes to its robustness and precision in detecting and localizing faces in diverse real-world scenarios.

Multi-task Cascaded Convolutional Networks (MT-CNN) are a widely used alternative approach for face detecion. MT-CNN is a face detection model developed by Zhang et al. The model is designed to perform multiple tasks in a cascaded manner, leading to accurate and efficient face detection in various real-world scenarios. MT-CNN consists of three stages: proposal network, refinement network, and output network. These stages work together to detect faces at different scales and make the detection process robust to variations in facial pose, lighting, expressions, and occlusions.

The proposal network generates candidate box regions containing faces, which are then refined by the refinement network to improve accuracy. Finally, the output network produces the final bounding box coordinates and facial landmarks. This multi-stage architecture enables MT-CNN to achieve precise and reliable face detection in complex environments, and has been extensively evaluated on benchmark WIDER FACE [26], demonstrating its effectiveness in detecting faces under various conditions.

The multi-task approach of MT-CNN allows it to simultaneously address the tasks of face detection, bounding box regression, and facial landmark localization. The architecture of MT-CNN, also allows it to effectively handle scale variations, occlusions, and pose variations, leading to precise and reliable face detection [29]. This makes MT-CNN a promising choice for real-world face detection applications where accuracy and speed are essential.

2.3.2 Landmark Detection Models

Once faces are detected, facial landmark detection models such as MobileFaceNets [30] are employed to identify key points on a face, such as the corners of the eyes, points on the edge of the lips, the edge of the face etc. These landmarks help in aligning faces before further analysis and are crucial for accurate emotion prediction.

MobileFaceNets is a highly efficient CNN designed specifically for real-time face verification on mobile and embedded devices, utilising less than 1 million parameters. By focusing on creating a balance between accuracy and computational efficiency, MobileFaceNets addresses the pressing need for reliable face verification technology that can operate within the constraints of mobile hardware. This is achieved through a meticulously structured CNN that significantly reduces the computational requirements without compromising the model's accuracy.

The novelty of MobileFaceNets lies in its design, which overcomes the limitations of common mobile networks in the context of face verification tasks. This includes the introduction of a global depthwise convolution (GDConv) layer as opposed to a global average pooling layer or a fully connected layer to generate a discriminative feature vector. This enables the network to assign varying levels of importance to different spatial positions in the face images, thus enhancing the model's accuracy and efficiency for face verification purposes.

MobileFaceNets was trained from scratch using the refined MS-Celeb-1M dataset [31] combined with the ArcFace loss function [32]. This approach ensured a robust training regime that adequately prepared the model to achieve high accuracy levels in real-world scenarios. The training process meticulously considered aspects such as weight decay, optimization methods (SGD with momentum), and learning rate adjustments to optimise performance further and ensure comprehensive learning from the extensive dataset.

When it comes to testing and performance evaluation, MobileFaceNets exhibited remarkable results on benchmark Labeled Faces in the Wild (LFW) and MegaFace [33], achieving an accuracy of 99.55% on LFW and a True Acceptance Rate (TAR) of 92.59% at a False Accept Rate (FAR) of 1e-6 on MegaFace. These results not only highlight the accuracy of MobileFaceNets compared to other state-of-the-art (SoTA) CNNs, such as NASNet [34] and ShuffleNet [35], but also its considerable speed-up in actual inference time on mobile devices, making it an innovative solution for efficient and reliable real-time face verification on mobile platforms.

MobileNets are a class of efficient, lightweight convolutional neural network architectures designed by Google for use on mobile and embedded devices. As such, this class of model architectures offers a good tradeoff between accuracy and performance for the use case discussed in this work. These networks leverage depthwise separable convolutions to build models that significantly reduce the computational burden, enabling faster operation with a minimal decrease in accuracy. MobileNets introduce two hyperparameters, width multiplier and resolution multiplier, allowing model builders to tailor the network size and computational requirements to their specific application needs, achieving a balance between latency and accuracy.

The novelty of MobileNets lies in their streamlined architecture, primarily based on depthwise separable convolutions. This method splits the filtering and combining steps of traditional convolutions [36] into separate layers, which drastically lowers the computational cost and model size. Depthwise separable convolutions, along with the introduction of the width and resolution multipliers, enable MobileNets to provide a modular and scalable approach to designing efficient neural networks that can be customized for a variety of mobile and embedded vision applications.

MobileNets were trained using the TensorFlow framework [37] and RMSprop, using asynchronous gradient descent [38]. The training process entailed less regularization and data augmentation techniques due to the reduced tendency of smaller models to overfit. Notably, MobileNets require careful adjustment of regularization, particularly a reduced or negligible weight decay for the depthwise filters, due to their limited number of parameters. These choices in training parameters and strategies were found crucial for optimizing the performance of MobileNets across different configurations.

Testing and performance evaluation of MobileNets demonstrated their effectiveness across a broad spectrum of applications, from ImageNet classification to object detection, fine-grained recognition, and geographical localization [39]. MobileNets not only achieve competitive accuracy compared to larger models but do so with substantially lower computational costs and model sizes. For example, when compared with conventional architectures like VGG16 [40] and GoogleNet [41] on the ImageNet dataset, MobileNets provided near-equivalent accuracy with significantly fewer parameters and mult-add operations, showcasing their suitability for resource-constrained environments while maintaining high performance.

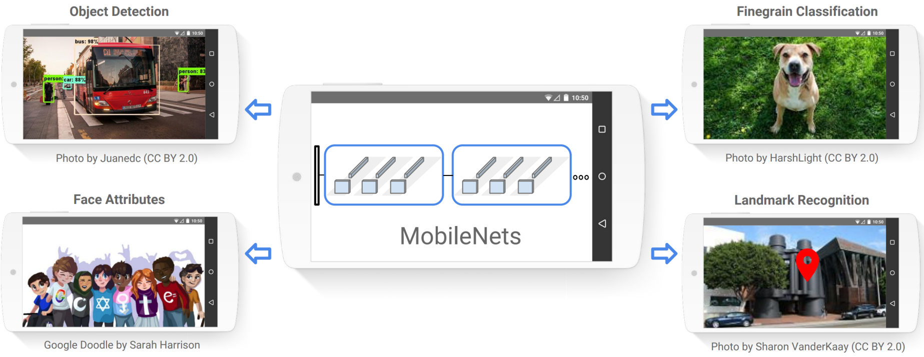
Object Detection
Photo by Juanedc (CC BY 2.0)
Face Attributes
Google Doodle by Sarah Harrison
MobileNets
Finegrain Classification
Photo by HarshLight (CC BY 2.0)
Landmark Recognition
Photo by Sharon VanderKaay (CC BY 2.0)

Figure 2.2: MobileNets includes a wide range of functions, including object detection, face attributes, finegrain classification, and landmark recognition. [42]

2.3.3 Face Pose Models

Understanding the orientation of the face, provided by models like Img2Pose, also aids in emotion analysis by adjusting for head movements and ensuring that emotional predictions are consistent regardless of how the face is positioned.

Img2pose, a face detection model developed by Albiero et al. in 2021 [43], is designed to perform simultaneous (one-shot) face detection and head pose estimation. This model aims to accurately detect faces and estimate the head pose in real-world scenarios, making it a valuable tool for applications such as biometrics, surveillance, and human-computer interaction.

One of the key aspects of img2pose is its ability to estimate the head pose in addition to facial landmark detection. This feature makes it particularly useful in scenarios where knowledge of the head's orientation is essential, such as in surveillance and human-computer interaction applications. By providing both face detection and head pose estimation, img2pose offers a comprehensive solution for understanding the spatial orientation of individuals within an image.

img2pose is trained on WIDER FACE dataset [26], which uses manually annotated five point facial landmarks on training faces from the Retina Face project [44]. The six degrees of freedom (6DoF) pose is obtained and then converted to global image frames.

The img2pose model was developed using PyTorch with ResNet-18 [45], utilizing stochastic gradient descent with dynamic learning rate adjustments based on validation loss. Due to the limitations of Euler angles for evaluating face poses with large yaw angles, the model was refined on the 300W-LP dataset [46] [47], which contains moderated yaw angles, to ensure accuracy. This refinement process involved a fixed learning rate and was completed in just 2 epochs, streamlining the training approach.

The model was tested on AFLW2000-3D [47] and BIWI [48] datasets. It outperformed existing state-of-the-art methods by directly processing full images without the need for pre-cropping or scaling. Achieving an mean absolute error in rotation (MAEr) of 3.913 and operating at 41 fps on the AFLW2000-3D dataset, and an MAEr of 3.786 at 30 fps for the BIWI dataset, the model demonstrated both high accuracy and efficiency. Further validation through an ablation study confirmed the combined loss functions significantly enhanced the model's head rotation predictions, showcasing the robustness and precision of the img2pose approach.

2.3.4 Action Unit (AU) Detection Models

Action Units (AUs) are specific movements of facial muscles that correlate with different emotions. Models like SVM (Support Vector Machine) and XGBoost are used to detect these AUs from facial landmarks. The detection of AUs allows for a fine-grained analysis of facial expressions.

2.3.5 Emotion Detection Models

With the face aligned and AUs detected, emotion detection models like ResMaskNet can classify the overall emotion of the face using Facial Expression Recognition. This model uses deep learning techniques to interpret the combined information from face detection, landmarks, and AUs to predict emotions such as happiness, sadness, anger, etc. The architecture enhances CNN performance by focusing on essential regions of the face, such as the eyes, nose, and mouth, which are crucial for recognizing emotions [49].

ResMaskNet is structured with four main Residual Masking blocks, each containing a Residual Layer and a Masking Block. The Residual Layers are part of the residual network architecture, and is primarily used for feature processing, wheres Masking Blocks produce weights for corresponding feature maps.

Evaluations of ResMaskNet on datasets FER2013 [50] and VEMO2020 (Vietnam Emotion) [51] [52] illustrate SOTA accuracy, outperforming well-known models such as VGG19 [40], ResAttNet56 [53], Densenet121 [54], Resnet152 [45], and Cbam_resnet50 [55].

2.3.6 Identity Detection Models

In scenarios where the identity of participants is also relevant, models like FaceNet can be used. This model provides a way to recognize and differentiate between participants, which can be particularly useful in personalized emotion analysis. [56]

FaceNet provides a 512-dimensional embedding representation of every face. This allows for accurate detection of the same person across different video frames or whole videos, even with complications such as faces being partially obscured, faces not fully present in the frame, or different lighting conditions.

The model uses a deep convolutional neural network to learn the embedding of face images. The similarity between two face images is assessed based on the square of the Euclidean distance between the corresponding normalized vectors in the 512-dimensional Euclidean space. The system used the triplet loss function as the cost function and introduced a new online triplet mining method. The system achieved an accuracy of 99.63% which is the highest score on Labeled Faces in the Wild dataset in the unrestricted with labeled outside data protocol. [57] [58]

2.4 AI Model Pipeline for Emotion Analysis

To achieve a comprehensive emotion analysis while considering the aforementioned accuracy-speed tradeoff and cost constraints, the AI models described above need to be combined into a pipeline in a particular way. [59]

The pipeline runs in a Docker container on a Runpod instance with a NVIDIA RTX A2000 GPU (6 GB GDDR6 memory, 8.0 TFLOPS single-precision performance, 15.6 TFLOPS RT Core performance), 35 GB RAM and 9 virtual CPU (vCPU) cores working at 2.7 GHz. At the time of writing, the cost of running this virtual instance was $0.14 USD per hour, making it very accessible not only to any organizations for day-to-day use for team meetings, but also for CCI to run large-scale team experiments and analyze hundreds of hours of video data.

In this setup, it takes about 1 s to analyze one video frame with 3 team members in a video call. Because human emotions don't change drastically more often than that, a sampling period of 1 s is sufficient to accurately analyze the team's development of emotions over time.

This particular setup was chosen to make it possible to launch the pipeline as real-time backend for *Moody* while at the same time enabling researchers to analyze long video recordings from meetings as a long-running job. In the backend mode, *Moody* asynchronously sends frames from the user's video call window to the instance running the analysis pipeline every 1 s. *Moody* then gets updates on team members' emotions asynchronously from the pipeline and shows the development of emotions in real time to the team members. Based on this real-time data, *Moody* is also able to give real-time actionable feedback to every individual user on how to improve their team interactions. This makes it possible for users to immediately implement that feedback into their interaction with others in the meeting, driving the team to best practices of communication — even if no one on the team is an expert in team communication.

In the long video analysis mode, one can upload a video to the Runpod instance, select the number of frames per second (FPS) and have a full analysis prepared after some time as a downloadable .csv file. Because it takes about 1 s to analyze 1 video frame, at the recommended sampling frequency of 1 FPS the processing time is about the same as the duration of the video. This mode enables detailed analysis of recorded meetings, providing valuable insights that can be used for further research and training purposes.

To illustrate each stage of the emotion analysis pipeline, consider this sample image frame from one of the video calls recorded at MIT CCI with consenting individuals (Figure 2.3). The emotion analysis pipeline selects a set number of equally spaced frames per second from a video and uses the AI models on it as described further.

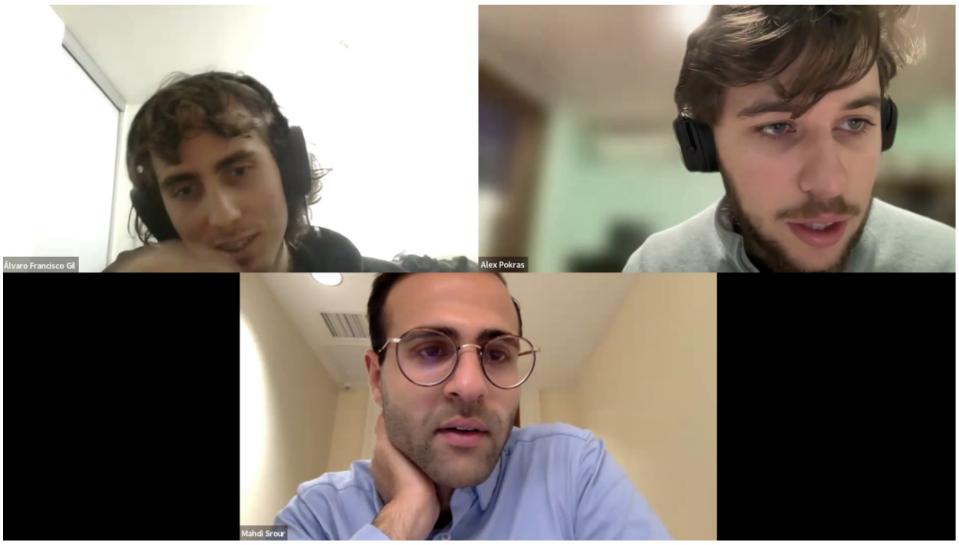

Figure 2.3: Example frame captured from a video call (Zoom software) with three team members.

To sum up, this multi-model approach leverages the strengths of each model, ensuring robust and accurate emotion analysis suitable for both real-time applications in video calls and offline video call data analysis. Each step adds a layer of refinement, making the system adept at handling the complexities of real-world video communication.

The following describes the stages of the analysis pipeline.

2.4.1 Extracting faces from the video frame using a face detection model

The initial step in the pipeline involves extracting faces from the video. RetinaFace emerged as the most promising solution for this task, successfully detecting the face of every participant in over 98 of video frames from a sample 1-hour team meeting recording we conducted. This high detection rate provides confidence in its sufficient effectiveness for our needs.

To achieve a speedup of 20-30% at a negligible loss of accuracy in the further stages of the analysis pipeline, the video data is compressed to 1/4 of its original size (1/2 along the X-axis and 1/2 on the Y-axis). RetinaFace outputs the bounding box coordinates for each face it detects in the video. The bounding box it outputs for each face was increased by 20% to account for movements of the head and eventual inaccuracies. The reason for this design decision was the empirical observation that it marginally increases the detection time of the further stages of the pipeline while at the same time including more important facial features, such as the whole edge of the face.

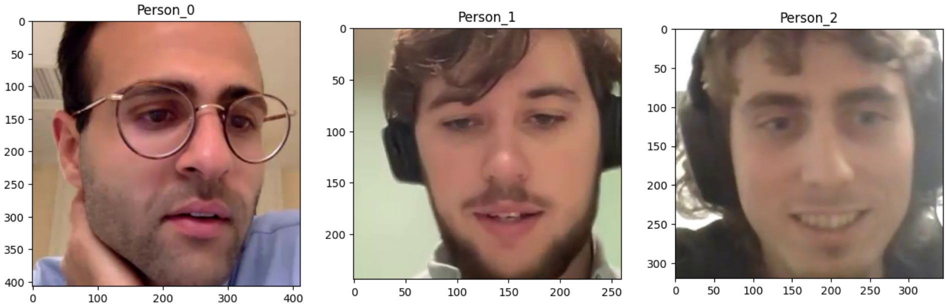

Figure 2.4: Face bounding boxes detected and cut from the example frame. Bounding box for each face widened by 20%. Frame image data compressed to 1/4 of its original size.

2.4.2 Identity detection

Because *Moody* needs to be as hands-free to use as possible and teammates can join or leave meetings without prior notice, it is necessary to implement functionality to map face frames to individuals on the team. This functionality is also crucial for providing real-time specific feedback to each team member individually.

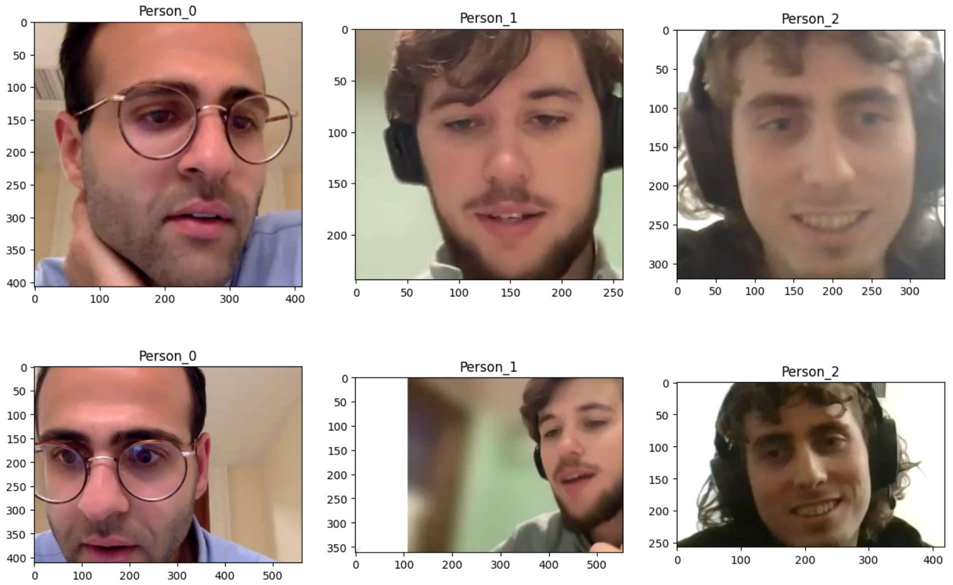

Figure 2.5: FaceNet allows for fast and accurate differentiation of team members even in the complex case of e.g. faces partially obscured or outside of the camera field of view.

The model used in the proposed implementation is FaceNet, specifically, the version with 512-dimensional embeddings. Those embeddings are stored in the .csv file that can be exported on request and labeled with teammates' names. This ensures be persistent across calls, thus elucidating team members' interactions over multiple meetings. This functionality can be utilized to empower researchers to explore long-term changes in team dynamics, e.g. over longer-term projects spanning multiple video calls over the course of several days to several months or years.

2.4.3 Detecting facial landmarks and analyzing AUs

A classic method for emotion recognition is using an advanced ML model to find facial landmarks (important points on one's face) on an image and feeding their coordinates into a simpler ML model that correlates those to action unit activation. [60]

Action units (AUs) are a classical approach used in the analysis of facial expressions to understand emotions. Prior to the rise of deep learning techniques, action units were developed based on domain knowledge about how different facial muscle movements correspond to specific emotions. Ekman and Friesen introduced the Facial Action Coding System in 1978, which provided a structured methodology for identifying and categorizing these action units. [61] [62] This system has been widely used in various fields such as psychology, neuroscience, and social sciences to study emotional responses through facial expressions. By observing and measuring the activation of specific action units, researchers can gain insights into the underlying emotional states of individuals.

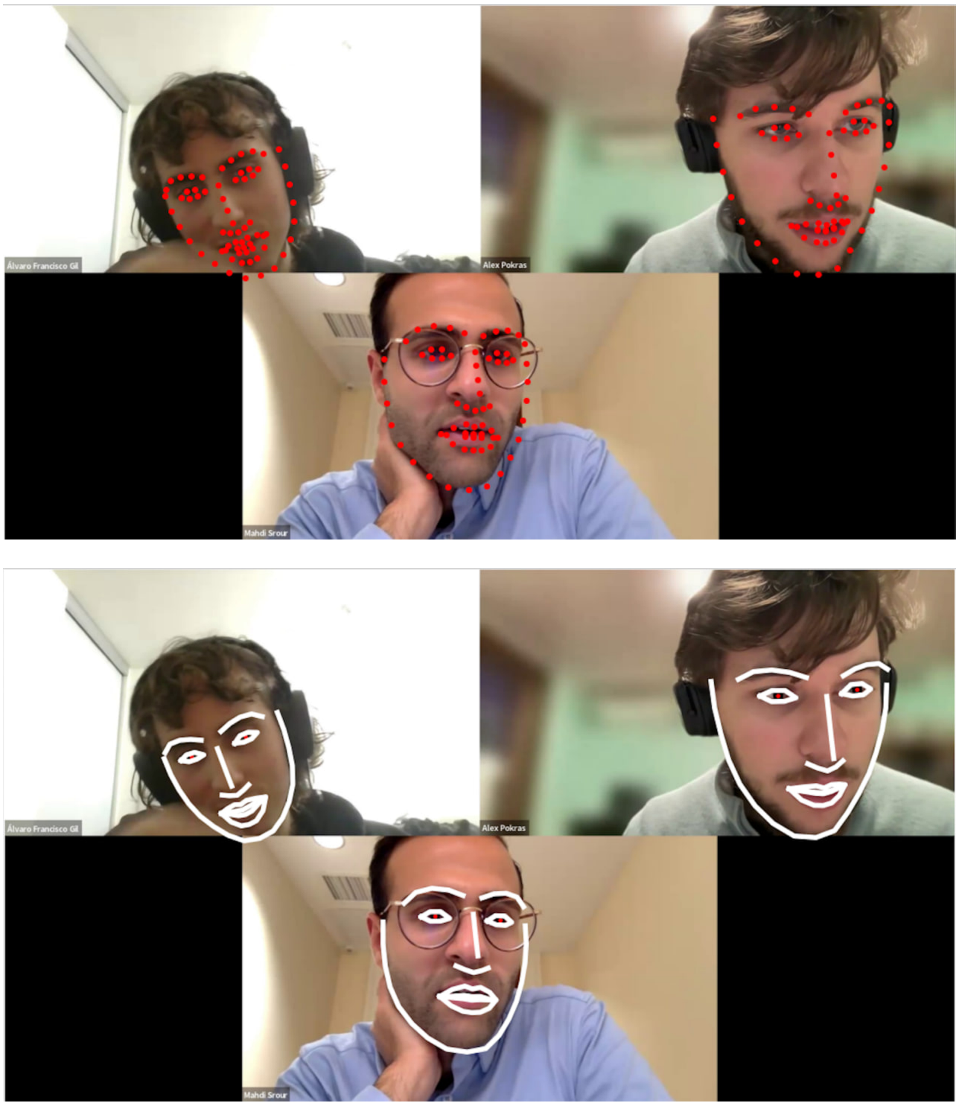

Figure 2.6: Facial landmark recognition.
Top: 68 key facial landmarks are identified using the img2pose model.
Bottom: Spline curve graphed along the points of the landmarks for visualization.

In the proposed analysis pipeline, the img2pose model is used to find 68 facial landmarks and a pretrained XGBoost model to correlate them to 43 AUs. A pretrained SVM is then used to recognize emotions based on AUs. [63]

After some testing, the approach combining XGBoost for AU detection and SVM for emotion detection turned out to provide unsatisfactory results — marginally better than random guessing. Consequently, it was then decided to use modern DL models for emotion detection, even at an increased compute cost. However, the facial landmark recognition model proved to be useful for the speaker diarization task, discussed later in this work.

2.4.4 Detecting emotions

According to the Emotion Facial Action Coding System (FACS) developed by Paul Ekman and Wallace Friesen, there are six emotions that are considered to be culturally universal: *anger, disgust, fear, happiness, sadness, and surprise*. Most DL models are trained for recognizing the FACS emotions on datasets that have been labeled for these six emotions as well as the *neutral* emotion. [64] [65]

In the analysis pipeline, Residual Masking Network (ResMaskingNet) is used. This model provides the most optimal speed-accuracy tradeoff at an affordable cost: it is a small enough model to run on a single NVIDIA RTX A2000 GPU at 1 FPS while fulfilling the cost requirement of ca. $0.01 USD per minute (depending on the dynamically calculated prices based on availability on Runpod — refer to Section 2.4). To achieve such performance, predictions for every face detected in the extraction stage are conducted concurrently on the same GPU. The model produces a confidence score 0-1 for each one of the six FACS emotions. This confidence score is directly used for both further analysis and real-time feedback for team members.

anger
disgust
fear
happiness
sadness
surprise
neutral

Figure 2.7: Legend for the six FACS emotions and the *neutral* state. This color scheme is used in all the following diagrams, so this legend is omitted for brevity.

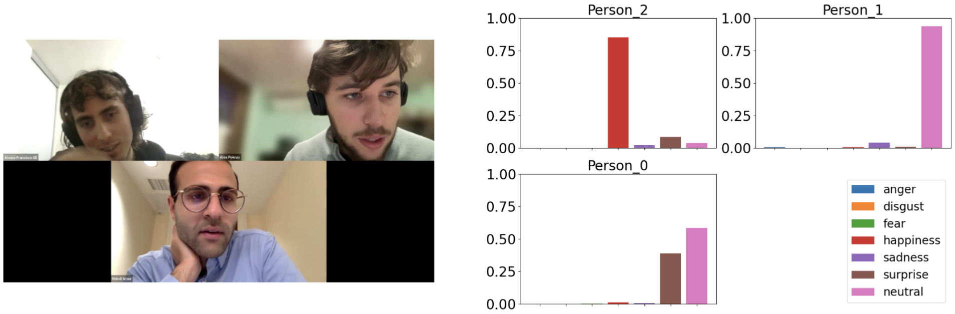

Figure 2.8: Emotions detected by ResMaskingNet in the example frame.

3 Speaker Diarization and Interruption Analysis with AI

3.1 Motivation

In this chapter, we will explore the importance of speaker diarization and interruption analysis in understanding team dynamics and productivity. Speaker diarization is the process of identifying who is speaking at different points in a conversation or meeting. It is a notoriously difficult task for automatic systems, but recent advancements in speech recognition technology using deep learning have made it more feasible. [66]

In the context of this work, speaker diarization is used to determine who is currently speaking in a video and analyze other teammates' emotional response to them. This is important for understanding group dynamics and communication patterns within a team. Some questions that can be answered are thus: How does the distribution of speaking time among team members impact productivity? How does each particular team member respond to each other one speaking? What are the patterns of such emotional response among all team members and how do they impact the well-being of the team?

Interruption analysis, on the other hand, involves identifying instances where one speaker interrupts another. Analyzing the impact of these interruptions brings insights on the flow of conversation and overall team dynamics.

By integrating speaker diarization and interruption analysis with emotion detection, this work aims to provide insights into how communication patterns within teams can affect productivity and teamwork. The ability to accurately identify speakers and analyze other teammates' emotional responses as well as incorprating interruptions data can provide valuable data for understanding power dynamics, conversational dominance, and the overall effectiveness of team communication. Additionally, by incorporating emotion detection, we can delve deeper into the emotional dynamics within the team. This analysis can help identify potential communication challenges, areas of improvement, and strategies for enhancing team collaboration.

A potentially promising improvement would be to add a speech recognition model such as OpenAI's Whisper to the system. [67] This would enable the researchers to analyze the sense, not just the sequences, of what is being said in the meeting. However, speech recognition was explicitly omitted in this work due to insurmountable privacy concerns for the use cases planned by MIT CCI. As such, this remains an open direction for future research.

3.2 Speaker Diarization Pipeline

For speaker diarization, the Python library *pyannote-audio* is used [68] [69]. At the time of writing, the version 3.1 pretrained pipeline offers state-of-the-art performance on various speaker diarization datasets.

3.2.1 Evaluation

The diarization error rate (DER) is a metric commonly used for speaker diarization evaluation. It is calculated as follows:

$$\text{DER} = \frac{\text{false alarm} + \text{missed detection} + \text{confusion}}{\text{total}}$$

where false alarm is the duration of non-speech incorrectly classified as speech, missed detection is the duration of speech incorrectly classified as non-speech, confusion is the duration of one speaker confused for another, and total is the total speech duration summed up among all speakers. [69]

Of all the datasets *pyannote-audio* has been trained and evaluated on, the AMI meeting corpus is the most relevant to the use case discussed in this work [70]. *pyannote-audio* achieves a 18.8% diarization error rate (DER) on the AMI dataset.

3.2.2 Pretrained Pipelines

The pyannote library offers pretrained pipelines that are ready to use out-of-the-box for various diarization tasks. These pipelines are trained on a wide range of data and are optimized for general use. The version 3.1 of pyannote.audio, as mentioned, includes state-of-the-art pretrained models that have shown significant improvements in performance metrics across several benchmarks, including the AMI meeting corpus.

The output of speaker diarization pipelines consists of timestamps when each speaker starts and ends talking. These can be visualized as segments along the time axis.

The following are the stages of a pyannote-audio pretrained speaker diarization pipeline.

1. **Speech Activity Detection (SAD)**: This component detects whether a segment of audio contains speech or not. This is crucial for efficiency of the pipeline as it reduces the amount of audio data that needs to be processed in subsequent steps.

2. **Speaker Change Detection (SCD)**: This component identifies points in the audio stream where the speaker changes. This helps in segmenting the audio into homogeneous segments that likely contain speech from only one speaker.

3. **Overlapped Speech Detection**: This component detects segments where multiple speakers are talking simultaneously. Handling overlaps is one of the most challenging aspects of speaker diarization. However, it is also informative because interruption analysis can provide valuable insights on team interaction.

4. **Speaker Embedding**: This involves extracting speaker-specific features from segments of speech which are then used to cluster segments belonging to the same speaker. For reasons listed in 2.4.2, this is crucial for ensuring accurate and streamlined operation.

5. **Clustering**: The extracted speaker embeddings are clustered into group segments belonging to the same speaker across the audio stream.

The speaker diarization pipeline outputs timestamps of each person starting and stopping talking. These timestamps can be interpreted and visualized as segments when each particular person is speaking. Figure 3.1 demonstrates how such segments intertwine and overlap in a sample 120-second audio snippet.

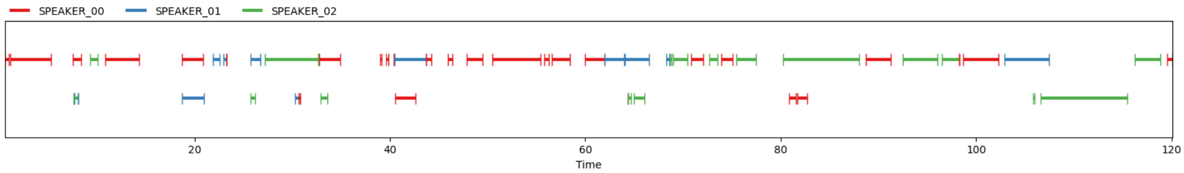

Figure 3.1: Output of the speaker diarization pipeline on an example 120-second audio snippet from the meeting.

3.2.3 Optimization

The components of the pyannote pipeline are optimized jointly in an end-to-end manner. This means that the entire pipeline, from speech detection to clustering, is trained together to minimize the overall diarization error rate (DER).

This approach also ensures high efficiency of the pipeline. The setup discussed in 2.4 takes around 1 minute to perform speaker diarization on a 1-hour video.

3.2.4 Interruption Analysis

Because the pipeline includes Speaker Change Detection and Overlapped Speech Detection components, it can be used for interruption analysis in addition to speaker diarization. Under the hood, the interruptions are defined as an overlap of segments from the speaker diarization pipeline. This allows for analyzing when and by whom each person gets interrupted.

Notably, interjections such as *yeah*, *hm* etc. are not considered interruptions. These so-called productive interjections are not necessarily well correlated to team dynamics. Instead, for the purposes of the discussed system, a heuristic rule is used when interruptions are only registered when one speaker starts talking while another speaker is already talking and speaks for over 2 s in total.

The cumulative number of interruptions among all team members for the whole duration of the video is also calculated. It can be indicative of the engagement of team members and the liveliness of the conversation.

3.3 Combining Speaker Diarization and Emotion Analysis

To combine speaker diarization and emotion analysis, the output from the speaker diarization system is used together with each team member's emotions sampled every 1 s from the emotion analysis pipeline. The timestamps provided by the speaker diarization system are used to align the emotions of each team member with a speech segment of a certain (possibly the same) team member.

The result can be visualized as a matrix of emotion reactions of each team member to each team member (possibly themselves) speaking developing over time.

3.3.1 Mapping of speakers to faces in the video

To make Figure 3.2 more informative for further research, it is beneficial to identify which person in the video is which speaker. To achieve this, we find facial landmarks as discussed in 2.4.3 and analyze the average movement of the points corresponding to the mouth (points 48-67 in Figure 3.3) for each person. Because the movement of the mouth can occur due to face moving or turning into the frame, the size of the faces is normalized, they are centered around the centerpoint of the mouth, and trigonometric transformations are used to account for roll, pitch, and yaw.

By correlating the relative movement of each person's mouth in the video to the segments that the speaker diarization pipelines, it is possible to correlate speakers in the audio track to persons in the video. This last step truly splices the results of the emotion analysis pipeline with the speaker diarization pipeline and allows for most informative insights. The result can be visualized as a matrix of emotion reactions of each team member to each team member speaking while considering their identities — developing over time or aggregated throughout the entire video.

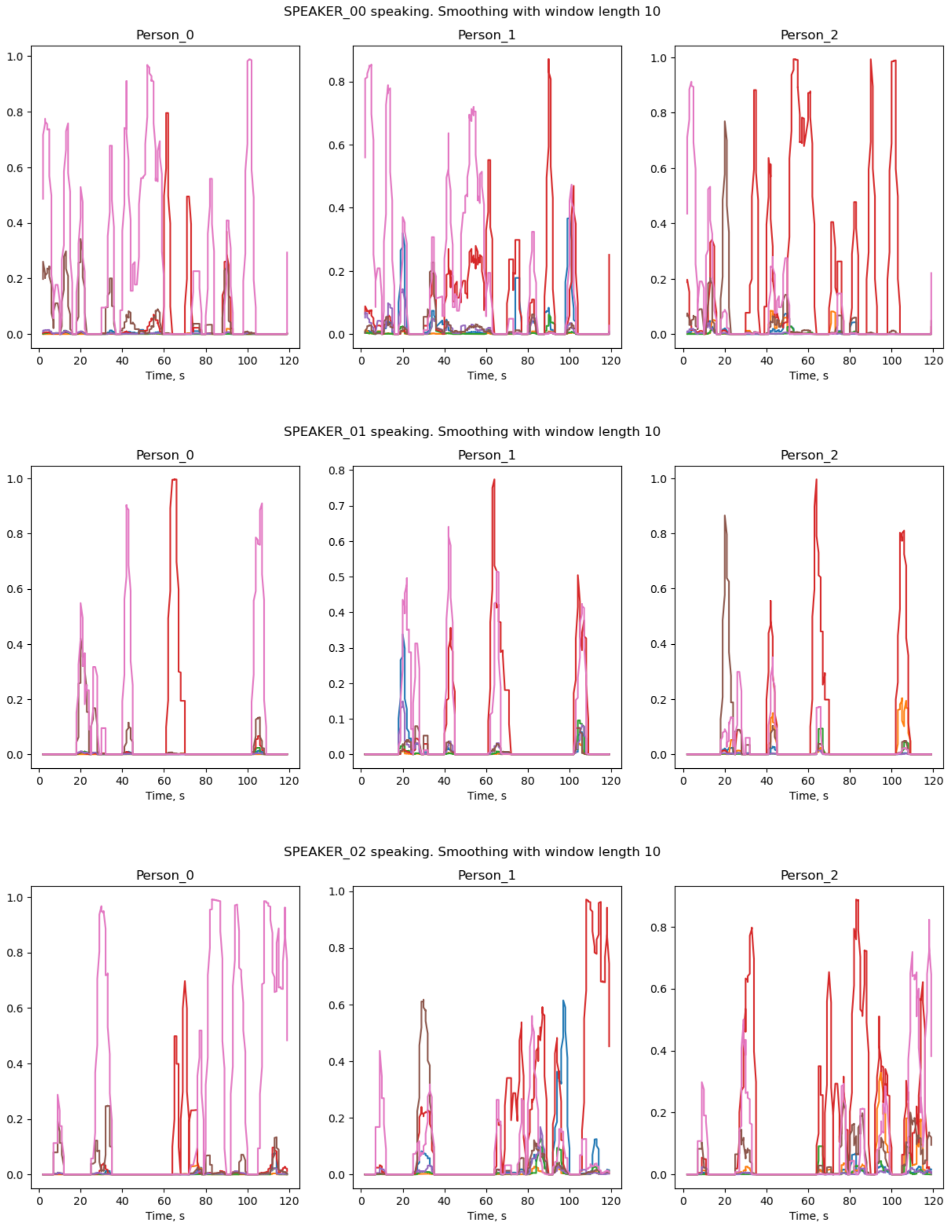

Figure 3.2: The emotional responses of each team member to each speaker (possibly themselves) over time in the same example 120-second video snippet as above. Smoothing with a moving average with a window length of 10 measurements is applied to reduce noise and highlight overall patterns. Legend: see Figure 2.7

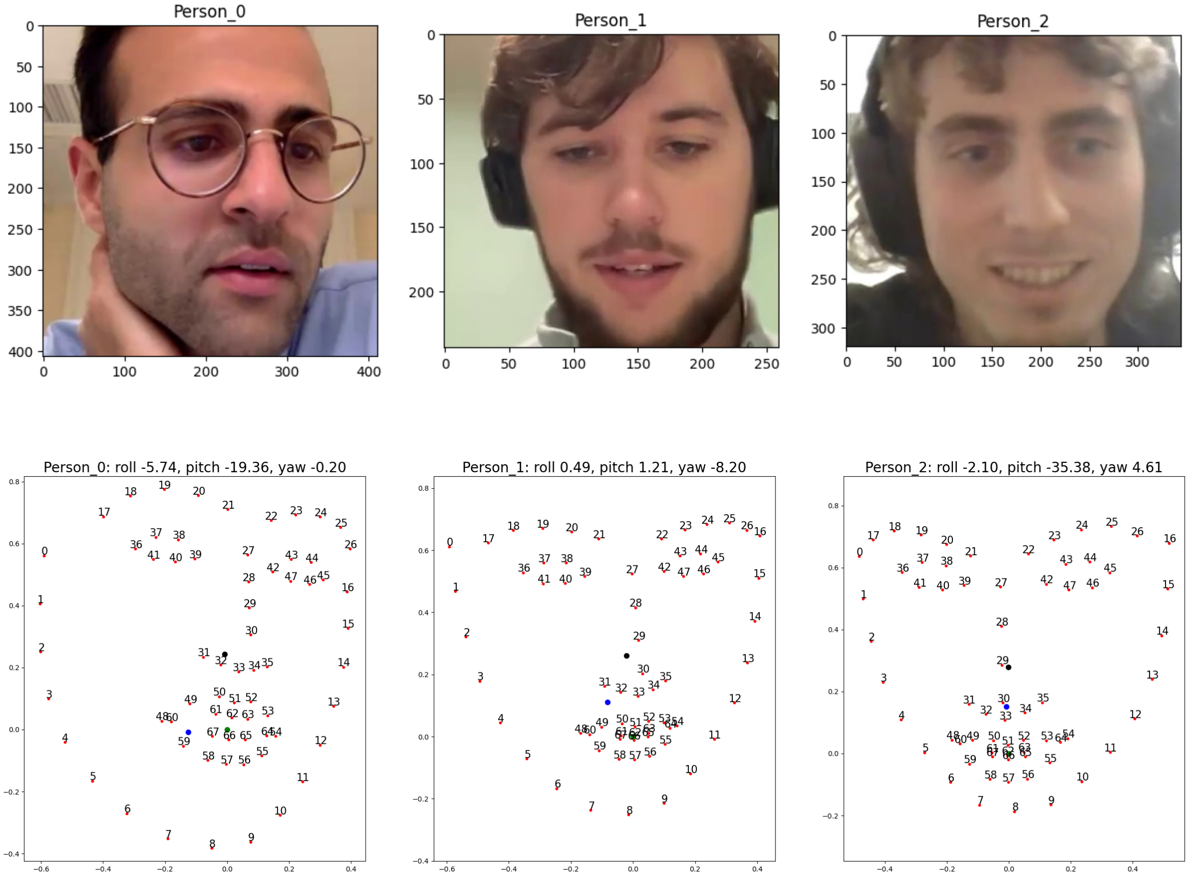
Person_0
Person_1
Person_2
0
50
100
150
200
250
300
350
400
Person_0: roll -5.74, pitch -19.36, yaw -0.20
Person_1: roll 0.49, pitch 1.21, yaw -8.20
Person_2: roll -2.10, pitch -35.38, yaw 4.61
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0.2
0.4

Figure 3.3: Faces from the example video frame and the output of the landmark analysis model.
Center points: black dot for the mean of the points on the edge of the face (points 0-16), green dot for the mean of mouth points (points 48-67), blue dot for the mean of all points.

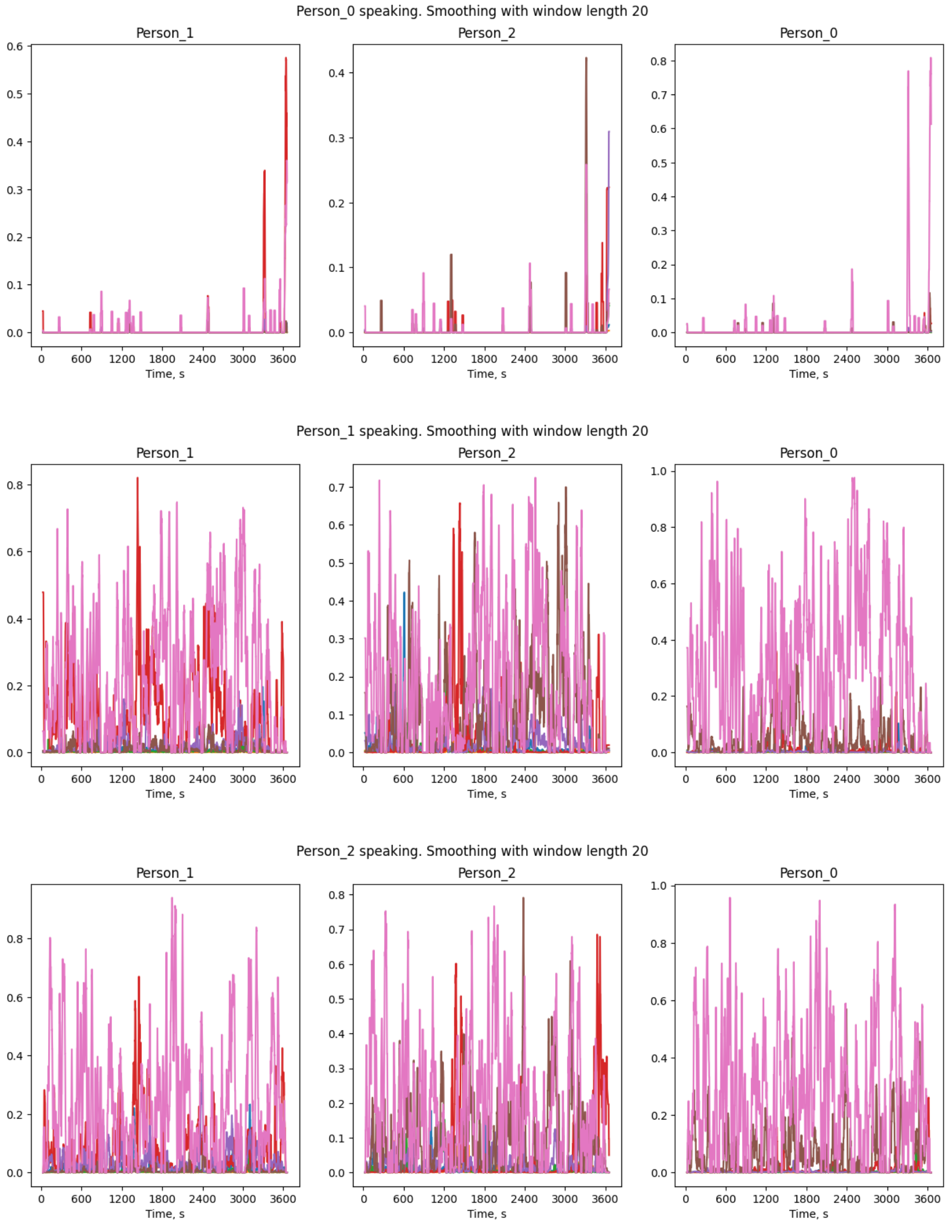

Figure 3.4: The emotional responses of every person in the video to every person talking over the entire 1-hour video call. Legend: see Figure 2.7
Smoothing with a moving average with a window length of 20 measurements is applied to reduce noise and highlight overall patterns.
Notably, the temporal development of team members' emotions over such long calls is quite difficult to interpret. Figure 3.5 offers an alternative data presentation.

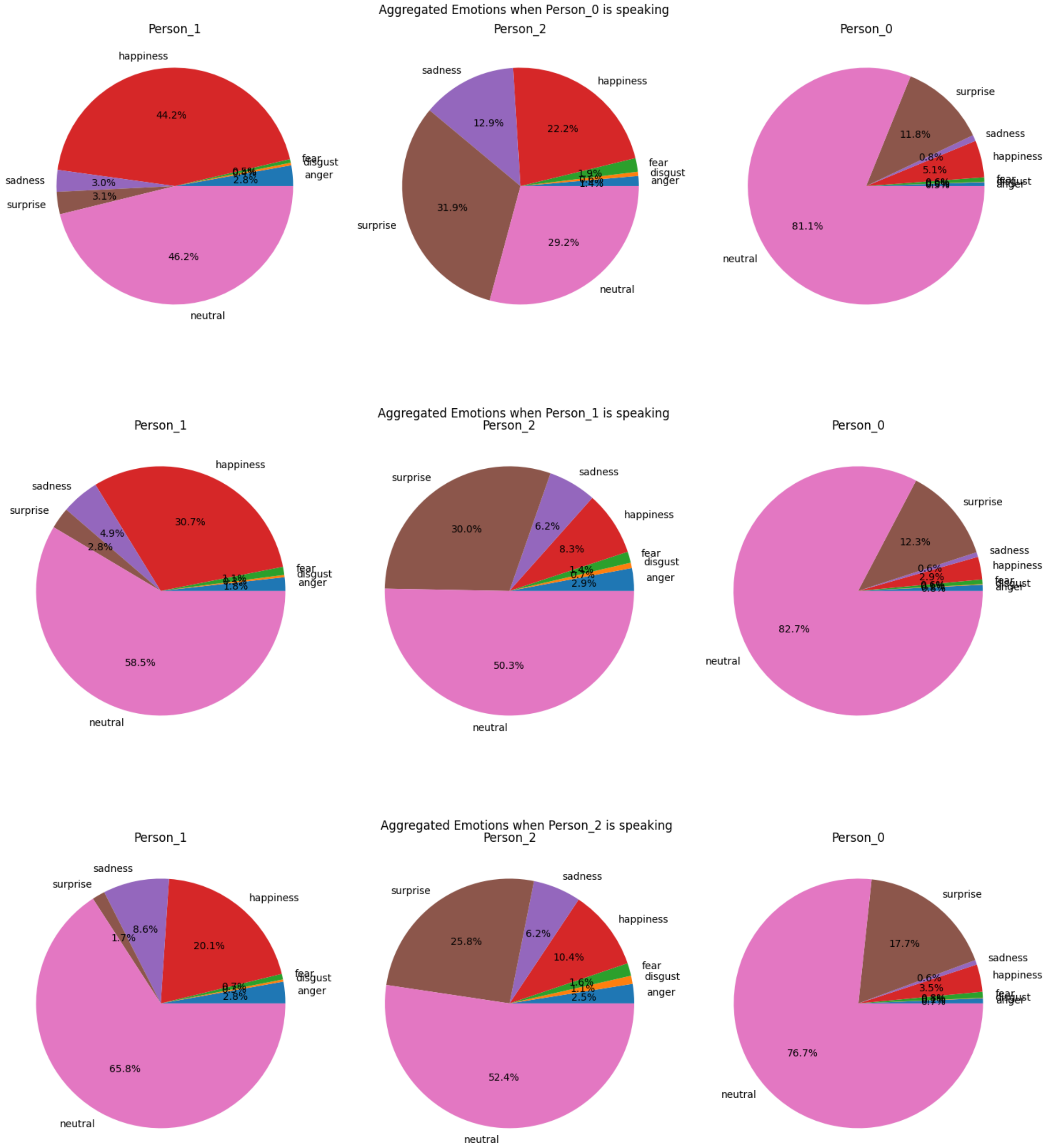

Figure 3.5: The aggregated emotional responses of every person in the video to every person talking over the entire 1-hour video call.

4 Conclusion

The proposed methodologies for analyzing team dynamics, comprising both emotion detection and speaker diarization pipelines, represent a significant step forward in understanding complex human emotion interactions in team settings. The combination of emotion analysis through advanced deep learning models and the use of speaker diarization to identify both talking segments and interruptions has yielded a system capable of providing deep insights into the impact of emotional entanglement on team productivity.

This sophisticated analysis is made easy and actionable through the user-friendly Moody interface, designed to be easy enough to use even for nonexperts in team intelligence. The Moody interface provides real-time feedback based on the team members' emotions and speaking patterns, advising users on making immediate adjustments to their communication strategies. By offering on-the-fly insights, Moody fosters a self-improving team environment where members can actively enhance their interpersonal dynamics and contribute to a more productive and emotionally intelligent workplace.

Moreover, the backend of the analysis system is robustly deployed in a Docker container on the Runpod platform, which efficiently handles the computational demands. At a cost of approximately $0.01 per minute of video analyzed at the recommended sampling frequency of 1 FPS, this setup is both cost-effective and scalable. The deployment serves dual purposes: it powers real-time analysis for Moody and supports asynchronous analysis for extended video recordings. This flexibility is particularly beneficial for academic researchers, who can use a series of provided Jupyter notebooks to analyze prerecorded videos and get deeper insights.

In summary, this work bridges the gap between advanced AI emotion detection models, academic research on the impact of emotional entanglement on team productivity, and practical application in organizational settings. By combining emotion and speech analysis pipelines with real-time feedback and scalable backend processing, the system is an easy-to-use tool for improving team productivity and emotional intelligence. For researchers, it provides a series of tools to numerically analyze the impact of emotional state on team productivity — something that previously was only done empirically.

List of Figures

Bibliography

[1] J. B. Bernerth, J. M. Beus, C. A. Helmuth, and T. L. Boyd, *The more the merrier or too many cooks spoil the pot? a meta-analytic examination of team size and team effectiveness*, Apr. 2023. [Online]. Available: `https://doi.org/10.1002/job.2708`.

[2] D. Rodríguez, M. Á. Sicilia, E. A. García, and R. Harrison, *Empirical findings on team size and productivity in software development*, Mar. 2012. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0164121211002366`.

[3] *Diffusion of responsibility*, Jan. 2016. [Online]. Available: `https://psychology.iresearchnet.com/social-psychology/group/diffusion-of-responsibility/`.

[4] *Diffusion of responsibility*, Dec. 2013. [Online]. Available: `https://web.archive.org/web/20131230093304/http://en.wikipedia.org/wiki/Diffusion_of_responsibility`.

[5] N. Gamero, B. González-Anta, V. Orengo, A. Zornoza, and V. Peñarroja, *Is team emotional composition essential for virtual team members' well-being? the role of a team emotional management intervention*, Apr. 2021. [Online]. Available: `https://doi.org/10.3390/ijerph18094544`.

[6] F. Wang, W. Liu, C.-D. Ling, P. Fan, and Y. Chen, *Combating team hopelessness: How and why leader interpersonal emotion management matters*, Apr. 2022. [Online]. Available: `https://doi.org/10.1111/peps.12508`.

[7] T. Chamorro-Premuzic, *Does diversity actually increase creativity?*, Jun. 2017. [Online]. Available: `https://hbr.org/2017/06/does-diversity-actually-increase-creativity`.

[8] *Why diverse teams are smarter*, Nov. 2016. [Online]. Available: `https://hbr.org/2016/11/why-diverse-teams-are-smarter`.

[9] C. S. Nam, J. B. Lyons, H. S. Hwang, and S. Kim, *The process of team communication in multi-cultural contexts: An empirical study using bales' interaction process analysis (ipa)*, Sep. 2009. [Online]. Available: `https://doi.org/10.1016/j.ergon.2009.03.004`.

[10] M. Hoegl, "Smaller teams–better teamwork: How to keep project teams small", *Business horizons*, vol. 48, no. 3, pp. 209–214, May 2005. DOI: `10.1016/j.bushor.2004.10.013`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0007681304001120`.

[11] D. Rodríguez, M. Á. Sicilia, E. A. García, and R. Harrison, "Empirical findings on team size and productivity in software development", *Journal of systems and software/The Journal of systems and software*, vol. 85, no. 3, pp. 562–570, Mar. 2012. DOI: 10.1016/j.jss.2011.09.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121211002366.

[12] L. Fried, "Team size and productivity in systems development bigger does not always mean better", *Journal of information systems management/The Journal of information systems management*, vol. 8, no. 3, pp. 27–35, Jan. 1991. DOI: 10.1080/07399019108964994. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/07399019108964994.

[13] R. Sethi, D. C. Smith, and C. W. Park, *Cross-functional product development teams, creativity, and the innovativeness of new consumer products*, Feb. 2001. [Online]. Available: https://doi.org/10.1509/jmkr.38.1.73.18833.

[14] J. C. Bradbury, *Determinants of revenue in sports leagues: An empirical assessment*, Sep. 2018. [Online]. Available: https://doi.org/10.1111/ecin.12710.

[15] N. M. Ashkanasy and A. D. Dorris, *Emotions in the workplace*, Mar. 2017. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-orgpsych-032516-113231.

[16] C. K. W. D. Dreu and L. R. Weingart, "Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis.", *Journal of applied psychology*, vol. 88, no. 4, pp. 741–749, Jan. 2003. DOI: 10.1037/0021-9010.88.4.741. [Online]. Available: https://psycnet.apa.org/doiLanding?doi=10.1037%2F0021-9010.88.4.741.

[17] T. K. I. Adham, *Conflict resolution in team: Analyzing the of conflicts and best skills for resolution*, Aug. 2023. DOI: 10.36347/sjet.2023.v11i08.001. [Online]. Available: https://doi.org/10.36347/sjet.2023.v11i08.001.

[18] P. A. Gloor, *Swarm creativity: Competitive advantage through collaborative innovation networks*, Jan. 2006. [Online]. Available: http://oplaunch.com/swarm_creativity.pdf.

[19] I. Yang, *When team members meet in a new team: An exploration of team development*, Nov. 2013. [Online]. Available: https://doi.org/10.3233/hsm-130794.

[20] M. T. Khan and S. Khalid, "Sentiment analysis for health care", *International journal of privacy and health information management*, vol. 3, no. 2, pp. 78–91, Jul. 2015. DOI: 10.4018/ijphim.2015070105. [Online]. Available: https://www.igi-global.com/gateway/article/142225.

[21] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos, *Densereg: Fully convolutional dense shape regression in-the-wild*, 2017. arXiv: 1612.01202 [cs.CV].

[22] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, *Joint 3d face reconstruction and dense alignment with position map regression network*, 2018. arXiv: 1803.07835 [cs.CV].

[23] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, *Wing loss for robust facial landmark localisation with convolutional neural networks*, 2018. arXiv: 1711.06753 [cs.CV].

[24] J. Wang, Y. Yuan, and G. Yu, *Face attention network: An effective face detector for the occluded faces*, 2017. arXiv: `1711.07246 [cs.CV]`.

[25] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis, *Ssh: Single stage headless face detector*, 2017. arXiv: `1708.03979 [cs.CV]`.

[26] S. Yang, P. Luo, C. C. Loy, and X. Tang, *Wider face: A face detection benchmark*, 2015. arXiv: `1511.06523 [cs.CV]`.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection*, 2018. arXiv: `1708.02002 [cs.CV]`.

[28] R. Girshick, *Fast r-cnn*, 2015. arXiv: `1504.08083 [cs.CV]`.

[29] X. Jiang, T. Gao, Z. Zhu, and Y. Zhao, “Real-time face mask detection method based on yolov3”, *Electronics*, vol. 10, p. 837, Apr. 2021. DOI: `10.3390/electronics10070837`.

[30] S. Chen, Y. Liu, X. Gao, and Z. Han, *Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices*, 2018. arXiv: `1804.07573 [cs.CV]`.

[31] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, *Ms-celeb-1m: A dataset and benchmark for large-scale face recognition*, 2016. arXiv: `1607.08221 [cs.CV]`.

[32] J. W. Davis, C. Menart, M. Akbar, and R. Ilin, *A classification refinement strategy for semantic segmentation*, 2018. arXiv: `1801.07674 [cs.CV]`.

[33] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, ISSN: 1939-3539. DOI: `10.1109/tpami.2021.3087709`. [Online]. Available: `http://dx.doi.org/10.1109/TPAMI.2021.3087709`.

[34] G. Pleiss, D. Chen, G. Huang, T. Li, L. van der Maaten, and K. Q. Weinberger, *Memory-efficient implementation of densenets*, 2017. arXiv: `1707.06990 [cs.CV]`.

[35] X. Zhang, X. Zhou, M. Lin, and J. Sun, *Shufflenet: An extremely efficient convolutional neural network for mobile devices*, 2017. arXiv: `1707.01083 [cs.CV]`.

[36] L. SIfre and S. Mallat, *Rigid-motion scattering for texture classification*, 2014. arXiv: `1403.1687 [cs.CV]`.

[37] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, 2016. arXiv: `1603.04467 [cs.DC]`.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *Rethinking the inception architecture for computer vision*, 2015. arXiv: `1512.00567 [cs.CV]`.

[39] T. Weyand, I. Kostrikov, and J. Philbin, "Planet - photo geolocation with convolutional neural networks", in *Lecture Notes in Computer Science*. Springer International Publishing, 2016, pp. 37–55, ISBN: 9783319464848. DOI: 10.1007/978-3-319-46484-8_3. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46484-8_3.

[40] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: 1409.1556 [cs.CV].

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, 2014. arXiv: 1409.4842 [cs.CV].

[42] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, 2017. arXiv: 1704.04861 [cs.CV].

[43] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, *Img2pose: Face alignment and detection via 6dof, face pose estimation*, 2021. arXiv: 2012.07791 [cs.CV].

[44] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild", Jun. 2020, pp. 5202–5211. DOI: 10.1109/CVPR42600.2020.00525.

[45] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].

[46] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge", Dec. 2013, pp. 397–403. DOI: 10.1109/ICCVW.2013.59.

[47] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, Jan. 2019, ISSN: 1939-3539. DOI: 10.1109/tpami.2017.2778152. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2017.2778152.

[48] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Gool, "Random forests for real time 3d face analysis", *Int. J. Comput. Vision*, vol. 101, no. 3, pp. 437–458, Feb. 2013, ISSN: 0920-5691. DOI: 10.1007/s11263-012-0549-0. [Online]. Available: https://doi.org/10.1007/s11263-012-0549-0.

[49] Y. Fan, J. C. K. Lam, and V. O. K. Li, *Multi-region ensemble convolutional neural network for facial expression recognition*, 2018. arXiv: 1807.10575 [cs.CV].

[50] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, *Challenges in representation learning: A report on three machine learning contests*, 2013. arXiv: 1307.0414 [stat.ML].

[51] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–I. DOI: `10.1109/CVPR.2001.990517`.

[52] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild", *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019. DOI: `10.1109/TAFFC.2017.2740923`.

[53] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450–6458. DOI: `10.1109/CVPR.2017.683`.

[54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. DOI: `10.1109/CVPR.2017.243`.

[55] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module", in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

[56] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering", *CoRR*, vol. abs/1503.03832, 2015. arXiv: `1503.03832`. [Online]. Available: `http://arxiv.org/abs/1503.03832`.

[57] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments", University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.

[58] G. B. H. E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures", University of Massachusetts, Amherst, Tech. Rep. UM-CS-2014-003, May 2014.

[59] J. H. Cheong, E. Jolly, T. Xie, S. Byrne, M. Kenney, and L. J. Chang, *Py-feat: Python facial expression analysis toolbox*, 2023. arXiv: `2104.03509 [cs.CV]`.

[60] K. Khabarlak and L. Koriashkina, "Fast facial landmark detection and applications: A survey", *Springer Science+Business Media*, vol. 22, no. 1, e02–e02, Apr. 2022. DOI: `10.24215/16666038.22.e02`. [Online]. Available: `https://arxiv.org/abs/2101.10808v2`.

[61] P. Ekman and W. V. Friesen, "Facial action coding system", *PsycTESTS Dataset*, Jan. 1978. [Online]. Available: `https://doi.org/10.1037/t27734-000`.

[62] P. Ekman and W. Friesen, "Facial action coding system: Manual", Jan. 1998. [Online]. Available: `https://www.semanticscholar.org/paper/Facial-Action-Coding-System%3A-Manual-Ekman-Friesen/161130a1ed058e920fb36be69726ccfe21a93c2c`.

[63] J. H. Cheong, T. Xie, S. Byrne, and L. J. Chang, "Py-feat: Python facial expression analysis toolbox", *Affective Science*, vol. 4, no. 4, pp. 781–796, Aug. 2023. DOI: `https://doi.org/10.1007/s42761-023-00191-4`. [Online]. Available: `https://doi.org/10.1007/s42761-023-00191-4`.

[64] A. Ortony and T. J. Turner, "What's basic about basic emotions?", *Psychological Review*, vol. 97, no. 3, pp. 315–331, Jan. 1990. [Online]. Available: `https://doi.org/10.1037/0033-295x.97.3.315`.

[65] P. Ekman, "Are there basic emotions?", *Psychological Review*, vol. 99, no. 3, pp. 550–553, Jan. 1992. [Online]. Available: `https://doi.org/10.1037/0033-295x.99.3.550`.

[66] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, *A review of speaker diarization: Recent advances with deep learning*, 2021. arXiv: `2101.09624 [eess.AS]`.

[67] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv: `2212.04356 [eess.AS]`.

[68] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization", in *Proc. INTERSPEECH 2023*, 2023.

[69] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe", in *Proc. INTERSPEECH 2023*, 2023.

[70] I. Mccowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska Masson, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus", *Int'l. Conf. on Methods and Techniques in Behavioral Research*, Jan. 2005.