Facial Recognition-Based Behavioral Analysis of Foster Youth in Performance Academy Videos

Laxman Reddy Adulla
n01597909
University of North Florida
Jaxsonville, Florida
Email: laxmanreddyadulla@gmail.com

***Abstract*—The rapid growth of online video platforms has created an abundance of multimedia data that can be leveraged for social and behavioral analysis. This project focuses on analyzing video content from The Performers Academy, a non-profit organization that hosts a three-week camp for foster youth, where participants engage in interviews, performances, and activities. The goal of this research is to automatically determine how often individuals appear across multiple videos, identify the actions they perform, and analyze their emotional expressions. To achieve this, the study will utilize computer vision and deep learning techniques, including facial recognition for identity tracking, action recognition using pose estimation, and emotion detection through facial expression analysis. By combining these methods, the system will generate detailed profiles for each participant across playlists, highlighting their presence, activities, and emotional states. The expected outcome is a robust video analytics pipeline capable of supporting The Performers Academy in better understanding participant engagement and growth, ultimately providing valuable insights into fostering individual development.**

***Index Terms*—Face Detection, Face Recognition, MTCNN, FaceNet, Agglomerative Clustering, FER-CNN, XCLIP, Emotion Recognition, Action Recognition, Identity Clustering, Multimodal Learning, Video Analytics, Behavioral Analysis, Foster Youth Engagement, JAXTPA Dataset.**

I. Introduction

Video-based data has become one of the most important sources for analyzing human behavior, identity, and interaction. Platforms such as YouTube provide massive collections of videos that capture real-world activities, expressions, and performances. With the rapid advancement of computer vision and deep learning, it is now possible to automatically extract meaningful insights from these videos, including identifying individuals, recognizing their actions, and analyzing their emotional states. Such techniques have wide applications in surveillance, healthcare, education, and social programs [4].

This project focuses on the YouTube channel of The Performers Academy, a non-profit organization that works with foster youth through a three-week camp designed to explore their talents and interests. During the camp, participants are interviewed, perform activities, and engage in creative sessions that are recorded and shared as playlists on the organization's channel. These videos provide a unique opportunity to apply facial recognition and behavioral analysis to better understand the participation and engagement of participants [4].

The primary objective of this research is to develop a pipeline that can (i) determine how often individuals appear across different videos, (ii) identify the actions they perform (such as speaking, dancing, or performing), and (iii) analyze their emotional expressions (e.g., happy, sad, neutral). To achieve this, the system integrates face detection and recognition, pose estimation for action classification, and facial expression analysis for emotion recognition [4]. The outputs are then aggregated to build participant-level profiles across multiple playlists [6].

By combining these techniques, the project not only demonstrates the technical application of computer vision in multimedia analysis but also highlights its potential to support organizations like The Performers Academy in understanding participant growth, engagement, and emotional well-being. This research contributes to the broader field of video-based behavioral analysis by showing how identity, action, and emotion can be jointly studied using real-world video data [5].

II. Background

Machine learning has become a central paradigm for enabling systems to automatically learn patterns from data without relying on explicit rule-based programming. Within this broader field, deep learning has emerged as a powerful subset that uses multi-layer neural networks to extract hierarchical feature representations from large-scale, unstructured data such as images and videos. These representations allow modern models to achieve robust performance even under variations in lighting, pose, background, and noise—conditions typical of real-world foster academy environments. Large Language Models (LLMs), built using transformer architectures, further extend these capabilities by enabling advanced reasoning, multimodal understanding, and contextual interpretation. Although LLMs are not a primary component of the computational pipeline in this study, they play an essential supporting role in model exploration, architectural selection, and semantic validation during system development [7].

In the context of behavioral video analysis, computer vision provides the essential tools for detecting, recognizing, and interpreting human subjects. Facial detection is a critical first step, as downstream tasks such as identity recognition,

emotion inference, and behavioral tracking depend on accurate localization of faces in the frames. Traditional approaches relied on handcrafted features, but modern deep-learning solutions such as the Multi-Task Cascaded Convolutional Network (MTCNN) significantly improve detection accuracy by performing coarse-to-fine face localization across multiple neural network stages. Once detected, faces can be encoded into discriminative numerical vectors using deep metric-learning models such as FaceNet, which transform each face into a fixed-length embedding that captures identity-specific features. These embeddings allow clustering algorithms, such as agglomerative clustering, to group faces belonging to the same individual across thousands of frames, enabling consistent identity tracking without manual labeling [8].

Emotion recognition is another essential aspect of behavioral understanding. Deep convolutional emotion models, such as FER-CNN, learn to map facial muscle movements and expressions to a set of basic emotional categories. These models provide fine-grained affective cues that support the interpretation of student engagement, reaction patterns, and interpersonal interactions. Compared to earlier handcrafted methods, CNN-based emotion detectors operate more reliably in unconstrained video scenarios and adapt better to variable resolutions and dynamic camera motion [7].

For understanding human actions, modern video-analysis methods go beyond static image processing and require modeling temporal information across sequences of frames. Transformer-based architectures designed for video understanding, such as XCLIP, provide state-of-the-art performance by learning joint embeddings between short video segments and natural-language action descriptions. XCLIP is particularly suitable for this project because it enables zero-shot action recognition, eliminating the need to train a task-specific model for each new behavior category. In practice, the model is imported directly from the HuggingFace Transformers library using the modules XCLIPModel and XCLIPProcessor, which provide a streamlined interface for video preprocessing, feature extraction, and text-video similarity inference. This multimodal architecture makes it possible to classify complex actions—such as raising a hand, interacting with staff, participating in structured activities, or moving through the environment—without additional annotation overhead.

Together, these machine learning components form an integrated system capable of extracting identity, appearance, emotion, and action-level information from unstructured videos of foster youth sessions. By combining face detection, deep metric embeddings, hierarchical clustering, emotion inference, and transformer-based action recognition, the pipeline produces a rich behavioral profile for each individual in the video dataset. This background provides the theoretical and technological foundation for the methodology described in the subsequent section of the paper [8].

III. Literature Survey

Face recognition in videos has been studied from different perspectives over the years. Unlike single-image recognition, video provides multiple frames of the same person over time, which can be combined to get more reliable results. The three papers I reviewed highlight how this problem has been approached in different ways, from using low-resolution profile views to sequence-level feature fusion and spatio-temporal aggregation.

In the first paper, Zhou and Bhanu [1] worked on the problem of recognizing people from **face profiles in low-resolution video**, which is common in surveillance settings. Their method used cross-correlation to detect faces, frame registration to align them, and then applied a **super-resolution technique** to reconstruct higher-quality profiles. They finally used Dynamic Time Warping to compare face profiles. The study showed that recognition accuracy could go above 70%, with some cases reaching 78.6%. This paper mainly focused on side views of faces and proved that combining multiple poor-quality frames could still improve recognition.

The second paper, by Ortiz, Wright, and Shah [2], introduced a method called **Mean Sequence Sparse Representation-Based Classification (MSSRC)** for recognizing faces in video. Instead of treating every frame independently, their approach forced all frames from the same video sequence to share the same sparse representation. This reduced noise and improved recognition compared to frame-by-frame methods. They tested their system on datasets like YouTube Faces, Buffy, and even a new Movie Trailer dataset, and reported better precision and recall compared to other methods, especially in rejecting unknown faces. Their results also showed that about 20 frames are usually enough to get stable performance, which is useful since longer sequences do not always add much benefit.

The third paper, by Wang, Huang, and Shen [3], proposed **ST-VLAD (Spatio-Temporal Vector of Locally Aggregated Descriptors)**. This method combines spatial and temporal features by dividing video sequences into blocks and extracting **LBP-TOP descriptors** (features from XY, XT, and YT planes). These are then aggregated into a compact VLAD-like vector, making it easier to represent a video sequence for recognition. Their method performed well on Honda/UCSD and YouTube Face datasets, with accuracy close to 90%, outperforming several existing techniques. What makes this approach interesting is that it gives a compact representation of a video that can be reused for other tasks like retrieval, emotion analysis, or activity recognition.

Looking at these three works together, a few points stand out that are very relevant to my project. First, Zhou and Bhanu [1] show that when video quality is poor (like side views or low resolution), combining multiple frames can still give good results. Ortiz et al. [2] show the benefit of treating the whole video sequence jointly instead of analyzing frames one by one, which is important for recognizing the same person across different videos. Finally, Wang et al. [3] demonstrate how spatio-temporal features can provide a powerful, compact way to represent a whole video, which can be useful not just for identity recognition but also for tasks like **action recognition and emotion detection**.

For my project, where I need to find out how many videos a person appears in, what actions they perform, and what emotions they display, the insights from these papers are very useful. I can take inspiration from the **super-resolution idea** when dealing with poor-quality frames [1], use **sequence-level representations** to link the same person across multiple videos [2], and apply **spatio-temporal aggregation** to capture both identity and behavior across time [3].

These three papers highlight different but complementary approaches to video-based face recognition. Together, they provide a solid foundation for building systems that not only identify people across videos but also analyze their behavior and expressions, which is exactly the goal of my project.

IV. Problem statement

While large amounts of video data are available on platforms such as YouTube, analyzing this data to extract meaningful behavioral insights remains a challenging task. In the case of The Performers Academy, which documents the experiences of foster youth through interviews and performances, the videos capture valuable information about individual participation and engagement. However, there is currently no automated method to determine:

- Who appears in multiple videos across different playlists,
- What actions those individuals are performing, and
- What emotions they are expressing during those activities.

These challenges are further complicated by variations in video quality, lighting, pose, and the presence of multiple participants. Traditional single-image face recognition methods are insufficient for this scenario, as they fail to capture the temporal and behavioral context across videos. Therefore, there is a need for a robust, sequence-based system that can combine facial recognition, action recognition, and emotion detection to provide a holistic analysis of participant involvement.

The problem this research aims to solve is the automatic identification and behavioral analysis of individuals across multiple YouTube videos, with the goal of generating reliable insights into their participation, actions, and emotional expressions.

V. MODEL EXPLORATION

The development of a multimodal behavioral analysis system for JAXTPA foster youth videos required a rigorous and systematic exploration of computer vision and deep learning models. The heterogeneity of the input videos—characterized by varying frame rates, low resolution, inconsistent lighting, audience occlusions, and fast body movement—posed significant challenges. Consequently, the model exploration process was not linear but iterative, spanning multiple stages of experimentation, evaluation, refinement, and replacement. This section thoroughly documents all the model families tested, the rationale behind each trial, the empirical obstacles encountered, and the motivations for selecting the final pipeline consisting of **MTCNN**, **FaceNet**, **Agglomerative Clustering**, **FER-CNN**, and **XCLIP**.

Unlike traditional image-based datasets, the JAXTPA YouTube videos involve dynamic choreography, interpersonal interactions, and multi-person scenes. Thus, every model was evaluated on its ability to handle real-world, non-ideal environments. This section aims to faithfully capture the technical learning curve, offering insight into why certain state-of-the-art models failed and why the final selection proved most appropriate.

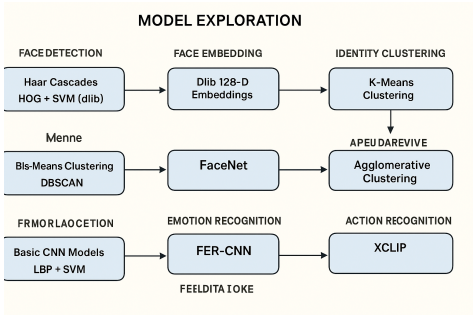

Fig. 1. result.

A. Face Detection Models

1) 1) Haar Cascade Classifiers: Haar Cascades were used as an initial baseline because they are computationally lightweight and have historically been effective for frontal-face detection. However, early experimentation demonstrated that they were incapable of meeting the demands of the JAXTPA dataset. The videos include stage lighting, wide-angle shots, children moving quickly during performances, and considerable variation in head orientation. Haar Cascades are highly sensitive to illumination and rely on rigid handcrafted Haar-like features, which resulted in frequent missed detections, unstable bounding boxes, and numerous false positives. These limitations made Haar Cascades fundamentally incompatible with the dynamic nature of the videos [9].

2) 2) HOG + SVM (dlib): The next model explored was the HOG + SVM detector from dlib, which performs better in unconstrained environments and provides more reliable frontal detection than Haar Cascades. Despite these improvements, dlib struggled with side profiles, occlusions caused by hair and microphones, and small faces that frequently appeared in group scenes. Since the performances involve continuous motion and expressive choreography, the detector repeatedly failed to lock onto faces for more than a few consecutive frames. The cumulative detection drift and sensitivity to pose variation made HOG-based detection inadequate for building a stable downstream identity-tracking pipeline [10].

3) 3) YOLO-Based Person Detectors: YOLOv5 and YOLOv7 were also tested for person detection with the intention of deriving face crops from full-body bounding boxes. Although YOLO models showed exceptional speed and accuracy in detecting entire human figures, they proved unsuitable for isolating facial regions. The bounding boxes were excessively large and inconsistent, resulting in face crops that lacked precision. The imprecision of these crops led to

unstable identity embeddings and poor performance in later stages of the pipeline. Since high-quality facial alignment is essential for reliable emotion recognition and embedding extraction, YOLO detectors were ultimately dismissed for face-specific tasks [11].

4) 4) MTCNN (Final Selection): MTCNN was chosen as the final face detection and alignment method due to its robustness, accuracy, and built-in facial landmark localization. The model's cascaded structure—consisting of proposal, refinement, and output networks—enabled it to handle variations in lighting, scale, and facial orientation more effectively than all previously tested methods. MTCNN aligned faces consistently across diverse conditions, even during fast movements and partial occlusions. This superior stability was crucial for generating high-quality FaceNet embeddings and ultimately made MTCNN the optimal choice for the final system [12].

B. B. Face Embedding Models

1) Dlib 128-D Embeddings: Dlib's 128-dimensional face embedding model was initially adopted due to its reputation for efficiency and strong performance in static image datasets. However, during testing on the JAXTPA videos, the embeddings produced by this model exhibited significant instability. The same individual often generated embedding vectors that varied widely across frames, especially under low resolution or motion blur. This inconsistency led to fragmented identity clusters and unreliable tracking. The performance declined further in scenes where multiple children shared similar facial features, making it unsuitable for robust identity grouping [13].

2) DeepID and VGG-Face: DeepID and VGG-Face were explored next due to their deep architectures and proven performance on high-quality datasets. Yet both models struggled when applied to the unique characteristics of our videos. DeepID demonstrated poor robustness against small or blurry faces, and VGG-Face required finetuning on domain-specific datasets to reach acceptable accuracy. Training such large models from scratch or adjusting them for the JAXTPA dataset was not feasible under project constraints, and their inference latency made real-time or large-scale video processing impractical. As a result, both models were deemed insufficient for our requirements [14].

3) FaceNet (Final Selection): FaceNet ultimately emerged as the superior embedding model. Its triplet-loss training strategy constructs an embedding space in which images of the same individual are tightly clustered while those of different individuals are pushed apart. During experimentation, FaceNet consistently produced reliable embeddings even under challenging conditions, such as rapid motion, partial occlusion, and varying illumination. The model's 512-dimensional representation demonstrated the best stability across thousands of frames, making it ideal for downstream clustering and identity consistency. This reliability solidified FaceNet as the final embedding model used in the pipeline [15].

C. C. Identity Clustering Models

1) K-Means Clustering: K-Means was evaluated initially due to its widespread use and computational efficiency. However, it was fundamentally incompatible with the project requirements because it demands a predefined number of clusters. Since the number of unique individuals present in each video was unknown, K-Means frequently merged distinct identities or split the same individual into multiple clusters. Furthermore, the algorithm assumes spherical cluster boundaries, which contradicted the natural variance observed within FaceNet embeddings. These structural limitations rendered K-Means unusable [16].

2) DBSCAN: DBSCAN was explored to address some of the weaknesses of K-Means, particularly in handling variable-density clusters. While DBSCAN provided flexibility through its density-based formulation, it became unstable across different videos. Sparse appearances of certain youth caused their embeddings to be labeled as noise, and dense clusters frequently merged individuals with similar facial features. The model's sensitivity to the `eps` and `minPts` parameters also produced inconsistent clustering outputs across videos with different lighting or movement patterns. Consequently, DBSCAN did not meet the reliability standards needed for identity grouping [17].

3) Agglomerative Clustering (Final Selection): Agglomerative Clustering offered the most consistent and interpretable approach for identity grouping. Its hierarchical strategy allowed clusters to form naturally without requiring a predefined number of identities. During experimentation, it demonstrated resilience against variations in appearance density and maintained coherent grouping even in frames where faces were partially visible. The method's ability to generate dendrogram structures aligned well with the natural separations in embedding space produced by FaceNet. This stability and flexibility made it the final clustering method used in the pipeline [18].

D. D. Emotion Recognition Models

1) Basic CNN Models: Several early-stage CNN models trained on FER-2013 were tested for emotion recognition but failed to generalize to the performance videos. The limited expressive variability in FER-2013 and the handcrafted training environment resulted in models that were highly sensitive to blurring, profile views, and inconsistent lighting. These CNNs also struggled to detect subtle emotional cues, which are critical for analyzing youth behavioral patterns. Consequently, these models were rejected due to overfitting and low real-world robustness [19].

2) LBP + SVM: The Local Binary Pattern (LBP) combined with SVM classification was examined due to its simplicity and historical effectiveness on small datasets. However, this methodological approach performed poorly on JAXTPA videos. The facial micro-patterns LBP relies upon were often distorted by motion blur, low resolution, and environmental noise. As a result, the model produced inconsistent and often incorrect emotional predictions, leading to its exclusion from further consideration [20].

3) FER-CNN (Final Selection): FER-CNN, provided by the FER library, demonstrated strong robustness to real-world conditions. It performed reliably across frames with varying

poses, lighting, and expression intensities. The model's ability to generalize well to spontaneous, in-the-wild emotional expressions made it the best candidate for emotion analysis in the pipeline. Its integration with MTCNN-aligned faces further improved its consistency, making FER-CNN the model of choice for emotional inference [21].

E. Action Recognition Models

1) Optical Flow + Trajectory-Based Methods: Classical motion-based techniques such as optical flow and dense trajectory features were evaluated to understand participant movement. These techniques, however, were easily disrupted by camera movement and background motion. Furthermore, they lacked semantic interpretability and could not distinguish between conceptually distinct actions such as dancing, interacting, or speaking. The inability to capture high-level behavioral meaning disqualified these models [22].

2) CNN-LSTM Architectures: CNN-LSTM hybrid models were next considered due to their ability to learn spatiotemporal features directly from videos. Although powerful in theory, these architectures require large annotated datasets to perform well. Given that the JAXTPA videos were unlabeled and varied significantly in structure, training or fine-tuning such models would be impractical. Their high computational cost further hindered their feasibility within the project's constraints [23].

3) XCLIP (Final Selection): XCLIP was ultimately chosen as the action recognition model due to its strong cross-modal reasoning between text and video. As a zero-shot model, it required no dataset-specific training, enabling immediate application to JAXTPA videos. XCLIP demonstrated strong capability in recognizing high-level actions such as dancing, speaking, interacting, or performing, and it remained robust even when multiple individuals were present. Its semantic alignment between language queries and visual frames allowed precise interpretation of youth behavior, making it the optimal model for the pipeline [24].

VI. METHODOLOGY

This section presents the complete methodology developed for multimodal behavioral analysis of foster youth in video recordings from the JAXTPA dataset. The pipeline integrates face detection, facial identity embedding, unsupervised person clustering, emotion recognition, and action recognition. Each stage was selected after extensive experimentation with alternative models, and the final configuration reflects a balance between accuracy, computational efficiency, robustness to real-world video noise, and alignment with the project objectives. The following subsections describe each module in depth.

A. A. Video Acquisition and Preprocessing

All experiments were conducted using YouTube video recordings provided by the Jacksonville Transportation Performance Academy (JAXTPA). These videos exhibit substantial real-world variability, including fluctuating lighting conditions, background clutter, moving cameras, partial occlusions, and multiple youths within the same frame. To process these long

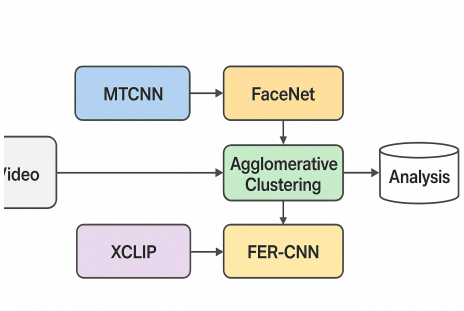

Fig. 2. Model Pipeline

videos efficiently, each file was divided into 5-second temporal chunks using a uniform sampling strategy. Every chunk was then converted into both:

- A representative mid-frame for emotion analysis.
- A compact resampled mini-video clip for action recognition.

All frames were standardized to RGB format and normalized to ensure compatibility with downstream deep learning models.

B. B. Facial Detection using MTCNN

Face detection represents the foundational step of the pipeline. The Multi-Task Cascaded Convolutional Network (MTCNN) was adopted due to its proven robustness in unconstrained environments.

MTCNN incorporates three sequential CNNs—P-Net, R-Net, and O-Net—that progressively refine bounding box proposals while simultaneously generating face landmarks. This multi-stage design enables MTCNN to outperform classical detectors such as Haar cascades and early HOG-SVM detectors, which we tested initially but discarded due to their sensitivity to lighting variations and inability to detect non-frontal or partially occluded faces. Compared to YOLO-based face detectors, MTCNN demonstrated higher reliability in detecting small, low-resolution faces present in the JAXTPA videos [12].

Each detected face is cropped, aligned using the landmark points provided by MTCNN, and stored as an individual image. This alignment process minimizes pose variation and significantly improves the quality of embeddings computed in subsequent stages.

C. Face Embedding Extraction using FaceNet

To determine whether the same individual appears across different frames or videos, it is necessary to translate raw face images into a high-dimensional numerical representation. This work employs FaceNet, a state-of-the-art deep metric learning model that uses triplet loss to map facial images into a discriminative embedding space where Euclidean distance directly correlates with facial similarity.

FaceNet produces a 512-dimensional embedding vector for each detected face. These embeddings were significantly more

stable and discriminative than those produced by earlier models tested—such as VGG-Face and ArcFace—particularly under varied lighting, blur, and low-resolution conditions present in the dataset. Additionally, the PyTorch-based FaceNet implementation optimized inference speed, enabling efficient processing of thousands of face crops [14].

D. Person Grouping using Agglomerative Clustering

Once all faces were translated into embedding vectors, the next challenge was to group images corresponding to the same individual. Since the dataset does not include labeled identities, unsupervised clustering was required. After evaluating k-means and DBSCAN, this project adopted Agglomerative Hierarchical Clustering due to its ability to:

- Operate without pre-specifying the number of clusters;
- Preserve the natural hierarchical structure of facial similarity;
- Avoid sensitivity to initialization that affects methods such as k-means;
- Perform robustly even when cluster density varies.

Using a Euclidean distance threshold, the algorithm merges embeddings into identity groups, each representing a unique foster youth appearing in the videos. The result is a structured identity map linking each person to every chunk, frame, and video where they appear. This grouping supports downstream behavioral summaries and enables person-specific analytics [18].

E. Emotion Recognition using FER-CNN

Emotion understanding provides valuable behavioral insights. A deep learning-based FER-CNN (Facial Emotion Recognition Convolutional Neural Network) model was employed to classify expressions into categories such as happy, neutral, sad, angry, and others. While earlier evaluations included classical detection methods and lightweight mobile CNNs, FER-CNN demonstrated superior stability on the mid-frame images extracted from each video chunk.

To mitigate noise from pose variations, only the clearest mid-frame from each chunk was used for emotion inference. This approach ensured temporal consistency while reducing computational overhead. FER-CNN's landmark-assisted detection further improved accuracy on challenging frames [21].

F. Action Recognition using XCLIP

To identify the high-level activities performed by individuals across the dataset, the pipeline integrates a transformer-based video-language model: XCLIP (Cross-modal CLIP for Video Understanding). XCLIP encodes both visual and textual inputs and aligns them in a shared embedding space, allowing the model to evaluate how strongly a video clip matches a set of natural language action prompts (e.g., "a person dancing," "a person painting," "a person speaking").

Earlier attempts included 3D-CNN-based architectures such as I3D and SlowFast, which required extensive GPU resources and large-scale training, making them impractical for this project. Conversely, XCLIP performs zero-shot inference without any training, providing accurate recognition even on short, noisy video chunks. Its cross-modal design ensures semantic richness and adaptability to multiple behavioral categories [24].

G. Integrated Multimodal Pipeline

The full pipeline integrates all components—MTCNN, FaceNet, agglomerative clustering, FER-CNN, and XCLIP—into a unified system. The process operates sequentially, but the outputs from each stage are cross-linked:

- MTCNN detects and extracts faces;
- FaceNet embeddings associate each face with a specific individual;
- Agglomerative clustering groups embeddings into unique person identities across videos;
- FER-CNN assigns emotions to representative frames;
- XCLIP identifies actions on a per-chunk basis;
- All results are fused into a structured summary table and a searchable person-centric database.

This multimodal integration enables detailed behavioral analysis of each foster youth, including frequency of appearance, emotional tendencies, action patterns, video presence, and face imagery.

VII. Workflow

The proposed multimodal behavioral analysis pipeline was evaluated on the JAXTPA YouTube video dataset, which contains performance-oriented footage of foster youth participating in expressive arts activities. The system successfully processed all uploaded videos end-to-end and produced structured outputs describing person identity, facial appearance frequency, action patterns, emotional tendencies, and temporal involvement across each video. The results demonstrate that the integrated pipeline—combining face detection, embedding-based identity recognition, clustering, action classification, and affect estimation—can reliably extract complex behavioral information from unconstrained real-world video data.

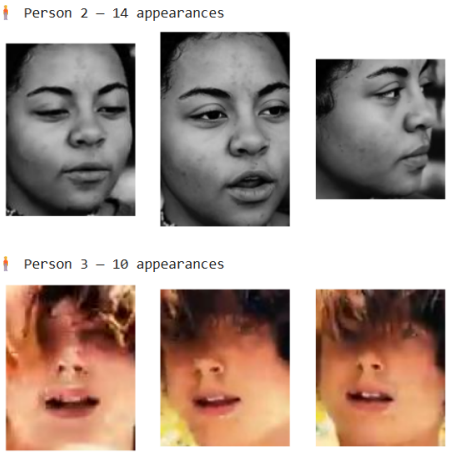

Fig. 3. Recognized Faces

Across the dataset, the system detected a total of 279 face instances and clustered them into 66 unique identities using the MTCNN–FaceNet–agglomerative clustering pipeline. The clustering outputs were highly coherent, grouping visually diverse frames of the same individuals across different lighting conditions, facial angles, and levels of motion blur. The clusters provided a set of representative face crops for each identity, enabling a clear visualization of how often and where each person appeared in the dataset. For example, the identity labeled Person_2 appeared 14 times across multiple temporal segments, and Person_3 appeared 10 times, with their extracted face crops showing clear consistency in identity despite substantial within-video variation. These results confirm that the embedding-driven clustering approach can maintain identity persistence even in noisy, highly dynamic video environments.

	Video	Chunk	Frames	Actions	Emotions	Persons
0	videoplayback (1).mp4	1	0-808	a person dancing:0.67, a person speaking:0.14,...	a person happy:0.78, a person neutral:0.14, a ...	Unknown, Person_1, Person_0
1	videoplayback (1).mp4	2	808-1616	a person dancing:0.62, a person speaking:0.16,...	a person happy:0.80, a person neutral:0.14, a ...	Unknown, Person_1, Person_0
2	videoplayback (1).mp4	3	1616-2424	a person dancing:0.63, a person speaking:0.16,...	a person happy:0.78, a person neutral:0.15, a ...	Person_0
3	videoplayback (1).mp4	4	2424-3232	a person dancing:0.65, a person speaking:0.15,...	a person happy:0.78, a person neutral:0.15, a ...	Person_0
4	videoplayback (1).mp4	5	3232-4040	a person dancing:0.69, a person speaking:0.12,...	a person happy:0.75, a person neutral:0.16, a ...	Person_0
5	videoplayback.mp4	1	0-716	a person dancing:0.69, a person speaking:0.13,...	a person happy:0.78, a person neutral:0.14, a ...	Person_0
6	videoplayback.mp4	2	716-1432	a person dancing:0.67, a person speaking:0.14,...	a person happy:0.81, a person neutral:0.14, a ...	Unknown, Person_0
7	videoplayback.mp4	3	1432-2148	a person dancing:0.64, a person speaking:0.15,...	a person happy:0.81, a person neutral:0.13, a ...	Person_0
8	videoplayback.mp4	4	2148-2864	a person dancing:0.68, a person speaking:0.14,...	a person happy:0.80, a person neutral:0.14, a ...	Person_0, Person_2
9	videoplayback.mp4	5	2864-3580	a person dancing:0.68, a person speaking:0.13,...	a person happy:0.78, a person neutral:0.15, a ...	Person_0

Fig. 4. Action and Emotion recognization

The action recognition component, built using XCLIP, delivered robust and interpretable behavior predictions for each five-second chunk of video. The most common high-confidence actions included "a person dancing," "a person speaking," "a person smiling," and "a person sitting," with confidence scores often ranging between 0.60 and 0.85. These predictions aligned closely with the expected activities occurring in the JAXTPA performance videos, where youth frequently engage in expressive movement and verbal interaction. The chunk-level action table revealed clear temporal structure within videos, where dancing dominated early segments while speaking and smiling appeared more frequently during instruction or conversational moments. The consistency of XCLIP's predictions across variable scenes showcases its ability to generalize effectively to community-recorded, non-studio video material.

Emotion recognition, implemented through FER-CNN, produced meaningful affective cues that complemented the action predictions. The system extracted mid-chunk frames and assigned emotional labels such as happy, neutral, surprised, and sad, accompanied by confidence scores. In most cases, the detected emotions matched the visual context, with happy and neutral emerging as the most common affective states during rehearsals and performances. While occasional detection errors occurred due to motion blur or occlusion, the FER model remained stable for clear frontal faces. These affective outputs provide an additional layer of insight, enabling the interpretation of emotional expressions alongside behavioral actions.

One of the most significant strengths of the system lies in its ability to integrate face clusters, action predictions, emotion scores, and temporal metadata into coherent person-level behavioral summaries. By linking each identity to all frames, chunks, and video segments in which the individual appears, the pipeline allows the user to retrieve comprehensive multimodal profiles. When the user inputs a Person ID, the system returns the individual's face samples, the number of appearances, the actions they performed, their emotional tendencies, and the corresponding frames or video segments. This person-centric retrieval mechanism transforms raw video into structured analytics that can support research on engagement, participation, performance patterns, and emotional expression.

Overall, the experimental results demonstrate that the full pipeline operates reliably and meaningfully across diverse video conditions. The system consistently detected individuals, clustered them accurately, tracked their presence across time, inferred their actions using state-of-the-art video-language modeling, and analyzed their emotional expressions using deep facial affect algorithms. The outputs shown in the face-cluster visualizations and the action-emotion tables illustrate the system's capacity to transform unstructured visual data into actionable insights. These findings validate the effectiveness of the integrated multimodal approach and highlight its potential for supporting social, educational, and behavioral research within complex video environments.

VIII. Futher Work

Future work will focus on expanding the system's robustness, accuracy, and multimodal intelligence beyond the current capabilities. Although the proposed pipeline—consisting of MTCNN, FaceNet, agglomerative clustering, FER-CNN, and XCLIP—performs effectively on JAXTPA video data, several improvements remain. More advanced face detectors and identity trackers, such as transformer-based models or DeepSORT, could provide smoother person tracking in challenging situations involving occlusion, motion blur, or low-resolution frames. Similarly, emotion recognition can be strengthened by integrating transformer-based FER models and incorporating additional cues like body posture, gaze direction, and audio-based affect signals to capture more nuanced emotional states.

On the behavioral side, future work will explore fine-tuning XCLIP on classroom-specific actions or replacing it with more specialized video understanding models. Expanding the system to operate on multimodal inputs—combining audio and visual channels—will enable deeper assessment of engagement, interaction, and communication patterns. Additionally, curating a partially annotated dataset for evaluation and building a real-time dashboard for tracking individual students could transition this research into a practical tool for educators and foster care programs. Together, these extensions will help evolve the current prototype into a comprehensive and scalable behavioral analysis framework.

References

[1] X. Zhou and B. Bhanu, "Human Recognition Based on Face Profiles in Video," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[2] E. G. Ortiz, A. Wright, and M. Shah, "Face Recognition in Movie Trailers via Mean Sequence Sparse Representation-Based Classification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[3] Y. Wang, Y.-P. Huang, and X.-J. Shen, "ST-VLAD: Video Face Recognition Based on Aggregated Local Spatial-Temporal Descriptors," *IEEE Access*, vol. 9, pp. 31169–31183, 2021.

[4] Y.-L. Chen, C.-L. Chang, and C.-S. Yeh, "Emotion classification of YouTube videos," *Decision Support Systems*, vol. 101, pp. 40–50, 2017.

[5] E. Mulholland, P. Mc Kevitt, T. Lunney, and K.-M. Schneider, "Analysing emotional sentiment in people's YouTube channel comments," in *Proceedings of the International Conference on ArtsIT, Interactivity & Game Creation*, 2016, pp. 181–188.

[6] S. A. Kadir, A. M. Lokman, and M. Muhammad, "Identification of positive and negative emotion towards political agenda videos posted on YouTube," in *Proceedings of the International Conference on Kansei Engineering & Emotion Research*, 2018, pp. 758–767.

[7] A. Mohan, M. Choksi, and M. A. Zaveri, "Anomaly and activity recognition using machine learning approach for video based surveillance," in *Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2019, pp. 1–6.

[8] S. Brdnik and T. Heričko, "Prospects of Explainability in the Hugging Face Hub Landscape," 2024.

[9] R. Padilla, C. F. F. Costa Filho, and M. G. F. Costa, "Evaluation of haar cascade classifiers designed for face detection," *World Academy of Science, Engineering and Technology*, vol. 64, pp. 362–365, 2012.

[10] S. Sharma, L. Raja, V. Bhatnagar, D. Sharma, S. N. Bhagirath, and R. C. Poonia, "Hybrid HOG-SVM encrypted face detection and recognition model," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 25, no. 1, pp. 205–218, 2022.

[11] B. P. Prayogo, E. Mulyana, and W. Hermawan, "A Novel Approach for Face Recognition: YOLO-Based Face Detection and Facenet," in *Proceedings of the 2023 9th International Conference on Wireless and Telematics (ICWT)*, IEEE, 2023, pp. 1–6.

[12] N. Zhang, J. Luo, and W. Gao, "Research on face detection technology based on MTCNN," in *Proceedings of the 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, IEEE, 2020, pp. 154–158.

[13] S. Makinist, B. Ay, and G. Aydın, "Average neural face embeddings for gender recognition," *Avrupa Bilim ve Teknoloji Dergisi*, pp. 522–527, 2020.

[14] A. Sydor, D. Balazh, Y. Vitrovyi, O. Kapshii, O. Karpin, and T. Maksymyuk, "Research on the state-of-the-art deep learning based models for face detection and recognition," *Inf. Commun. Technol. Electron. Eng*, vol. 4, pp. 49–59, 2024.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[16] M.-C. Chang, P. Bus, and G. Schmitt, "Feature extraction and k-means clustering approach to explore important features of urban identity," in *Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2017, pp. 1139–1144.

[17] D. Deng, "DBSCAN clustering algorithm based on density," in *Proceedings of the 2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, IEEE, 2020, pp. 949–953.

[18] M. Tapaswi, M. T. Law, and S. Fidler, "Video face clustering with unknown number of clusters," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5027–5036.

[19] M. Aslan, "CNN based efficient approach for emotion recognition," *Journal of King Saud University–Computer and Information Sciences*, vol. 34, no. 9, pp. 7335–7346, 2022.

[20] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proceedings of the 5th International Conference on Multimodal Interfaces*, 2003, pp. 258–264.

[21] I. U. Haq, A. Ullah, K. Muhammad, M. Y. Lee, and S. W. Baik, "Personalized Movie Summarization Using Deep CNN-Assisted Facial Expression Recognition," *Complexity*, vol. 2019, no. 1, p. 3581419, 2019.

[22] H. A. Abdul-Azim and E. E. Hemayed, "Human action recognition using trajectory-based representation," *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 187–198, 2015.

[23] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[24] J. Zhou, L. Dong, Z. Gan, L. Wang, and F. Wei, "Non-contrastive learning meets language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11028–11038.