A Hybrid Movies Recommendation System Based on Demographics and Facial Expression Analysis using Machine Learning

Mohammed Balfaqih
Department of Computer and Network Engineering, College of Computer Science and Engineering
University of Jeddah, Jeddah, 23890, Saudi Arabia

***Abstract*—Cinemas and digital platforms offer an extensive array of content requiring tailored filtering to cater to individual preferences. While recommender systems prove invaluable for this purpose, conventional movie recommendations tend to emphasize specific attributes, leading to a reduction in overall accuracy and reliability. Notably, the extraction process of facial temporal attributes exhibits a suboptimal level of accuracy, thereby influencing the classification of attributes and the overall accuracy of the recommendation system. This article introduces a hybrid recommender system that seamlessly integrates collaborative filtering and content-based methodologies. The system takes into account crucial factors such as age, gender, emotion, and genre attributes. Films undergo an initial categorization based on genre, with a subsequent selection of the most representative genres to ascertain group preferences. Ratings for these selected movies are then predicted and organized in descending order. Employing Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, the system achieves real-time extraction of facial attributes, particularly enhancing the accuracy of emotion attribute extraction through sequential processing. The CNN model demonstrates a commendable 55.3% accuracy score, the LSTM model excels with a 59.1% score, while the combined CNN and LSTM models showcase an impressive 60.2% accuracy. The performance of the recommendation system is rigorously evaluated using standard metrics, including precision, recall, and F1-measure. Results underscore the superior performance of the proposed system across various testing scenarios compared to the established benchmark. Nevertheless, it is noteworthy that the precision of the benchmark marginally surpasses the proposed system in the age groups of 8-14 and 15-24.**

***Keywords*—Recommender system; movies recommendation; emotion prediction; k-means clustering; deep learning**

I. Introduction

Recommender systems play a pivotal role in facilitating user exploration and item selection within the expansive choices available on web or electronic platforms. These systems employ advanced strategies, including content-based, collaborative, and hybrid approaches [1-4], to systematically filter and prioritize information, delivering clients tailored and pertinent data. Content-based filtering, a component of these systems, tailors recommendations based on individual preferences, effectively addressing the challenging cold start problem. Concurrently, collaborative filtering relies on user similarities or machine learning algorithms to recommend items. Hybrid recommender systems integrate both content-based and collaborative approaches, synergistically optimizing performance and honing the precision of recommendations. The amalgamation of these techniques empowers recommender systems to furnish users with suggestions that are not only personalized but also highly accurate [5-8].

A consumer's emotional state influences their decision-making process. Emotion is an unconscious mental state that arises spontaneously, accompanied by physiological and psychological changes in human organs and tissues, such as heart rate, facial expression, and the brain [9]. However, the recommendation process typically neglects the viewer's emotional state due to the intricate interplay of physiological signals with emotions, making subtle emotional expressions easily misunderstood. Prior research has predominantly focused on understanding user emotions through various means, such as ratings, comments, and helpfulness votes, among others [10, 11].

The Affective Video Recommender System (AVRS) has emerged as a prominent research area within recommender systems, diverging from traditional text, image, and speech emotion recognition. Focused on analyzing emotional states within videos and discerning emotions in distinct scenes [12], AVRS strategically recommends video content to viewers based on identified emotional states. An insightful study revealed gender-based disparities in movie preferences, with men exhibiting variations between mood and movie choices, while women tend to demonstrate a more congruent pattern [13]. To provide personalized recommendations in theaters or movie platforms devoid of recorded user data, the incorporation of face recognition becomes imperative. This facilitates the identification of attributes such as age, gender, and emotion estimation.

Within the domain of pattern recognition and computer vision, face recognition involves the identification of individuals by assessing distances between key facial points or the angles formed by facial components [14]. The creation of an efficient face recognition system necessitates considerations for speed, accuracy, scalability for system updates, and improvements in subject recognition. Fundamental face recognition approaches include holistic matching, feature-based (structural) methods that analyze local features, and hybrid techniques combining both holistic and feature

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-21-DR-47). The author, therefore, acknowledges with thanks the University of Jedah technical and financial support.

extraction methodologies [15]. The culmination of these efforts results in the development of a functional and practical face recognition system.

An exhaustive review of literature on movie recommendation systems has identified two pivotal research challenges. Firstly, existing solutions for tracking and extracting facial temporal attributes exhibit substantial computational complexity and diminished accuracy, impacting attribute classification and diminishing the precision of recommendation systems. Secondly, prevalent movie recommendation systems concentrate on specific attributes, leading to a decline in overall accuracy and reliability [16-23]. As an extension of our previous work in [24], this study seeks to elevate the accuracy and efficacy of movie recommendations by crafting a precise face recognition system capable of discerning age, gender, and emotion from video data. Additionally, it discerns the most accurate deep learning models by incorporating multiple attributes, including user emotion and gender.

- A hybrid movies recommendation system based on demographics and facial expression analysis using machine learning. The system effectively captures user age, gender, and emotion in real-time through CNN and LSTM models to solve information overload and data sparsity problems.
- The system ensures the suitability for both new and existing users, providing effective movie recommendations, regardless of the presence of historical records or ratings. Movies are initially grouped by genre, selecting the most representative of preferred genres as the group's choice. Ratings for these movies are then predicted and listed in descending order.
- The study conducts experiments using real-world facial expressions data to demonstrate the excellent performance of the proposed methodology in the existing recommendation approach. It also identifies facial expressions as a crucial factor in movie recommendations.

The paper's structure is organized as follows: Section II addresses the research background, encompassing recommendation systems and face attribute recognition techniques, along with an exploration of related works and distinctions from existing systems. Section III provides a comprehensive overview of the proposed hybrid movie recommendation system, detailing the components of face attribute extraction, movie clustering, and recommendations. Section IV delves into the implementation setup and dataset usage, while Section V extensively discusses the findings derived from the proposed system. Finally, the paper concludes, summarizing the key insights and contributions.

II. Research Background

In this section, we explore the primary techniques employed by the proposed system, encompassing recommendation systems and facial attribute recognition. We also delve into recent research in these areas. In the final sub-section, we examine existing systems focused on integrating facial attributes recognition with movie recommendations, highlighting their limitations, and identifying research gaps.

A. Recommendation Systems

Recommendation systems, designed to address the information overload issue, filter pertinent information based on a user's interests, preferences, or observed behavior for a specific item [25, 26]. This problem arises as data volume increases, hindering effective decision-making [27]. The recommendation systems aim to provide meaningful user-specific recommendations for various items or products [28].

In the domain of movie recommendations, a multitude of recommendation systems has been explored in academic literature. Reddy S. et al. [16] proposed a framework akin to collaborative filtering techniques, integrating genre similarity and content-based filtering to enhance personalized recommendations. User feedback on films and genres significantly influences categorization, contributing to the customization of recommendations. Katarya Rahul [17] developed a hybrid recommender system utilizing the MovieLens dataset, incorporating the k-means clustering algorithm with bio-inspired artificial bee colony optimization.

Taking a distinct approach, [18] introduced an object-based collaborative filtering method, delving into the user's item rating matrix to establish connections among items for personalized recommendations. The author in [19] devised an efficient Graph Convolutional Network (GCN) algorithm, merging random walks and graph convolutions to generate embeddings. The author in [20] presented a comprehensive hybrid recommender system that integrates collaborative filtering via the Singular Value Decomposition (SVD) algorithm, a content-based system, and a fuzzy expert system. This expert system evaluates movie significance based on factors such as average rating and the number of ratings.

In contrast, [21] proposed a dynamic weighted hybrid recommender system, adapting the blend of collaborative filtering (CF) and content-based filtering (CBF) dynamically, deviating from fixed weights. Film-Conseil employs a machine learning algorithm to assess consumer advisory capability, replacing explicit movie scores [22]. Additionally, a movie recommendation system utilizes inductive learning, a machine learning method that effectively reduces sparsity and enhances scalability through experiments [23]. Nevertheless, current recommendation solutions exhibit limitations, lacking consideration for user emotions and demographics. Consequently, there is an imperative need to accurately and efficiently capture real-time user facial expressions.

B. Face Attributes Recognition Techniques

Feature detection and image matching are crucial tasks in machine vision, with varying computational efficiency and accuracy depending on the chosen feature detector and descriptor extraction algorithm. It is essential to select an appropriate algorithm for specific feature matching tasks [29].

In recent research, a significant focus has been placed on real-time analysis of age, gender, and emotional states. A pivotal study [30] employed a Convolutional Neural Network (CNN) to achieve a remarkable 95 percent accuracy in age and

gender identification using the IMDB-WIKI dataset. Simultaneously, the CNN demonstrated a 66 percent accuracy in emotion detection, leveraging the FER dataset. An alternative approach [31] utilized an artificial neural network, achieving a commendable 70.5 percent accuracy in age and gender recognition.

Imane et al. [32] introduced an innovative paradigm integrating the HAAR cascade and CNN for facial recognition, incorporating the FER2013 dataset for normalization and emotion detection. The integration of K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms resulted in a 70% accuracy rate. Rajesh et al. [33] contributed to the field with a real-time emotion detection system based on a nine-layer CNN, demonstrating an approximate 90 percent accuracy in categorizing seven distinct emotions.

Additionally, a method proposed by [34] deployed the Local Binary Pattern (LBP) classifier, achieving impressive results with a score of 94.39 percent using the CK+ dataset and 92.22 percent with the JAFFE dataset. Lastly, the method introduced by [34] implemented Exploratory Data Analysis (EDA), SVM, and demographic classification strategies, achieving an outstanding 99 percent accuracy and efficiency.

C. Related Works

Several systems focused on integrating facial attributes detection with items recommendations have been proposed in the literature. The authors in [36] introduced a video recommendation system centered on emotion detection, with the potential to address various conditions through a focus on human emotions and cognition. This system suggests YouTube videos based on captured emotions in images, videos, or webcam feeds, considering emotion intensity. For instance, individuals expressing happiness receive recommendations for funny videos, while those displaying sadness are guided to motivational content. The system has achieved an average emotion detection accuracy of 56%. For the same purpose, a dataset was created in [37] with five classes and then compared with an alternative dataset, showcasing state-of-the-art performance. The results revealed that, apart from CNN and DenseNet201, VGG16, InceptionV3, and MobileNetV2 exhibited superior accuracy compared to the collected dataset, affirming the excellence of our dataset over the alternative one. The primary limitations of these works involve accuracy concerns and a lack of consideration for user age and gender, which may not fully align with users' preferences.

Haar cascade and Local Binary Patterns Histogram (LBPH) algorithms were utilized in [38] for face detection, feature extraction, and emotion detection. Emotion detection relies on the FER 2013 dataset, while age and gender detection use the Adience dataset. For web application development, they implemented the Django framework. The video recommendation system adopts a content-based approach, customizing recommendations based on detected emotions, age, and gender. These personalized suggestions are sourced from the internet, prioritizing videos aligned with the target emotion, popularity, or user interactions. Video categories encompass music videos, motivational speeches, quotes, movies, cartoons, humor, action, and lifestyle.

To automate the process of identifying users and deducing their preferences from their content feedback of TV applications, a solution based on face detection and recognition services was proposed in [39]. Demographic characteristics (age and gender) classified the user, addressing the cold start problem. Smiles and emotions detected served as automatic feedback during content consumption. Accurate results were achieved with a frontal view of the face, while deviations from this angle and suboptimal lighting conditions could hinder face detection and recognition, particularly if parts like the eyes or mouth were not clearly visible. In [40], an innovative approach was introduced to address the lack of affective data for newly added videos on platforms like YouTube. It used reinforcement learning and deep bidirectional recurrent neural networks to process videos, gather affective annotations, and integrate emotion and affective intensity aspects, refining as user feedback was collected. Both implicit and explicit interactions, including facial expressions in real-time video streams, were tracked to train personalized reinforcement learning models for short-term affective behavior learning. This approach also highlighted the value of sequencing videos in different contexts to understand long-term affective trends using context-aware features. Its effectiveness was tested in experiments on two diverse video datasets.

An advertising video recommendation process was introduced in [41], leveraging computer vision and deep learning to gauge users' emotional responses to ads in real time by analyzing their facial expressions. This involved a CNN-based predictive model for rating predictions and a real-time SIFT algorithm-based similarity model to identify users with similar preferences. Instead of relying on users' historical records, the approach continuously updated a dynamic user profile based on real-time facial expression changes. Experimental tests using food advertising videos showcased the superiority of this method compared to conventional approaches like random recommendations, average ratings, and traditional collaborative filtering, offering improved recommendations for both existing and new users in the realm of advertising video recommendations.

Table I summarizes the existing movies/video recommendation systems based on face feature extraction. While there have been previous studies on the integration of videos/movies recommendation systems with face attributes extraction techniques, there are still gaps in the literature that call for further research. It can be concluded that the primary limitations to be addressed in this study pertain to the suboptimal accuracy of recommendation systems due to inadequate data, and the complexity of data analysis, which results in elevated computational costs.

TABLE II. A SUMMARY OF THE MOST RELATED MOVIES RECOMMENDATION SYSTEMS

Attributes	Customer detection	Purpose	Age	Gender	Emotion	Wild environment
Bokhare, A., & Kothari, T., 2023 [1].	Image	Video recommendation	Not consider	Not consider	7 categories	Not considered
Elias, T., et al., 2022 [2].	Image	Movies recommendation	Not considered	Not consider	5 categories	Not considered
Babanne, V., et al., 2020 [3].	Image	Video recommendation	Age groups	Considered	7 categories	Not considered
De Pessemier, T., 2016 [4].	Image	User's non-verbal affective feedback	Accurate ages	Considered	7 categories	Considered
Tripathi, A., et al., 2019 [5].	Video	User's non-verbal affective feedback	Age group	Considered	3 categories	Considered
Kim, G., et al., 2021 [6].	Video	Video advertisement recommendation	Age groups	Considered	-	Not considered
Proposed system	Video & image	Movies recommendation	Age groups	Considered	7 categories	Considered

III. PROPOSED HYBRID MOVIES RECOMMENDATION SYSTEM

In this section, an overall description of the proposed hybrid movies recommendation system. The outline of the proposed system framework is illustrated in Fig. 1. The proposed system will be implemented on a mobile/web platform for use in movie theaters or personal digital devices with embedded cameras. The system is composed of three individual modules, i.e., Face attributes extraction module, Movies clustering model based on movies attributes, and recommendation system. The first module is used to extract the demographic and emotion attributes of the users, while the second module clusters the movies into 19 different genres based on their attributes. The outputs from these modules are utilized by the recommendation system module to provide movie cluster recommendations tailored to specific user groups, considering age, gender, and emotion. Furthermore, the list of movies is organized based on the predicted ratings for all movies within the users' groups. The next subsections describe each module in detail.

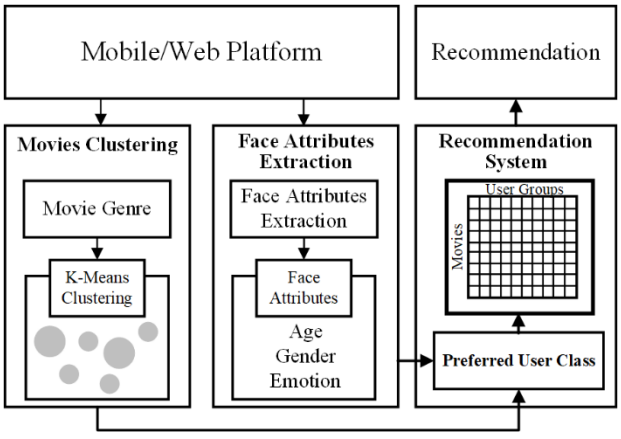

Fig. 1. A block diagram of the proposed system.

A. Face Attributes Extraction

The facial attribute extraction module is delineated into two discrete constituents: the CNN-based attributes extraction model and the LSTM-based attributes extraction model, as depicted in Fig. 2. This system facilitates real-time extraction of facial attributes, encompassing age, gender, and emotion. Positioned as a video-based application, its efficacy in capturing emotion attributes is heightened within a temporal sequence [42].

The CNN-based attributes extraction model concentrates predominantly on non-temporal features, with its output assuming a pivotal role in the ultimate synthesis. Notably, the Inception Net [43] and DenseNet [44] architectures are harnessed for their exceptional performance within this model. In contrast, the LSTM-based attribute extraction model is deployed for extracting vital expressive features, utilizing the VGG architecture to scrutinize emotional nuances within the temporal sequence. The symbiotic alignment of LSTM with the challenge's objectives manifests in competitive outcomes. The prognostications of facial attributes from the employed models are amalgamated by assigning weights to each method based on their performance metrics on the Acted Facial Expression in Wild (AFEW) validation set. The datasets utilized in this proposed system will be expounded upon in Section IV. The emotions considered for classification encompass Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

In both models, the initial step involves resizing the image to different scales, creating an image pyramid, which serves as input for the subsequent three-stage cascaded framework outlined in [45]. For face detection, candidate facial windows and their associated bounding box regression vectors are acquired using a fully convolutional network referred to as the proposal network (P-Net). Calibration of the candidates is performed based on the estimated bounding box regression vectors, followed by the application of non-maximum suppression (NMS) to consolidate highly overlapping candidates. Moving to the second stage, all candidates are directed through another CNN known as the refine network (R-Net), which serves to further eliminate a significant number of erroneous candidates, refine bounding boxes through regression, and perform NMS for accuracy. The final stage mirrors the second stage, with a focus on face regions that receive more supervision, leading to the network outputting the positions of facial landmarks.

1) CNN-based attributes extraction model: As shown in Fig. 2, the diagram of the model is divided into three main sections: Frames feature extraction, Frame-level feature aggregation, and classification. In frames feature extraction,

four networks are fine-tuned for the prediction of individual static images, specifically Inception V3, DenseNet121, DenseNet161, and DenseNet201. These networks were employed due to their efficient feature extraction and high image recognition accuracy. Inception-V3 is a popular CNN model known for its deep architecture and unique inception modules, enabling efficient feature extraction across multiple scales. It is optimized for CPU and GPU usage and has achieved accuracy rates exceeding 78.1% on ImageNet. On the other hand, DenseNet models are characterized by dense interconnections between layers. They vary in the number of layers, with DenseNet201 being the deepest. DenseNet models excel in image classification and feature extraction, delivering state-of-the-art results.

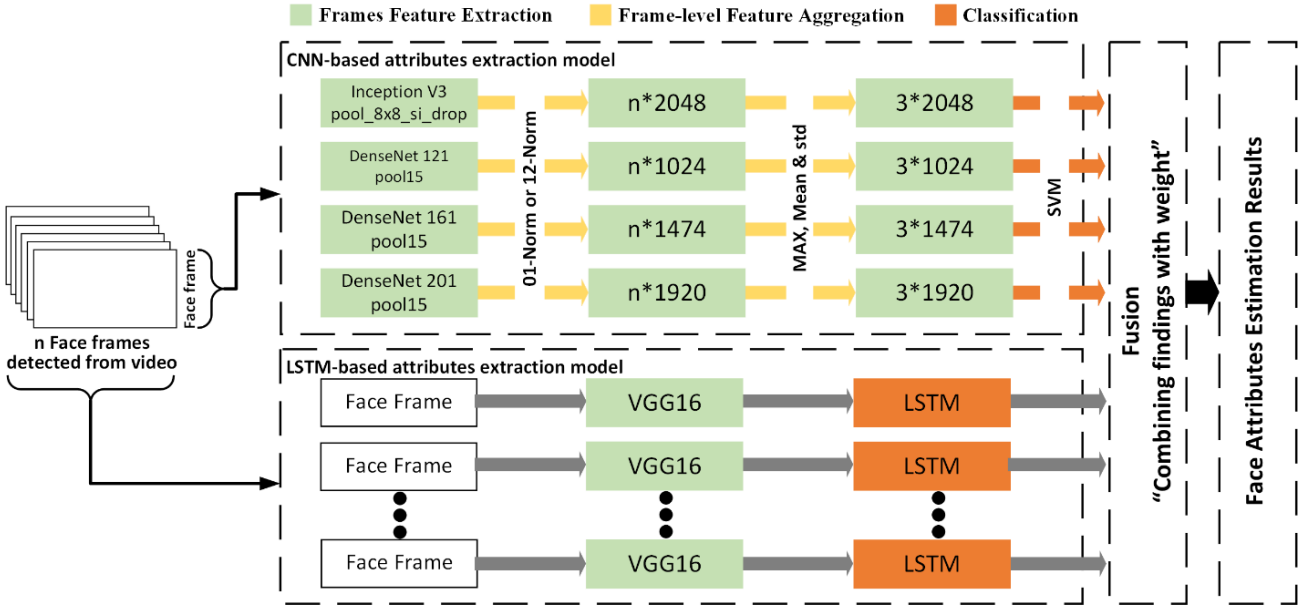

Fig. 2. Face attributes extraction module.

On the Real-world Affective Faces (RAF) validation test, these networks achieved accuracy scores of 82.74%, 83.84%, 83.25%, and 79.73%, with corresponding feature dimensions of 2048*3, 1024*3, 1474*3, and 1920*3, respectively. Subsequently, fine-tuned models extract features from the final layers of aligned faces, using them as the foundational representation. The Number of Features section shows the number of features extracted by each CNN layer. However, since the feature dimension in each video is directly linked to the number of detected faces and the layer dimension, normalization is used to standardize feature dimensions. The Video Features section shows the video features extracted by the CNN layers.

For frame-level feature aggregation, two normalization methods are applied separately to the features extracted by various CNN models from each aligned face in video frames using mean, max, and standard deviation. The video feature tripled compared to the initial CNN extraction. Following that, the RootSIFT and a normalization method ranging from [0,1] are applied to process the original feature [46]. Here's the calculation process:

$$F_i^{l2} = \frac{|F_i^v|}{\sum_j^n (F_j^v)^2} \quad (1)$$

$$F_i^{01} = \frac{F_i^v - \min(F^v)}{max(F^v) - min(F^v)} \quad (2)$$

Where F^v is the original feature extracted from a video, F_i^v is the index of feature in the feature vector. $max(F^v)$ and $min(F^v)$ stand for the maximum and minimum values in the video feature vector. F_i^{l2} and F_i^{01} are the features we normalized by l2-norm and 01-norm.

The classification step shows the output of the SVM. Linear SVMs were trained with various extracted features by the two normalization methods and four networks. Parameters were evaluated using 5-fold cross-validation. The results on the AFEW validation set for different features, based on SVM models, are notably lower than those in the RAF static image set. This difference can be attributed to the richer information in video clips regarding the expression process.

2) LSTM-based attributes extraction model: For face detection, the model is trained using VGG facial features on the AFEW training dataset. Maintaining stable face tracking is essential for optimal model performance in the analysis of the time sequence. Given that most video frames contain at least one face, the face detection threshold is set to a lower value to capture more faces while minimizing errors. A comprehensive set of face landmarks is leveraged to precisely locate the primary character's face, typically the one with the largest facial area. Additionally, a larger window is employed to capture finer details from regions like the forehead and chin.

For frame feature extraction, VGG-16 architecture is employed due to its exceptional quality of the FER2013 dataset in grayscale and its ample collection of meticulously selected faces. VGG-16 stands out for its depth, comprising 16 weight layers, uniform architecture with small 3x3 convolutional

filters, robustness, and high accuracy in image classification. The pre-trained VGG-Face Model is used on the FER2013 emotion dataset to fine tune the network and enhance its performance [14]. The images are resized to 224x224 in grayscale. Through the application of data augmentation techniques, 70.96% accuracy on the validation dataset is achieved which considered high level of performance.

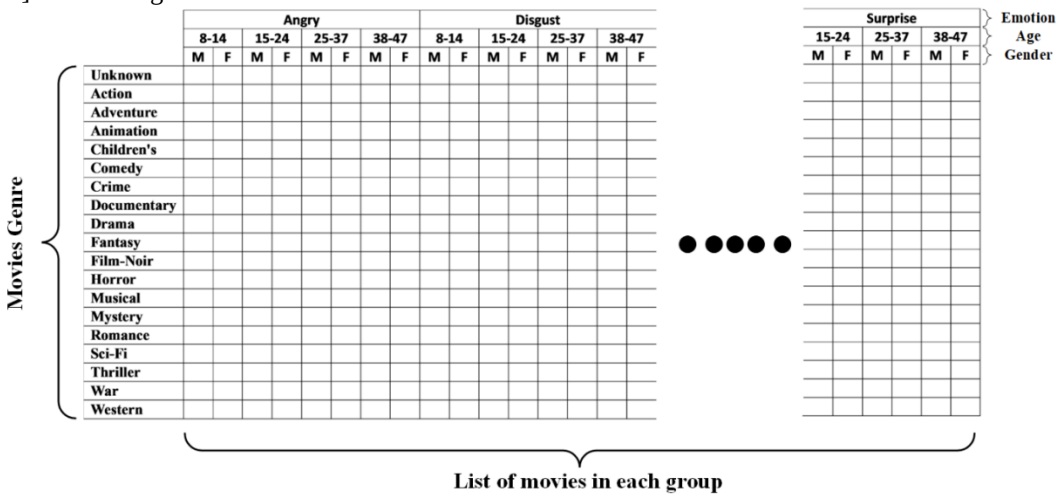

Fig. 3. Preferred movies based group clusters.

Finally, the classification is done using classic LSTM architecture, incorporating memory cell, input gate, output gate, and forget gate, was chosen. The implementation strategy aligns with [16]. LSTMs excel at handling long-term dependencies in sequential data, making them suitable for tasks like time series forecasting. Training videos have been segmented into 16-frame clips. An essential data augmentation step is the overlapping of clips by 8 frames, a well-established and effective technique for both training and testing. Additionally, mirror and multi-scale methods are employed. Temporal features from continuous video frames of facial expressions are extracted by the LSTM layer, utilizing a single LSTM layer with 128 embedding outputs. Notably, the final emotion prediction accuracy on the AFEW validation dataset achieves 46.21%. In contrast, utilizing an LSTM-256 layer instead of 128 leads to a slightly reduced accuracy of 43.07% in the validation dataset.

B. Movies Clustering based on Movie Attributes

Film characteristics will be derived from the Movielens dataset [47] to undergo system testing. Each film is delineated by its unique movie ID, title, release date, IMDb URL, and is classified across 19 genres, encompassing unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, and Western. To systematically categorize these films into distinct groups, the k-means algorithm will be employed—an unsupervised machine learning method.

This algorithm extracts insights from datasets through vectors, operating independently of labeled outcomes. It establishes a predetermined number (k) of cluster centroids within the dataset, computing the distances between each object and these centroids. Objects are subsequently assigned to the nearest cluster based on these distances, and the averages of all clusters are recalculated iteratively until the criterion function is satisfied. Attribute similarity is assessed using the Euclidean similarity approach, wherein objects with analogous attributes exhibit smaller dissimilarity distances, while those with disparate attributes display larger dissimilarity distances.

For a matrix X with i quantitative variables, the Euclidean distance d between two features, x_1 and x_2, can be computed as

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{1n}, x_{2n})^2} \quad (3)$$

The determination of the optimal K value in K-Means involves executing the algorithm with different k values and assessing variance. The selected K value corresponds to the point where variance is minimized. Variance, in this context, is computed as the cumulative sum of distances between each centroid and the items within its assigned cluster. The process includes plotting the variance against various K clusters, enabling the identification of the elbow point. The elbow point signifies the threshold at which the reduction in variance becomes notably stagnant. This allows for a systematic evaluation of K values based on variance metrics, providing a quantitative basis for determining the most suitable number of clusters in the K-Means clustering algorithm.

C. Movies Recommendation

Individuals are stratified based on attributes such as age group, gender, and emotional states to furnish tailored movie recommendations, as delineated in Fig. 3. The designated age cohorts encompass 8-14, 15-24, 25-37, and 38-47 years. Subsequently, the movies selected by each user are allocated to corresponding clusters. The cluster exhibiting the utmost representation of favored movie genres is identified as the group's preference. This determination relies on the cluster with the highest prevalence among all movie clusters, exemplified in a group with movie clusters [1, 1, 1, 2, 3, 3, 4, 1, 1], where cluster 1 is acknowledged as the favored group cluster.

Following this, the algorithm illustrated in Fig. 4 is employed to prognosticate movie ratings within each cluster. The system employs the Singular Value Decomposition (SVD) technique [47], a renowned Collaborative Filtering method. SVD, a matrix factorization method grounded in linear algebra, dissects a real matrix *X* into three matrices: *U*, *S*, and *V*. The SVD outcome encompasses matrix *U*, signifying user vectors, and matrix V^T, signifying movie vectors, with the singular values of *X* on the diagonal of matrix *S*, as depicted in the equation.

Algorithm: Ratings_prediction_of_movies

g: users' group (i.e., specific age, gender, and emotion)
m_g: a movie selected by a user group
m_c: a cluster of movies created by clustering module
C: representative movies cluster to specific user's group
m_i: a movie in the representative movies cluster
$\hat{X}$: rating of recommended movies
U: user vectors matrix
V^T: movies vectors matrix
S: diagonal matrix containing the singular values
m_p: movies preference for each *C*
Begin
for each users group *g* **do**
    selected movies $m_g \in m_c$ are listed
    identify *C* for each *g* by calculating maximum select movies cluster m_c
end for
for each movie $m_i \in C$ **do**
    rate the movies using $\hat{X} \approx U.S.V^T$
    list the movies preference m_p
end for
return m_p
End

Fig. 4. Algorithm of ratings prediction of the movies in each cluster.

$$X = U \times S \times V^T \tag{4}$$

SVD is selectively applied solely to movies in the favored group cluster, crafting a matrix where rows symbolize users in the same gender, age, and emotion category as active users, and columns signify chosen movies in a descending order. Matrix values denote the frequency of a specific movie being chosen. Ultimately, recommended movies with the highest ratings are ascertained through the dot product of *U*, *S*, and V^T, as elucidated in the equation.

$$\hat{X} \approx U.S.V^T \tag{5}$$

IV. Implementation Setup and Datasets

The computational framework is developed in Python and executed using Jupyter Notebook. Experimental procedures are conducted on a MacBook Pro featuring an Intel Core i7 processor and 8GB of RAM. Convolutional Neural Network (CNN) models, employed for facial attributes extraction, undergo pre-training utilizing the RAF and FER2013 datasets. The FER2013 dataset encompasses 28,709 training images, 3,589 validation images, and 3,589 testing images. Similarly, the RAF dataset comprises 12,271 training samples and 3,068 testing samples. Additionally, the AFEW dataset is applied for video clips emotion recognition, serving as a dynamic temporal facial expressions data repository derived from cinematic contexts. This dataset incorporates 957 samples, spanning six expression classes, and features neutral, natural head pose movements, occlusions, and a diverse array of subjects representing varied races, genders, ages, and other demographic characteristics. The Movielens 100k dataset is employed to facilitate the training and evaluation of the proposed system. Comprising 100,000 ratings across a 1 to 5 scale, the dataset involves 943 users rating 1682 movies, with each user contributing assessments for a minimum of 20 movies. The performance evaluation of the system involves the execution of 150 experimental observations.

V. Results and Discussion

This section first presents the outcomes of implementing a face attributes extraction model, followed by a movie recommendation model. The results of implementing face attributes extraction using CNN and LSTM based models individually and collaboratively are presented. Variations in sample sizes across classes highlight differences in category significance. To address this, class weights are employed, scaling scores by the square root of sample numbers, improving model performance in easily distinguishable categories. Model weights were determined through experiments on the validation set. The CNN-based model achieved a 55.3% score, while the LSTM-based model reached 59.1%. The combination of CNN and LSTM models achieved an impressive 60.07% accuracy which represents the overall accuracy of predicting the characteristics of the user correctly. However, the prediction accuracy of age, gender, and emotion are 86.3%, 87%, and 85%, respectively. Table II summarizes the accuracy findings of the proposed face attributes extraction model. Gender estimation errors mostly occurred among younger individuals aged 8-14. This may be due to the inherent difficulty in accurately predicting gender in children. On the other hand, age estimation exhibited a relatively high number of prediction errors in individuals aged [19-20]. This can be explained by the distinctiveness of aging patterns among individuals, which also varies based on gender.

Fig. 5. Preferred movies based group clusters.

In Fig. 5, an illustrative overview of movie recommendations incorporating estimated age, gender, and emotion is showcased. The assessment of movie recommendations' performance is undertaken, drawing comparisons with the methodology introduced in a prior study [32]. Assessment metrics include precision, recall, and F1-score. Precision, indicating the accuracy of movie predictions, is calculated as the ratio of true positive predictions to all positive predictions [49]. Recall, representing the model's

ability to predict occurrences, is computed as true positive predictions divided by actual positive predictions [49]. The F1-score, a harmonic mean of precision and recall, is determined using the specified formula [49].

The outcomes of these performance metrics are succinctly presented in Table II, underscoring the superior efficacy of the proposed system across various scenarios. While the benchmark system exhibits slightly higher precision in the 8-14 age group, the proposed system outperforms in recall and F1-score. Notably, the precision in the 15-24 age group reaches around 0.903, surpassing the proposed system's precision of approximately 0.862. Furthermore, in the 25-37 age group, the system achieves an average precision, recall, and F1-score of 0.873, 0.894, and 0.88, respectively.

TABLE III. THE ACCURACY OF THE PROPOSED FACE ATTRIBUTES EXTRACTION MODEL

		Emotion							Average (%)
Gender	**Age**	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**	
Male	**8-14**	55.42	53.12	56.45	56.34	53.22	56.65	57.37	55.51
	15-24	58.34	56.23	57.40	62.35	61.44	56.32	58.43	58.64
	25-37	63.21	60.75	61.21	64.23	64.23	64.15	63.75	63.07
	38-47	62.56	60.34	60.13	64.56	64.25	65.43	64.35	63.08
Female	**8-14**	55.40	54.02	56.34	55.64	52.95	56.67	56.77	55.39
	15-24	58.02	55.97	57.82	62.15	61.64	55.81	58.82	58.60
	25-37	64.51	59.63	62.19	62.43	63.73	64.03	63.82	62.90
	38-47	62.77	62.43	59.94	64.02	65.02	64.86	64.85	63.41
Average (%)		60.02	57.81	58.93	61.46	60.81	60.49	61.02	60.07

While the proposed system demonstrates relatively high accuracy, it may not be suitable for users who cover their faces (e.g., wearing Hijab), as facial features cannot be extracted under such circumstances. To address this limitation, the system could enhance its capabilities by integrating demographic feature extraction through voice analysis. Furthermore, the computational efficiency and complexity of the face attribute extraction module could be enhanced by leveraging the advantages of the You Only Look Once (YOLO) algorithm due to its superior efficiency and suitability for real-time applications [50].

VI. CONCLUSION

Film venues and digital platforms provide an extensive array of cinematic options, often overwhelming consumers faced with a multitude of choices. To address this issue and enhance the efficiency of the decision-making process for clients, sophisticated recommendation systems have been devised. Existing movie recommendation systems, characterized by intricate computational processes, exhibit a notable deficiency in accurately tracking and extracting temporal facial attributes. This limitation compromises the precision of attribute classification, consequently diminishing the overall accuracy of the recommendation system. Furthermore, prevalent systems tend to overlook various attributes, contributing to a reduction in overall accuracy and dependability. This study introduces a novel hybrid recommender system that seamlessly integrates collaborative filtering and content-based methodologies. The system takes into account diverse attributes such as age, gender, emotion, and genre to optimize movie recommendations. Initially, movies undergo categorization based on genre, with a preference determined by selecting the most representative genres. Ratings for these films are subsequently predicted and presented in descending order. Employing Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, the system conducts real-time extraction of facial attributes, enhancing accuracy in emotion attribute extraction by ensuring sequential extraction of attributes. Results from the system's implementation reveal its superior performance compared to the benchmark system across various test scenarios. However, the benchmark system exhibits marginally higher precision in the age groups of 8-14 and 15-24. Despite the proposed system's commendable accuracy, it faces limitations when users conceal their faces, necessitating improvements through the integration of demographic feature extraction via voice analysis and optimization of the computational efficiency of the face attribute extraction module utilizing the YOLO algorithm.

REFERENCES

[1] M. Elmisery, & D. Botvich, “Agent based middleware for private data mashup in IPTV recommender services,” In In 2011 IEEE 16th international work- shop on computer aided modeling and design of communication links and networks (CAMAD), pp. 107–111, 2011.

[2] B. Barragáns-Martínez, E. Costa-Montenegro, and J. Juncal-Martínez, “Developing a recommender system in a consumer electronic device,” Expert Systems with Applications, vol. 42, no. (9), pp. 4216–4228, 2015.

[3] M. Balfaqih, W. Jabbar, M. Khayyat, and R. Hassan, “Design and development of smart parking system based on fog computing and internet of things,” Electronics, vol. 10, no. 24, pp. 3184, 2021.

[4] A. Subasi, M. Balfaqih, Z. Balfagih, and K. Alfawwaz, “A comparative evaluation of ensemble classifiers for malicious webpage detection,” Procedia Computer Science, 194, pp. 272-279, 2021.

[5] M. Balfaqih, and S. A. Alharbi, “Associated Information and Communication Technologies Challenges of Smart City Development,” Sustainability, vol. 14, no. 23, pp. 16240, 2022.

[6] V. Shepelev, A. Glushkov, I. Slobodin, and M. Balfaqih, “Studying the Relationship between the Traffic Flow Structure, the Traffic Capacity of Intersections, and Vehicle-Related Emissions,” Mathematics, vol. 11, no. 16, pp. 3591, 2023.

[7] Yusof, M. H. M., Mokhtar, M. R., Zain, A. M., & Maple, C. "Embedded feature selection method for a network-level behavioural analysis detection model," International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, 2018.

[8] M. M. H. Khan, E. W. K. Loh, and P. T. Singini, "Stabilization of tropical residual soil using rice husk ash and cement," International journal of applied environmental sciences, vol. 11, no. 1, pp. 73-87, 2016.

[9] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, (2018). "A review of emotion recognition using physiological signals," Sensors, vol. 18, no. 7, pp. 2074, 2018.

[10] S. Roy, and S. C. Guntuku, "Latent factor representations for cold-start video recommendation," In Proceedings of the 10th ACM conference on recommender systems, pp. 99-106, 2016.

[11] C. Orellana-Rodriguez, E. Diaz-Aviles, and W. Nejdl, "Mining affective context in short films for emotion-aware recommendation," In Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 185-194, 2015.

[12] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional LSTM," IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 680-688, 2019.

[13] M. B. Devlin, L. T. Chambers, and C. Callison, "Targeting mood: Using comedy or serious movie trailers," Journal of Broadcasting & Electronic Media, vol. 55, no. 4, pp. 581-595, 2011.

[14] C. A. Hansen, "Face Recognition," Institute for Computer Science University of Tromso, Norway, 2009.

[15] R. Jafri, and H. R. Arabnia, "A survey of face recognition techniques," journal of information processing systems, vol. 5, no. 2, pp. 41-68, 2009.

[16] S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok, and B. Venkatesh, "Content-Based Movie Recommendation System Using Genre Correlation. In: Smart Intelligent Computing and Applications," Smart Innovation, Systems and Technologies. Springer, Singapore, pp. 391–397, 2019.

[17] R. Katarya, "Movie recommender system with metaheuristic artificial bee. Neural Computing and Applications," vol. 30, no. 6, pp. 1983-1990, 2018.

[18] L. T. Ponnam, S. D. Punyasamudram, S. N. Nallagulla, and S. Yellamati, "Movie recommender system using item based collaborative filtering technique," In In 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), pp. 1–5, 2016.

[19] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, J. Leskovec, "Graph Convolutional Neural Networks forWeb-Scale Recommender Systems," In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18. Association for Computing Machinery, London, United Kingdom, pp. 974–983, 2018.

[20] B. Walek, V. Fojtik, "A hybrid recommender system for recommending relevant movies using an expert system," Expert Systems with Applications, vol. 158, pp. 113452, 2020.

[21] H. Q. Do, T. H. Le, and B. Yoon, "Dynamic Weighted Hybrid Recommender Systems," In 2020 22nd International Conference on Advanced Communication Technology (ICACT), pp. 644-650, 2020.

[22] P. Perny, and J. D. Zucker, "Preference-based search and machine learning for collaborative filtering: the "film-conseil"movie recommender system," Informa- tion, Interaction, Intelligence, vol. 1, no. 1, pp. 9–48, 2001.

[23] P. Li, and S. Yamada, "A movie recommender system based on inductive learning," In IEEE conference on cybernetics and intelligent systems, pp. 318–323, 2004.

[24] M. Balfaqih, A. Altwaim, A. A. Almohammedi, and M. H. M. Yusof, "An Intelligent Movies Recommendation System Based Facial Attributes Using Machine Learning," In 2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA), pp. 1-6, 2023.

[25] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," Egyptian informatics journal, vol. 16, no. 3, pp. 261-273, 2015.

[26] K. Haruna, M. Akmar Ismail, S. Suhendroyono, D. Damiasih, A. C. Pierewan, H. Chiroma, T. Herawan, "Context-aware recommender system: A review of recent developmental process and future research direction," Applied Sciences, vol. 7, no. 12, pp. 1211, 2017.

[27] J. Gantz, and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, vol. 2007, no. 2012, 1-16, 2012.

[28] P. Melville, and V. Sindhwani, "Recommender systems. Encyclopedia of machine learning, vol. 1, pp. 829-838, 2010.

[29] E. Karami, S. Prasad, and M. Shehata, "Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images," arXiv preprint arXiv:1710.02726, 2017.

[30] M. J. Uddin, P. C. Barman, K. T. Ahmed, S. A. Rahim, A. R. Refat, and M. Abdullah-Al-Imran, "A convolutional neural network for real-time face detection and emotion & gender classification," SR Journal of Electronics and Communication Engineering, vol. 15, no. 3, pp. 37-46, 2020.

[31] T. R. Kalansuriya, and A. T. Dharmaratne, "Neural network based age and gender classification for facial images," The International Journal on Advances in ICT for Emerging Regions, vol. 7, no. 2, 2014.

[32] I. Lasri, A. R. Solh, and M. El Belkacemi, "Facial emotion recognition of students using convolutional neural network," In 2019 third international conference on intelligent computing in data sciences (ICDS), pp. 1-6, 2019.

[33] G. A. Rajesh Kumar, and R. K. K. G. Sanyal, "Facial Emotion Analysis using Deep Convolutional Neural Network," 2017 International Conference on Signal Processing and Communication (ICSPC), pp. 369-374, 2017.

[34] S. L. Happy, and A. Routray, "Automatic facial expression recognition using features of salient facial patches," IEEE transactions on Affective Computing, vol. 6, no. 1, pp. 1-12, 2014.

[35] R. Azarmehr, R. Laganiere, W. S. Lee, C. Xu, and D. Laroche, "Real-time embedded age and gender classification in unconstrained video," In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 57-65, 2015.

[36] A. Bokhare, and T. Kothari, "Emotion Detection-Based Video Recommendation System Using Machine Learning and Deep Learning Framework," SN Computer Science, vol. 4, no. 3, pp. 215, 2023.

[37] T. Elias, U. S. Rahman, and K. A. Ahamed, "Movie Recommendation Based on Mood Detection using Deep Learning Approach," In 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1-6, 2022.

[38] V. Babanne, M. Borgaonkar, M. Katta, P. Kudale, and V. Deshpande, "Emotion based personalized recommendation system," Int. Res. J. Eng. Technol.(IRJET), vol. 7, pp. 701-705, 2020.

[39] T. De Pessemier, D. Verlee, and L. Martens, "Enhancing recommender systems for TV by face recognition," In 12th international conference on web information systems and technologies (WEBIST 2016), vol. 2, pp. 243-250, 2016.

[40] A. Tripathi, T. S. Ashwin, and R. M. R. Guddeti, "EmoWare: A context-aware framework for personalized video recommendation using affective video sequences," IEEE Access, vol. 7, pp. 51185-51200, 2019.

[41] G. Kim, I. Choi, Q. Li, and J. Kim, "A CNN-based advertisement recommendation through real-time user face recognition," Applied Sciences, vol. 11, no. 20, pp. 9705, 2021.

[42] C. Liu, T. Tang, K. Lv, and M. Wang, "Multi-feature based emotion recognition for video clip," In Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 630-634, 2018.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826, 2016.

[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017.

[45] K. Zhang, Z. Zhang, Z., Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, 2016.

[46] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video, arXiv preprint arXiv:1711.04598, 2017.

[47] F. M. Harper, and J. A. Konstan, "The movielens datasets: History and context," Acm transactions on interactive intelligent systems (tiis), vol. 5, no. 4, pp. 1-19, 2015.

[48] M. G. Vozalis, and K. G. Margaritis, "Using SVD and demographic data for the enhancement of generalized collaborative filtering," Information Sciences, vol. 177, no. 15, pp. 3017-3037, 2007.

[49] Z. Omary and F. Mtenzi, "Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning," International Journal for Infonomics (IJI), vol. 3, no. 3, pp. 314-325, 2010 .

[50] K. S. Dixit, M. G. Chadaga, S. S. Savalgimath, G. R. Rakshith, and M. N. Kumar, "Evaluation and evolution of object detection techniques YOLO and R-CNN," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S3, 2019.