Content recommendation system based on facial expression recognition using swin transformer and Haar Cascade

KAZI MAHATHIR RAHMAN, Brac University, Bangladesh

In the digital era, personalized content recommendation systems play a pivotal role in enhancing user experience and engagement across various online platforms. Traditional methods often face challenges in capturing intricate patterns within vast datasets efficiently. In this study, we propose a novel approach to content recommendation leveraging the Swin Transformer architecture coupled with a Hierarchical Cascade mechanism. The Swin Transformer, known for its ability to model long-range dependencies effectively, is employed to capture complex relationships among items and users in the recommendation system. Additionally, we introduce a Hierarchical Cascade framework, which enables a multi-stage recommendation process, thereby refining the suggestions iteratively. By cascading recommendations through multiple levels of granularity, from broad interests to specific preferences, the proposed method enhances the relevance and diversity of recommended content. We evaluate our approach on real-world datasets and demonstrate its superiority over baseline methods in terms of recommendation accuracy, diversity, and coverage. Our findings suggest that integrating Swin Transformer with a Hierarchical Cascade offers a promising avenue for advancing content recommendation systems, paving the way for more personalized and engaging user experiences in online platforms. Our model achieved 48% accuracy on swin transformer model.

Additional Key Words and Phrases: Swin Transformer, Haar Cascade, python, CNN, FER, Facial Recognition

ACM Reference Format:
Kazi Mahathir Rahman . 2024. Content recommendation system based on facial expression recognition using swin transformer and Haar Cascade. 1, 1 (May 2024), 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Facial Emotion Recognition (FER) is a subfield of computer vision that focuses on automatically detecting and classifying human emotions from facial expressions. It aims to analyze facial features like eyes, eyebrows, mouth, and wrinkles to understand the emotional state of a person. In 2019, a self-supervised learning method used various facial priors to estimate facial movements in videos.[14]. A new music player was introduced that uses machine learning to analyze a user's facial expressions and select music based on the detected emotions.[2] Another system utilizes machine learning to analyze facial features and classify emotions from a person's face.[7] The era of CNN, every machine-learning method was converted to deep learning. A CNN-based deep learning method was designed to efficiently recognize emotions from facial expressions.[8] In 2022, a convolutional neural network (CNN) based system for improved facial expression recognition, where features are extracted using Discrete Wavelet Transform (DWT) for better performance. [1] ResNet16-based model is used for recognizing complex emotions involving combinations of basic expressions on a person's face.[16] After introducing the transformer model, a system that recommends music based on a user's current mood, potentially utilizing techniques like sentiment analysis or physiological data. [13] Some multi-modal models were

Author's address: Kazi Mahathir Rahman, Brac University, Dhaka, Bangladesh, kazi.mahathir.rahman@g.bracu.ac.bd.

Manuscript submitted to ACM

introduced in early 2023. A deep learning architecture that combines information from multiple modalities (likely visual and potentially audio) for real-time detection of facial expressions and recognition of emotions. [17] After the release of the vision transformer model, a different variant of the transformer was introduced to the world. Swin Transformer, a recent advancement in deep learning models, for facial expression recognition. Swin Transformers offer advantages like efficient processing compared to traditional methods, potentially leading to improved recognition accuracy.[9] But The paper has provided abundant inspiration for our work and is widely regarded as the state-of-the-art in our field. SwinFace, a novel deep learning model based on the Swin Transformer architecture that tackles four facial analysis tasks simultaneously: face recognition (identifying individuals), expression recognition (detecting emotions), age estimation (predicting age), and attribute estimation (determining characteristics like gender). [15] In this paper, we propose a novel recommendation system leveraging advanced computer vision techniques, specifically the SWIN Transformer and Haar Cascade classifiers, for analyzing facial expressions. The system aims to enhance user experience and engagement in various applications by accurately interpreting facial cues and providing tailored recommendations. The SWIN Transformer model serves as the backbone for capturing spatial dependencies in facial features, while Haar Cascade classifiers facilitate robust facial detection and landmark localization.

2 DATASET

2.1 RAF-DB

The Real-world Affective Faces Database (RAF-DB) is a large-scale dataset designed for Facial Expression Recognition (FER).[18] It offers several advantages for researchers working in this field due to its:

Real-world nature: Images are sourced from the internet, reflecting the variability of real-world facial expressions compared to controlled lab settings. Diversity: The dataset encompasses a wide range of subjects with varying ages, ethnicities, genders, head poses, lighting conditions, and occlusions (e.g., glasses, facial hair). Rich annotations: Each image is labelled with basic or compound emotions by multiple annotators, providing robust ground truth data.

Fig. 1. Demo of RAF-DB dataset

. The RAF-DB dataset comprises 29,672 images sourced from the internet, encompassing a wide range of real-world facial expressions. These images are meticulously labeled with both basic emotions (surprise, anger, fear, joy, sadness, disgust, neutral) and 12 compound emotions. Additionally, each image is enriched with annotations like facial landmark locations, bounding boxes, subject demographics, and baseline emotion recognition outputs, providing a comprehensive resource for facial expression recognition research.

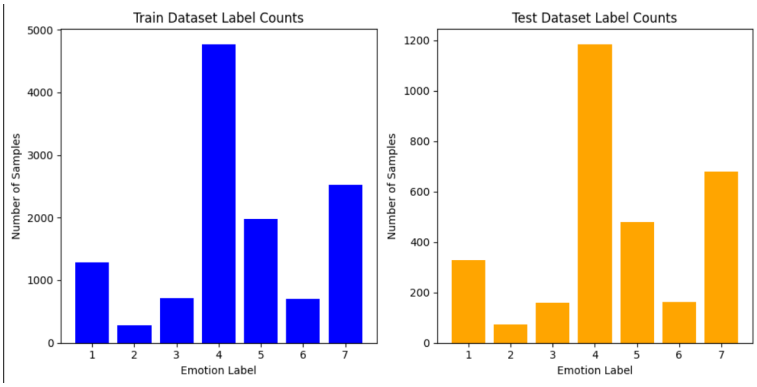

Fig. 2. Total sample of RAF-DB

2.2 CK+

The Cohn-Kanade dataset (CK+), developed by Jeffrey Cohn and Takeo Kanade [12], is another prominent dataset widely used in Facial Expression Recognition (FER) research.

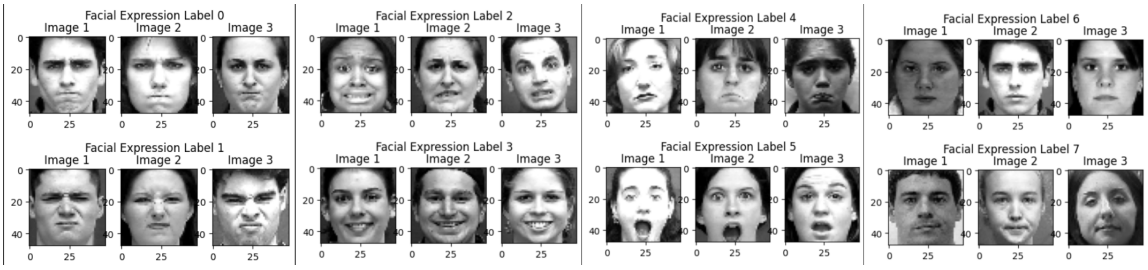

Fig. 3. Total sample of CK+

The Cohn-Kanade dataset (CK+) offers a collection of approximately 1,000 video sequences for facial expression recognition research. These videos capture subjects' facial expressions in a controlled laboratory environment. Unlike RAF-DB's internet-sourced images, CK+ allows researchers to study facial expressions under consistent lighting and pose conditions. Each video sequence is meticulously annotated with labels for six basic emotions: anger, contempt, disgust, fear, happiness, and sadness. Additionally, the dataset provides facial landmark locations, pinpointing key points on the face that move during emotional expressions. This combination of video data, emotion labels, and facial landmark annotations makes CK+ a valuable resource for researchers investigating the dynamics of facial expressions in a controlled setting.

2.3 FER2013

The Facial Expression Recognition 2013 (FER2013) dataset is another widely used resource for research in facial expression recognition (FER).[4] It offers a balance between size and manageability.

The FER2013 dataset offers a middle ground between size and manageability for facial expression recognition research. It consists of approximately 35,000 grayscale images (48x48 pixels) sourced from the internet. These images are divided

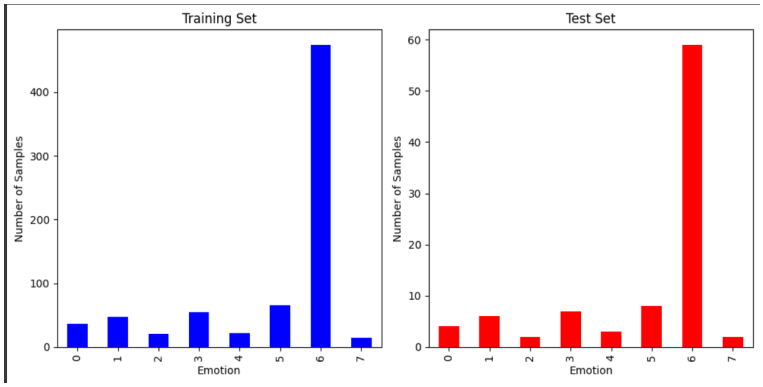

Fig. 4. Total sample of CK+

Fig. 5. Demo of FER2013 dataset

into training, validation, and test sets, with roughly equal representation for seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, neutral). While lacking the controlled setting of CK+, FER2013's larger size and inherent variability in pose, lighting, and ethnicity make it valuable for exploring real-world FER applications.

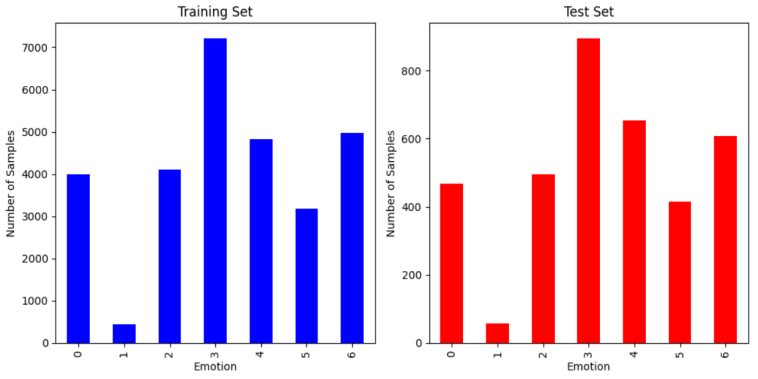

Fig. 6. Total sample of FER2013

Table 1. Comparison of RAF-DB, CK+, and FER2013 Datasets

Feature	RAF-DB	CK+	FER2013
Number of Images	30,000	1002	35,887
Number of Subjects	8,000	123	14,381
Number of Emotions	7	8	7
Image Resolution	224x224	256x256	48x48
Image Type	RGB	Gray Scale	Gray Scale
Labeling Method	Manual by annotators	Posed expressions induced by stimuli	Crowdsourcing

3 METHOD

3.1 Haar Cascades

Haar cascades offer a machine learning approach for real-time face detection within images or video frames, making them a crucial first step in FER systems.[19] This method relies on identifying specific features within an image using rectangular filters called Haar-like features. These features can be simple shapes like edges, lines, or center-surround patterns, mathematically represented as:

Haar-like feature (f): A function that calculates the difference between the sum of pixel intensities within rectangular regions in an image. These regions can be adjacent (e.g., edge) or overlapping (e.g., center-surround). Here's a formula for a simple edge feature:

$$f(x, y) = \sum_{i \in A1} wi(x, y) - \sum_{i \in A2} wi(x, y) \tag{1}$$

x, y: Coordinates of the top-left corner of the feature in the image. A1, A2: Rectangular regions within the feature. wi(x, y): Pixel intensity at location (x, y) within the image.

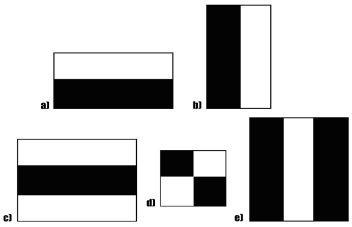

Fig. 7. Feature extracting filters

AdaBoost Algorithm: The AdaBoost algorithm acts as a classifier selection tool. It iteratively analyzes the training data (positive images containing faces and negative images without faces) and selects a small subset of the most informative features from the pool. This selection focuses on features that effectively differentiate faces from non-faces. Mathematically, AdaBoost builds a strong classifier by combining multiple weak classifiers (Haar-like features) with weighted coefficients:

$$F(x, y) = \sum_{t=1}^{T} \alpha_t ht(x, y) \tag{2}$$

F(x, y): The final strong classifier for face detection. T: Total number of weak classifiers (Haar-like features) selected by AdaBoost. αt: Weight assigned to each weak classifier based on its discriminative power. ht(x, y): Output of the t-th weak classifier (either +1 for "face" or -1 for "non-face").

Cascade Classifier Construction: The selected features are arranged in a cascade structure. Each stage in the cascade uses a subset of features. If an image region fails a certain number of stages (i.e., the classifier output falls below a threshold), it's classified as "not face" and discarded, significantly improving efficiency. Images that pass all stages are considered potential faces.

3.2 Swin transformer

In the realm of computer vision, convolutional neural networks (CNNs) have long been the dominant architecture for tasks like image classification and object detection. However, recent advancements have seen the rise of vision transformers,[3] a powerful alternative that leverages self-attention mechanisms. Among these transformers, the Swin Transformer[11] stands out for its hierarchical design. This approach breaks down an image into smaller patches, processes them locally using self-attention, and then progressively merges information across these patches to capture both local and global features. This strategy empowers Swin Transformers to achieve high accuracy while maintaining efficient computation, making them a promising advancement for various computer vision tasks.

3.2.1 Patch Embedding: In Swin Transformer, the process of patch embedding serves as the initial step in transforming an input image into a format suitable for processing by the transformer architecture. This step involves breaking down the image into smaller, non-overlapping patches, akin to dividing a puzzle into manageable pieces. Each patch is then linearly projected into a lower-dimensional embedding space, effectively transforming it into a vector representation. This embedding process encapsulates the local visual information contained within each patch, compressing it into a more compact and abstract form that the subsequent transformer layers can work with. By representing image patches as embeddings, Swin Transformer enables the model to treat images as sequences of tokens, facilitating the application of self-attention mechanisms to capture spatial relationships and dependencies across different regions of the image. Thus, patch embedding serves as a crucial bridge between raw pixel data and the hierarchical processing of visual information within the transformer architecture, enabling effective image understanding and classification.

3.2.2 Patch Partitioning: Patch partitioning is a critical preprocessing step that involves dividing the input image into smaller, non-overlapping patches or tiles. Unlike traditional approaches that use fixed-size patches, Swin Transformer employs a hierarchical patch partitioning strategy, dividing the image into patches at multiple scales or levels. This hierarchical approach allows the model to capture information at various spatial resolutions and effectively handle long-range dependencies across different regions of the image. By breaking down the image into patches, Swin Transformer transforms the spatial information in the image into a structured format that can be efficiently processed by the subsequent transformer layers. This patch partitioning process forms the foundation for enabling the model to understand and analyze images at different scales and resolutions, facilitating effective image understanding and classification.

3.2.3 Shifted Window-based Self-Attention: Shifted window-based self-attention is a variant of the self-attention mechanism used in transformers like Swin Transformer, designed to enable interactions between patches that are spatially distant. In this approach, instead of processing patches sequentially, a shifting operation is applied to the self-attention mechanism, allowing each patch to attend not only to its immediate neighbors but also to patches that

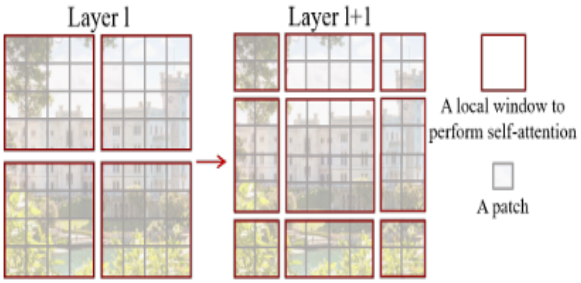

Fig. 8. Patch Partition

are spatially shifted by a certain distance. By incorporating this shifting operation, shifted window-based self-attention facilitates capturing long-range dependencies across different regions of the input image, effectively addressing the challenge of modeling interactions between distant patches. This mechanism enhances the model's ability to capture spatial relationships and dependencies across the entire image, enabling more effective representation learning and feature extraction in vision tasks.

3.2.4 Multi-stage Architecture. The multi-stage architecture in Swin Transformer embodies a strategic organization of hierarchical transformer blocks to process image data at different scales and resolutions progressively. Each stage consists of a sequence of transformer layers, with the number of layers typically decreasing as the spatial resolution decreases. At the initial stages, where patches are relatively large, the model captures low-level details and local features. As the data progresses through subsequent stages, the spatial resolution decreases while the complexity and abstraction of the features increase, allowing the model to extract high-level semantics and capture long-range dependencies. This multi-stage design enables Swin Transformer to efficiently process input images of varying complexities and sizes, providing a scalable and effective framework for a wide range of computer vision tasks, from fine-grained image classification to dense semantic segmentation.

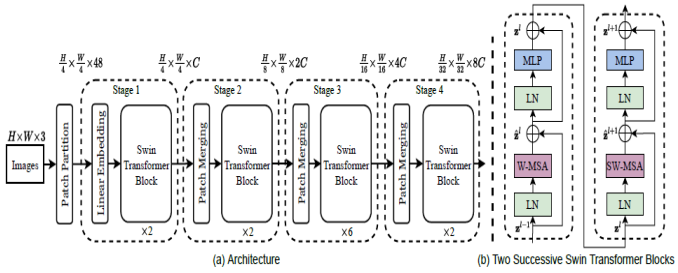

Fig. 9. Swin transformer model

3.2.5 Patch Merging: Patch merging is a crucial mechanism employed to aggregate information from smaller patches into larger ones, facilitating the model's ability to capture long-range dependencies across different regions of the input image. After processing through multiple hierarchical transformer blocks, the spatial resolution of the feature maps decreases, resulting in larger patches representing higher-level semantic information. Patch merging involves combining

neighboring patches into larger patches by reshaping and merging the feature maps spatially. This process effectively reduces the number of patches and increases the receptive field, allowing the model to capture more global context and semantic information. By incorporating patch merging, Swin Transformer enhances its capacity to understand complex visual scenes and perform tasks such as object recognition and image classification with improved accuracy and efficiency.

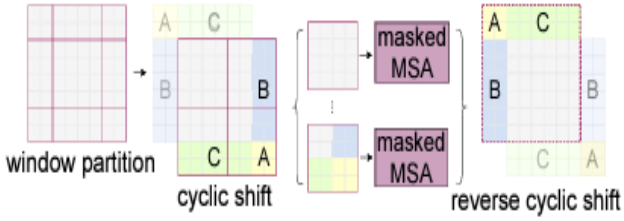

Fig. 10. Swin transformer model

3.2.6 Feed-Forward Network (FFN):. In Swin Transformer, the feed-forward network (FFN) serves as a crucial component within each transformer block, responsible for processing the extracted features and enabling non-linear transformations. This FFN typically consists of two linear layers separated by a non-linear activation function such as ReLU (Rectified Linear Unit). The FFN takes the output of the self-attention mechanism as input and applies a series of affine transformations, followed by element-wise non-linear activations, to model complex relationships and capture higher-level representations. Through this process, the FFN enhances the model's capacity to learn intricate patterns and abstract features from the input data, contributing to its ability to perform tasks such as image classification, object detection, and semantic segmentation effectively.

3.2.7 Residual Connection and Layer Normalization: Residual connections and layer normalization are fundamental techniques employed within each transformer block to facilitate stable and efficient training. Residual connections, inspired by the residual networks (ResNets)[5] in convolutional neural networks, involve adding the input to the output of a transformer layer before passing it through the subsequent layer. This mechanism enables the gradients to flow more smoothly during training, alleviating the vanishing gradient problem and promoting deeper model architectures. Layer normalization is applied after each residual connection, normalizing the activations across the feature dimensions. This normalization technique helps stabilize the training process by reducing the internal covariate shift and accelerating convergence. Together, residual connections and layer normalization contribute to the stability, efficiency, and effectiveness of Swin Transformer, enabling robust learning and performance across various computer vision tasks.

3.2.8 Classification Head: The classification head in Swin Transformer represents the final component of the model responsible for producing the desired output predictions, typically in the form of class probabilities for image classification tasks. This head is attached to the top of the transformer backbone and consists of one or more fully connected layers followed by a softmax function. The fully connected layers serve to map the high-level features extracted by the transformer backbone into a format suitable for classification. The softmax function then converts the output of these layers into a probability distribution over the possible classes, enabling the model to make predictions. Through the

classification head, Swin Transformer transforms the learned representations into actionable insights, allowing it to accurately classify input images into predefined categories.

3.2.9 Model Training: In the evaluation of our Content Recommendation System (CRS), we utilized a pre-trained model of the Haar Cascade classifier for face detection and trained the Swin Transformer model using the RAF-DB dataset, a widely recognized benchmark dataset for facial expression recognition. Our experimental results demonstrated a achieved accuracy of 48% in facial expression recognition tasks. While this accuracy level represents a significant milestone, it also suggests avenues for further improvement. The performance of the Swin Transformer model may be influenced by various factors, including the complexity and diversity of facial expressions within the dataset, as well as potential limitations in the model architecture or training methodology

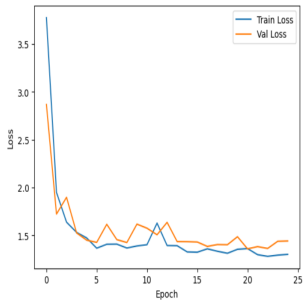

Fig. 11. Swin Transformer model Training curve

4 PROJECT ARCHITECTURE

We present a comprehensive framework for a Content Recommendation System (CRS) leveraging cutting-edge computer vision methodologies and deep learning architectures. Our system is designed to offer personalized content suggestions based on real-time facial expression analysis. The process begins with the application of the Haar Cascade classifier, a robust method for face detection renowned for its accuracy and speed. Haar Cascade efficiently identifies facial regions within images or video frames, laying the foundation for precise facial feature extraction.

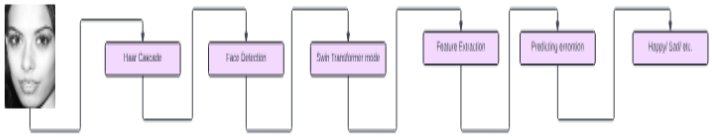

Fig. 12. Work flow

These localized facial regions are then passed through the Swin Transformer model, a recent advancement in the field of computer vision, known for its hierarchical and non-local self-attention mechanisms. By leveraging Swin

Transformer's capabilities, our system achieves remarkable accuracy in recognizing facial expressions, enabling a nuanced understanding of users' emotional states. Finally, based on the detected expressions, our CRS dynamically suggests a curated playlist of videos tailored to the user's mood and preferences. This seamless integration of Haar Cascade and Swin Transformer not only enhances the accuracy and efficiency of facial expression recognition but also empowers the CRS to deliver engaging and personalized content recommendations, thus advancing the frontier of content recommendation systems in the digital landscape.

5 RESULT

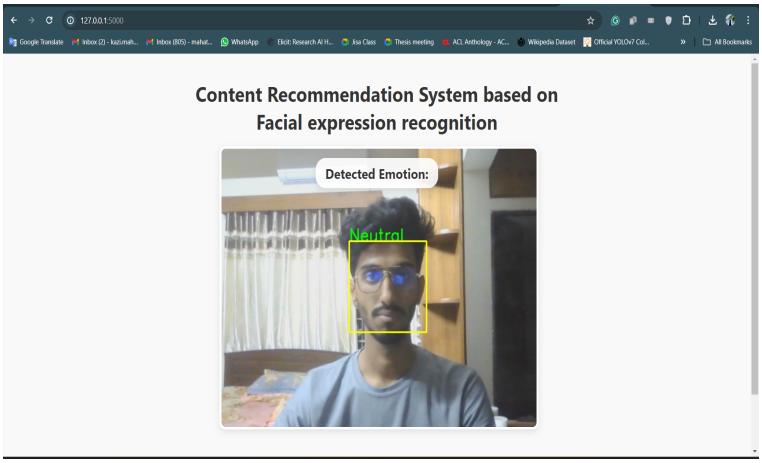

Fig. 13. Home page of Web Application

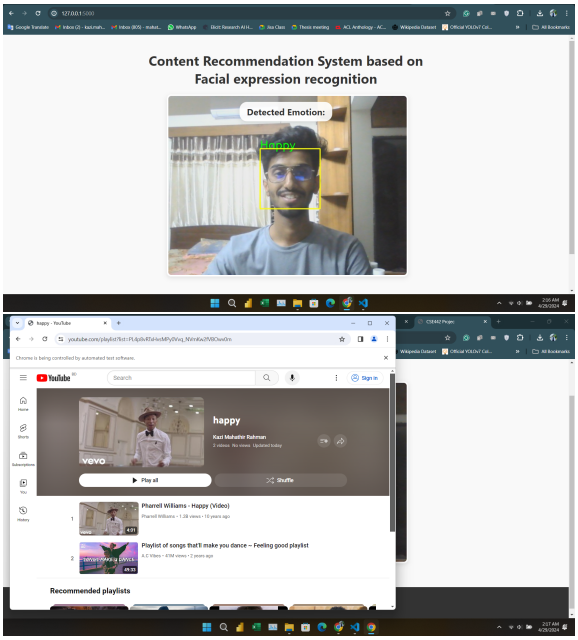

Fig. 14. Happy emotion

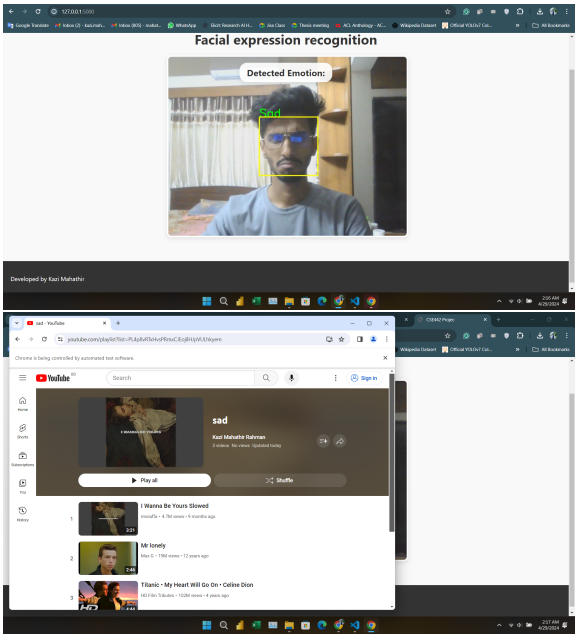

Fig. 15. Happy emotion

6 LIMITATION

- Our model is unable to detect the micro expression which can a good insight for the recommendation system. Micro-expressions, by their very nature, are extremely brief and subtle facial movements that convey authentic emotional responses. Micro-expressions occur spontaneously and often involuntarily, making them inherently difficult to capture and analyze accurately. Factors such as low resolution or frame rate in the input video data, limited training data capturing a wide range of micro-expressions, and the presence of noise or occlusions can all contribute to the model's inability to detect these subtle facial cues reliably. Additionally, the interpretation of microexpressions requires a high level of sensitivity and contextual understanding, as different individuals may express the same emotion through varying micro-expressions.[6] Our model may lack the necessary granularity and robustness to distinguish between subtle variations in micro-expressions and accurately infer the underlying emotional states. As a result, the recommendation system may miss out on valuable insights into users' true emotional responses to content, leading to less personalized and potentially less effective recommendations.
- The performance of your system heavily relies on the quality and diversity of the training data. If the training dataset is limited in size or biased towards certain demographics, the system's ability to generalize to new, unseen data may be compromised.[10]
- While Swin Transformer is known for its efficiency in handling large images, the real-time performance of the system could be a concern, especially if it needs to process high-resolution video streams in real-time.
- Facial recognition technology, when deployed in public spaces or used without individuals' explicit consent, raises profound privacy concerns. This technology can capture and analyze biometric data, such as facial features, without individuals' knowledge or permission. As a result, individuals may feel that their privacy is being compromised, as their movements and interactions in public spaces can be tracked and monitored without their consent. The use of facial recognition technology also raises significant ethical implications. There are concerns about the potential for misuse or abuse of this technology, including mass surveillance, profiling, and discrimination. For example, certain demographic groups, such as racial minorities or marginalized communities, may be disproportionately targeted or surveilled by facial recognition systems, leading to unjust outcomes and reinforcing existing biases and inequalities. Facial recognition systems also raise concerns about data security and the protection of biometric information. Biometric data, once captured, can be highly sensitive and can potentially be used for nefarious purposes if it falls into the wrong hands. Ensuring robust data security measures, such as encryption, access controls, and secure storage practices, is crucial to mitigate the risk of unauthorized access or data breaches.

REFERENCES

[1] Ridha Ilyas Bendjillali, Mohammed Beladgham, Khaled Merit, and Abdelmalik Taleb-Ahmed. 2019. Improved facial expression recognition based on DWT feature for deep CNN. *Electronics* 8, 3 (2019), 324.

[2] S Deebika, KA Indira, et al. 2019. A machine learning based music player by detecting emotions. In *2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Vol. 1. IEEE, 196–200.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[4] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang,

and Yoshua Bengio. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks* 64 (2015), 59–63. https://doi.org/10.1016/j.neunet.2014.09.005 Special Issue on "Deep Learning of Representations".

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[6] Sakshi Indolia, Swati Nigam, Rajiv Singh, Vivek Kumar Singh, and Manoj Kumar Singh. 2023. Micro Expression Recognition Using Convolution Patch in Vision Transformer. *IEEE Access* 11 (2023), 100495–100507. https://doi.org/10.1109/ACCESS.2023.3314797

[7] J Jayalekshmi and Tessy Mathew. 2017. Facial expression recognition and emotion classification system for sentiment analysis. In *2017 International Conference on Networks & Advances in Computational Technologies (NetACT)*. IEEE, 1–8.

[8] Asad Khattak, Muhammad Zubair Asghar, Mushtaq Ali, and Ulfat Batool. 2022. An efficient deep learning technique for facial emotion recognition. *Multimedia Tools and Applications* 81, 2 (2022), 1649–1683.

[9] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. 2022. Facial expression recognition with swin transformer. *arXiv preprint arXiv:2203.13472* (2022).

[10] Bei Li, Yi Jing, Xu Tan, Zhen Xing, Tong Xiao, and Jingbo Zhu. 2023. TranSFormer: Slow-fast transformer for machine translation. *arXiv preprint arXiv:2305.16982* (2023).

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *CoRR* abs/2103.14030 (2021). arXiv:2103.14030 https://arxiv.org/abs/2103.14030

[12] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[13] Ankita Mahadik, Shambhavi Milgir, Janvi Patel, Vijaya Bharathi Jagan, and Vaishali Kavathekar. 2021. Mood based music recommendation system. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume* 10 (2021).

[14] Zhuang Peng, Boyi Jiang, Haofei Xu, Wanquan Feng, and Juyong Zhang. 2023. Facial optical flow estimation via neural non-rigid registration. *Computational Visual Media* 9, 1 (2023), 109–122.

[15] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. 2023. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[16] Khadija Slimani, Yassine Ruichek, and Rochdi Messoussi. 2022. Compound facial emotional expression recognition using cnn deep features. *Engineering Letters* 30, 4 (2022), 1402–1416.

[17] KV Sridhar and Sitaram Thripurala. 2023. Real-Time Facial Emotion Detection System Using Multimodal Fusion Deep Learning Architecture. In *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*. IEEE, 1–6.

[18] Shanghua Sun, Yichen Liu, Xilin Yan, Timothy Chua, and Lee-Feng Cheong. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2. 5902–5910.

[19] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Vol. 1. I–I. https://doi.org/10.1109/CVPR.2001.990517

Received 28 April 2024; revised 28 April 2024; accepted 28 April 2024