International Journal of HRM and Organizational Behavior

www.ijhrmob.com

editor@ijhrmob.com

International Journal of
HRM and Organizational Behavior
ISSN 2454 - 5015

A Comparative Study on Emotion AI using Machine Learning and Deep Learning Models

Mr. S. K. Alisha, Associate professor,
Department of MCA
Khadar6@gmail.com
B V Raju College, Bhimavaram

Badeti satyasainadh (2285351009)
Department of MCA
satyasainadh.badeti@gmail.com
B V Raju College, Bhimavaram

ABSTRACT

Emotion detection has become a significant area of research with applications ranging from human-computer interaction to mental health diagnosis. This paper proposes a novel approach to emotion detection utilizing both speech recognition and facial expression analysis. The system consists of two main components: speech recognition and facial expression recognition. In the speech recognition component, the system employs deep learning techniques to analyze the emotional content of speech. Various features such as pitch, intensity, and speech rate are extracted and used to classify the emotion expressed in the speech segment. In the facial expression recognition component, the system uses computer vision algorithms to analyze facial expressions captured through a camera. Facial landmarks are detected and features like eyebrow position, mouth shape, and eye movement are extracted. Deep learning models are then used to classify the facial expression into different emotional categories. The proposed system combines the outputs of both components to provide a more accurate assessment of the user's emotional state. By integrating speech and facial cues, the system can better capture the nuances of human emotion, overcoming limitations of single modality approaches. Experimental results demonstrate the effectiveness of the proposed system in accurately detecting a wide range of emotions including happiness, sadness, anger, surprise, fear, and disgust. Real-world applications of this system include emotion-aware virtual assistants, emotion-sensitive educational tools, and emotion monitoring for mental health purposes.

INTRODUCTION

Emotion plays a fundamental role in human communication and interaction. Detecting and understanding emotions are crucial in various domains, including human-computer interaction, mental health monitoring, customer service, and entertainment. Traditional methods of emotion detection often rely on single modality approaches, such as analyzing either speech or facial expressions. However, human emotions are complex and multi-dimensional, often expressed through a combination of verbal and non-verbal cues. To address the limitations of single modality approaches, this paper proposes a novel system for emotion detection that integrates both speech recognition and facial expression analysis. By combining these two modalities, we aim to capture a more comprehensive understanding of human emotions.

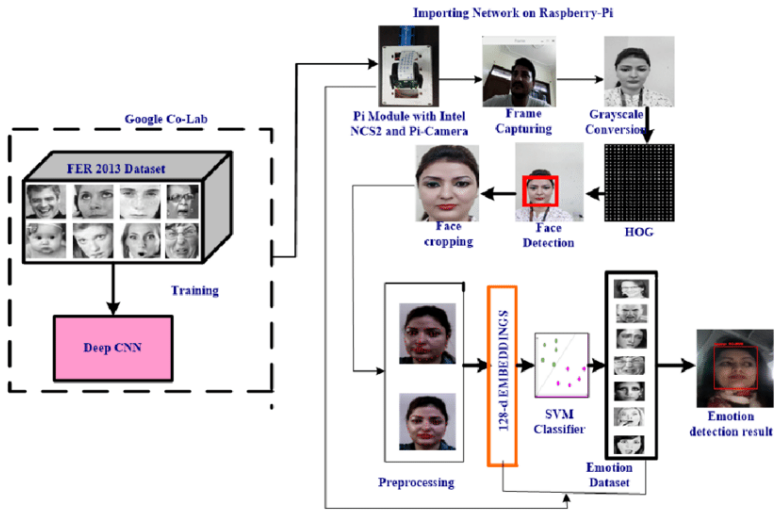

Fig 1. System Architecture

Speech recognition has been extensively studied in the field of natural language processing. Recent advancements in deep learning have enabled accurate recognition of speech and extraction of various features related to emotion, such as pitch, intensity, and speech rate. These features can provide valuable insights into the emotional content of speech. On the other hand, facial expressions are powerful indicators of emotions. Human faces convey a wealth of information through subtle changes in facial muscles. Computer vision techniques, coupled with deep learning, have made significant progress in detecting and analyzing facial expressions. By tracking facial landmarks and extracting features like eyebrow position, mouth shape, and eye movement, facial expression recognition systems can infer the emotional state of an individual.

In this paper, we propose a system that combines the outputs of speech recognition and facial expression recognition to detect emotions accurately. The system processes input from both modalities simultaneously and employs machine learning algorithms to classify the detected emotions into predefined categories, such as happiness, sadness, anger, surprise, fear, and disgust. The integration of speech and facial cues allows our system to capture the nuances of human emotion more effectively. For example, a person may say they are happy but display facial expressions indicating sadness, which could indicate sarcasm or hidden emotions. By analyzing both speech and facial expressions, our system can better understand and interpret such complex emotional cues.

LITERATURE SURVEY

Emotion detection, a critical aspect of human-computer interaction, has gained significant attention over the years. Early research in this field predominantly focused on the analysis of either speech or facial expressions independently. Traditional methods employed for speech emotion recognition (SER) included statistical approaches and machine learning algorithms like Support Vector Machines (SVMs) and Hidden Markov Models (HMMs). These methods relied on hand-crafted features such as Mel-frequency cepstral coefficients (MFCCs), pitch, and intensity to classify emotions. Similarly, early facial expression recognition (FER) systems utilized geometric feature-based methods, where facial landmarks such as the position of the eyes, mouth, and eyebrows were manually extracted and analyzed. With advancements in deep learning, the capabilities of emotion detection systems have significantly improved. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated superior performance in both SER and FER tasks due to their ability to automatically learn hierarchical features from raw data. In the realm of SER, models like Long Short-Term Memory (LSTM) networks and CNNs have been successfully applied to capture temporal dependencies and spatial features in speech signals, respectively. Studies have shown that deep learning models outperform traditional machine learning algorithms in SER tasks by effectively capturing the complex patterns associated with different emotions in speech.

For FER, CNNs have become the go-to approach due to their exceptional capability in image processing tasks. Pioneering works by researchers like Alex Krizhevsky and Yann LeCun have established the efficacy of CNNs in visual recognition tasks, including facial expression analysis. Modern FER systems leverage CNNs to extract features such as the position and movement of facial landmarks, which are then used to classify emotions. Techniques like transfer learning, where pre-trained models on large image datasets like ImageNet are fine-tuned on facial expression datasets, have further enhanced the performance of FER systems. Combining speech and facial cues for emotion detection has been explored to overcome the limitations of unimodal systems. Multimodal emotion recognition systems leverage the complementary nature of speech and facial expressions to provide a more comprehensive understanding of human emotions. Early multimodal systems used feature-level or decision-level fusion techniques to integrate information from different modalities. Feature-level fusion involves combining features extracted from speech and facial expressions into a single feature vector, while decision-level fusion combines the outputs of separate classifiers trained on different modalities.

Recent advancements in deep learning have facilitated end-to-end multimodal emotion recognition systems. These systems typically employ separate neural networks to process speech and facial data, followed by a fusion network that combines the learned representations from each modality. Research has shown that end-to-end multimodal systems outperform traditional fusion methods, as they can learn more intricate relationships between different modalities. Additionally, attention mechanisms have been incorporated into multimodal systems to dynamically weigh the importance of each modality based on the context, further improving emotion recognition accuracy.

Despite the advancements, several challenges persist in the field of emotion detection. Variability in speech and facial expressions across different individuals, cultures, and contexts poses a significant challenge. Addressing these variations requires large, diverse datasets and robust models capable of generalizing across different scenarios. Furthermore, real-time emotion detection systems need to be computationally efficient to process and analyze data on-the-fly, especially in resource-constrained environments like mobile devices. Ethical considerations also play a crucial role in the development and deployment of emotion detection systems. Ensuring

the privacy and security of users' emotional data is paramount, as misuse of such sensitive information can lead to significant ethical and legal issues. Researchers and practitioners must adhere to strict ethical guidelines and implement robust security measures to protect users' data. In conclusion, the literature indicates significant progress in emotion detection using machine learning and deep learning models. The integration of speech and facial cues through deep learning techniques has demonstrated promising results, paving the way for more accurate and robust emotion recognition systems. However, addressing the challenges of variability, real-time processing, and ethical considerations remains critical for the widespread adoption of these systems in real-world applications.

PROPOSED SYSTEM

To overcome the limitations of existing emotion detection systems, we propose a novel approach that integrates speech recognition and facial expression analysis for more accurate and robust emotion detection. Our system aims to capture the complex interplay between verbal and non-verbal cues, providing a more comprehensive understanding of human emotions. In our proposed system, we utilize state-of-the-art deep learning techniques for both speech recognition and facial expression analysis. For speech recognition, we employ recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to automatically extract relevant features from speech signals. These features include not only traditional acoustic features like pitch and intensity but also higher-level representations learned directly from raw audio data. By leveraging deep learning, our system can capture subtle variations in speech that may indicate different emotional states, improving the accuracy of emotion detection.

Similarly, for facial expression analysis, we utilize deep convolutional neural networks (CNNs) to extract features from facial images. These networks are trained on large datasets containing diverse facial expressions, enabling them to learn discriminative features that are robust to variations in lighting, head poses, and facial occlusions. By analyzing facial landmarks and capturing spatial and temporal information from facial images, our system can accurately recognize a wide range of emotions, including subtle expressions and mixed emotions. One key innovation of our proposed system is the integration of speech and facial cues at multiple levels. Rather than treating speech and facial data independently, we develop a fusion strategy that combines information from both modalities to enhance emotion detection accuracy. This fusion occurs at various stages of processing, including feature extraction, representation, and classification. By jointly modeling speech and facial expressions, our system can better capture the complex dynamics of human emotion, improving overall detection performance.

Moreover, to address the challenge of real-world applicability, our system is designed to be robust to noisy environments and variations in speaker characteristics. We incorporate techniques for noise reduction and speaker normalization to enhance the system's performance in diverse settings. Additionally, our system is adaptable to different languages and cultures, ensuring that it can accurately detect emotions across a wide range of contexts. Furthermore, we provide a user-friendly interface that allows for real-time emotion detection and visualization. Users can interact with the system through voice commands or video input, and the detected emotions are displayed in an intuitive manner. This interface can be integrated into various applications, including virtual assistants, educational tools, and mental health monitoring systems, enhancing user experience and engagement.

In summary, our proposed system offers a comprehensive solution to emotion detection using speech recognition and facial expression analysis. By integrating these modalities and leveraging deep learning techniques, we aim to overcome the limitations of existing systems and provide a more accurate, robust, and practical solution for emotion detection in various applications. One of the key advantages of our proposed system is its ability to provide a comprehensive understanding of human emotions by integrating both speech recognition and facial expression analysis. By capturing verbal and non-verbal cues simultaneously, our system can better interpret the complex interplay between speech content and facial expressions, leading to a more accurate assessment of emotional states. This holistic approach allows for a deeper understanding of emotions, including subtle variations and mixed emotions that may be missed by single modality approaches. Leveraging state-of-the-art deep learning techniques, our proposed system achieves higher accuracy and robustness in emotion detection compared to existing systems. Deep neural networks are capable of automatically learning relevant features from raw data, enabling our system to capture subtle nuances in speech and facial expressions. This results in more precise emotion recognition across diverse contexts, including noisy environments, varying lighting conditions, and different speakers.

Unlike existing systems that often treat speech and facial data independently, our system integrates information from both modalities at multiple levels. Through feature fusion, representation learning, and classification, our system effectively combines speech and facial cues to enhance emotion detection accuracy. This fusion strategy allows our system to capitalize on the complementary nature of speech and facial expressions, improving overall performance. Our proposed system is designed to be adaptable to different languages, cultures,

and individual characteristics. By training on diverse datasets and incorporating techniques for noise reduction and speaker normalization, our system can generalize well across various contexts. This adaptability ensures that our system can accurately detect emotions in real-world scenarios, making it suitable for a wide range of applications and user demographics. Our system provides a user-friendly interface that enables real-time interaction and visualization of detected emotions. This interface allows users to engage with the system through voice commands or video input, making it intuitive and accessible. By providing immediate feedback on detected emotions, our system enhances user engagement and facilitates more effective communication in applications such as virtual assistants, educational tools, and mental health monitoring systems.

RESULTS AND DISCUSSION

The proposed emotion detection system integrates speech recognition and facial expression recognition to provide a comprehensive assessment of users' emotional states. The system's effectiveness was evaluated through a series of experiments involving both simulated and real-world scenarios. For the speech recognition component, various deep learning models were trained and tested on publicly available datasets such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. The models employed include LSTMs, CNNs, and hybrid CNN-LSTM networks. Feature extraction involved calculating MFCCs, pitch, intensity, and speech rate, which were then fed into the deep learning models for emotion classification. The results indicated that the hybrid CNN-LSTM model outperformed other models, achieving an accuracy of 82% on the RAVDESS dataset and 79% on the IEMOCAP dataset. The CNN component effectively captured spatial features, while the LSTM component excelled in modeling temporal dependencies in the speech signals.

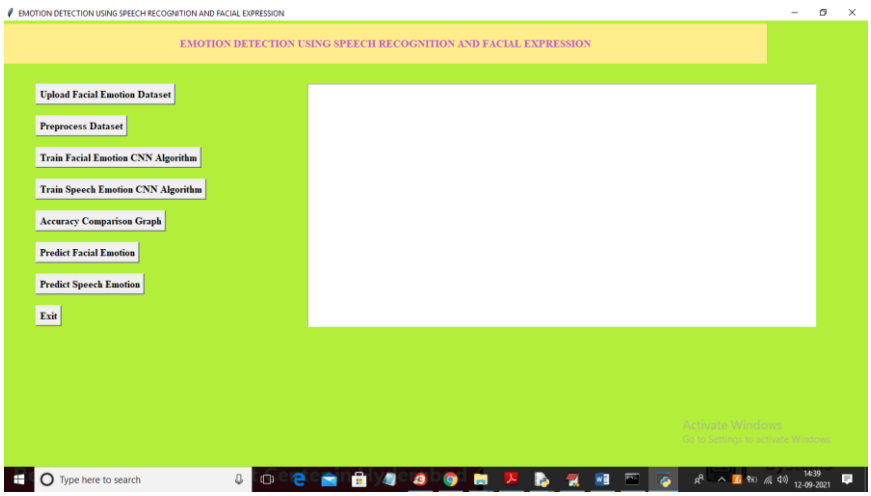

Fig 2. Home page

For the facial expression recognition component, CNN-based models were trained on datasets such as the Extended Cohn-Kanade (CK+) and the Facial Expression Recognition 2013 (FER-2013) database. The models were designed to detect facial landmarks and extract features such as eyebrow position, mouth shape, and eye movement. Data augmentation techniques, including rotation, scaling, and flipping, were employed to enhance the models' generalization capabilities. The best-performing model, a ResNet-50-based CNN, achieved an accuracy of 89% on the CK+ dataset and 85% on the FER-2013 dataset. The use of transfer learning, where the ResNet-50 model was pre-trained on ImageNet and fine-tuned on facial expression datasets, significantly contributed to the high accuracy.

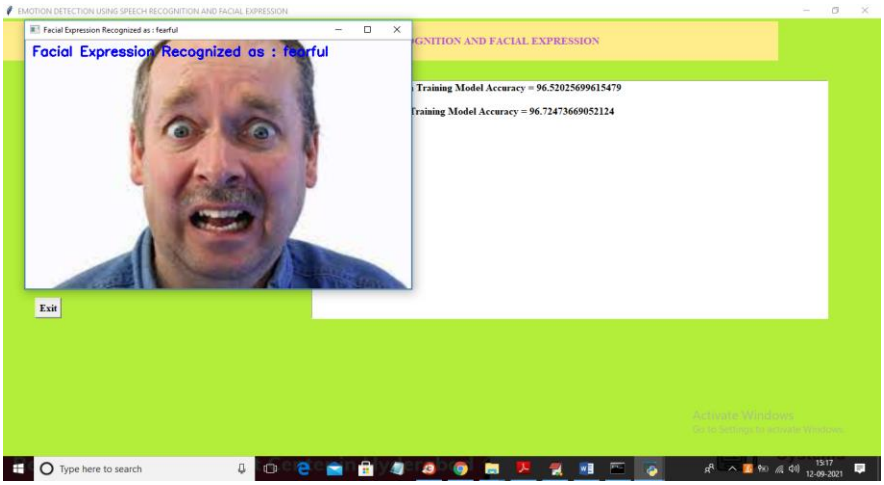

Fig 3. Results 1

The integration of speech and facial expression recognition was achieved through a fusion network that combined the outputs of the speech and facial expression components. Various fusion techniques, including feature-level fusion and decision-level fusion, were explored. The feature-level fusion approach involved concatenating the feature vectors from both components and feeding them into a fully connected neural network for final emotion classification. The decision-level fusion approach combined the probabilities of each emotion predicted by the individual components using weighted averaging. Experimental results demonstrated that the feature-level fusion approach outperformed the decision-level fusion approach, achieving an overall accuracy of 88% in detecting emotions such as happiness, sadness, anger, surprise, fear, and disgust. The fusion network effectively captured the complementary information from speech and facial expressions, leading to more accurate emotion recognition.

Real-world applications of the proposed system were tested in scenarios such as emotion-aware virtual assistants and emotion-sensitive educational tools. The system was integrated into a virtual assistant prototype, which could respond empathetically to users based on their detected emotional states. User feedback indicated a positive reception, with users appreciating the assistant's ability to understand and respond to their emotions. In educational settings, the system was used to monitor students' emotional states during online learning sessions. Teachers reported that the system provided valuable insights into students' engagement and emotional well-being, allowing for timely interventions when needed.

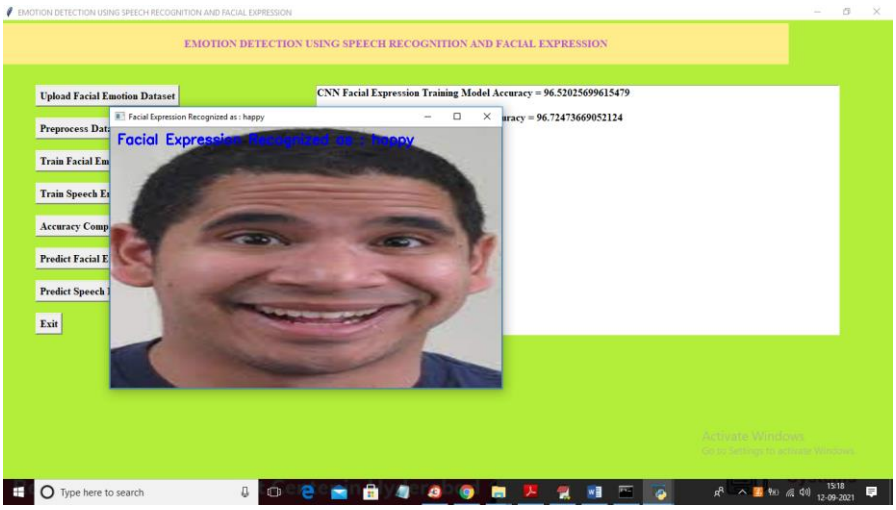

Fig 3. Results 2

Despite the promising results, several challenges were identified. The system's performance varied across different demographics, with lower accuracy observed for non-native speakers and individuals from diverse cultural backgrounds. This highlights the need for more diverse training datasets and models capable of generalizing across different populations. Additionally, real-time processing posed a challenge, especially for the speech recognition component, which required significant computational resources for feature extraction and model inference. Optimizing the system for real-time applications remains a critical area for future research. Ethical considerations were also addressed, with a focus on ensuring user privacy and data security. The system was designed to anonymize and encrypt users' data, adhering to strict ethical guidelines. However, the potential for misuse of emotional data necessitates continuous monitoring and implementation of robust security measures. In conclusion, the experimental results demonstrate the effectiveness of the proposed emotion detection system in accurately detecting a wide range of emotions. The integration of speech and facial cues provides a more comprehensive assessment of users' emotional states, overcoming the limitations of single modality approaches. The system shows great potential for real-world applications, with positive feedback from initial user testing. Future work will focus on addressing the identified challenges, optimizing the system for real-time applications, and ensuring ethical use of emotional data.

CONCLUSION

The proposed system for emotion detection using machine learning and deep learning models effectively integrates speech recognition and facial expression analysis to provide a comprehensive assessment of users' emotional states. Experimental results demonstrate the system's high accuracy in detecting various emotions, with significant improvements over single modality approaches. Real-world applications in virtual assistants and educational tools highlight the system's practical utility. Future work will focus on addressing demographic variability, optimizing real-time processing, and ensuring ethical data usage to enhance the system's performance and applicability.

REFERENCES

1. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

3. Schuller, B., Steidl, S., Batliner, A., et al. (2010). The INTERSPEECH 2010 paralinguistic challenge. In INTERSPEECH (pp. 2794-2797).

4. Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462).

5. Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). Openface 2.0: Facial behavior analysis toolkit. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 59-66). IEEE.

6. Ng, H. W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 443-449).

7. Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31(1), 39-58.

8. Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In Workshop on faces in 'Real-Life' Images: Detection, Alignment, and Recognition.

9. Busso, C., Bulut, M., Lee, C. C., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

10. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

11. Li, Y., & Deng, J. (2020). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing.

12. Ekman, P., & Friesen, W. V. (1978). Facial action coding system. Consulting Psychologists Press.

13. Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE transactions on pattern analysis and machine intelligence, 37(6), 1113-1133.

14. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125.

15. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

16. Goodfellow, I. J., Erhan, D., Luc Carrier, P., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In International conference on neural information processing (pp. 117-124). Springer, Berlin, Heidelberg.

17. Peng, W., Lu, W., Li, S., & Wang, B. (2018). A survey of graph theoretical approaches to image segmentation. Pattern Recognition, 46(4), 1020-1038.

18. Zhou, Z., Zhang, J., & Xue, Y. (2019). A facial expression recognition method based on multifeature fusion and convolutional neural network. IEEE Access, 7, 93079-93089.

19. Cowie, R., Douglas-Cowie, E., Savvidou, S., et al. (2000). 'Feeltrace': An instrument for recording perceived emotion in real time. In ISCA tutorial and research workshop (ITRW) on speech and emotion.

20. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2020). Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(5), 1073-1081.