Emotion Recognition: A Survey of Face, Speech, and Multimodal Approaches

www.surveyx.cn

Abstract

The survey paper explores the interconnected fields of emotion recognition, focusing on face, speech, and multimodal emotion recognition, alongside affective computing and audiovisual emotion recognition. These domains, integral to artificial intelligence and machine learning, aim to accurately interpret human emotions through modalities such as facial expressions, vocal tones, and physiological signals. The survey highlights the significance of emotion recognition in enhancing human-computer interaction, with applications extending to healthcare, marketing, and mental health diagnostics. It underscores the interdisciplinary nature of the field, involving psychology, computer science, and natural language processing, which enriches the theoretical framework and drives methodological innovations. The paper delves into the historical development of emotion recognition, transitioning from unimodal to multimodal approaches, and emphasizes the role of advanced techniques like deep learning and meta-learning. Key challenges include integrating diverse data sources and addressing biases in model performance. The survey also examines practical applications in domains such as social robotics and HCI, while addressing ethical concerns related to privacy and bias. The conclusion highlights ongoing advancements in multimodal fusion strategies and the potential of novel modalities, advocating for future research to enhance dataset diversity and model interpretability. By refining methodologies and exploring emerging trends, the field promises to develop systems that are more accurate, reliable, and ethically responsible, capable of interpreting complex human emotions across various contexts.

1 Introduction

1.1 Significance of Emotion Recognition

Emotion recognition is fundamental to artificial intelligence (AI) and machine learning, enhancing human-computer interaction by enabling systems to perceive and respond to human emotions [1]. This capability is critical in empathetic domains like healthcare, where AI can detect patient emotions, facilitating early mental health interventions [2]. For older adults, emotion recognition technologies are essential for effective interactions with virtual coaches, requiring tailored approaches to meet their specific needs [3].

The transformative potential of emotion recognition extends across various sectors, including marketing, gaming, and surveillance [4]. In marketing, understanding consumer emotions can refine strategies, while in gaming, it enhances user engagement by adapting to players' emotional states. In surveillance, analyzing emotional cues aids in threat identification [1]. The complexity of human emotions demands advanced methodologies like meta-learning to improve the adaptability and performance of recognition systems [5].

Moreover, emotion recognition is vital for interpersonal emotional well-being. Systems designed to recognize emotions in couples' interactions can promote emotional health through timely interventions

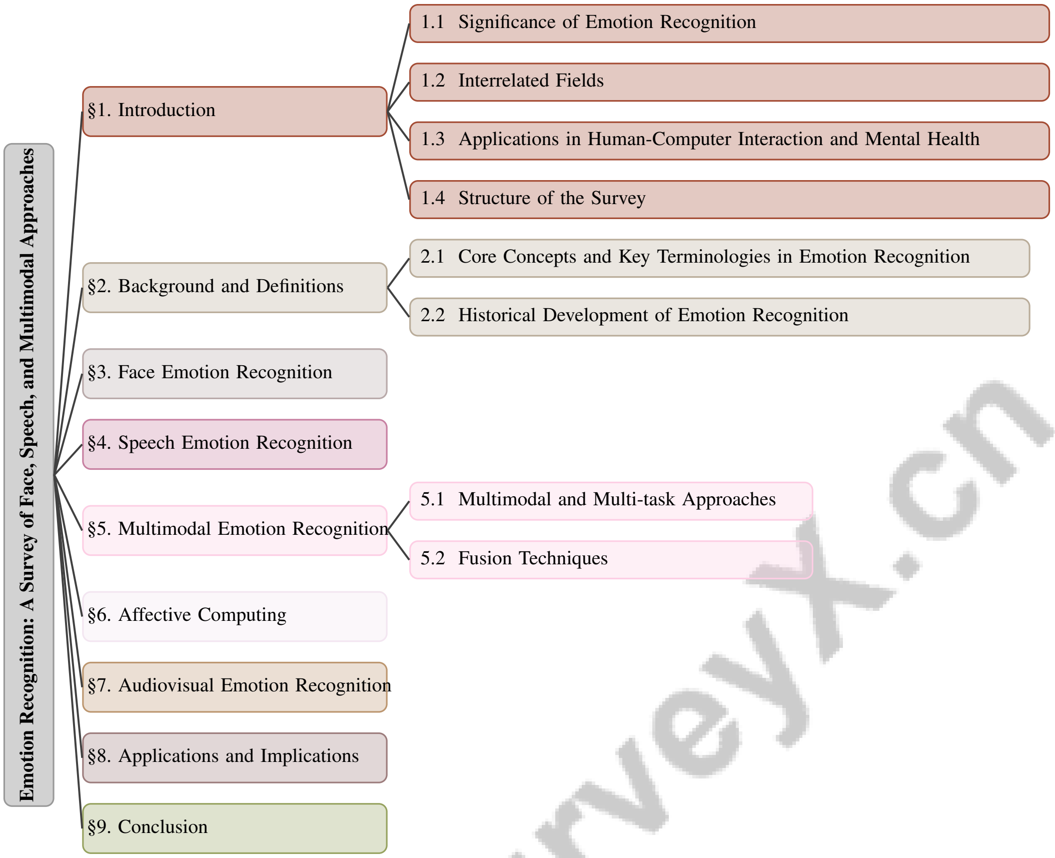

Emotion Recognition: A Survey of Face, Speech, and Multimodal Approaches
§1. Introduction
1.1 Significance of Emotion Recognition
1.2 Interrelated Fields
1.3 Applications in Human-Computer Interaction and Mental Health
1.4 Structure of the Survey
§2. Background and Definitions
2.1 Core Concepts and Key Terminologies in Emotion Recognition
2.2 Historical Development of Emotion Recognition
§3. Face Emotion Recognition
§4. Speech Emotion Recognition
§5. Multimodal Emotion Recognition
5.1 Multimodal and Multi-task Approaches
5.2 Fusion Techniques
§6. Affective Computing
§7. Audiovisual Emotion Recognition
§8. Applications and Implications
§9. Conclusion

Figure 1: chapter structure

[6]. Challenges in automatic emotion recognition, particularly in enhancing accuracy through diverse feature integration, underscore the need for robust AI capable of navigating human emotional intricacies [7]. The field's expansion into text-based recognition further highlights its importance, bridging a knowledge gap in understanding emotions expressed through text [8].

As AI and machine learning technologies evolve, accurately recognizing and interpreting emotions will remain crucial for their successful integration into daily life. Ongoing research and development in this area promise to enhance human-machine interaction, making systems more intuitive and responsive to users' emotional needs [9].

1.2 Interrelated Fields

Emotion recognition is an interdisciplinary field intersecting psychology, computer science, and natural language processing (NLP), each offering unique insights and methodologies. The integration with psychological frameworks is essential for understanding emotional expressions, particularly in contexts like pain, where emotional responses are closely linked to psychological states [10]. This collaboration facilitates the development of structured corpora capturing emotional narratives, enhancing the classification and understanding of complex emotions through components such as BEHAVIOR, FEELING, THINKING, and TERRITORY [11].

In computer science, emotion recognition has advanced significantly through AI technologies, particularly with real-time emotion detection systems that utilize machine learning and affective computing [4]. These advancements have led to complex emotion recognition systems (CERS) that integrate facial expressions, EEG, and ECG signals, providing a comprehensive understanding of human emotions beyond basic recognition [5]. The synergy between psychological insights and computational frameworks is further evident in multimodal systems, which enhance detection accuracy by combining acoustic, lexical, and visual modalities [6].

AI-generated, for reference only.

NLP plays a crucial role in deciphering emotions conveyed through text, particularly in understanding the causes of emotions in conversations, which is vital for applications in human-computer interaction and psychotherapy [12]. Sophisticated models capable of decoding fine-grained emotional categories from language enrich computational models with psychological insights [13].

The integration of emotion recognition with computer vision and multimodal analysis is essential for capturing the full spectrum of human emotions. This comprehensive approach involves fusing text, audio, and video data, facilitating a deeper understanding of the dynamic and context-dependent nature of emotions [14]. Datasets such as FindingEmo, capturing emotional content in social settings, exemplify the confluence of AI and psychology, providing rich data to improve recognition systems [15].

The intersection of emotion recognition with these diverse fields not only enriches theoretical frameworks but also drives innovation in methodologies and applications, paving the way for more accurate and context-aware emotion recognition technologies.

1.3 Applications in Human-Computer Interaction and Mental Health

Emotion recognition technologies are integral to advancing human-computer interaction (HCI), enhancing systems' ability to interpret and respond to users' emotional states, thereby improving user experience and satisfaction [1]. The integration of diverse data sources, including audio, visual, and textual inputs, is crucial for capturing the complexity of real-world emotional expressions, which are inherently multimodal [16]. This multimodal approach, exemplified by systems like the Emolysis toolkit, ensures contextual alignment across modalities, facilitating accurate emotion recognition and interaction capabilities.

Incorporating non-verbal cues, such as body gestures, further enhances interaction capabilities by providing a more holistic understanding of user emotions [17]. The combination of visual, audio, and linguistic information yields improved results compared to single-source approaches, underscoring the effectiveness of multimodal integration in affective state detection, particularly in complex environments like multi-party conversations [18].

In mental health diagnostics, emotion recognition systems offer transformative potential by enabling precise identification of emotional cues through facial expressions, speech, and physiological signals [19]. Multimodal frameworks that integrate diverse data sources, such as speech and text, have demonstrated enhanced reliability and accuracy in diagnostic tools, evidenced by their state-of-the-art performance [19]. The use of semi-supervised multimodal models further highlights the potential of these technologies in mental health applications by effectively utilizing unlabeled data to improve recognition performance [20].

Wearable technologies, such as smartwatches equipped with sensors, represent significant advancements in enhancing HCI by predicting user emotions through real-time physiological data processing. These innovations improve HCI and offer promising applications in mental health by enabling continuous monitoring of emotional states, crucial for early intervention and support [1]. Additionally, the implementation of artificial empathy in AI systems enhances therapeutic contexts by allowing for more effective emotional analysis, demonstrating practical applications of emotion recognition in therapy [10].

Exploring contextual factors affecting emotion recognition systems underscores the importance of understanding the nuances of emotional expressions in various scenarios. This contextual awareness is vital for developing systems capable of accurately interpreting emotions in real-world settings, enhancing their applicability in both HCI and mental health diagnostics. The ongoing evolution of emotion recognition technologies promises to transform human-computer interaction and mental health landscapes, offering innovative solutions that cater to users' and patients' emotional needs [3].

1.4 Structure of the Survey

This survey is structured to comprehensively explore emotion recognition technologies, beginning with an introduction that outlines the significance and interdisciplinary nature of the field, along with its applications in human-computer interaction and mental health. The introduction sets the stage for subsequent sections by highlighting the transformative potential of emotion recognition across various domains.

The survey progresses with a detailed background section elucidating core concepts and key terminologies, offering a historical perspective on the evolution of emotion recognition methodologies. This foundational knowledge is critical for understanding the advancements discussed in later sections.

Following the background, the survey delves into specific modalities of emotion recognition, starting with face emotion recognition. This section examines the role of computer vision and deep learning in analyzing facial expressions, addressing the challenges and opportunities inherent in this research area. Subsequently, the survey explores speech emotion recognition, focusing on audio signal processing techniques and recent advancements propelling the field forward.

The discussion then transitions to multimodal emotion recognition, highlighting the integration of multiple data sources to enhance recognition accuracy. This section emphasizes the benefits and challenges of multimodal and multi-task approaches, as well as the fusion techniques employed to effectively combine different modalities.

Affective computing is explored next, providing an overview of systems that interpret and respond to human emotions, followed by a discussion of their applications across various domains. This section underscores the practical implications of affective computing in enhancing emotion recognition technologies.

The survey comprehensively examines audiovisual emotion recognition by reviewing cutting-edge methodologies and recent advancements utilizing both audio and visual data to enhance emotion detection. It highlights the significance of affective video content analysis (AVCA) as a vital area within affective computing, focusing on the challenges of video feature extraction, expression subjectivity, and multimodal feature fusion. The survey discusses various emotion representation models and evaluates both unimodal approaches, such as facial expression and posture recognition, and multimodal strategies incorporating innovative fusion techniques like attention mechanisms and factorized bilinear pooling. Additionally, it outlines performance evaluation standards and identifies future research directions, including applications in human-computer interaction and emotional intelligence [21, 22].

The survey concludes with a comprehensive discussion on the applications and implications of emotion recognition technologies, highlighting their practical uses across various fields such as healthcare—where they can aid in early detection of mental health issues—and education—by helping to identify student frustrations. It also addresses the ethical considerations associated with deploying these technologies, including privacy concerns and the potential for misuse, thereby emphasizing the need for responsible implementation to enhance human-machine interactions while safeguarding individual rights [23, 2, 24, 25, 6]. This comprehensive structure ensures a thorough understanding of the current state and future directions of emotion recognition research.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Core Concepts and Key Terminologies in Emotion Recognition

Emotion recognition involves identifying and analyzing human emotions through modalities such as facial expressions, vocal tones, textual content, and physiological signals. This section details key terminologies, focusing on face emotion recognition (FER), speech emotion recognition (SER), and multimodal emotion recognition, each offering distinct methodologies and insights into emotional expressions.

Face emotion recognition (FER) utilizes computer vision techniques to analyze facial features, particularly the eyes and mouth, categorizing emotions into anger, disgust, fear, happiness, sadness, surprise, and neutrality [4]. The variability of expressions and potential for multiple emotion labels per image pose challenges, addressed by dynamic models analyzing image sequences [17].

Speech emotion recognition (SER) leverages vocal intonations and acoustic features for emotion classification using machine learning [26]. Challenges include robust feature extraction from speech signals, especially in low-resourced languages with limited annotated data [27], and operation in noisy environments complicating recognition [28].

Multimodal emotion recognition enhances classification accuracy by integrating data sources like facial expressions, speech, and physiological signals [29]. This approach captures temporal dependencies and correlations across modalities, addressing continuous emotional state recognition [16]. Combining text, audio, and video offers a comprehensive view of emotional behavior [9], though ineffective inter-modal relationship capture remains a challenge due to inadequate fusion techniques [30].

Affective computing expands emotion recognition by incorporating physiological signals and focusing on emotional dimensions like valence and arousal [31]. Traditional methods relying on manually extracted features face challenges necessitating sophisticated approaches using both labeled and unlabeled data. Key obstacles include privacy violations, lack of consent, data bias, and potential emotional manipulation [1].

Integrating cognitive psychological theories with advanced computational techniques is crucial for improving emotion recognition technologies. Multimodal emotion recognition utilizing personal, scene, and semantic features is vital for accuracy enhancement [7]. Recognizing human emotions in text remains challenging due to vagueness and context-dependence [8].

In healthcare surveillance systems, accurate emotion recognition through various modalities is critical, necessitating advanced methodologies for enhanced performance [2]. The field is characterized by diverse methodologies and challenges, with research aimed at improving detection accuracy and robustness across modalities.

2.2 Historical Development of Emotion Recognition

Emotion recognition technologies have evolved significantly from unimodal to multimodal approaches, driven by the need to enhance accuracy and reliability across contexts. Early systems focused on single modalities like facial expressions or speech, facing challenges due to variability from speaker characteristics and environmental noise [26]. The limited accuracy in uncontrolled environments highlighted the need for robust methodologies [4].

The shift to multimodal emotion recognition marked a pivotal advancement, integrating multiple data sources to capture emotional complexity. This transition addressed single-signal method limitations, with physiological data and facial expressions integration improving accuracy [32]. Methodologies evolved to include data collection, preprocessing, model training using machine learning and deep learning, and evaluation metrics, reflecting a structured approach [5].

Advancements focus on overcoming challenges related to feature representation and model performance degradation, particularly in multimodal fusion [29]. Ensemble and hybrid methods, combining keyword-based, learning-based, and hybrid approaches, enhance emotion classification effectiveness and accuracy [8].

Despite advancements, challenges persist, such as the scarcity of labeled multimodal datasets and limitations in speech emotion recognition systems, especially in low-resourced languages [27]. The absence of comprehensive benchmarks for evaluating models and augmentation strategies across datasets remains critical, impacting performance and generalizability [33].

The historical trajectory of emotion recognition technologies underscores the importance of integrating diverse data sources and refining methodologies to capture human emotion complexity. As research progresses, the focus remains on developing systems that are accurate and adaptable to the dynamic nature of emotional expressions across contexts and individuals [9].

In recent years, the field of face emotion recognition (FER) has garnered significant attention due to its potential applications in various domains, including human-computer interaction and affective computing. To elucidate the complexities of this domain, Figure 2 illustrates the hierarchical structure of FER, highlighting the role of computer vision and deep learning. This figure categorizes key aspects such as convolutional neural networks (CNNs), advanced deep learning architectures, and affective video content analysis. Additionally, it details the challenges and opportunities in contextual integration and multimodal data fusion, providing a comprehensive overview of the current landscape in FER research. By integrating this visual representation, we can better understand the intricate relationships and evolving trends that characterize the advancement of FER technologies.

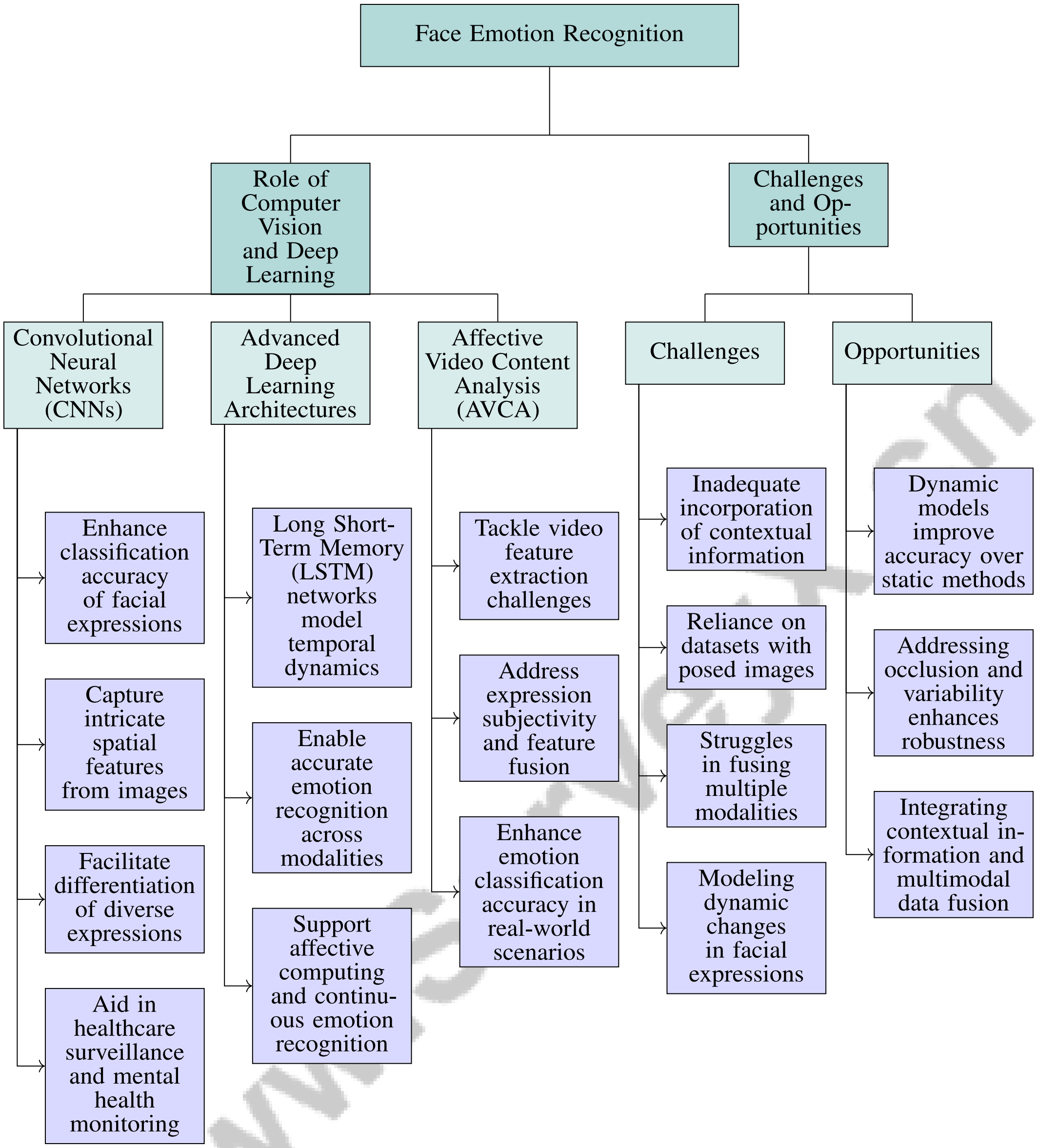

Face Emotion Recognition
Role of Computer Vision and Deep Learning
Challenges and Opportunities
Convolutional Neural Networks (CNNs)
Advanced Deep Learning Architectures
Affective Video Content Analysis (AVCA)
Challenges
Opportunities
Enhance classification accuracy of facial expressions
Capture intricate spatial features from images
Facilitate differentiation of diverse expressions
Aid in healthcare surveillance and mental health monitoring
Long Short-Term Memory (LSTM) networks model temporal dynamics
Enable accurate emotion recognition across modalities
Support affective computing and continuous emotion recognition
Tackle video feature extraction challenges
Address expression subjectivity and feature fusion
Enhance emotion classification accuracy in real-world scenarios
Inadequate incorporation of contextual information
Reliance on datasets with posed images
Struggles in fusing multiple modalities
Modeling dynamic changes in facial expressions
Dynamic models improve accuracy over static methods
Addressing occlusion and variability enhances robustness
Integrating contextual information and multimodal data fusion

Figure 2: This figure illustrates the hierarchical structure of face emotion recognition (FER), highlighting the role of computer vision and deep learning, and detailing the challenges and opportunities in advancing FER research. It categorizes the key aspects of CNNs, advanced deep learning architectures, and affective video content analysis, alongside the challenges and opportunities in contextual integration and multimodal data fusion.

3 Face Emotion Recognition

3.1 Role of Computer Vision and Deep Learning

The integration of computer vision and deep learning has significantly advanced face emotion recognition (FER), with Convolutional Neural Networks (CNNs) playing a pivotal role in enhancing the classification accuracy of facial expressions by capturing intricate spatial features from images [4]. These networks extract subtle emotional cues, facilitating the differentiation of diverse expressions. In healthcare surveillance, deep learning techniques have improved emotion recognition accuracy, aiding mental health monitoring by accurately identifying patient emotions [2]. CNNs, with their

AI-generated, for reference only.

ability to process large datasets and learn complex features, are essential for developing reliable real-time systems.

The evolution of emotion recognition technologies is characterized by frameworks that categorize research into machine learning, deep learning, and meta-learning, highlighting deep learning's contribution to advancing complex emotion recognition systems (CERS) [5]. Beyond static analysis, advanced deep learning architectures like Long Short-Term Memory (LSTM) networks model temporal dynamics of facial expressions, providing nuanced understanding of emotional transitions over time. LSTMs capture long-range dependencies, significantly contributing to affective computing and enabling accurate emotion recognition across modalities, including visual and auditory signals [34, 35]. Temporal modeling is crucial for continuous emotion recognition applications, where the dynamic nature of human emotions is a significant factor.

The synergy between computer vision and deep learning in FER underscores their transformative impact on emotion recognition. Researchers employ sophisticated algorithms and models, encompassing unimodal and multimodal approaches, to advance human emotion understanding through facial analysis. Recent studies emphasize integrating advanced techniques in affective video content analysis (AVCA) to tackle challenges such as video feature extraction, expression subjectivity, and feature fusion. The use of deep learning and state-of-the-art visual and temporal networks enhances emotion classification accuracy in real-world scenarios. Investigations into specific facial features, particularly the eyes and mouth, further elucidate their influence on emotion recognition, collectively pushing the boundaries of automated emotion detection for improved human-computer interaction and emotional intelligence applications [23, 36, 37, 22].

3.2 Challenges and Opportunities

Face emotion recognition (FER) research encounters significant challenges that impede its effectiveness in real-world applications. A primary issue is the inadequate incorporation of contextual information, which limits the accurate interpretation of emotions based solely on facial expressions [7]. This limitation is compounded by reliance on datasets with posed images captured in controlled environments, resulting in unreliable predictions in dynamic, real-world scenarios [4].

Current FER methodologies often struggle to effectively fuse information from multiple modalities, such as visual, audio, and linguistic cues, constraining continuous emotion recognition systems crucial for capturing the fluid nature of emotional expressions [16]. Modeling dynamic changes in facial expressions presents a challenge, as most state-of-the-art neural networks inadequately account for uncertainty, making them less suited for nuanced emotion recognition tasks [38].

Despite these challenges, there are promising opportunities for advancement in FER research. Techniques that capture dynamic changes in facial expressions, such as dynamic models, have demonstrated improved accuracy over static methods [17]. Approaches addressing occlusion and variability in facial expressions can significantly enhance the robustness and applicability of FER systems in diverse environments.

The ongoing development of FER technologies must focus on integrating contextual information and improving multimodal data fusion to overcome existing limitations. By addressing inherent challenges and leveraging opportunities for methodological advancements, FER research can progress toward developing precise and dependable emotion recognition systems. These systems will be better equipped to operate in real-world scenarios, enhancing human-machine interactions across various applications, including healthcare, education, and conversational AI. Recent studies underscore the importance of integrating state-of-the-art deep learning techniques and multimodal data, such as visual and audio inputs, to improve emotion classification accuracy, particularly in dynamic environments where emotions are expressed in nuanced ways [39, 37, 24].

4 Speech Emotion Recognition

4.1 Feature Extraction and Analysis

Speech emotion recognition (SER) relies heavily on the extraction and analysis of vocal intonations and acoustic features, such as pitch, intensity, and spectral properties, to discern emotional states [26]. Traditional methods, which focus solely on these features, often fall short in capturing the full complexity of emotions, necessitating more advanced approaches. Recent developments have introduced innovative feature extraction techniques that enhance the capabilities of emotion recognition systems.

Neural Automatic Speech Recognition (ASR) systems have proven effective in extracting features, enabling the prediction of valence and arousal values, thus highlighting the potential of neural networks in improving emotion recognition [40]. Concurrently, Support Vector Machines (SVM) have shown promise in emotion classification based on speech signal features, improving detection accuracy [26]. The creation of a natural code-mixed speech emotion dataset, which incorporates word-level valence, arousal, and dominance (VAD) values, illustrates the advantages of integrating diverse data sources to enhance recognition performance [41].

Data augmentation strategies have been pivotal in improving model performance, as evidenced by benchmarks demonstrating the interaction between model architectures and augmentation techniques, underscoring the necessity of robust data preprocessing in SER [33]. A two-stage process that separately processes acoustic and text features before integrating them in an SVM for final predictions further exemplifies the effectiveness of combining multiple feature types for enhanced emotion recognition [42].

Moreover, the development of language-specific models using multiple pre-trained speech models and a multi-domain training approach with a multi-gating mechanism addresses the challenges of multilingual speech emotion recognition, emphasizing the need for tailored approaches in diverse linguistic contexts [27]. These advancements reflect ongoing efforts to refine feature extraction methodologies, aiming to improve the accuracy and reliability of emotion recognition systems across various domains. By leveraging advanced neural architectures and innovative extraction techniques, researchers continue to push the boundaries of speech emotion recognition.

4.2 Recent Advancements and Future Directions

The field of speech emotion recognition (SER) has seen significant advancements through the integration of innovative methodologies and models, enhancing the precision and robustness of emotion detection systems. Mel-frequency cepstral coefficients (MFCCs) have outperformed linear predictive coding coefficients (LPCCs) in feature extraction, achieving high recognition rates with LDC datasets, which underscores the importance of effective feature selection in capturing emotional nuances [26].

Multimodal approaches, such as the MDRE model, have demonstrated superior performance in emotion classification, achieving higher accuracies and reducing misclassification of neutral emotions compared to audio-only models, illustrating the potential of integrating multiple modalities to enhance emotion recognition capabilities. Additionally, aligning emotional information in multimodal fusion has significantly improved performance, as evidenced by experiments on the IEMOCAP corpus [29].

Data augmentation techniques, including speed perturbation and copy-paste augmentation, have been beneficial in improving model performance across various architectures, particularly convolutional models [33]. These strategies reflect ongoing efforts to refine data preprocessing techniques to bolster the robustness of emotion recognition systems.

Advanced neural network architectures, such as fine-tuning Wav2vec 2.0 for emotion-dependent phonetic units, have significantly enhanced recognition accuracy, especially with broad phonetic classes [40]. The introduction of a two-stage late-fusion approach has further improved dimensional emotion recognition performance compared to single-modality methods [42].

Despite these advancements, challenges remain in enhancing the robustness of emotion classification in noisy environments and addressing the complexities of real-world emotional expressions. Future research should focus on optimizing attention mechanisms and integrating additional modalities to further enhance emotion recognition capabilities [43]. Expanding datasets to include a wider variety of speakers and emotional contexts, alongside exploring multimodal approaches that integrate speech with other forms of emotional expression, are crucial areas for future exploration. By addressing these challenges and leveraging the latest advancements, SER research can continue to evolve, developing more accurate and reliable systems across diverse contexts and modalities.

5 Multimodal Emotion Recognition

Multimodal emotion recognition is crucial for capturing the complexity of human emotions through the integration of various data modalities. This section examines innovative approaches, particularly focusing on multimodal and multi-task strategies that enhance the robustness of emotion detection systems and address the challenges associated with fusing diverse modalities. The subsequent subsection explores these strategies in detail, examining their contribution to advancing emotion recognition technologies.

5.1 Multimodal and Multi-task Approaches

Multimodal emotion recognition advances affective computing by integrating auditory, visual, and textual signals to capture emotional complexity, leveraging the complementary nature of these modalities to enhance detection accuracy and robustness [3]. This integration improves predictive performance and enables sophisticated emotion recognition [44]. However, challenges arise in effectively fusing diverse modalities while managing asynchrony and redundancy, compounded by dataset size and variations in emotional expression due to factors such as age and gender [2]. Approaches like FoalNet emphasize aligning emotional information across modalities before fusion to improve recognition performance [29]. Transformer-based models, such as the modified HTS-AT combined with R(2+1)D CNN for video analysis, capture interactions between audio and visual features, addressing challenges in emotion recognition from limited labeled datasets [45].

Multi-task learning frameworks predict stress and emotion simultaneously, using inter-modal attention mechanisms to weigh different modalities [46]. The DFF-ATMF model integrates multi-feature and multi-modality fusion, leveraging complementary information from audio and text modalities to improve sentiment analysis accuracy [47]. The joint cross-attention mechanism captures complementary relationships between audio and visual features, enhancing emotion recognition [43].

Incorporating physiological signals like ECG and EDA with traditional modalities such as audio and video has proven effective in improving emotion recognition accuracy [31]. Integrating scene and semantic features with personal characteristics further enhances multimodal emotion recognition, offering a holistic view of emotional expressions [7].

Recent advancements include systems like Emolysis, a toolkit for multimodal emotion analysis that processes synchronized video input to map group-level emotion, valence, and arousal. The Human-In-The-Loop (HITL) framework integrates human cognitive processes in multimodal pain recognition, highlighting the potential of hybrid systems incorporating human insights [3]. Future research should continue exploring multimodal approaches to improve automatic emotion annotation and address existing limitations [32]. By overcoming data integration challenges and enhancing model generalization, researchers can develop more accurate systems capable of interpreting human emotions across diverse contexts.

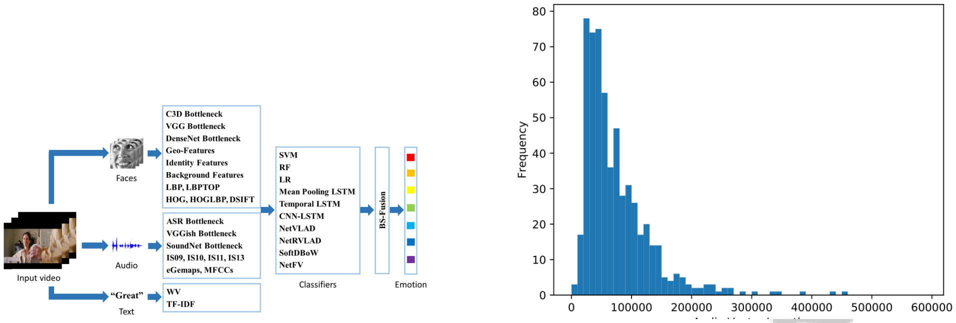
5.2 Fusion Techniques

The integration of diverse modalities in emotion recognition is key to enhancing detection accuracy and robustness. Various fusion methodologies combine data from different sources, each contributing uniquely to recognition performance. Cross-attention mechanisms effectively model intra- and inter-modal relationships, as demonstrated by the joint cross-attention model, which captures complementary relationships between audio and visual features to enhance emotion recognition [43].

The DFF-ATMF model uses parallel branches for audio and text modalities with attention mechanisms to improve feature representation and fusion, showing strong performance in multimodal sentiment analysis [47]. FoalNet employs cross-attention to integrate audio and video emotional information, enhancing modal fusion quality and overall recognition accuracy [29].

Multimodal emotion recognition methods that fuse audio and visual data significantly enhance recognition performance. Palmero et al. demonstrate the effectiveness of leveraging the complementary nature of different modalities [3]. Adaptive fusion techniques, such as those used in the AMuSE framework, enhance performance by utilizing instance-specific data integration, allowing more tailored emotion detection [18].

Future research should explore richer feature sets and sophisticated fusion techniques for combining modalities. Continuous refinement of recognition methodologies and investigation of deep learning trends are essential for improving system accuracy and efficiency. This progress is crucial for nuanced interpretations of human emotions across contexts, such as healthcare for psychological analysis and education for identifying student frustrations. Recent advancements, including Transformer-based models and data augmentation techniques, enhance emotion detection from textual data and improve robustness against noise in speech, crucial for emotion-aware dialogues and addressing challenges in emotion recognition in conversation [33, 23, 48, 24].



(a) A diagram illustrating a multi-modal emotion recognition system[49]

(b) The histogram shows the frequency distribution of audio vector lengths.[47]

Figure 3: Examples of Fusion Techniques

In Figure 3, the integration of multiple modalities, such as visual, audio, and textual data, enhances emotion recognition system performance. A diagram outlines a multi-modal emotion recognition system that processes input video through pre-processing steps to extract relevant features, including face analysis and audio processing. This system leverages the strengths of different data types to improve classification accuracy. A histogram shows the frequency distribution of audio vector lengths, highlighting data variability and range, with a peak around 700,000 followed by a drop-off, indicating data concentration in a specific range. These visual representations underscore the importance of fusion techniques in multi-modal emotion recognition, illustrating how various data streams are integrated to enhance the system's ability to discern emotional cues [49, 47].

6 Affective Computing

6.1 Overview of Affective Computing

Affective computing is a vital domain within artificial intelligence, focusing on systems that interpret and respond to human emotions through diverse data modalities. This field utilizes multimodal data, including physiological signals and facial expressions, to enhance emotion recognition accuracy [50]. The integration of 3D facial data with physiological signals has shown superior performance in emotion recognition tasks.

Advanced machine learning techniques, such as Bayesian Neural Networks, are critical in affective computing for incorporating uncertainty in predictions, thereby handling ambiguous data more effectively [38]. Additionally, ensemble methods combining transformer models for topic classification with SVM and Random Forests for emotional state classification have improved the interpretative capabilities of these systems [51].

Co-attention mechanisms further enhance affective computing by leveraging the strengths of multiple modalities while mitigating individual limitations, leading to a nuanced understanding of emotional states [16]. These systems are instrumental in recognizing pain and expressing empathy, deepening the understanding of human emotions [10]. Research on emotional interactions, especially within couples, has provided foundational insights into emotional exchanges, advancing emotion recognition capabilities [6].

Benchmarking frameworks are crucial for evaluating emotion recognition systems, facilitating the extraction of emotion-cause pairs from conversations and advancing research in emotion cause analysis [52]. Recent advancements underscore the importance of multimodal approaches that integrate physical data, such as textual, audio, and visual inputs, with physiological signals like EEG and ECG. This fusion enhances the accuracy of emotion recognition, benefiting applications across human-computer interaction, entertainment, education, and driving safety [53, 54]. Continued refinement of these methodologies and exploration of emerging trends promise further enhancements in affective computing systems, leading to more nuanced interpretations of human emotions across various contexts.

6.2 Applications of Affective Computing

Affective computing is transforming diverse domains by fundamentally altering how systems interpret and respond to human emotions. In healthcare, these systems enhance patient monitoring and mental health diagnostics by leveraging multimodal data, such as facial expressions and physiological signals, for accurate emotion assessments critical for early intervention and treatment [2]. Future research should aim to develop robust systems capable of generalizing across various environments and emotional expressions, alongside exploring novel fusion techniques for audio-visual recognition.

In human-computer interaction (HCI), affective computing enhances user experience by enabling systems to dynamically adjust to users' emotional states through emotion recognition and sentiment analysis, utilizing both unimodal and multimodal data encompassing textual, audio, visual, and physiological signals. This adaptability improves user engagement and broadens the applicability of affective computing in entertainment, education, and safe driving [54, 22]. Virtual assistants and social robots benefit from enhanced emotion recognition capabilities, leading to more intuitive and empathetic interactions.

In education, affective computing employs advanced emotion recognition technologies to assess student engagement and customize educational content. By using multimodal approaches, such as analyzing facial expressions and textual sentiment, educators gain deeper insights into students' emotional states, facilitating more effective teaching strategies and responsive learning environments [13, 23, 24, 22, 8]. This integration enhances the student experience and promotes emotional intelligence in educational contexts.

Affective computing also plays a crucial role in the entertainment and gaming industries, where systems adapt to players' emotional states, enhancing immersive experiences. By dynamically adjusting game scenarios based on players' emotions, these systems improve engagement and deliver tailored experiences, using advanced emotion recognition techniques that analyze behavioral, cognitive, and contextual components of emotions [55, 22, 11].

Despite advancements in emotion detection from non-textual sources, text-based emotion detection remains an emerging field with significant potential for improvement [8]. Investigating the roles of specific facial features and their implications for computer vision algorithms and therapeutic interventions presents promising research avenues [36].

In speech emotion recognition (SER), establishing comprehensive benchmarks is crucial for advancing the field by providing thorough evaluations of both traditional and modern feature extraction techniques [56]. These benchmarks are essential for enhancing the accuracy of emotion recognition systems across various applications.

Affective computing is increasingly shaping fields such as human-computer interaction, entertainment, education, and safety applications by offering cutting-edge solutions that improve human engagement with machines. Recent advancements in emotion recognition and sentiment analysis leverage both unimodal and multimodal data, including physical inputs like text, audio, and visual signals, alongside physiological indicators such as EEG and ECG. By integrating these data types, affective computing enhances the accuracy of emotional state recognition, addressing challenges like interpreting concealed emotions through conventional physical cues. This systematic review highlights key developments, emotion models, and databases in affective computing, emphasizing its broad applicability and potential for future innovations, including improved fusion strategies and the creation of benchmark datasets [57, 54]. By focusing on refining methodologies and exploring new applications, researchers can further leverage the potential of affective computing to develop systems that are more responsive and attuned to human emotions.

7 Audiovisual Emotion Recognition

7.1 State-of-the-Art Approaches and Achievements

Recent progress in audiovisual emotion recognition underscores the effectiveness of integrating audio and visual data to enhance system accuracy and robustness. For instance, Vielzeuf et al.'s model achieved a leading accuracy of 60.64

The incorporation of attention mechanisms has further improved the performance of audiovisual emotion recognition systems. By focusing on the most salient features within audio and visual streams, these mechanisms enhance the ability to detect subtle emotional expressions, crucial in real-world scenarios where emotions are conveyed through complex interactions among various modalities, including text, audio, and visual cues [13, 23, 58]. Despite technological advancements, challenges remain in ensuring that emotion recognition systems generalize effectively across diverse environments and populations, particularly in applications such as healthcare, education, and customer support. Factors such as background noise, language variability, and diverse emotional expressions complicate this task [41, 23, 59, 24]. Future research should focus on refining fusion techniques for better integration of audio and visual data and exploring unsupervised and semi-supervised learning methods to address the limitations posed by scarce labeled datasets. Continued innovation and methodological refinement are crucial for advancing audiovisual emotion recognition, ultimately leading to more effective and reliable emotion detection systems across various applications.

8 Applications and Implications

8.1 Practical Applications in Various Domains

Emotion recognition technologies are becoming increasingly significant in domains such as social robotics and human-computer interaction (HCI). In social robotics, these systems enable robots to interpret and respond to emotional cues, enhancing interaction quality and user satisfaction [60]. The ability to personalize and provide context-aware interactions is essential for engaging user experiences, with models achieving near-state-of-the-art accuracy using single datasets in real-time [4].

In HCI, emotion recognition systems support adaptive interfaces that dynamically respond to users' emotional states, fostering intuitive interactions. The integration of video and text data enhances emotion detection accuracy, underscoring the efficacy of multimodal strategies [14]. These technologies show promise in applications like mental health diagnostics, where accurate emotion detection can aid early intervention and treatment [61].

Moreover, emotion recognition is vital in scenarios requiring low latency and privacy, such as non-speech systems that operate without verbal input [35]. Recognizing emotions in natural, code-mixed speech is particularly relevant in customer care contexts [41], and implementations in voicemail applications further demonstrate their utility [26].

In the automotive sector, multimodal emotion recognition enhances in-car speech interactions, utilizing transfer learning for text-based classifiers [62]. These systems perform robustly in noisy environments, guided by benchmarks that inform future research, emphasizing model architecture and data augmentation strategies [63, 33].

Future research should integrate additional modalities, such as facial expressions and voice, while considering dynamic emotion labeling for a comprehensive understanding of emotional states. Enhancing models to withstand noisy data and exploring new applications will enable more responsive and empathetic systems across diverse fields [18].

8.2 Ethical and Societal Implications

The deployment of emotion recognition technologies raises significant ethical and societal challenges, particularly concerning privacy, bias, and model generalization across diverse contexts. Privacy issues are critical in sensitive applications involving children and older adults, necessitating ethical considerations in therapeutic settings. Edge computing offers a potential solution by keeping sensitive data localized and secure, addressing privacy concerns associated with cloud-based systems [1].

AI-generated, for reference only.

Bias in emotion recognition systems is a major concern, often resulting from data scarcity and imbalances in minority emotion categories, which can distort model performance and lead to unfair outcomes. The reliance on specific lexicons and datasets, such as NRC-VAD and IEMOCAP, may introduce cultural biases, affecting model generalizability across different contexts. Addressing biases related to race and culture is essential for developing inclusive and representative datasets [31].

Challenges in accurately detecting emotions under variable conditions, such as poor audio quality and transcription errors due to code-switching, highlight the limitations of current methodologies [41]. The complexity of multimodal data and potential biases in emotion categories can hinder model generalization and performance, necessitating improved experimental protocols and comprehensive datasets that reflect real-world complexities [5].

Current studies often face limitations, including small datasets for complex emotions and the need for extensive training data due to the increased number of emotion-dependent models [40]. The dependency on high-quality text transcriptions and challenges posed by linguistic variations in emotional expression further complicate these issues [42].

To address these ethical and societal challenges, future research should focus on developing inclusive and adaptive models that generalize across diverse contexts while ensuring privacy and fairness. Enhancing the reliability and ethical deployment of emotion recognition systems will facilitate their responsible integration into applications such as healthcare for psychological analysis and education for understanding student emotions, ultimately improving emotional well-being across various settings. These advancements are crucial for tackling current challenges in emotion recognition, particularly in analyzing real-world conversational data and gaining insights into individual emotional states, especially in intimate relationships [6, 24].

9 Conclusion

Advancements in emotion recognition have been significantly propelled by the integration of multimodal approaches and sophisticated computational techniques. These innovations have enhanced the precision and resilience of emotion detection systems, adeptly capturing the intricacies of human emotions through diverse data modalities such as facial expressions, vocal tones, and physiological signals. The efficacy of multimodal fusion strategies, particularly in complex environments, underscores the importance of synthesizing audio and visual data to optimize emotion recognition outcomes.

Frameworks like EMERSK and AMuSE exemplify the potential of unified systems and adaptive analyses, achieving cutting-edge performance and providing insightful visualizations of model predictions, thus setting a foundation for future explorations in multimodal systems. The exploration of novel modalities, including thermal imagery, presents promising avenues for facial emotion recognition, despite challenges related to data availability.

However, the field still faces hurdles, notably in recognizing less common emotions and ensuring model robustness across diverse datasets. Future research should focus on expanding datasets to encompass a broader range of emotions and refining model architectures to boost accuracy. The integration of scene and semantic features with individual characteristics has shown substantial improvements in emotion recognition, indicating that further investigation in this area could be fruitful. Techniques such as transfer learning and Transformers have proven effective, highlighting the critical role of multimodal approaches in enhancing recognition accuracy.

The ongoing innovation within emotion recognition research holds the potential to broaden the applicability of these technologies, facilitating the creation of more nuanced systems capable of interpreting the subtleties of human emotions in varied contexts. Future research should prioritize enhancing model interpretability and addressing ethical concerns, particularly in sensitive areas like facial emotion recognition for vulnerable populations. By advancing methodologies and exploring new trends, the field can unlock further potential in emotion recognition technologies, fostering systems that are not only more precise and reliable but also ethically sound in their application.

References

[1] Siddique Latif, Hafiz Shehbaz Ali, Muhammad Usama, Rajib Rana, Björn Schuller, and Junaid Qadir. Ai-based emotion recognition: Promise, peril, and prescriptions for prosocial path, 2022.

[2] Marwan Dhuheir, Abdullatif Albaseer, Emna Baccour, Aiman Erbad, Mohamed Abdallah, and Mounir Hamdi. Emotion recognition for healthcare surveillance systems using neural networks: A survey, 2021.

[3] Cristina Palmero, Mikel deVelasco, Mohamed Amine Hmani, Aymen Mtibaa, Leila Ben Letaifa, Pau Buch-Cardona, Raquel Justo, Terry Amorese, Eduardo González-Fraile, Begoña Fernández-Ruanova, Jofre Tenorio-Laranga, Anna Torp Johansen, Micaela Rodrigues da Silva, Liva Jenny Martinussen, Maria Stylianou Korsnes, Gennaro Cordasco, Anna Esposito, Mounim A. El-Yacoubi, Dijana Petrovska-Delacrétaz, M. Inés Torres, and Sergio Escalera. Exploring emotion expression recognition in older adults interacting with a virtual coach, 2023.

[4] Akash Saravanan, Gurudutt Perichetla, and K. S. Gayathri. Facial emotion recognition using convolutional neural networks, 2019.

[5] Javad Hassannataj Joloudari, Mohammad Maftoun, Bahareh Nakisa, Roohallah Alizadehsani, and Meisam Yadollahzadeh-Tabari. Complex emotion recognition system using basic emotions via facial expression, eeg, and ecg signals: a review, 2024.

[6] George Boateng, Elgar Fleisch, and Tobias Kowatsch. Emotion recognition among couples: A survey, 2022.

[7] Zhifeng Wang and Ramesh Sankaranarayana. Using scene and semantic features for multi-modal emotion recognition, 2023.

[8] Shiv Naresh Shivhare and Saritha Khethawat. Emotion detection from text, 2012.

[9] Minoo Shayaninasab and Bagher Babaali. Multi-modal emotion recognition by text, speech and video using pretrained transformers, 2024.

[10] Can ai detect pain and express pain empathy? a review from emotion recognition and a human-centered ai perspective.

[11] Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. Emotion recognition based on psychological components in guided narratives for emotion regulation, 2023.

[12] Fanfan Wang, Heqing Ma, Jianfei Yu, Rui Xia, and Erik Cambria. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations, 2024.

[13] Felix Casel, Amelie Heindl, and Roman Klinger. Emotion recognition under consideration of the emotion component process model, 2022.

[14] Yuntao Shou, Tao Meng, Wei Ai, Nan Yin, and Keqin Li. A comprehensive survey on multi-modal conversational emotion recognition with deep learning, 2023.

[15] Laurent Mertens, Elahe' Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. Findingemo: An image dataset for emotion recognition in the wild, 2024.

[16] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3, 2022.

[17] Nikunj Bajaj, Aurobinda Routray, and S L Happy. Dynamic model of facial expression recognition based on eigen-face approach, 2013.

[18] Naresh Kumar Devulapally, Sidharth Anand, Sreyasee Das Bhattacharjee, Junsong Yuan, and Yu-Ping Chang. Amuse: Adaptive multimodal analysis for speaker emotion recognition in group conversations, 2024.

[19] Ziyu Ma, Fuyan Ma, Bin Sun, and Shutao Li. Hybrid mutimodal fusion for dimensional emotion recognition, 2021.

[20] Jingjun Liang, Ruichen Li, and Qin Jin. Semi-supervised multi-modal emotion recognition with cross-modal distribution matching, 2020.

[21] Hengshun Zhou, Debin Meng, Yuanyuan Zhang, Xiaojiang Peng, Jun Du, Kai Wang, and Yu Qiao. Exploring emotion features and fusion strategies for audio-video emotion recognition, 2020.

[22] Junxiao Xue, Jie Wang, Xuecheng Wu, and Qian Zhang. Affective video content analysis: Decade review and new perspectives, 2024.

[23] Maude Nguyen-The, Guillaume-Alexandre Bilodeau, and Jan Rockemann. Leveraging sentiment analysis knowledge to solve emotion detection tasks, 2021.

[24] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances, 2019.

[25] Fei Ma, Yucheng Yuan, Yifan Xie, Hongwei Ren, Ivan Liu, Ying He, Fuji Ren, Fei Richard Yu, and Shiguang Ni. Generative technology for human emotion recognition: A scope review, 2024.

[26] Manas Jain, Shruthi Narayan, Pratibha Balaji, Abhijit Bhowmick, Rajesh Kumar Muthu, et al. Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*, 2020.

[27] Zihan Wang, Qi Meng, HaiFeng Lan, XinRui Zhang, KeHao Guo, and Akshat Gupta. Multilingual speech emotion recognition with multi-gating mechanism and neural architecture search, 2022.

[28] Xi Ma, Zhiyong Wu, Jia Jia, Mingxing Xu, Helen Meng, and Lianhong Cai. Study on feature subspace of archetypal emotions for speech emotion recognition, 2016.

[29] Qifei Li, Yingming Gao, Yuhua Wen, Cong Wang, and Ya Li. Enhancing modal fusion by alignment and label matching for multimodal emotion recognition, 2024.

[30] R Gnana Praveen, Eric Granger, and Patrick Cardinal. Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention, 2022.

[31] Kyle Ross, Paul Hungler, and Ali Etemad. Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data, 2020.

[32] Jing Zhang, Yong Zhang, Suhua Zhan, and Cheng Cheng. Ensemble emotion recognizing with multiple modal physiological signals, 2020.

[33] Ravi Shankar, Abdouh Harouna Kenfack, Arjun Somayazulu, and Archana Venkataraman. A comparative study of data augmentation techniques for deep learning based emotion recognition, 2022.

[34] Diogo Cortiz. Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra, 2021.

[35] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks, 2017.

[36] Martin Wegrzyn, Maria Vogt, Berna Kireclioglu, Julia Schneider, and Johanna Kissler. Mapping the emotional face. how individual face parts contribute to successful emotion recognition. *PloS one*, 12(5):e0177239, 2017.

[37] Carl Norman. Ai in pursuit of happiness, finding only sadness: Multi-modal facial emotion recognition challenge, 2019.

[38] Maryam Matin and Matias Valdenegro-Toro. Hey human, if your facial emotions are uncertain, you should use bayesian neural networks!, 2020.

[39] Rim EL Cheikh, Hélène Tran, Issam Falih, and Engelbert Mephu Nguifo. A comparative study of emotion recognition methods using facial expressions, 2022.

[40] Jiahong Yuan, Xingyu Cai, Renjie Zheng, Liang Huang, and Kenneth Church. The role of phonetic units in speech emotion recognition, 2021.

[41] N V S Abhishek and Pushpak Bhattacharyya. "we care": Improving code mixed speech emotion recognition in customer-care conversations, 2023.

[42] Bagus Tris Atmaja and Masato Akagi. Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm, 2022.

[43] R. Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro Koerich, Simon Bacon, Patrick Cardinal, and Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition, 2024.

[44] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE spoken language technology workshop (SLT)*, pages 112–118. IEEE, 2018.

[45] Ohad Cohen, Gershon Hazan, and Sharon Gannot. Multi-microphone and multi-modal emotion recognition in reverberant environment, 2024.

[46] Fei Tao, Gang Liu, and Qingen Zhao. An ensemble framework of voice-based emotion recognition system for films and tv programs, 2018.

[47] Feiyang Chen, Ziqian Luo, Yanyan Xu, and Dengfeng Ke. Complementary fusion of multi-features and multi-modalities in sentiment analysis, 2019.

[48] Rajib Rana. Emotion classification from noisy speech - a deep learning approach, 2016.

[49] Zheng Lian, Ya Li, Jianhua Tao, and Jian Huang. Investigation of multimodal features, classifiers and fusion methods for emotion recognition, 2018.

[50] Diego Fabiano, Manikandan Jaishanker, and Shaun Canavan. Impact of multiple modalities on emotion recognition: investigation into 3d facial landmarks, action units, and physiological data, 2020.

[51] Hoang Manh Hung, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee. Variants of bert, random forests and svm approach for multimodal emotion-target sub-challenge, 2020.

[52] Imen Trabelsi, Dorra Ben Ayed, and Noureddine Ellouze. Improved frame level features and svm supervectors approach for the recogniton of emotional states from speech: Application to categorical and dimensional states, 2014.

[53] Tzyy-Ping Jung, Terrence J Sejnowski, et al. Multi-modal approach for affective computing. In *2018 40th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 291–294. IEEE, 2018.

[54] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. A systematic review on affective computing: Emotion models, databases, and recent advances, 2022.

[55] Joost Broekens. Modeling the experience of emotion, 2009.

[56] Adrian Bogdan Stânea, Vlad Striletchi, Cosmin Striletchi, and Adriana Stan. An analysis of large speech models-based representations for speech emotion recognition, 2023.

[57] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022.

[58] A. Sutherland, S. Magg, C. Weber, and S. Wermter. Analyzing the influence of dataset composition for emotion recognition, 2021.

[59] Haiyang Sun, Fulin Zhang, Yingying Gao, Zheng Lian, Shilei Zhang, and Junlan Feng. Mfsn: Multi-perspective fusion search network for pre-training knowledge in speech emotion recognition, 2024.

[60] Jiehui Jia, Huan Zhang, and Jinhua Liang. Bridging discrete and continuous: A multimodal strategy for complex emotion detection, 2024.

[61] Mijanur Palash and Bharat Bhargava. Emersk – explainable multimodal emotion recognition with situational knowledge, 2023.

[62] Zexu Pan, Zhaojie Luo, Jichen Yang, and Haizhou Li. Multi-modal attention for speech emotion recognition, 2020.

[63] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.