A Systematic Literature Review of Multimodal Emotion Recognition

Yeni Dwi Rahayu1*, Lutfi Ali Muharrom2, Ika Safitri Windiarti3, Auratania Hadisah Sugianto4
1-2Department of Informatics Engineering, Faculty of Engineering, Universitas Muhammadiyah Jember, Indonesia
3Information Technology Study Programme, Universiti Muhammadiyah Malaysia

Abstract.
Purpose: This literature review aims to identify Multimodal Emotion Recognition (MER) in depth and breadth by analyzing the topics, trends, modalities, and other supporting sources discussed in research over the years and between 2010 and 2022. Based on the screening analysis, a total of 14,533 articles were analyzed to achieve this goal.
Methods: This research was conducted in 3 (three) phases, including Planning, Conducting and Reporting. The first step was defining the research objectives by searching for systematic reviews with similar topics to this study, then reviewing them to develop research questions and systematic review protocols for this study. The second stage is to collect articles according to a pre-determined protocol, selecting the articles obtained and then conducting an analysis of the filtered articles in order to answer the research questions. The final stage is to summarize the results of the analysis so new findings from this research can be reported.
Result: In general, the focus of MER research can be categorized into two issues, namely the object background and the source or modality of emotion recognition. When looking at the object background, most of the 55% to support emotion recognition with a health background, especially brain function decline, 34% based on age, 10% based on gender, 1% data collection situation and a small portion of less than 1% related to ethnic culture. In terms of the source of emotion recognition, research is divided into electromagnetic signals, voice signals, text, photo/video and the development of wearable devices. Based on the above results, there are at least 7 scientific fields that discuss MER research, namely health, psychology, electronics, grammar, communication, socio-culture and computer science.
Novelty: MER research has the potential to develop further. There are still many areas that have received less attention, while the ecosystem that uses them has grown massively. Emotion recognition modalities are numerous and diverse, but research is still focused on validating the emotions of each modality, rather than exploring the strengths of each modality to improve the quality of recognition results.

Keywords: Emotion Recognition, Modalities, Research Topics

This work is licensed under a Creative Commons Attribution 4.0 International License.

INTRODUCTION

Emotion recognition may assist humans to understand themselves, understand others, as well as improve the overall quality of life [1]. Therefore, the study of emotion recognition continues to grow significantly [2]. Emotion recognition itself is a complex phenomenon that involves psychological, physiological, and social aspects. Psychological aspects include how emotions are processed and interpreted by the brain, while physiological aspects involve the physiological changes that occur in the body when a person experiences emotions, such as heart rate and endocrine gland activity. On the other hand, the social aspect relates to the way emotions are expressed and influenced by the social and cultural environment. Therefore, to understand emotions well, a holistic understanding is required and involves various fields of science, such as psychology, neuroscience, biology, anthropology and sociology [3].

The complexity of emotion recognition demands that emotion recognition computing research also involves various aspects of modalities to improve the quality of emotion recognition. This phenomenon has been recognized by the emergence of various studies on the topic of multimodal emotion recognition [4], [5] with challenges that are still wide open. Ya Li defines the challenges of multimodal emotion recognition around datasets, variations in recognition sources (audio, image, video), cultural influences and

*Corresponding author.
Email addresses: yenidwirahayu@unmuhjember.ac.id (Rahayu), lutfi.muharom@unmuhjember.ac.id (Muharom), ika.windiarti@unam.edu.my (Windiarti), auratania036@gmail.com (Sugianto)
DOI: 10.15294/sji.v10i2.43792
optimization methods [6]. While Priyanka categorizes the modalities for recognizing emotions into three, namely psychological, behavioral and brain signals [5].

Previous literature study research has actually mapped emotion recognition which resulted in a variety of optimization methods, emotion distribution, several open datasheets, and challenges that are still wide open [7]–[10]. However, considering the complexity of emotion recognition and the increase in multimodal emotion recognition research, it is necessary to review what fields of science are involved in this emotion recognition research. This research will investigate possible fields of science and answer some research questions in the field of multimodal emotion recognition in Table 1.

Table 1. Research questions on literature review

<table>
<thead>
<tr>
<th>Number</th>
<th>P population</th>
<th>T intervention</th>
<th>C comparison</th>
<th>O outcome</th>
<th>T time</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Any science</td>
<td>Emotion recognition, multimodal emotion recognition, Multimodal Physiological Signal, Multimodal Emotion Recognition Database</td>
<td>No comparison</td>
<td>fields of science that can be involved to improve the quality of emotion recognition</td>
<td>All year</td>
<td>What fields of science are involved in emotion recognition?</td>
</tr>
<tr>
<td>2.</td>
<td>Any science</td>
<td>Emotion recognition, multimodal emotion recognition, Multimodal Physiological Signal, Multimodal Emotion Recognition Database</td>
<td>No comparison</td>
<td>the most significant journals in the multimodal emotion recognition field</td>
<td>All year</td>
<td>Which journal is the most significant multimodal emotion recognition journal?</td>
</tr>
<tr>
<td>3.</td>
<td>Any science</td>
<td>Emotion recognition, multimodal emotion recognition, Multimodal Physiological Signal, Multimodal Emotion Recognition Database</td>
<td>No comparison</td>
<td>the most active and influential researchers who contributed so much on the research area of multimodal emotion recognition</td>
<td>All year</td>
<td>Who are the most active and influential researchers in the multimodal emotion recognition field?</td>
</tr>
<tr>
<td>4.</td>
<td>Any science</td>
<td>Emotion recognition, multimodal emotion recognition, Multimodal Physiological Signal, Multimodal Emotion Recognition Database</td>
<td>No comparison</td>
<td>research topics and trends in multimodal emotion recognition</td>
<td>2019 - 2023</td>
<td>What kind of research topics are selected by researchers in the multimodal emotion recognition field?</td>
</tr>
</tbody>
</table>

Recent works that have contributed to Multimodal Emotion Recognition (MER) and Emotion recognition (ER) are discussed in this paper, not only in computing but also in other fields. With the research questions answered in this paper, we expect that the results of this study will contribute to more optimized possibilities for the development of MER research.

METHODS

This research was conducted in 3 (three) phases, including Planning, Conducting and Reporting based on original guidance proposed in [11] and some stages modified with motivation from [10], [12], [13]. The first step was defining the research objectives by searching for systematic reviews with similar topics to this study, then reviewing them to develop research questions and systematic review protocols for this study. The second stage is to collect articles according to a pre-determined protocol, selecting the articles obtained and then conducting an analysis of the filtered articles in order to answer the research questions. The final stage is to summarize the results of the analysis so new findings from this research can be reported. Conducting phase is depicted in the chart in Figure 1.
Article collection was conducted using publish or perish software by Scopus indexing source. Article collection was carried out on April 4, 2023 with several parameters with the results in Table 2.

Table 2. Collected articles

<table>
<thead>
<tr>
<th>Number</th>
<th>Keywords (search terms)</th>
<th>Indexing</th>
<th>Year</th>
<th>Number Of Articles</th>
<th>number of citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Emotion Recognition (ER)</td>
<td>Scopus</td>
<td>2019 - 2023</td>
<td>1000</td>
<td>40.381</td>
</tr>
<tr>
<td>2.</td>
<td>Emotion Recognition (ER)</td>
<td>Scopus</td>
<td>All year</td>
<td>200</td>
<td>152.398</td>
</tr>
<tr>
<td>4.</td>
<td>Multimodal Emotion Recognition (MER)</td>
<td>Scopus</td>
<td>All year</td>
<td>200</td>
<td>34.706</td>
</tr>
<tr>
<td>5.</td>
<td>Multimodal Physiological Signal (MPS)</td>
<td>Scopus</td>
<td>2019 - 2023</td>
<td>618</td>
<td>5.793</td>
</tr>
<tr>
<td>6.</td>
<td>Multimodal Physiological Signal (MPS)</td>
<td>Scopus</td>
<td>All year</td>
<td>200</td>
<td>22.041</td>
</tr>
<tr>
<td>7.</td>
<td>Multimodal Emotion Recognition Database (MERD)</td>
<td>Scopus</td>
<td>2019 - 2023</td>
<td>150</td>
<td>1.721</td>
</tr>
<tr>
<td>8.</td>
<td>Multimodal Emotion Recognition Database (MERD)</td>
<td>Scopus</td>
<td>All year</td>
<td>200</td>
<td>12.875</td>
</tr>
</tbody>
</table>

Total number of Articles 3,420
The 3,420 datasets obtained were then processed with the following steps:

1. Merging of datasets result:
   a) 4 terms in a particular year are merged, namely Emotion Recognition, Multimodal Emotion Recognition, Multimodal Physiological Signal, and Multimodal Emotion Recognition Database are merged.
   b) 3 terms all year merged, namely Multimodal Emotion, Recognition, Multimodal Physiological Signal, Multimodal Emotion Recognition Database

2. With the results of the search with separate terms, there are several duplicate datasets that need to be eliminated.

3. Metadata was added to all data by Mendeley and then checked for incomplete author names or missing journal sources, so that articles with inappropriate metadata were eliminated.

After processing the articles, the results of the data selection can be seen in Table 3.

### Table 3. Filtered datasets for 2019 - 2023

<table>
<thead>
<tr>
<th>Number</th>
<th>Number of Articles after merging</th>
<th>Number of Articles after duplicate elimination</th>
<th>Number of Articles after data completeness evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2,620</td>
<td>2,255</td>
<td>2203</td>
</tr>
</tbody>
</table>

### Table 4. Scopus all year datasets

<table>
<thead>
<tr>
<th>Number</th>
<th>Keywords (search terms)</th>
<th>Number of Articles after merging</th>
<th>Number of Articles after duplicate elimination</th>
<th>Number of Articles after data completeness evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Emotion Recognition</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Multimodal Emotion, Recognition, Multimodal Physiological Signal, Multimodal Emotion Recognition Database</td>
<td>600</td>
<td>511</td>
<td>511</td>
</tr>
<tr>
<td>2.</td>
<td>Total Articles</td>
<td>800</td>
<td>711</td>
<td>711</td>
</tr>
</tbody>
</table>

### RESULT AND DISCUSSION

#### Significant Journal Publications

Although the Publish or Perish search is limited to 200 articles per search, the number can be used to measure trends in article distribution. However, it can be seen in Figure 2 that ER publications peaked in 1995-2010 while MER publications were rampant from 2010 until now. On the other hand, Figure 3 shows that the number of ER citations also reached its peak, but the number of MER citations is still stable and tends not to peak. This shows that research on MER which is specific research from ER still has the potential to grow.

![Trend Number ER vs MER+MERDB+MPS](image)

Figure 2. “Multimodal emotion recognition” vs “emotion recognition” article over the years
While the distribution of Figure 4 shows the distribution of where the journal was published and it shows in detail the quality of the journal through the Schimago Journal Rank (SJR) value and Q categories (Q1-Q4).

Table 5. Detail quality of journals

<table>
<thead>
<tr>
<th>Number</th>
<th>Journal</th>
<th>Number of Articles</th>
<th>SJR</th>
<th>QCategory</th>
<th>Publication Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>IEEE Transactions on Affective Computing</td>
<td>95</td>
<td></td>
<td>Q1</td>
<td>Journals</td>
</tr>
<tr>
<td>2.</td>
<td>IEEE Access</td>
<td>85</td>
<td></td>
<td>Q1</td>
<td>Journals</td>
</tr>
<tr>
<td>3.</td>
<td>Sensors (Switzerland)</td>
<td>75</td>
<td></td>
<td>Q2</td>
<td>Journals</td>
</tr>
<tr>
<td>4.</td>
<td>ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings</td>
<td>40</td>
<td>0.057638889</td>
<td></td>
<td>Conferences and Proceedings</td>
</tr>
<tr>
<td>5.</td>
<td>Multimedia Tools and Applications</td>
<td>37</td>
<td></td>
<td>Q1</td>
<td>Journals</td>
</tr>
<tr>
<td>6.</td>
<td>Biomedical Signal Processing and Control</td>
<td>31</td>
<td></td>
<td>Q1</td>
<td>Journals</td>
</tr>
<tr>
<td>7.</td>
<td>Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH</td>
<td>26</td>
<td>0.478472222</td>
<td></td>
<td>Conferences and Proceedings</td>
</tr>
<tr>
<td>8.</td>
<td>Information Fusion</td>
<td>22</td>
<td></td>
<td>Q1</td>
<td>Journals</td>
</tr>
<tr>
<td>9.</td>
<td>Applied Sciences (Switzerland)</td>
<td>20</td>
<td></td>
<td>Q1</td>
<td>Journals</td>
</tr>
<tr>
<td>10.</td>
<td>Frontiers in Psychology</td>
<td>18</td>
<td></td>
<td>Q1</td>
<td>Journals</td>
</tr>
</tbody>
</table>

Most Active and Influential Researchers

Based on the results of the article processing obtained, it is possible to identify 10 researchers who certainly make a significant contribution to the field of emotion recognition and, more specifically, to multimodal emotion recognition. Figure 5 shows the distribution of names and their article contributions either as first
authors or non-first authors. The ten names are Bjorn W. Schuller, Maja Pantic, Shrikanth Shri Narayanan, Bao Liang Lu, Jianhua Tao, Florian Eyben, Erik Cambria, Zixing Zhang, Emily Mower Provost and Nicu Sebe.

![Researcher](image)

**Figure 5. Influential researchers and a number of studies**

**Research Topic in Multimodal Emotion Recognition**

Multimodal emotion recognition is claimed to improve the quality of emotion recognition, because each modality has its own characteristics and can complement each other [14], [15]. Based on the analysis that has been done, the research focuses on 2 main topics as follows:

1) Exploration of Recognized object context divided into by ethnicity, gender, age, health condition by situation at the time of data collection.
2) Exploration of Recognition Sources/Modalities consisting of electrical signals, wearable device, audio/speech, Text, and Video/Image.

Females can discern anger and "distinguish" emotions better, so the emotions displayed by females are more accurate than those displayed by males [16]–[25]. Other studies show different results when combined with age, girls, young women, adult women and so on [18], [19], [26], [27]. These age and gender parameters will greatly affect the validation of emotions being learned and analyzed.

Health parameters have also emerged as a factor to consider in emotion recognition. Most emotion recognition research explores diseases that attack the brain as a possible factor to consider, such as autism [18], [28]–[36], stress [37]–[48], Alzheimer's, Parkinson's [49], [50], Huntington's, and schizophrenia [31], [51]–[53]. It is mentioned that Huntington's disease will interfere with the recognition of negative emotions, especially anger, disgust, and fear [54]–[56]. Patients with affective decline who attack their brains best identify their emotions through physiological signals, including electroencephalography (EEG), electrocardiogram (ECG), photoplethysmography (PPG), and respiratory (RA), to classify PD and healthy control (HC) [57]. This is in line with other research that showed weaknesses in facial emotion recognition in these patients due to impairments in facial emotion mimicking [58].

The context in which emotion recognition data is collected is another factor that should be considered in validating emotion recognition. Unfortunately, there are still few studies that explore this side. There are at least 3 situations studied including job interviews, assessment interviews and forensics interviews/interrogations as well as other interviews with specific purposes [59]–[64]. Another challenge in this field is the attempt to circumvent emotions during certain situations, especially interviews [65].

Based on the exploration of the objects to be recognized for emotions, it shows that most of the focus is on the health background of the object, then age, gender, a small proportion discusses the situation when the data was taken and very little is related to ethnic background. The composition of the research is shown in Figure 6.
To capture the source of emotions reflected by objects has also been explored, one of which is through electrical signals reflected by the body. Some research explores electrical signals in the brain/electroencephalography (EEG) [50], [66],–[143], in the heart/electrocardiogram (ECG) [88], [144]–[147], in skin tissue/plhotoplethysmography (PPG) [145], [148], [149] and respiratory tract [150], [151]. These electrical signals are effective for recognizing human emotions. So, although the signals captured are the same, the researchers explored tools to capture these signals in simpler tools by utilizing simple and wearable tools on the body. These tools are developed through modification of virtual reality headsets [99], [152], wrist/bracelet [153]–[156], chip [157], [158], flexible printed circuit board [159], [160] and textile [161]. These various wearable device developments show that emotion recognition research has a strong market and benefits. Many of these devices are being developed for mass production, so the pace of research in this field will certainly accelerate.

The computational side of exploration through Audio/Speech, Text and Video/Image is also very much developed. With speech datasets, researchers have analyzed the signal spectrogram [162]–[164] or the conversation transcript [15], [165]. Of course, the modalities obtained are different from the features that can be analyzed. Speech signals are more feature-rich than conversation transcripts (text [166]). Various information such as prosodic, spectral, voice quality, and features based on Teager energy operator can be analyzed from speech signals [167]. Classical classification methods are used (Decision Tree, and SVM [168]–[171]) or deep learning (CNN, DNN, RNN) [40], [42], [73], [108], [112], [130], [172]–[188] and deep learning with enrichment [189].

Emotion recognition analysis features through video/image are not as many as speech. There are at least two groups categorized, namely faces and body gestures [190]. Faces are the most popular, even a layman with his experience of knowing objects is very likely to easily recognize the emotions of objects. This fact is the reason why emotion recognition research through faces has developed stronger than other modalities, this is in terms of the availability of data sets and the level of accuracy produced [191]. The Facial Action Coding System (FACS) was developed to describe the judgement of cues that are used in the facial cue judgement approach. The FACS is a classification system of human facial expressions. Its origins can be traced back to its original development by [192] and revised in [193]. The revised version outlines 32 distinct facial muscle movements referred to as Action Units (AUs), as well as 14 supplementary Action Descriptors (ADs) which account for factors such as head position, gaze direction, and various actions including jaw thrust, blowing, and biting.

Mind Map
The comprehensive mind map depicted in Figure 7 illustrates the findings of the systematic literature review conducted on Multimodal Emotion Recognition. Mind maps are a useful tool for examining the connections between concepts and components of an argument, as well as for generating solutions to problems. By presenting a holistic view of all pertinent issues and analyzing options in the context of the larger picture,
Mind maps offer a new perspective on information [194]. They also facilitate the logical organization of data and the incorporation of new knowledge. In this particular study, the mind map serves as a visual representation of the outcomes of the systematic literature review on Multimodal Emotion Recognition.

Figure 7. Mind map of the SLR on multimodal emotion recognition

Future Challenges and Research Opportunities

Modalities exploration: According to Figure 7, there are many modalities that can be filled, but the distribution of investigation of each modality has not yet reached the maximum conclusion on emotion recognition, especially in relation to the context that is still widely explored and its contribution to the daily life.

Dataset availability: It is of high priority to improve datasets in this area. In relation to the small number of discussions, there is also a need for a variety of datasets based on modalities of influence.

Interdisciplinary connection: In this research, it was found that there are at least 7 fields of science involved, it would be very meaningful if there is a mutual agreement so that the contribution of each field of science can be aligned and support each other to produce a roadmap for the development of emotion recognition that is more advanced and useful.

CONCLUSION

This literature review aims to identify multimodal emotion recognition in depth and breadth by analyzing topics, trends, modalities, and other supporting sources discussed in research throughout the years and between 2010-2023. Based on the filtering analysis, a total of 14,533 articles were analyzed to achieve the goal. The conclusions generated in this literature review were conducted through a systematic method including Preparation, Implementation and Reporting.

In general, the focus of MER research can be categorized into two issues, object background and emotion recognition sources or modalities. When looking at the background of the object, most 55% to support emotion recognition with a health background, especially brain function decline, 34% based on age, 10% by gender, 1% data collection situation and a small portion of less than 1% related to ethnic culture. Referring to the source of emotion recognition, research is distributed on electromagnetic signals, voice signals, text, photo/video and wearable device development. Based on the above results there are at least 7 fields of science that discuss MER research, namely the fields of Health, Psychology, electronics, grammar, communication, socio-culture, and computing.

MER research has the potential to develop further. There are still many areas that have received less attention, while the ecosystem that uses them has grown massively. Emotion recognition modalities are numerous and diverse, but research is still focused on validating the emotions of each modality, rather than exploring the strengths of each modality to improve the quality of recognition results.
ACKNOWLEDGEMENT
This research was supported by Universitas Muhammadiyah Jember through Riset Pemula Stimulus Scheme. We thank our colleagues who provided insight and expertise that greatly assisted the research, although they may not agree with all the interpretations/conclusions of this paper.

REFERENCES


Y. Wang, W. B. Jiang, R. Li, and B. L. Lu, “Emotion Transformer Fusion: Complementary Representation Properties of EEG and Eye Movements on Recognizing Anger and Surprise,”
Acoustic Measurements

N. Siddiqui and R. H. M. Chan, “Multimodal Pre


Scientific Journal of Informatics, Vol. 10, No. 2, May 2023