Affect Representation Schemes Used in Affective Video Content Analysis

A Systematic Literature Review

Ashika Chakravorty

Supervisor: Bernd Dudzik
Responsible Professor: Chirag Raman

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Ashika Chakravorty[1]
Final project course: CSE3000 Research Project
Thesis committee: Chirag Raman[1], Bernd Dudzik[1], Cynthia Liem[1]

Abstract

Affective Video Content Analysis aims to automatically analyze the intensity and type of affect (emotion or feeling) that are contained in a video and are expected to arise in users while watching that video. This study aims to provide a systematic overview of various affect representation schemes utilised by researchers in the field of Affective Video Content Analysis and the reasons behind their choice. The main objectives of the study were to investigate the diversity of affect representation scheme types, their popularity over time, the basis of their selection, and the relationship between input data sources, in terms of direct and implicit analysis, and scheme types. Following the PRISMA guidelines to conduct a systematic literature review, a total of 45 papers were included in the study which were original journals and conference proceedings in the field of Affective Video Content Analysis and that were related only to Affective Movie Content Analysis published in English after 2008. Papers concerning Video Emotion Recognition were excluded from the review. The findings reveal that dimensional, categorical, and combined approaches are commonly used in this field, with the dimensional approach based on valence and arousal being the most prevalent. However, there is no significant trend in the popularity of affect representation scheme types over time. The study highlights the lack of clear motivation and explicit justifications for scheme selection, emphasizing the need for transparency and the inclusion of psychological theories as a basis for scheme choices. Additionally, the study found that audiovisual data was the most commonly used input compared to physiological signals and visible behavioural data. The study acknowledges its limitations, including time constraints and single researcher involvement, and suggests allocating more time and involving multidisciplinary teams for comprehensive insights.

1 Introduction

The affective content of a certain video, such as movies [1] or advertisement clips, can be described as the intensity and type of affect (emotion or feeling) that are contained in the video and are expected to arise in users while watching that video [2]. Since emotion is an important component in classifying and retrieving videos [3], Affective Video Content Analysis (AVCA) is a research topic that has attracted increasingly more attention in recent years. As stated by Wang et al [3], the objective of AVCA is to automatically assign emotional tags to each video clip based on its affective content. It is a process of automatically analysing the emotional content of videos, through the use of computational algorithms to extract and identify the affective states of individuals in the video or recognise emotions elicited by the video.

AVCA is crucial for organizing video collections and aiding users in quick and efficient video retrieval. With the increasing popularity of video consumption to fulfil emotional needs, such as alleviating boredom, the need for categorizing videos based on their affective content has grown [3]. AVCA goes beyond traditional content-based video analysis by identifying videos capable of evoking specific emotions in users, rather than solely focusing on the main event within the video [3]. Thus, AVCA has several applications, such as in mood-based video indexing [4] [5], video summarization [6], personalized content recommendations [7], and efficient movie visualization and browsing [8]. This interdisciplinary field bridges the gap between psychology and video analysis and opens new avenues for leveraging affective content to enrich video-related applications.

In psychology, the concepts of emotions and affect have been studied as distinct but related constructs. An emotion is a conscious affective state defined by cognitive appraisal, while affect is a broader term that refers to the experience of feeling, emotion or mood [9]. It encompasses a wider range of subjective experiences and can include both conscious and unconscious states. Finding a universal solution for accurately predicting the emotions experienced by most individuals while watching videos is an exceptionally challenging task due to the highly subjective nature of emotional experiences [1].

A large number of work has been proposed in the literature to tackle this highly challenging task of recognizing the affective content of videos. Previous research in AVCA has predominantly focused on inferring the affective content of videos directly from their audiovisual features [3]. This approach is known as direct video affective content analysis. However, more recent research has explored an alternative approach called implicit video affective content analysis, which leverages users' spontaneous nonverbal responses, such as facial recordings and physiological responses [10]. While previous reviews, such as the work by Wang et al. [3], have discussed the integration of video content, users' nonverbal

responses, and emotional descriptors in AVCA, there has not been an in-depth examination of the specific ARS used by researchers and their relation to the two focuses of AVCA.

As demonstrated by Baveye et al [1], the field of AVCA can be divided into two: affective movie content analysis (AMCA) and video emotion recognition (VER). AMCA involves examining the video itself as the stimulus for investigating emotions. On the other hand, VER aims to automatically estimate the expressed emotion of an agent in a video recording, often concerning an emotion induced by a stimulus. VER models facilitate affective interaction between humans and computers, while AMCA models concentrate on analyzing emotions within the video stimuli [1]. This paper will delve into the different affect representations utilized only in AMCA.

In the field of affective computing, the way that affect prediction systems extract the affective content is by using a particular scheme internally for representing affect states, i.e. Affect Representation Scheme (ARS). These schemes define how various affective states are represented for computational modelling, describing the space of different categories, attributes or qualities in terms of which a system discriminates when making predictions. In a practical sense, they define what different affective states exist and can be detected. Several different affect representation schemes have been proposed in the literature to capture the affective content of videos. These representations can be broadly categorized into two major approaches: categorical and dimensional. The categorical approach assigns videos to predefined discrete emotion labels, such as Ekman's basic emotions: anger, disgust, fear, happiness, sadness, and surprise [11]. While the dimensional approach represents emotions along continuous dimensions such as Russell's Valence-Arousel model [12], with valence indicating the positive or negative nature of emotion, and arousal representing the level of intensity or activation associated with the emotion. These approaches provide distinct perspectives on characterizing the emotional content of videos.

The technological research on AVCA exhibits a significant variation in the ARS adopted by different systems. In some cases, ad-hoc representations have been developed without any basis in established psychological theories or representations have been combined to achieve the desired functionality of systems [13]. As a result, there is no consensus among experts in the field regarding the superiority of any particular scheme. Moreover, in practice, the choice of affect representation may often depend on the needs of a specific application, modelling approach, or the costs and availability of relevant data. AVCA is a complex task as emotions are subjective and can vary across individuals, cultures, and contexts [14]. The accuracy and effectiveness of systems that rely on the analysis of affective video content can be significantly influenced by the way affective states are represented and measured. Given the limited knowledge of ARS used in AVCA, conducting researching and comparing different ARS can help to identify the most effective and efficient ways of integrating schemes and improving the accuracy and reliability of prediction systems in AVCA.

The purpose of this review is to provide a systematic overview of the various affect representation schemes that have been utilised by researchers in AVCA and the reasons behind their choice. By examining and analysing these diverse representation schemes, the aim is to gain deeper insights into the challenging task of predicting affective content in videos. The main research question answered in this paper is: "How are various affect representation schemes currently used in Affective Video Content Analysis?". To foster evidence-based design and research this paper addresses eight sub-questions. To foster evidence-based design and research, this paper addresses eight sub-questions. The following Table 1 presents the eight sub-questions along with their respective motivations.

Table 1: Targeted affective states by included studies in terms of emotions, mood, attitude and violence.

No.	Research Question	Motivation
RQ1	*What different types of input data are used in affective video content analysis?*	By exploring the different types of input data used in AVCA helps identify the primary focus of AVCA systems (direct or implicit analysis). Analysing the relation between the input data and ARS can give valuable insight into its influences on the choice and design of ARS.
RQ2	*What types of affective states have been targeted by prediction systems for affective video content analysis?*	By examining the targeted affective states can provide insights into the range of affective dimensions considered in AVCA. This question encompasses emotions, moods, and feelings, which are distinct but related constructs within the realm of affect.

RQ3	*What different affect representation schemes have been used for affective video content analysis, and if so, what is the motivation for this particular scheme?*	By investigating the various ARS used can help understand the different approaches and perspectives taken to represent and model affective content in videos. Examining the motivations behind each scheme can provide insights into the underlying theories or considerations that influenced their adoption.
RQ4	*Are systems using more than one emotion representation scheme simultaneously, and if so, what is their motivation for doing so?*	By exploring the motivations for the use of multiple ARS in AVCA can help in understanding the different approaches taken to combine different schemes and their potential benefits in enhancing AVCA.
RQ5	*Are the majority of affect representation schemes used based on psychological theory?*	By investigating the basis of existing psychological theories in the selection of ARS, can help provide insights into how AVCA systems represent affect and if they are accurately captured.
RQ6	*Are there differences in affect representation schemes used in the two focuses of AVCA systems: direct and implicit analysis of affective content in videos?*	By comparing ARS used in direct and implicit analysis of affective video content can help distinguish between the approaches and maybe help to identify potential similarities or differences in ARS selection.
RQ7	*Are there differences in the popularity of schemes used for modelling different affective states in video content analysis?*	By analyzing the popularity of schemes for modelling different affective states can provide insights into the preferences and trends in AVCA systems. It can reveal which schemes are more commonly employed.
RQ8	*Has the popularity of specific affect representation schemes changed over time?*	By investigating the temporal changes in ARS popularity can allow understanding of the evolution of research in AVCA. It can reveal shifts in adoption of certain schemes.

The paper uses the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 [15] [16] standard for reporting the study and is organized as follows. Section 2 of this paper presents the research methods employed and outlines the systematic literature review process conducted. It describes the approach taken to gather and analyse relevant literature in a systematic and rigorous manner. Section 3 presents the results obtained from the analysis of the collected studies, offering comprehensive insights into the synthesized data by addressing the sub-research questions and exploring the different affect representation schemes utilized in AVCA. Following the results, Section 4 delves into responsible research, critically reflecting on the ethical aspects of this line of research. Section 5, provides a discussion on the validity of the research and any potential limitations, while also highlighting the challenges that future works in this field may encounter. Finally, Section 6 offers a conclusion summarizing the main findings of the study and offers closing remarks.

2 Methodology

A systematic literature review (SLR) was used in the study as a methodological approach to comprehensively explore the current state of knowledge in the specific domain of interest. The systematic approach was selected to ensure transparency, rigour, and the ability to replicate the study, as it aimed to identify key studies and conduct a thorough review [17]. The PRISMA approach was adopted, which is an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses [15]. The following steps were performed: developing a research question and its sub-research questions, setting inclusion and exclusion criteria for the search, constructing a search strategy, defining the search terms and search syntax, deciding on search engines, an iterative process of screening and selection of papers based on eligibility criteria, the final selection of papers, performing the data extraction, and the synthesis and analysis of included studies. The design of the study is described in Sections 2.1 - 2.5 and the results of the search are provided in Section 2.6.

2.1 Eligibility Criteria

In selecting studies for inclusion, a set of predefined inclusion criteria was established. Only original research papers published in journals and conference proceedings written in English were included. A time constraint was implemented to make the SLR feasible and manageable. Specifically, papers published after 2008 were considered within the scope of the review, limiting the timeframe to the most recent 15 years. This decision aimed to reduce the volume of literature and focus on the most up-to-date research in the field. Additionally, the papers were limited to the field of Computer Science, and required to concern emotion recognition or affect prediction specifically within the context of AVCA. To maintain the focus on AMCA, only papers relate to AMCA were included

Conversely, a set of exclusion criteria was also applied to refine the selection process and exclude studies that did not meet the predefined criteria. Papers that only focused on VER techniques, such as only facial or speech emotion recognition from videos, were excluded. Next, papers that solely proposed algorithms for affect retrieval, without presenting the choice of affect representation scheme, were excluded. Additionally, review articles were not included in the review to maintain a focus on original research studies. Moreover, studies with incomplete or inaccessible full text were excluded to ensure that the selected papers could be thoroughly analysed

These criteria were carefully designed to ensure that the selected studies were directly relevant to the research questions and of sufficient quality for meaningful analysis. However, we also acknowledged the potential value of studies that provided unique insights or findings, even if they did not fully meet all the predefined criteria. In such cases, we agreed to include a limited number of reviews and studies that brought interesting contributions to the field of AVCA. The process of including these studies and retrieving them will be discussed in the next section.

2.2 Search Engines

In order to conduct a comprehensive search for relevant studies, we made the decision to utilize three well-known scientific databases: Scopus [1], IEEE Xplore Digital Library [2], and Web of Science (Core Collection) [3]. These databases were selected based on their popularity and wide coverage of scientific literature in the field of AVCA. Initially, Google Scholar[4] was also considered but found it generated too many irrelevant results that would have made the review process more challenging, so it was not included in our study.

2.3 Search Strategy

Before conducting the formal search, we performed scoping searches using the selected databases to gain insights into the papers that would be suitable for inclusion in the review. This preliminary step allowed for the refinement of the search strategy and the identification of relevant keywords and terms to be employed in subsequent searches. The key search terms (keywords) were identified and further refined until a final search strategy was established, which was used to conduct the literature search for each individual database.

The keywords used in the search were derived by breaking down the research question into two main concepts: affect representation and affective video content analysis. For each concept, a set of alternative search terms was generated to ensure a comprehensive search. After defining the search terms for each concept, separate search queries were formulated. Finally, these individual queries were combined to create the final search query, which was used to retrieve relevant studies for the review. The final search query appears as follows:

```
("affect*" OR "emotion" OR "affect* representation" OR "affect recognition"
OR "emotion recognition" OR "mood" OR "feeling" OR "affect* prediction" OR
"emotion prediction" OR "affect* estimation")
AND
("video" OR "video content analysis" OR "video abstraction" OR "video con-
tent representation" OR "video content modelling" OR "movie*" OR "film*" OR
"video analysis" OR "video retrieval")
```

The process of searching for relevant studies involved an iterative approach, where keywords were continuously added or removed to refine the search strategy. By carefully adjusting the keywords, the search was fine-tuned to ensure a balance between the relevance and manageability of the retrieved studies.

In the beginning, an initial analysis was done to determine the most appropriate search field to obtain a feasible number of records for the SLR. Three options were considered: searching in all fields, in title-abstract-keywords only, and in the title only. The search was conducted in May 2023 across all selected scientific databases. The number of records obtained for each database and search field in the initial search is presented in Table 2. After careful consideration, we chose to focus on the title field as

[1] https://www.scopus.com/
[2] https://ieeexplore.ieee.org/
[3] https://www.webofscience.com/
[4] https://scholar.google.com/

it provided a manageable number of records (866) for analysis. This approach allowed us to maintain a feasible number of studies to examine, compared to other fields that returned a significantly higher number of results.

Table 2: Number of results obtained by searching databases by different search fields.

		Search Field		
		All fields	**Title-Abstract-Keywords**	**Title**
Database	**IEEE Xplore**	8,475	79	264
	Scopus	30,076	4,581	215
	Web of Science	7,438	119	387
	Total	45,989	4,779	866

Since each database has its own search engine and query format, slight modifications were made to the search query to accommodate these requirements. For example, in IEEE Xplore, the field name had to be added to each keyword, while the process was simpler for Scopus and Web of Science. The specific search syntax used for each database can be found in the Appendix A.

2.4 Selection Process

The selection of papers for inclusion in the review was conducted manually, based solely on searching the titles of the papers. An individual reviewer, namely the author, carried out this process due to time constraints. Following the main literature search, a literature screening was conducted using the strategies outlined in the book by Boland et al [18].

The screening process involved several key steps. Initially, the search results were imported into reference management software, specifically Mendeley, to efficiently store, organize, and manage the records. Following this, duplicates were identified and removed from the search results by manually reviewing them. By removing duplicates, we ensured that the subsequent screening and selection phases were conducted on a clean and non-repetitive set of records. The screening and selection of studies were then carried out in two stages. In stage 1, titles and abstracts were screened to determine their relevance to the research topic. Subsequently, the full-text papers of all potentially eligible references were obtained. In stage 2, full-text papers were screened and selected based on predetermined inclusion and exclusion criteria. This rigorous selection process led to the final set of included studies, and the results of these selected studies are presented in Section 2.6. The papers that passed the screening phase underwent a detailed analysis in order to extract relevant information.

After completing the first stage of screening, it was decided to conduct the second stage of screening and data extraction in parallel. This approach was chosen to expedite the process while ensuring the inclusion of an adequate number of papers for a comprehensive review. To maintain systematicity and avoid selection bias, papers screened during the second stage were chosen randomly. Papers that focused solely on VER or that were not relevant to AMCA were excluded from the screening process. However, if a paper met the eligibility criteria and provided all the relevant information, it was included in the review, and the data was extracted immediately. The subsequent section outlines the process of data extraction in detail.

2.5 Data Extraction Process

For data extraction, an elaborate spreadsheet was prepared, serving as a tool to extract the key findings from each paper. The references were exported to a spreadsheet, and specific columns were added to address each research question of interest. The following information was extracted from each paper:

- Input Data: the type of data used by the system for emotion recognition
- Targeted Affective States: the specific affective states targeted by the system, for example, emotions or mood
- Affective Representation Scheme (ARS): the type of ARS employed by the system and the motivation behind its use

- Single or Multiple ARS: whether the system utilized one ARS or multiple ARS simultaneously, along with the reasoning behind this choice
- Year of Publication: the publication year of the paper, allowing for an examination of changes in the popularity of ARS over time
- Basis of ARS: whether the ARS used by the system was based on psychological theory or not
- AVCA Type: whether the system was based on AMCA or VER
- Direct or Implicit Analysis: whether the system focused on the direct or implicit analysis of affective content in videos.

It was also noted whether the paper addressed intended, induced or expected emotions. Additionally, any interesting observations from each included paper were noted in a designated "notes" section of the form. If applicable, each paper was also labelled as a database or review which provided further categorization and contextual information for the review. This approach of screening and data extraction ensured that relevant papers were included and their information could be systematically recorded and analysed for further synthesis.

2.6 Search Results

In the study selection process, a total of 866 records were obtained through search queries on selected scientific databases, as mentioned in Sections 2.3 and 2.4. However, during the transfer of these records to the reference management software, four of them could not be imported due to unknown reasons. After removing duplicates from the imported records, a total of 587 unique papers remained. These papers then underwent a screening process where titles and abstracts were reviewed. From this screening, 407 papers were excluded, leaving 180 papers for full-text retrieval. In the next stage of screening, 176 papers were selected for further analysis for eligibility. Due to time constraints, not all papers could be screened, and out of the 176 papers, 63 were selected for screening at random. Ultimately, 18 papers were excluded from the review due to their focus on VER, video games, or lack of relevance to AMCA. After the final stage of screening, a total of **45** papers were included in the review for analysis and reporting. Among these papers, two were identified as review papers [1] [3] and four as database papers [19] [20] [21] [22]. The results from this phase of SLR are presented in Figure 1.

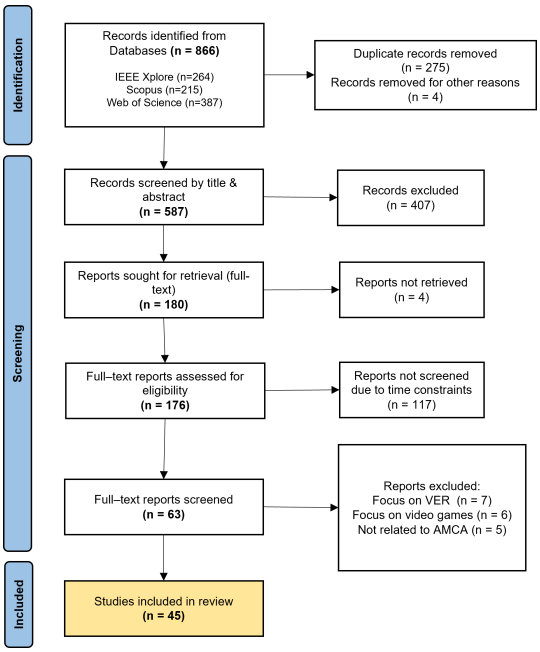

Figure 1: PRISMA flow diagram [15] for the systematic review of affect representation schemes in affective video content analysis

3 Results

The results of this study are presented in five subsections. Section 3.1 describes the different approaches of ARS used in AVCA (RQ3, RQ4). Section 3.2 delves into the specific affective states that are commonly considered in AVCA research (RQ2). Section 3.3 explores whether psychological theories inform the design and selection of affective representation schemes (RQ5). Section 3.4 describes how researchers use different types of input data to infer the affective content of videos (RQ1, RQ6). Finally, in Section 3.5 the overall popularity of different types of ARS used in AVCA and their use over time is presented (RQ7, RQ8).

3.1 Affect Representation Schemes used in AVCA

In the analysis of the included papers, several observations were made regarding the different affect representation schemes utilized for AVCA. These schemes aim to capture and represent the affective content present in videos or elicited by videos. The following subsections provide an overview of the observed types of ARS used in the literature, categorized into two major approaches: categorical and dimensional. The motivation behind the selection of each scheme is also explored, shedding light on the considerations and goals guiding their implementation in AVCA.

Categorcial Approach

The categorical emotions approach, also commonly referred to as the discrete emotions approach, is particularly helpful in differentiating emotions based on discrete categorical distinctions, especially when analysing facial expressions [1]. As highlighted in Table 3, numerous discrete categorizations of emotions have been utilised by AVCA systems to represent emotions. One widely adopted categorization was proposed by Ekman, who identified six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [11]. Another categorization, suggested by Plutchik, organized emotions into eight basic emotions arranged in four pairs of polar opposites: joy-sadness, anger-fear, trust-distrust, and surprise-anticipation [23]. Parrott's list of emotions [24] introduced a hierarchical approach to emotion classification, containing six primary emotions: love, joy, surprise, anger, sadness and fear, that further expand into secondary and tertiary emotions. An interesting representation adopted by [25] is the seven primary-process emotions: seeking, play, care, fear, grief, rage, and lust, which is based on affective neuroscience research by Panksepp [26]. Some papers represented emotions in only two ways, categorizing emotions as either positive or negative [27] [28], or focused on distinguishing between slapstick and no slapstick that identifies slapstick comedy which is one of the categories of humour [4].

Several observations were made based on the analysis of the papers. Firstly, a significant majority of the papers used Ekman's basic emotions as the chosen ARS. However, it was observed that most studies utilized only a subset of these emotions, often including the neutral affect state. Additionally, it was noted that different papers used varying terminologies to describe similar emotions. For example, terms like 'joy' and 'happiness' were used interchangeably, despite their potential distinctions in their meanings in terms of emotion or mood [29].

Table 3: Categorical Affect Representation Schemes used in Affective Video Content Analysis Systems

Scheme based on	Id	Affect Representation Scheme	Papers	# papers
Ekman's basic emotions	6N	happiness, surprise, fear, sadness, anger, disgust + neutral	[30]	9
	5A	happiness, surprise, fear, sadness, anger	[7]	
	5D	joy, surprise, fear, sadness, disgust	[31] [32]	
	4FN	happiness, sadness, fear, angry + neutral	[33]	
	4SN	happiness, sadness, surprise, anger + neutral	[34]	
	3N	happiness, sadness, anger + neutral	[6]	
	3	joy, fear, sadness	[35] [36]	
Plutchik's basic emotions	PL-8	joy, surprise, fear, sadness, anger, disgust, trust, anticipation	[37] [38]	3

Continued on next page

Table 3: Categorical Affect Representation Schemes used in Affective Video Content Analysis Systems (Continued)

	PL-8N	joy, surprise, fear, sadness, anger, disgust, trust, anticipation + neutral	[39]	
Parrott's list of emotions	Pa6	joy, surprise, fear, sadness, anger, love	[40]	1
7 primary-process emotions	7PP	seeking, play, care, fear, grief, rage, lust	[25]	1
Misc	PoNe	positive or negative	[27] [28]	5
	SnS	slapstick or no slapstick	[4]	
	M-15	happiness, joy, surprise, sadness, disgust, annoyance, curiosity, worry, excitement, affection, interest, boredom, enthusiasm, empathy, hostility	[41]	
	M-5	exciting, fearful, tense, sad, relaxing	[19]	
Total				19

Dimensional Approach

The dimensional emotions approach proposes that emotions can be described using two or more dimensions. This perspective suggests that emotions are not limited to discrete categories but can instead be represented along continuous dimensions. In the field of AVCA a significant number of papers have utilised the dimensional approach proposed by Russel et al. [12], which divided divides emotion into a 3D continuous spaces: valence, arousal, and dominance (VAD). Valence (positive vs. negative) measures the degree of pleasure or unpleasantness, representing the 'good feeling' or 'bad feeling' associated with the emotion. Arousal measures the level of activation or excitement associated with an emotion, ranging from passive to active or excited states. Dominance reflects the controlling or dominant nature of the emotion, ranging from submissive to dominant. However, measuring dominance can be challenging due to its complexity in the annotation of emotions and its difficulty in computationally predicting them [42]. As a result, it is often omitted AVCA. This omission has led to the widespread adoption of a two-dimensional approach focusing on valence and arousal (VA), as highlighted in Table 4.

Among the analysed papers, 18 of them utilized the VA dimensions as their ARS. These papers typically assigned a dual score or an average score of valence and arousal to represent the emotional content of the video clips. Interestingly, one of the papers employed a unique approach to directly assess the mood of films using three dimensions: hedonic tone (HT), energetic arousal (EA), and tense arousal (TA) [43]. They used these three continuous scales to represent the mood of a video clip, with the scales ranging from 'negative' to 'positive' (HT), 'sleepy' to 'energetic' (EA) and 'calm' to 'tense' (TA). Additionally, the midpoint of each scale was then marked as the neutral affect state.

Table 4: Dimensional Affect Representation Schemes used in Affective Video Content Analysis Systems

Scheme based on	**Id**	**Affect Representation Scheme**	**Papers**	**# papers**
Russell Thayer's VAD model	VA	Valence-Arousal	[5] [8] [20] [21] [22] [44] [45] [46] [47] [48] [49] [50] [51] [52] [42] [53] [54] [10]	20
	HETN	3D - HT, EA, TA + neutral	[43]	

Continued on next page

Table 4: Dimensional Affect Representation Schemes used in Affective Video Content Analysis Systems (Continued)

VAD	Valence-Arousal -Dominance	[55]

Note: HT = Hednoic Tone, EA = Energetic Arousal , TA = Tense Arousal

Combination of schemes

The categorical and dimensional definitions of emotions are interconnected, and it is possible to map categorical emotional states onto the dimensional space [3]. For example, a relaxed state corresponds to low arousal, while anger is associated with high arousal. Positive valence is aligned with a happy state, whereas negative valence is related to feelings of sadness or anger. This mapping of categorical emotions onto the valence-arousal space was introduced by Russell and visualized in the circumplex model of emotion [12]. Another model used in combination with discrete emotions was proposed by Mehrabian [56], called the pleasure-arousal-dominance (PAD) model, which is similar to the VAD model as pleasure is synonymous with valence.

A total of four papers used multiple ARS to represent emotions as presented in Table 5. It was observed that dimensional schemes were often used and subsequently divided into quadrants, which can be interpreted as a way of categorizing emotions.

Table 5: Combinational Affect Representation Schemes used in Affective Video Content Analysis Systems

Combination of Schemes	Id	Output	Papers	# papers
Mehrabian's PAD model + 5 discrete emotions + neutral	Co-M5N	PAD values	[57]	4
Mehrabian's PAD model + Russell's circumplex model + Plutchik's wheel	Co-MRP	One of the four quadrants of PA model	[58]	
Russell's VA model divided into 9 quadrants	Co-R9	One of the 5 quadrants out of 9 given discrete values - arousal, normal, unpleasant, relaxation, pleasant	[59]	
Russell's VA model divided into 4 quadrants	Co-R4	One of the four quadrants of VA model - NH, NL, PH, PL	[60]	

Note: PAD = Pleasure-Arousal-Dominance, NH = negative-high, NL = negative-low, PH = positive-high, PL = positive-low

Motivations behind Affect Representation Schemes selection

It was observed that out of the 43 papers analyzed, only 12 of them provided motivations for their choice of affect representation schemes in AVCA [5] [25] [58] [40] [35] [57] [52] [44] [43] [60] [50] [19]. This suggests that researchers often do not explicitly state their motivations for selecting a particular ARS.

For categorical approaches, Lin et al. [35] justified their choice of representing affect in terms of joy, sadness, and fear based on the belief that these emotions are fundamental in psychology. They excluded anger and surprise as they found them difficult to distinguish based on low-level film features. On the other hand, Thao et al. [19] argued that while the dimensional approach allows for modelling the diversity and complexity of emotions, it may struggle to effectively distinguish and represent certain emotions, such as nostalgia. They proposed that using a categorical representation of emotions could overcome this limitation. Saraswat et al. [40] adopted Parrott's list of emotions (Pa6) as a categorical representation of emotions because it is a hierarchical model that provided a good foundation for constructing a lexicon encompassing all emotion words derived from user movie reviews. The advantage of using the Pa6 model was that it facilitated capturing a range of distinct emotions expressed in unstructured text, effectively conveying a mix of different emotions [40].

For dimensional approaches, the choice of using VA model as the ARS was primarily motivated by its significant features and widespread popularity in AVCA [43] [5]. Researchers argued that complex

emotions can be effectively expressed by combining valence and arousal in different ways, as compared to discrete emotions [5]. They also highlighted that the VA model has been found to adequately describe human emotional responses to videos in various studies [50]. Another reason for choosing a dimensional approach over discrete emotions was presented by Soleymani et al. [52], who pointed out that while discrete emotions may have universal aspects, their labels can be interpreted differently across languages and cultures. Furthermore, Chan et al. [44] suggested that since affective states are subjective, users annotating videos may select different labels to describe the same affective state associated with the content, even when provided with a set of discrete affective labels. Ultimately, researchers motivated their choice of a dimensional approach, emphasizing that methods using a categorical approach have limited flexibility and arguing that models such as the VAD model can effectively represent the full range of emotions experienced by humans[42].

Papers also gave their reasons for the choice of using multiple ARS. Arifin [57] PAD model for affect representation by highlighting its potential to represent a large number of emotional states. Nemati [60] aimed to obtain ground truth for a PAD estimator by using five predefined emotions. They utilized these emotions as reference points to estimate the PAD values of video shots. By anchoring their estimation process to these known emotional states, they could establish a reliable and validated framework for assessing the PAD dimensions in their study. Canini [58] employed a combination of V and A labels to categorize video clips. They labelled spaces as 'negative-high' (NH), 'negative-low' (NL), 'positive-high' (PH), and 'positive-low' (PL). If a video clip had an arousal or valence value above or below 5, it was labelled as high or low arousal and as positive or negative respectively. The motivation for using an emotion wheel instead of direct ratings on the PA model was to simplify the self-assessment phase for users. By providing one or more emotional labels from the emotion wheel, users could express their emotional state more easily than by combining pleasure and arousal values.

3.2 Targeted Affective States in AVCA

A well-known work by Scherer [29] discusses the challenges in defining and measuring emotions. He highlights the difficulty in precisely defining emotions and reaching a consensus due to the various perspectives and theories surrounding the subject. Emotions are often misinterpreted as mood or feeling, although they differ in terms of duration, intensity, and underlying processes [29].

Scherer attempted to clarify the distinctions between different affective states. Emotions are intense and relatively brief experiences triggered by specific events or stimuli, involving subjective experience, physiological arousal, expressive behaviour, cognitive appraisal, and action tendencies. Feelings, on the other hand, are subjective experiences that arise from emotions and pertain to conscious awareness or perception. Moods are more generalized and longer-lasting affective states that are not typically triggered by specific events or stimuli, often persisting for extended periods. An attitude, on the other hand, refers to a lasting belief or predisposition towards something or someone, comprising cognitive (beliefs), affective (emotions), and behavioural (actions) components. Attitudes influence our thoughts, feelings, and behaviours towards the object of our attitude [29].

It was discovered that the majority of the papers focused on emotions, as indicated in Table 6. The following categorizations were made based on assumptions regarding affective states used by researchers:

- Emotions & Attitude: Scherer suggested considering 'love' as an interpersonal attitude with a strong positive affect component rather than an emotion [29].
- Emotions & Mood - The terms 'energetic', 'tense', 'exciting', and 'relaxing' are more likely to be associated with moods because they describe general and longer-lasting affective states rather than brief and intense experiences. Additionally, moods are characterized by relatively low intensity and can fluctuate without an immediate or obvious external cause, which aligns with the nature of the given terms.
- Emotions & Violence: Although 'violence' is not directly related to affect, it was utilized by Arifin et al. ([57] as one of the labels to represent emotions elicited by videos.

Table 6: Targeted affective states by included studies in terms of emotions, mood, attitude and violence.

Targeted affective states	Type of ARS	States selected to represent affect	Papers
emotions	~	~	[4]–[8], [10], [20]–[22], [25], [27], [28], [30]–[39], [41], [42], [44]–[55], [58]–[60]
emotions and attitude	categorical	**love**, joy, surprise, anger, sadness, fear	[40]
emotions and mood	dimensional	hedonic tone (HT), **energetic** arousal (EA), **tense** arousal (TA) + neutral	[43]
	categorical	**exciting**, fearful, tense, sad, **relaxing**	[19]
emotions and violence	combination	P-A-D values & sadness, **violence**, fear, happiness, amusement + neutral	[57]

3.3 Psychological Theory and Affect Representation Schemes

It was observed that only 46% of the papers mentioned the basis of their selected ARS. The majority of these papers were grounded in psychological theories [39] [6] [37] [31] [40] [38] [5] [43] [58] [59] [60] [8] [52] [19] [44] [20] [57] [48] [55] [42], with one exception.

Radeta et al. [25] based their selection on the 7 primary-process emotions (7PP) derived from affective neuroscience research conducted by Panksepp [26]. The authors justified their choice by highlighting the comprehensive nature of Panksepp's work, which identified key emotions based on empirical evidence. These emotions were found to be present across all mammals and were associated with specific neurotransmitters, precise brain areas, and observable behaviours [25].

3.4 Types of Input Data used in AVCA: Direct vs. Implicit Analysis

The input data used in AVCA can be divided by the two different focuses of AVCA: direct and implicit video affective content analysis [3]. Direct analysis aims to infer the affective content of a video directly from its audiovisual features. On the other hand, implicit analysis utilizes the user's spontaneous nonverbal responses, such as facial recordings, physiological responses, and visible behaviours, to capture the emotions evoked while watching a video. It was also observed that methods combining both direct and implicit analysis were used for affect recognition in AVCA. In the following tables Table 7, Table 8 and Table 9 the input data is presented, categorized into different types of ARS - categorical, dimensional, and combinational.

The following is an overview of input data used in AVCA:

- Direct analysis - This includes *audiovisual* data and *text* extracted from the movie script.
- Implicit analysis - This involves *physiological signals* such as electroencephalography (EEG), electrodermal activity (EDA), heart rate (HR), skin temperature (TEMP), blood pressure (BP), etc. as well as *visible behaviours* like facial expressions and body movements. *User annotations* of videos, which involve labelling videos based on the emotions evoked in users while watching, are also categorized here. *Text* data from movie reviews or user comments are also considered.

Any combination of these input data sources, whether from direct or implicit analysis, is categorized into the hybrid analysis - direct & implicit analysis.

Table 7: Input data used in categorical Affect Representation Schemes classified by direct or implicit analysis of affective video content.

Focus of AVCA	Input data	Papers	# papers

Continued on next page

Table 7: Input data used in categorical Affect Representation Schemes classified by direct or implicit analysis of affective video content. (Continued)

Direct	audiovisual	[33] [4] [39]	6
	video & text (movie dialogues)	[28]	
	video & film grammar	[35] [36]	
Implicit	EEG signals	[27]	8
	EDA, BVP, ACC, TEMP, HR signals	[41]	
	facial expression & HR	[34]	
	facial expression	[34]	
	user annotation	[25]	
	text (movie reviews)	[40] [38]	
	text (user comments) & user annotation	[37]	
Direct & implicit	audiovisual user annotation	[19]	5
	audiovisual & face and body movements	[6]	
	audiovisual & facial expressions	[32] [31]	
	text & user annotation	[30]	

Note: physiological signals - EEG = electroencephalography, EDA = electrodermal activity, BVP = blood volume pulse , ACC = accelerator , TEMP = skin temperature, HR = heart rate

Table 8: Input data used in dimensional Affect Representation Schemes classified by direct or implicit analysis of affective video content.

Focus of AVCA	**Input data**	**Papers**	**# papers**
Direct	audiovisual	[49] [50] [42] [53] [54] [10] [8]	9
	audiovisual & text (from movie's script)	[52]	
	audiovisual & user profile	[5]	
Implicit	EEG signals	[47]	3
	EDA & body movement	[48]	
	GSR & HF	[45]	
Direct & implicit	audiovisual & GSR, BP, RSP, TEMP, EMG, eye blinking rate	[51]	7
	audiovisual & user annotation	[43] [20] [22]	
	audiovisual & text (verbal emotion labels)	[44]	
	audiovisual, text (from movie script) & hand movement	[55]	
	audiovisual, text & user-annotation	[21]	
Implicit hyrbid	audiovisual EEG, EOG, ECG, EMG, GSR, RSP, TEMP, PLET	[46]	1

Note: physiological signals - EEG = electroencephalography, EDA = electrodermal activity, EOG = electroculogram, ECG = electrocardiograpgh, EMG = electromyography, GSR = galvanic skin resistance , RSP = respiration pattern, TEMP = skin temperature, PLET = plethysmograph, HF = heat flux, BP = blood pressue

Researchers have also provided motivations for combining different input data in AVCA. For example, Nemati et al. [60] highlighted that the combination of audiovisual data with text helps overcome limitations associated with using only audiovisual features. Solely relying on audiovisual features can lead to the loss of global relations in the data, while constructing high-level features can be time-consuming and problem-dependent. To address these challenges, incorporating user opinions and views about a video was found to provide valuable information[60]. Additionally, Srivastava et al. [28] emphasized the advantage of combining video and text, as relying solely on audio information can be noisy due to the variability in people's speech. Moreover, research has shown that textual cues can be more effective than acoustic cues in recognizing emotions [28].

Table 9: Input data used in combinational Affect Representation Schemes classified by direct or implicit analysis of affective video content.

Focus of AVCA	Input data	Papers	# papers
Direct	audiovisual	[57] [59]	2
Direct & implicit	video & user annotation	[58]	2
	audiovisual & text (social media users comments)	[60]	

Observations were made regarding the preference of analysis methods for different types of ARS. For categorical ARS shown in Table 7, implicit analysis was found to be preferred over direct analysis, and a combination of both direct & implicit analysis methods. In contrast, for dimensional ARS shown in Table 8, direct analysis was favoured over implicit analysis, as well as the combination of both methods. Interestingly, for combinational ARS shown in Table 9, an equal preference was observed for both direct and direct & implicit analysis approaches.

Overall, it was observed that all types of input data, whether obtained through direct or implicit analysis, were utilized in all types of ARS. However, no significant findings were reported regarding a specific relationship between the types of input data and the corresponding types of ARS used. This is likely because each AVCA system tends to employ unique and customized input data based on its specific function and requirements. Therefore, the selection of input data is largely driven by the design and goals of the individual AVCA systems rather than a predefined relationship between input data and ARS types.

3.5 Popularity of Affect Representation Schemes

The popularity of different ARS types varied in relation to the psychological theories they were associated with. As shown in Figure 2, among the ARS types, the dimensional approach, which focuses mainly on dimensions such as valence and arousal, was the most popular, accounting for 47% of the papers. This shows that a lot of researchers support the dimensional approach for understanding and representing emotions in AVCA.

In contrast, the categorical approach, which is based on theories like Ekman's basic emotions and Plutchik's basic emotions, had a slightly lower but still considerable popularity, accounting for 44% of the papers. However, there was a notable difference in the distribution of these theories. As shown in Figure 3, Ekman's basic emotions were mentioned in 21% of the papers, while Plutchik's basic emotions were referenced in only 7% of the papers. While Parott's list of emotions and the 7 primary-process emotions were both mentioned in 2% of papers each. This suggests a relatively limited adoption of these particular categorization schemes. Overall, these results show that distinct categories of emotions have had a significant influence on the understanding of emotions in AVCA, similar to the dimensional approach.

In Figure 3, other miscellaneous approaches accounted for 21% of the papers. These include various combinations, modifications, or alternative ARS not directly linked to specific psychological theories.

The contrast between the popularity of ARS types and theories used in AVCA indicates a greater preference for dimensional frameworks for representing emotions in the studies examined. This may suggest the growing recognition of the continuous nature and multidimensional aspects of emotions. Researchers

in AVCA seem to appreciate the practical advantages of the dimensional approach in capturing the complexity and diversity of emotional experiences.

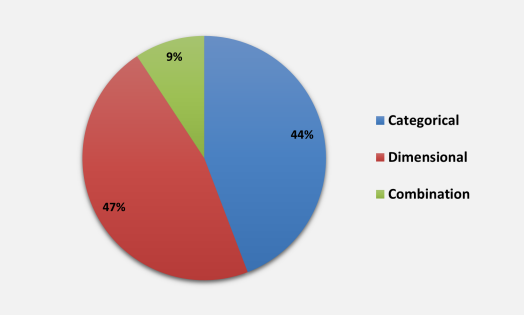

Figure 2: Popularity of Affect Representation Schemes based on its type

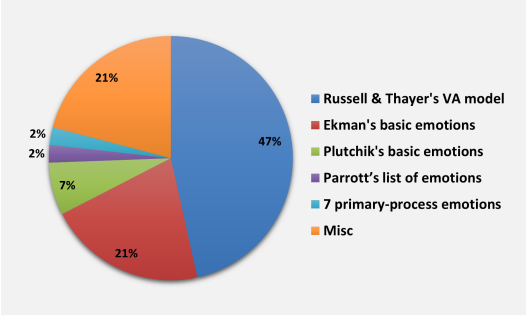

Figure 3: Popularity of Affect Representation Schemes based on its theory

Over the years, in the period from 2008 to 2023, the popularity of different ARS types has shown some variations. Figure 5 provides a breakdown of the popularity of ARS types in three-year intervals, while the scatter plot in Figure 4 illustrates the temporal analysis of ARS type popularity for each individual year.

Overall, the popularity of different types of ARS fluctuates over time, with dimensional ARS being consistently popular in most periods, followed by categorical ARS. However, the availability of limited data makes it difficult to fully access the popularity of combinational ARS. Additionally, it is important to note that the provided data is not comprehensive enough to fully capture the overall popularity of ARS in the field of AVCA.

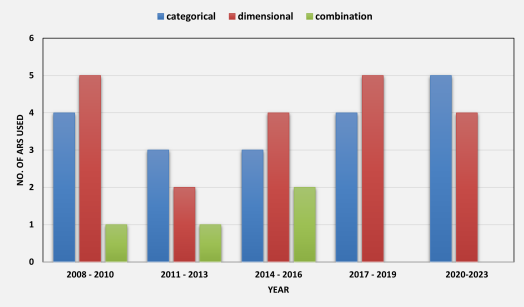

Figure 4: Popularity of Affect Representation Schemes over 15 years

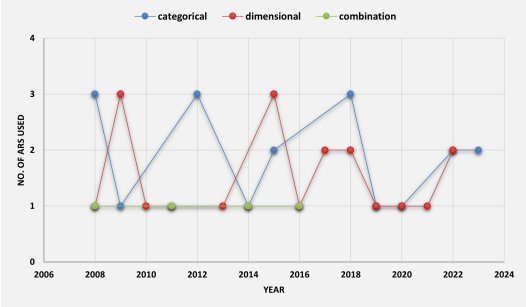

Figure 5: Temporal analysis of Affect Representation Schemes popularity

4 Responsible Research

Responsible research refers to the practice of conducting scientific investigations and scholarly activities in an ethical and accountable manner. In terms of ethical issues related to this study, several considerations arise regarding the methodology employed. As this study involved a rapid systematic literature review conducted by a single researcher, there is a potential for bias. Although efforts were made to maintain systematicity and reduce bias during the screening process by randomly selecting papers, the possibility of selection bias cannot be completely ruled out. Furthermore, biases may have influenced the interpretation of results and decisions made during the extraction of information from the selected papers, potentially impacting the quality of the review.

Ethical concerns also emerge regarding the topic of this review, which focuses on affective video content analysis aiming to automatically recognize emotions in videos or elicited by videos. It was observed that a majority of the included papers did not explicitly mention or provide justification for their choice of affect representation schemes (ARS). This lack of transparency poses an ethical issue as researchers should strive to clearly articulate the rationale behind their methodological choices. Without such transparency, the validity and reliability of the research may be compromised.

Additionally, it is important to acknowledge the inherent risks associated with affect recognition and prediction systems, which often rely on predefined affect states and paradigms that may not encompass

the full complexity of human emotions. These predefined affect states can be subject to interpretation variations across different languages and cultures. While some efforts have been made to address this issue, such as research on incorporating cultural differences into affect modeling [51][22], there remains a lack of comprehensive databases and extensive research in this area. This raises ethical concerns regarding the generalizability and applicability of affective video content analysis systems across diverse cultural contexts.

In conclusion, this review critically reflects on the ethical aspects of the research and acknowledges certain limitations and potential biases inherent in the methodology. The study also highlights the need for greater transparency and justification of ARS choices in affective video content analysis research. Moreover, the ethical challenges associated with predefined affect states and cultural variations warrant further attention and research efforts to ensure the responsible and culturally sensitive development of affect recognition and prediction systems.

5 Discussion

This section provides a general interpretation of the results found in answering the sub-research questions about the the various aspects of affect representation schemes used in affective video content analysis and sheds light on their differences and implications.

In this study, three main types of ARS were identified: categorical, dimensional, and a combination of both approaches or alternative methods. It is evident that each paper, except those utilizing Russell's valence and arousal model [12], uses distinct ARS to represent emotions in their respective systems. However, an interesting observation arises from the categorical approach, where researchers frequently employ different terminologies to describe emotions. Surprisingly, only a small number of papers explicitly motivate their choice of ARS, indicating a lack of justification or reporting the reasoning behind their selection.

In AVCA, the targeted affective states encompass emotions, mood, attitude, and an intriguing aspect, violence. Among these, the majority of papers (90%) focus on emotions, while only four papers delve into other affective states. Strikingly, out of these four papers, only one explicitly mentions mood [43], while the remaining three interpret the chosen affective states solely as emotions, despite the inherent differences highlighted by Scherer's work [29]. This highlights the tendency of researchers to misinterpret emotions as other affective states, further emphasizing the need for clarity and precision in defining and distinguishing these states.

An additional noteworthy observation is that only 46% of the papers mention the basis of their selected ARS, with all but one paper being grounded in psychological theories. The exceptional paper draws from affective neuroscience, specifically Panksepp's work. The fact that over 50% of the papers do not provide information about the foundation of their chosen ARS suggests that researchers may often overlook thorough exploration of the available ARS options and their alignment with psychological theories. These findings highlight the need for clear and transparent selection and justification of ARS in AVCA. By establishing a solid theoretical foundation and clearly explaining the reasons behind their choices, researchers can improve the reliability and comparability of their studies, advancing the understanding and use of AVCA.

In terms of input data used in AVCA the choices varied across different systems, with few similarities observed. When considering the different types of ARS, no significant trends were found in the use of input data based on direct or implicit analysis. This lack of association can be attributed to the tendency of each AVCA system to adopt unique and customized input data according to its specific function and requirements. However, it is worth noting that audiovisual data was the most commonly used input across all types of ARS, while the utilization of physiological signals and visible behavioural data varied among systems. These findings suggest a diverse range of techniques employed for affect recognition and prediction in videos.

In terms of the popularity of ARS, the data indicates a slight preference for the dimensional approach over the categorical approach, with only a few instances of combining ARS. Notably, Russell's VA model is widely utilized, appearing in 47% of the papers, which is a significant proportion compared to the utilization of Ekman's basic emotions (21%) or Plutchik's basic emotions (7%). Although both dimensional and categorical ARS exhibit similar levels of popularity, the VA model stands out as the most widely used. However, when examining the popularity of ARS types over the years from 2008 to 2023, there are only minor variations observed. The overall popularity of different ARS types fluctuated

over time without a significant trend. It is important to note that these observations are limited by the availability of data, and therefore do not provide a comprehensive overview of the popularity of ARS in the field of AVCA.

There are several limitations to acknowledge in this study. Firstly, it is important to note that this was a rapid systematic literature review conducted within a limited timeframe of ten weeks, due to which not all retrieved papers could be screened, which potentially limited the inclusion of relevant studies and the availability of data for more in-depth analysis. Additionally, the review was performed by a single researcher, potentially introducing bias and limiting the comprehensiveness of the study. Moreover, the quality of evidence included in the review could be a concern since it was not assessed by multiple reviewers. With additional time and a team of researchers, the review could have been more comprehensive and potentially reduced biases in the selection and analysis process. These limitations should be taken into consideration when interpreting the findings of this study, and future research with more extensive timeframes, collaborative efforts, and rigorous quality assessments would provide a more robust understanding of the topic.

6 Conclusions and Future Work

This study aimed to systematically review and examine the range of affect representation schemes (ARS) utilised in the field of Affective Video Content Analysis (AVCA) and investigate the underlying reasons for their selection. Overall, this study aimed to investigate the diversity of ARS types, their popularity over time, the basis of their selection, and the relationship between input data sources, in terms of direct and implicit analysis, and ARS types in AVCA.

Through a systematic literature review conducted following PRISMA guidelines [15], a total of 45 relevant papers were included in the study. This was done by searching three databases, namely Scopus, IEEE Xplore Digital Library, and Web of Science (Core Collection), using a set search strategy that allowed the inclusion of papers from original journals and conference proceedings in the field of AVCA after 2008 that were related only to affective movie content analysis published in English.

The findings revealed that dimensional, categorical, and combined approaches were the main types of ARS utilized in AVCA, with the dimensional approach being the most prevalent overall. However, the popularity of ARS types did not show a significant trend over time, indicating the dynamic nature of research in this field.

One important observation was the lack of clear motivation provided for the selection of ARS, with only 12% of the papers offering explicit justifications. Additionally, only 46% of the included papers mentioned a psychological theory as the basis for their chosen ARS. This highlights the need for researchers to provide more thorough justifications and enhance transparency in reporting their ARS choices, which would improve the reliability and comparability of studies in AVCA.

Furthermore, the study emphasized the significance of considering input data sources in AVCA. While audiovisual data was the most commonly used input, the utilization of physiological signals and visible behavioural data varied among different systems. This suggests a wide range of techniques being employed for affect recognition and prediction in video content.

However, it is crucial to acknowledge the limitations of this study, including the time constraints of a rapid systematic literature review and the involvement of a single researcher. Future research efforts should allocate more time and involve multidisciplinary teams to overcome these limitations and provide more comprehensive insights into the field. To build upon the findings of this study, future work could explore the impact of different affect representation schemes on the performance and accuracy of AVCA systems in predicting emotions. Additionally, future research should delve into the evaluation of video emotion recognition systems, which was not extensively covered in this study. Furthermore, the generalizability of AVCA systems across different video content domains and cultures should be researched. This involves investigating the impact of cultural factors on emotion recognition and adapting affect representation schemes for cross-cultural variations.

Overall, this study contributes to the understanding of ARS choices in AVCA and emphasizes the importance of providing clear justifications for their selection. By addressing the identified gaps and limitations, future studies can further advance the field of AVCA and promote the development of robust and reliable AVCA.

References

[1] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Transactions on Affective Computing*, vol. 9, pp. 396–409, 4 2018. DOI: 10.1109/TAFFC.2017.2661284.

[2] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005, ISSN: 1941-0077. DOI: 10.1109/TMM.2004.840618.

[3] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, pp. 410–430, 4 2015, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2015.2432791.

[4] J. H. French, "Automatic affective video indexing: Sound energy and object motion correlation discovery," IEEE, 2012, pp. 1–6, ISBN: 978-1-4673-1375-9. DOI: 10.1109/SECon.2012.6196925.

[5] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Transactions on Multimedia*, vol. 12, pp. 510–522, 6 2010, ISSN: 1520-9210. DOI: 10.1109/TMM.2010.2059634.

[6] M. Dammak, A. Wali, and A. M. Alimi, "Viewer's affective feedback for video summarization," *Journal of Information Processing Systems*, vol. 11, pp. 76–94, 1 2015, ISSN: 2092805X. DOI: 10.3745/JIPS.01.0006.

[7] Y. Soni, C. O. Alm, and R. Bailey, "Affective video recommender system," IEEE, 2019, pp. 1–5, ISBN: 978-1-7281-4352-1. DOI: 10.1109/WNYIPW.2019.8923087.

[8] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, "Utilizing affective analysis for efficient movie browsing," IEEE, 2009, pp. 1853–1856, ISBN: 978-1-4244-5653-6. DOI: 10.1109/ICIP.2009.5413590.

[9] J. Tarvainen, M. Sjoberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Transactions on Multimedia*, vol. 16, pp. 2085–2098, 8 2014, ISSN: 1520-9210. DOI: 10.1109/TMM.2014.2357688.

[10] Q. Wang, X. Xiang, J. Zhao, and X. Deng, "P2sl: Private-shared subspaces learning for affective video content analysis," IEEE, 2022, pp. 1–6, ISBN: 978-1-6654-8563-0. DOI: 10.1109/ICME52920.2022.9859902.

[11] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*. John Wiley Sons, Ltd, 1999, ch. 3, pp. 45–60, ISBN: 9780470013496. DOI: doi.org/10.1002/0470013494.ch3.

[12] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980. DOI: 10.1037/h0077714.

[13] G. Smith and J. Carette, "What Lies Beneath—A Survey of Affective Theory Use in Computational Models of Emotion," 2022. DOI: 10.36227/techrxiv.18779315.v2.

[14] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, "Investigating the influence of personal memories on video-induced emotions," in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '20, Association for Computing Machinery, 2020, pp. 53–61. DOI: 10.1145/3340631.3394842.

[15] *Prisma transparent reporting of systematic reviews and meta-analyses*, 2020. [Online]. Available: http://www.prisma-statement.org/.

[16] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021. DOI: 10.1136/bmj.n71.

[17] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," vol. 2, 2007. [Online]. Available: https://www.researchgate.net/publication/302924724.

[18] A. Boland, M. G. Cherry, and R. Dickson, *Doing a systematic review: A student's guide*. Sage Publications Ltd, 2017.

[19] H. T. P. Thao, G. Roig, and D. Herremans, "Emomv: Affective music-video correspondence learning datasets for classification and retrieval," *Information Fusion*, vol. 91, pp. 64–79, 2023, ISSN: 15662535. DOI: 10.1016/j.inffus.2022.10.002.

[20] Y. Baveye, J.-N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret, "A large video database for computational models of induced emotion," IEEE, 2013, pp. 13–18, ISBN: 978-0-7695-5048-0. DOI: 10.1109/ACII.2013.9.

[21] P.-C. Hsu, J.-L. Li, and C.-C. Lee, "Romantic and family movie database: Towards understanding human emotion and relationship via genre-dependent movies," IEEE, 2022, pp. 1–8, ISBN: 978-1-6654-5908-2. DOI: 10.1109/ACII55700.2022.9953881.

[22] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, pp. 43–55, 1 2015, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2015.2396531.

[23] R. Plutchik, "Chapter 1 - a general psychoevolutionary theory of emotion," in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds., Academic Press, 1980, pp. 3–33, ISBN: 978-0-12-558701-3. DOI: 10.1016/B978-0-12-558701-3.50007-7.

[24] W. G. Parrott, *Emotions in social psychology: Key Readings in Social Psychology.* Psychology Press, 2001.

[25] M. Radeta, Z. Shafieyoun, and M. Maiocchi, "Affective timelines: Towards the primary-process emotions of movie watchers: Measurements based on self-annotation and affective neuroscience," Universidad de los Andes, 2014, pp. 679–688. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84912079104&partnerID=40&md5=e1bcda6123bd8bbf117cb408399103e6.

[26] J. Panksepp, *Affective neuroscience: The foundations of human and Animal Emotions*. Oxford University Press, 1998.

[27] M. Kwon and M. Lee, "Emotion understanding in movie clips based on eeg signal analysis," vol. 7665 LNCS, 2012, pp. 236–243. DOI: 10.1007/978-3-642-34487-9_29.

[28] R. Srivastava, S. Yan, T. Sim, and S. Roy, "Recognizing emotions of characters in movies," 2012, pp. 993–996. DOI: 10.1109/ICASSP.2012.6288052.

[29] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005. DOI: 10.1177/0539018405058216.

[30] B. Iavarone and F. Dell'Orletta, "Predicting movie-elicited emotions from dialogue in screenplay text: A study on "forrest gump"," vol. 2769, CEUR-WS, 2020. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097882128&partnerID=40&md5=633b7d9b2a6beaabd7a0d417839a3033.

[31] D. N. Amali, A. R. A. Besari, A. R. Barakbah, and D. Agata, "Automatic user-video metrics creations from emotion detection," IEEE, 2018, pp. 663–668, ISBN: 978-1-5386-8402-3. DOI: 10.1109/EECSI.2018.8752750.

[32] D. N. Amali, A. R. Barakbah, A. R. A. Besari, and D. Agata, "Semantic video recommendation system based on video viewers impression from emotion detection," IEEE, 2018, pp. 176–183, ISBN: 978-1-5386-8079-7. DOI: 10.1109/KCIC.2018.8628592.

[33] L. Chen, G.-C. Chen, C.-Z. Xu, J. March, and S. Benford, "Emoplayer: A media player for video clips with affective annotations," *Interacting with Computers*, vol. 20, pp. 17–28, 1 2008, ISSN: 09535438. DOI: 10.1016/j.intcom.2007.06.003.

[34] T. Elias, U. S. Rahman, and K. A. Ahamed, "Movie recommendation based on mood detection using deep learning approach," Institute of Electrical and Electronics Engineers Inc., 2022. DOI: 10.1109/ICAECT54875.2022.9807654.

[35] X. Lin, X. Wen, Z. Lu, and Y. Sun, "Film affective content recognition based on fuzzy inference," vol. 2, IEEE, 2008, pp. 34–37, ISBN: 978-0-7695-3560-9. DOI: 10.1109/ISBIM.2008.92.

[36] X. Lin, X. Wen, Z. Lu, and W. Zheng, "Video affective content recognition based on film grammars and fuzzy evaluation," IEEE, 2008, pp. 264–267, ISBN: 978-0-7695-3556-2. DOI: 10.1109/MMIT.2008.87.

[37] C. Orellana-Rodriguez, E. Diaz-Aviles, and W. Nejdl, "Mining affective context in short films for emotion-aware recommendation," Association for Computing Machinery, Inc, 2015, pp. 185–194. DOI: 10.1145/2700171.2791042.

[38] M. Cohen-Kalaf, J. Lanir, P. Bak, and O. Mokryn, "Movie emotion map: An interactive tool for exploring movies according to their emotional signature," *Multimedia Tools and Applications*, vol. 81, pp. 14 663–14 684, 11 2022. DOI: 10.1007/s11042-021-10803-5.

[39] G. Irie, K. Hidaka, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Latent topic driving model for movie affective scene classification," 2009, pp. 565–568. DOI: 10.1145/1631272.1631357.

[40] M. Saraswat and S. Chakraverty, "Leveraging movie recommendation using fuzzy emotion features," vol. 799, Springer Verlag, 2018, pp. 475–483. DOI: 10.1007/978-981-10-8527-7_40.

[41] S. Derdiyok and F. P. Akbulut, "Biosignal based emotion-oriented video summarization," *Multimedia Systems*, 2023, ISSN: 0942-4962. DOI: 10.1007/s00530-023-01071-4.

[42] H. T. P. Thao, D. Herremans, and G. Roig, "Multimodal deep models for predicting affective responses evoked by movies," Institute of Electrical and Electronics Engineers Inc., 2019, pp. 1618–1627. DOI: 10.1109/ICCVW.2019.00201.

[43] J. Tarvainen, J. Laaksonen, and T. Takala, "Computational and perceptual determinants of film mood in different types of scenes," IEEE, 2017, pp. 185–192, ISBN: 978-1-5386-2937-6. DOI: 10.1109/ISM.2017.10.

[44] C. H. Chan and G. J. F. Jones, "An affect-based video retrieval system with open vocabulary querying," vol. 6817 LNCS, 2011, pp. 103–117. DOI: 10.1007/978-3-642-27169-4_8.

[45] X. Y. Chen and Z. Segall, "Xv-pod: An emotion aware, affective mobile video player," vol. 3, IEEE, 2009, pp. 277–281, ISBN: 978-0-7695-3507-4. DOI: 10.1109/CSIE.2009.982.

[46] S. Chen, S. Wang, C. Wu, Z. Gao, X. Shi, and Q. Ji, "Implicit hybrid video emotion tagging by integrating video content and users' multiple physiological responses," IEEE, 2016, pp. 295–300, ISBN: 978-1-5090-4847-2. DOI: 10.1109/ICPR.2016.7899649.

[47] Y. Ding, X. Hu, Z. Xia, Y.-J. Liu, and D. Zhang, "Inter-brain eeg feature extraction and analysis for continuous implicit emotion tagging during video watching," *IEEE Transactions on Affective Computing*, vol. 12, pp. 92–102, 1 2021, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2018.2849758.

[48] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun, "Films, affective computing and aesthetic experience: Identifying emotional and aesthetic highlights from multimodal signals in a social setting," *Frontiers in ICT*, vol. 4, JUN 2017. DOI: 10.3389/fict.2017.00011.

[49] I. Mironica, B. Ionescu, M. Sjöberg, M. Schedl, and M. Skowron, "Rfa at mediaeval 2015 affective impact of movies task: A multimodal approach," vol. 1436, CEUR-WS, 2015. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84989869543&partnerID=40&md5=824f72a1d0118e652ad552932c55e608.

[50] S. Sivaprasad, T. Joshi, R. Agrawal, and N. Pedanekar, "Multimodal continuous prediction of emotions in movies using long short-term memory networks," Association for Computing Machinery, Inc, 2018, pp. 413–419. DOI: 10.1145/3206025.3206076.

[51] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," IEEE, 2008, pp. 228–235. DOI: 10.1109/ISM.2008.14.

[52] M. Soleymani, J. J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," IEEE, 2009, pp. 1–7, ISBN: 978-1-4244-4800-5. DOI: 10.1109/ACII.2009.5349563.

[53] H. T. P. Thao, "Deep neural networks for predicting affective responses from movies," Association for Computing Machinery, Inc, 2020, pp. 4743–4747. DOI: 10.1145/3394171.3416517.

[54] M. P. Vlastelica, S. Hayrapetyan, M. Tapaswi, and R. Stiefelhagen, "Kit at mediaeval 2015 - evaluating visual cues for affective impact of movies task," vol. 1436, CEUR-WS, 2015. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84989935544&partnerID=40&md5=e5355050e73577d96b74356521bd7e28.

[55] L. Tian, M. Muszynski, C. Lai, *et al.*, "Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same?," vol. 2018-January, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 28–35. DOI: 10.1109/ACII.2017.8273575.

[56] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," vol. 14, Dec. 1996. DOI: 10.1007/BF02686918.

[57] S. Arifin and P. Y. K. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Transactions on Multimedia*, vol. 10, pp. 1325–1341, 7 2008, ISSN: 1520-9210. DOI: 10.1109/TMM.2008.2004911.

[58] L. Canini, S. Benini, and R. Leonardi, "Affective analysis on patterns of shot types in movies," 2011, pp. 253–258. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-83455168969&partnerID=40&md5=4034cb0059d5076e77d0d0ccab68eedd.

[59] J.-K. Chang and S.-T. Ryoo, "Real-time emotion retrieval scheme in video with image sequence features," *Journal of Real-Time Image Processing*, vol. 9, pp. 541–547, 3 2014, ISSN: 1861-8200. DOI: 10.1007/s11554-013-0366-x.

[60] S. Nemati and A. R. Naghsh-Nilchi, “Incorporating social media comments in affective video retrieval,” *Journal of Information Science*, vol. 42, pp. 524–538, 4 2016. DOI: 10 . 1177 / 0165551515593689.

[61] OpenAI. [Online]. Available: https://chat.openai.com/chat.

A Search Strategy for each Database

The following search syntac was used for each database.

A.1 IEEE Xplore Digital Library

((((”Document Title”: affect* OR ”Document Title”: ”emotion” OR ”Document Title”: ”affect* representation” OR ”Document Title” :”affect* recognition” OR ”Document Title”: ”emotion recognition” OR ”Document Title”: ”mood” OR ”Document Title”: ”feeling” OR ”Document Title”: ”affect* prediction” OR ”Document Title”: ”emotion prediction”) AND (”Document Title”: ”video” OR ”Document Title”: ”video content analysis” OR ”Document Title”: ”video abstraction” OR ”Document Title”: ”video content representation” OR ”Document Title”: ”video content modelling” OR ”Document Title”: ”movie” OR ”Document Title”: ”film” OR ”Document Title”: ”video analysis” OR ”Document Title”: ”video retrieval”))))

Filters Applied: Conferences, Journals, Early Access Articles, 2008 - 2023

A.2 Scopus

TITLE ((”affect*” OR ”emotion” OR ”affect* representation” OR ”affect recognition” OR ”emotion recognition” OR ”mood” OR ”feeling” OR ”affect* prediction” OR ”emotion prediction” OR ”affect* estimation”) AND (”video content analysis” OR ”video abstraction” OR ”video content representation” OR ”video content modelling” OR ”movie*” OR ”film*” OR ”video analysis” OR ”video retrieval”)) AND PUBYEAR ¿ 2007 AND PUBYEAR ¡ 2024 AND (LIMIT-TO (SUBJAREA , ”COMP”)) AND (LIMIT-TO (DOCTYPE , ”cp”) OR LIMIT-TO (DOCTYPE , ”ar”)) AND (LIMIT-TO (LANGUAGE , ”English”))

A.3 Web of Science (Core Collection)

(”affect*” OR ”emotion” OR ”affect* representation” OR ”affect recognition” OR ”emotion recognition” OR ”mood” OR ”feeling” OR ”affect* prediction” OR ”emotion prediction” OR ”affect* estimation”) AND (”video” OR ”video content analysis” OR ”video abstraction” OR ”video content representation” OR ”video content modelling” OR ”movie*” OR ”film*” OR ”video analysis” OR ”video retrieval”)

Filters applied: 2008-2023, Language: English, Research Areas: Computer Science, Document Types: Article, Proceeding Paper, Early Access & Excluding Review papers and Book Chapters

B Use of Large Language Models used in Paper

B.1 ChatGPT

ChatGPT [61] was used during the writing process of this report to assist in rephrasing certain sentences and fixing certain grammatical issues. The content and ideas presented in the report were solely developed by the researcher and the help taken from ChatGPT was limited to providing language support in specific sections of the report. They are as follows - Section 1 , Section 2, Section 5, Section 6. Additionally, help was taken from ChatGPT to provide instructions to format text and tables in LaTeX.

An example of a prompt given to ChatGPT:

”It was observed that only 46% of the papers mentioned the basis of their selected ARS, out of which all were based on psychological theories except one. fix the language”

Text generated by ChatGPT, 25th June, 2023:

”It was observed that only 46% of the papers provided a rationale for the basis of their selected affective representation schemes (ARS). The majority of these papers were grounded in psychological theories, with one exception.”

Text used in report: ”It was observed that only 46% of the papers mentioned the basis of their selected ARS. The majority of these papers were grounded in psychological theories, with one exception.”