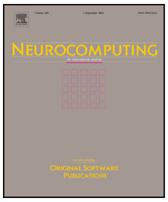
Neurocomputing 607 (2024) 128348

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Survey Paper

Multi-modal data clustering using deep learning: A systematic review

Sura Raya, Mariam Orabi, Imad Afyouni, Zaher Al Aghbari *

College of Computing and Informatics, University of Sharjah, United Arab Emirates

ARTICLE INFO

Communicated by M. Sato-Ilic

Keywords:
Multi-modal data
Clustering algorithms
Deep learning
Review article

ABSTRACT

Multi-modal clustering represents a formidable challenge in the domain of unsupervised learning. The objective of multi-modal clustering is to categorize data collected from diverse modalities, such as audio, visual, and textual sources, into distinct clusters. These clustering techniques operate by extracting shared features across modalities in an unsupervised manner, where the identified common features exhibit high correlations within real-world objects. Recognizing the importance of perceiving the correlated nature of these features is vital for enhancing clustering accuracy in multi-modal settings. This survey explores Deep Learning (DL) techniques applied to multi-modal clustering, encompassing methodologies such as Convolutional Neural Networks (CNN), Autoencoders (AE), Recurrent Neural Networks (RNN), and Graph Convolutional Networks (GCN). Notably, this survey represents the first attempt to investigate DL techniques specifically for multi-modal clustering. The survey presents a novel taxonomy for DL-based multi-modal clustering, conducts a comparative analysis of various multi-modal clustering approaches, and deliberates on the datasets employed in the evaluation process. Additionally, the survey identifies research gaps within the realm of multi-modal clustering, offering insights into potential future avenues for research.

1. Introduction

Clustering, an unsupervised learning technique, involves partitioning objects into groups based on shared characteristics, distinguishing them from objects in other groups [1]. This technique finds applications in various fields, such as medical diagnosing through image segmentation [2] or reports [3], stock market analysis [4], intelligent marketing [5], network analysis [6], and numerous other domains.

Traditional clustering methods rely on extracting features from single-modal data, such as textual, image, audio, or video data. However, in the era of web-level scraped multi-modal data, the proliferation of multi-sensory Internet of Things (IoT) devices, and the diversity of sensors, a rich source of multi-modal data has emerged. Multi-modal data refers to data from different types and contexts [7], such as audio-visual, visual-text, numeric-signal, etc. This abundance presents an opportunity to extract valuable insights across diverse fields [8]. Recognizing that common features exist across multi-modal data and are correlated, the cognitive community acknowledges the parallel with how the human brain processes correlated features simultaneously, enhancing object perception and recognition [9]. This underscores the significance of multi-modal clustering, as it establishes links between different data types, making the clustering model adaptive and robust.

Acquiring data labels for multi-modal data obtained from various sources can be challenging and expensive. Additionally, multi-modal data often exhibits complex relationships and patterns that may lack explicit labels [10]. While supervised techniques are crucial for specific tasks in multi-modal data analysis, they rely on labeled data and are best suited for targeted objectives.

On the contrary, clustering methods demonstrate greater adaptability to multi-modal data, providing a distinct perspective. They excel in uncovering exploratory insights, enabling a more comprehensive understanding of the intrinsic structure within the data. This becomes particularly valuable in scenarios where the data lacks clear labels or the objective is to unveil hidden patterns. Recent approaches [11–13] incorporate clustering within unsupervised, semi-supervised, and few-shot learning frameworks, facilitating realistic learning environments for open-world and big data challenges.

Moreover, clustering plays a significant role in unsupervised pretraining [14] and vice versa [15]. Notably, unsupervised pre-training has gained prominence for its ability to enhance the performance of supervised tasks without the reliance on expensive labeled data [16]. It involves learning valuable representations or features from unlabeled data without explicit guidance or supervision [17]. Recent attention has been given to pertaining leveraging multi-modal and multi-view data [18].

The exploration of clustering techniques for multi-modal data has garnered considerable attention from the research community in recent

* Corresponding author.
E-mail addresses: U16103218@sharjah.ac.ae (S. Raya), U19103914@sharjah.ac.ae (M. Orabi), iafyouni@sharjah.ac.ae (I. Afyouni), zaher@sharjah.ac.ae (Z. Al Aghbari).

https://doi.org/10.1016/j.neucom.2024.128348
Received 5 September 2022; Received in revised form 4 May 2024; Accepted 10 August 2024
Available online 14 August 2024
0925-2312/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

years [19]. The objective is to emulate the human brain's ability to uncover relations and similarities in the real world. DL techniques, inspired by human Neural Networks (NNs), have become increasingly prominent in clustering multi-modal data. Specifically, NNs have been applied to tasks such as object prediction [20], object detection [21], and enhancement of clustering performance [22].

Utilizing DL techniques, clustering algorithms extract features from various modalities, such as audio, visual, and text, to represent objects. Subsequently, during the parameter learning of the deep NN, clustering algorithms generate clusters based on the correlated features from different modalities; a process referred to as *Deep Clustering* (DC). DL techniques for feature extraction are categorized into AE-based [23], RNN-based [24], CNN-based [25], and GCN-based [26] methods.

Several DC frameworks integrate features from diverse modalities, as observed in [23,27–32]. These frameworks leverage different clustering algorithms, including K-means [23,27], k-medoids [22], agglomerative [33], and spectral [24]. However, it is noteworthy that the exploration of DC methods for multi-modal data remains incomplete.

General reviews on clustering techniques are found in [34,35]. However, specific discussions on clustering algorithms for single-modal data employing DC are detailed in [36,37]. In [38], a taxonomy of multi-modal clustering using traditional feature selection methods is introduced. Despite these contributions, the existing reviews lack comprehensiveness, and alternative taxonomies for DC are presented in [19, 39–41].

Notably, none of the prior works provides a taxonomy for Multi-Modal Deep Clustering (MMDC) that integrates all three crucial components: clustering, DL, and multi-modal data.

This paper stands as the pioneering work in surveying clustering techniques for multi-modal data through the lens of DL, termed herein as MMDC.

2. Key contributions

This survey paper makes significant contributions to the understanding of MMDC data through the following key aspects:

- **Comprehensive review:** This systematic review stands as the first comprehensive examination of techniques employed in MMDC, offering a thorough analysis of existing methodologies.
- **Innovative taxonomies:** Introducing three novel taxonomies tailored for the specific nuances of MMDC. These taxonomies categorize research based on clustering techniques, leveraged modalities, and involved mechanisms, providing a structured framework for understanding the landscape.
- **Detailed methodology comparison:** Conducting an in-depth comparison of various MMDC methods and the frameworks utilizing them. This analysis offers valuable insights into the diverse approaches, datasets used, underlying assumptions, mechanisms involved, and limitations.
- **Roadmap and gap identification:** The survey goes beyond a mere compilation and analysis, identifying the research roadmap and critical gaps in the domain of MMDC. This identification not only synthesizes the existing knowledge but also serves as a guide for future research, opening new avenues within this rapidly evolving field.

The subsequent sections of this paper are organized as follows: Section 3 delves into related works, offering insights into the existing literature. Section 4 is dedicated to show the followed method to conduct this review. Following this, Section 5 provides a concise background, introducing fundamental concepts related to the clustering of multi-modal data, DC, and MMDC. It also unveils the proposed taxonomies specifically tailored for MMDC clustering techniques, utilized modalities, and involved mechanisms. Furthermore, Section 6 meticulously explores the taxonomy of frameworks for MMDC. Moreover, Section 7 engages in a thorough discussion of various approaches within the realm of MMDC methods, along with outlining potential future lines of research. To conclude, Section 8 summarizes the findings of the review and guides for possible future directions for research.

3. Related work

While there are existing survey papers on clustering, multi-modal clustering, and DL-based clustering individually, there is a notable gap in the literature concerning survey papers specifically addressing DL-based clustering using multi-modal data. In light of this, the following section provides a comprehensive discussion of recent survey papers covering clustering in a general context and those focusing on multi-modalities.

3.1. *Clustering*

There is a limited number of surveys specifically dedicated to clustering techniques for multi-modal and monomodal data. Existing survey papers either focus on clustering multi-modal data without incorporating DL or delve into DL-based clustering for monomodal data.

Rai et al. [34] presented a comprehensive taxonomy of clustering techniques, providing an in-depth discussion of each method. Swarndeep Saket et al. [35] provided an overview of five types of partition-based clustering techniques, offering a comparative analysis of their advantages, disadvantages, complexities, and efficiency.

In the realm of monomodal data clustering, Sajana et al. [36] summarized clustering algorithms, including Partition-based, Hierarchical-based, Density-based, Grid-based, and Model-based clustering algorithms. However, it is important to note that their review primarily covered clustering methods proposed between 1973 and 2002.

3.1.1. *Deep clustering*

The architectures of DC techniques for unimodal data have been comprehensively examined by Min et al. [37]. In their work, the authors introduced a taxonomy of DC comprising four distinct categories: AE-based, Clustering Deep NN-based, Variational AE-based, and Generative Adversarial Network-based. Furthermore, they engaged in a detailed discussion and comparison of these techniques, summarizing their respective algorithms.

3.1.2. *Clustering multi-modal data*

Multi-modal clustering involves the unsupervised division of multi-modal data into distinct clusters [29].

Mehmood et al. [38] introduced clustering techniques based on the particle swarm optimization (PSO) algorithm, specifically focusing on PSO-based multi-modal partition-based clustering. The employed clustering methods encompassed k-means, k-harmonic means, and hybrid clustering. Notably, the survey lacked information regarding the specific dataset types and modalities utilized.

3.2. *Multi-modalities*

Ramachandram et al. [39] conducted a comprehensive review of research on the fusion of multi-modal data using DL methods. The survey encompassed various aspects, including applications, models, fusion structures, multi-modal regularization, fusion structure learning, optimization techniques, and the datasets employed. The covered research papers spanned the period from 2011 to 2017.

Bayoudh et al. [40] conducted a review of research papers between 2015 and 2020, exploring the role of DL methods in the fusion of multi-modal data. The authors presented different types of data fusions and highlighted datasets frequently used in research works during the period from 2011 to 2017.

Chen et al. [41] conducted a review of research papers focusing on deep multi-modal content understanding. Their discussion encompassed multi-modal applications, challenges associated with deep multi-modal learning, and the latest advancements in deep multi-modal feature learning. The reviewed research papers in this study were published between 2016 and 2019.

Similarly, Baltrušaitis et al. [19] proposed a taxonomy for multi-modal machine learning. They outlined challenges in multi-modal clustering using machine learning, including translation, data fusion, alignment, object representation, and co-learning. The authors delved into various applications such as speech recognition and synthesis, event detection, emotion and affect, media description, and multimedia retrieval. Despite investigating multi-modal data and DL, the surveys in this section do not specifically address clustering.

A distinctive feature of our present survey, in contrast to prior works, lies in its encompassing exploration of techniques that integrate three essential components: multi-modal data, clustering algorithms, and DL-based feature extraction and processing.

4. Systematic review methodology

This survey aims to address the following research questions:

RQ1 : What are the categories of multi-modal clustering using DL methods?

RQ2 : What datasets are used for evaluating the clustering results in the literature?

RQ3 : What characteristics of DL have been utilized?

RQ4 : What are the MMDC algorithms discussed in the literature?

RQ5 : What performance measures are provided in the literature for clustering techniques?

RQ6 : What are the gaps in this research area, and what future lines of research are suggested?

To address these questions, this review adheres to the Systematic Literature Review guidelines proposed by [42]. The search process focused on identifying publications related to multimodal data clustering using DL, limited to all published research before February 2024. The search terms used in IEEE, Springer, Elsevier, ACM, Wiley, AAAI, and Neural Information Processing Systems (NeurIPS) libraries included:

- “Deep Multi-modal Clustering”
- “DL with Multi-modal Data Clustering” OR “Multi-modal Data Clustering”
- “Deep Multi-modal Data” AND “Deep Clustering”
- “Deep Clustering” And “Multi-modal Data”

The initial search included more than a thousand research papers, which were manually filtered based on relevance, ending up with 39 papers. Notably, no publications on MMDC were found between 2011 and 2013, with a significant increase observed in the last few years, indicating a growing interest in this topic.

5. Multi-modal data deep clustering

To initiate this review, it is crucial to grasp the concept of MMDC. Thus, this section provides concise insights into multiple modalities, the clustering of multi-modal data, MMDC, and mechanisms used in MMDC.

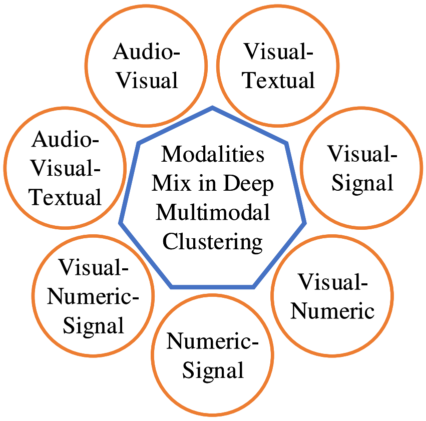


Fig. 1. Different modalities combination used in MMDC.

5.1. Multiple data modalities

The primary modalities encompass image, video, audio, text, numeric, and time series signals. Contemporary research often delves into the consideration of multiple modalities, typically involving two modalities, as depicted in Fig. 1. It is worth noting that audio and time series signal data share similarities in their temporal dependencies and sequential structure, yet they diverge in continuity, underlying features, and representation approaches. Audio signals, representing fluctuations in air pressure over time, are continuous and analog. In contrast, time series signals, encompassing diverse data types, can be either continuous or discrete and undergo feature extraction based on specific applications. While audio signals are often visualized as waveforms or spectrograms, time series signals are represented as sequential data points, typically depicted through line plots. These distinctions highlight their unique characteristics, designating them as distinct modalities. Throughout this article, the term “signal” specifically denotes time series signals.

Understanding the distinctions between multi-view (or multiview) and multi-modal data is crucial for navigating diverse data representations. In the realm of multi-view data, instances are portrayed through multiple high-level perspectives [43]. This can involve representing the same instance in various views of a single modality or generating multiple views from a singular source [44]. This stands in contrast to multi-modal data, where a single instance is originally represented across different modalities.

Additionally, the concept of mixed-modal data introduces a nuanced perspective. In this context, data is characterized by multiple modalities, but instances within the data are presented in a singular modality [45,46]. What sets multi-modal data apart is the emphasis on processing the relationships between distinct modalities’ representations of an instance; a crucial step not explicitly addressed in traditional mixed-modal data [47].

The focus of this work is a comprehensive exploration of the research domain dedicated to developing solutions for MMDC This involves exploring methods that tackle the complex task of blending information from different sources, exploring complex relationships, and helping to improve data analysis and mining through MMDC.

5.2. Clustering multi-modal data

Clustering serves the purpose of grouping data objects with similar features, aiding in the extraction of latent knowledge from a dataset, including object groupings and the determination of group numbers [48]. Various clustering techniques exist, such as the K-means clustering algorithm, which endeavors to group data points (e.g., objects) into

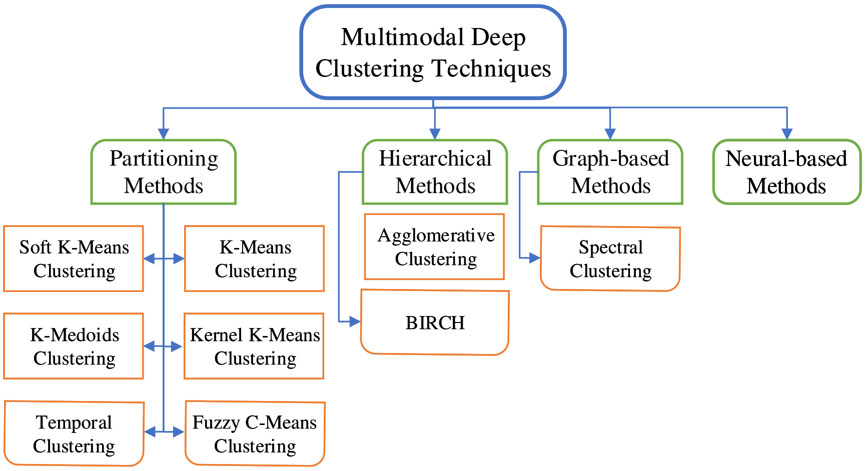


Fig. 2. MMDC methods.

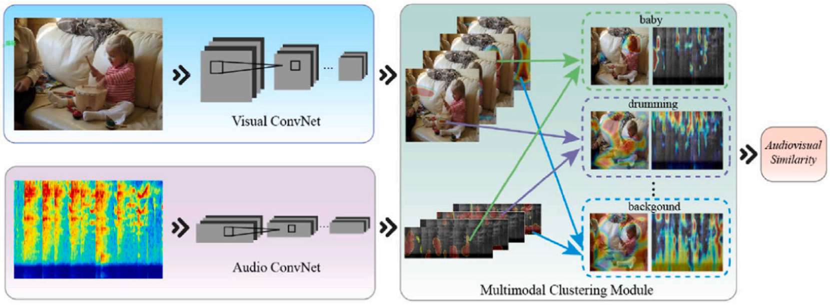


Fig. 3. MMDC using CNN [9].

k clusters while minimizing distances within the same cluster and maximizing distances between objects in different clusters [9].

When capturing data from different modalities related to the same scene, it is observed that common features among multi-modal data are often correlated [49]. Therefore, clustering multi-modal data attracted the attention of recent research. The different clustering methods used in the reviewed research are found in Fig. 2 and are explained in detail in Section 6.

5.3. Deep clustering multi-modal data

The integration of DL architecture with clustering is termed DC. NNs are interconnected models designed to mimic the functioning of the human brain. Deep NNs, an extension of traditional NNs, possess a deeper architecture characterized by a substantial number of hidden layers. The term “deep” indicates the increased depth compared to conventional NNs, allowing for more effective data representation through layer-by-layer extraction [23].

CNNs are pivotal in feature extraction from modalities like images and audio. CNN comprises three types of layers: Convolutional layers, Pooling layers, and Fully Connected layers (FC). The convolutional layer, the initial layer in CNN, extracts features from input data, creating a feature vector representing an object such as an image. The second layer, often the pooling layer, reduces inter-layer connections, operating independently on each feature map to minimize computational costs and serving as a link between the convolutional layer and the FC layer. The FC layer, positioned before the CNN output, incorporates weights, biases, and neurons to connect neurons across two layers.[1] Fig. 3 illustrates an instance of MMDC with CNN architecture [9].

Conversely, Deep RNNs enhance the performance of sequential data, including text and audio [50]. DL techniques have gained widespread popularity across various domains, including object recognition in images, speech-to-text conversion, content matching to user interests, and optimizing search results [50].

Another noteworthy model within the realm of DL is the AE, proven to enhance the accuracy of replicating input data [29]. AE consists of two main components: encoders and decoders. Encoders learn the latent representation of input data, while decoders master the reconstruction of input data from the acquired latent representation [27]. Fig. 4 provides an example of an AE-based application of MMDC.

In recent developments within DL for multi-modal data clustering, Graph Neural Networks have emerged as a prominent trend, with a specific focus on GCNs. These advanced techniques excel in deriving intricate graph representations from input data, effectively integrating the feature space across different modalities [26].

5.4. Involved mechanisms in deep multi-modal clustering

Different mechanisms are involved in approaches for solving deep multi-modal clustering problems, as depicted in Fig. 5. These mechanisms either represent tools used to gain specific benefits or assumptions consequences.

- The clustering process initiates with the extraction of features from diverse modalities. Following this, the fusion of common features takes place, paving the way for the subsequent clustering operation on the feature representations. Data representation fusion can be classified into three main types [40]:
  - Early fusion: This involves merging representations from different modalities into unified representations prior to

[1] https://www.upgrad.com/blog/basic-cnn-architecture/

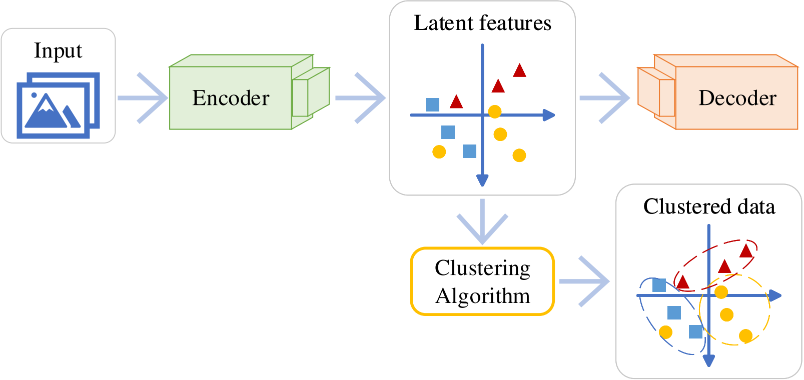


Fig. 4. MMDC using AE.

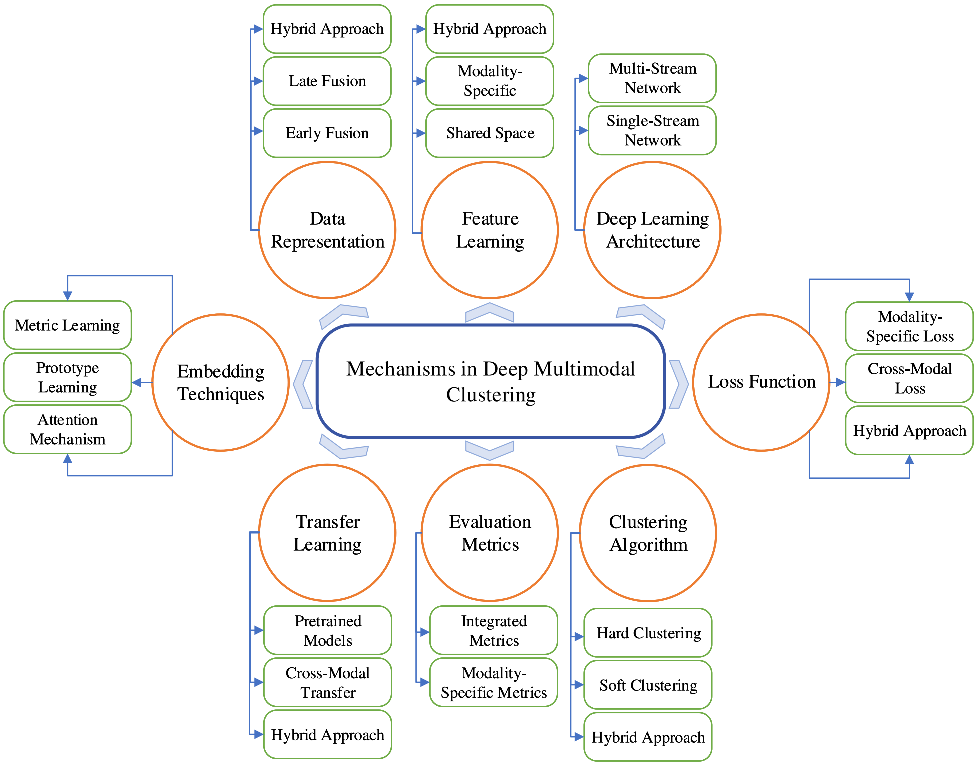


Fig. 5. Mechanisms in Deep Multi-modal Clustering.

the clustering process [11]. While early fusion promotes cross-modal learning, it has the drawback of expanding the representation space significantly.

– Late fusion or decision fusion: In this approach, features are extracted independently for each modality [12]. The processing of distinct representations is carried out separately, and, at times, individual clustering is performed. The fusion then occurs at a later stage. This simplifies the model training process and allows for modality-specific data extraction. However, there is a risk of overlooking relationships between modalities.
– Hybrid fusion or intermediate fusion: This approach combines elements of both early and late fusion. Multiple fusions are applied to address various tasks within the model [26].

Each fusion strategy offers unique advantages and trade-offs, making the choice dependent on the specific requirements and characteristics of the multi-modal data being processed.

- The learned feature representations in multi-modal learning can be categorized into the following scenarios:
  – Shared space learning: A shared space is learned during the model training phase to map data from different modalities into a common feature space, promoting modality-invariant representations [26]. This scenario explores shared patterns and dependencies between modalities, crucial in situations where there are high dependencies, such as audio and visual dependencies in videos [11].
  – Modality-specific learning: Separate models are trained for each modality, and features are mapped to distinct spaces. This allows the capture of knowledge specific to each modality, aiming to obtain unique features in each modality [51]. This approach is valuable when different modalities inherently embed unique and independent features, as seen in tasks like clustering multi-modal web data [51].

  - Hybrid approach: A hybrid approach combines both shared space learning and modality-specific learning. Some approaches pair related embeddings of different modalities after modality-specific learning [12].

  These scenarios provide a spectrum of choices, allowing researchers to tailor their approach based on the characteristics of the data and the goals of the multi-modal learning task.
- Multi-modal data clustering using DL models requires tailored architectures to suit the specific needs and tasks. The DL architectures for MMDC can be categorized as follows:
  - Single-stream networks: A single neural network path processes multi-modal data. Inputs from different modalities are fed into the same neural network model, traversing a unified path [26]. This simplified architecture is suitable when joint processing is viable.
  - Multi-stream networks: Multiple neural networks are employed, each dedicated to processing a specific modality [52]. This approach allows for customized processing of different modalities, particularly when they provide complementary information.
- DL model training necessitates robust loss functions for effective optimization. The design of loss functions exhibits two main trends in the literature:
  - Modality-specific loss: Loss functions are customized for individual modalities to guide the learning process [52]. Models employing dedicated pipelines for various modalities train each deep learning pipeline using its specific loss function. This approach aligns with the goals of multi-stream networks.
  - Cross-modal loss: Cross-modal loss functions capture relationships between instances of different modalities [26]. When a shared feature space is employed, cross-modal loss becomes valuable.
  - Hybrid approach: A hybrid approach integrates both modality-specific and cross-modal loss types. Different modal-specific and cross-modal losses can be combined into a single loss function for optimization [53].
- In addition to representations and deep learning model design, clustering methods exhibit distinct characteristics utilized for various applications.
  - Soft clustering: In soft clustering, instances are assigned to multiple clusters with probabilistic memberships. Techniques like fuzzy C-means [21] and soft K-means [30] are applied when clusters are not mutually exclusive. For instance, patients can have multiple health conditions [54].
  - Hard clustering: Approaches like k-means clustering, Kernel k-means [55], and K-medoids [22] assume hard clustering. In this method, instances are assigned to a single cluster based on the highest probability, leading to mutually exclusive cluster assignments. This is useful in scenarios such as speaker recognition [52].
  - Hybrid approach: A hybrid approach integrates different clustering types for various modalities [56], offering a flexible and tailored strategy.
- When assessing the performance of MMDC models, two primary evaluation approaches are observed:
  - Integrated metrics: Clustering performance is evaluated by considering all modalities collectively [51].
  - Modality-specific metrics: In addition to assessing the overall MMDC model performance, some models are evaluated based on the individual performance of each modality [57].
- In addition to the fundamental assumptions and characteristics essential for an MMDC model, literature explores additional mechanisms, including transfer learning and embedding techniques. Two prominent transfer learning approaches are discussed:
  - Pre-trained models: Modality-specific pre-trained models are commonly utilized for extracting features [58]. The objective is to enhance model performance by transferring knowledge from models extensively trained.
  - Cross-modal transfer learning: In this approach, knowledge learned in one modality is transferred to another during training. The goal is to improve the model's performance in a specific modality, particularly when data is limited or absent in that modality but abundant in another [53].
  - Hybrid approaches: Both pre-trained models and cross-modal transfer learning approaches can be seamlessly combined for enhanced performance [52].
- The literature explores the following embedding techniques:
  - Metric learning: Involves learning a distance function to measure the similarity between points. The data is mapped to a space where similar instances are brought closer together, while dissimilar instances are pushed farther apart [33].
  - Prototype learning: Focuses on learning representative prototypes or centroids for clusters, often in the form of central points. These prototypes serve as key reference points during clustering [58].
  - Attention mechanisms: Empowers the model to selectively attend to specific features or regions, emphasizing important information during the learning process [59].

6. Deep clustering of multi-modal data methods

This section presents the review body and the taxonomy of MMDC. Fig. 2 illustrates the taxonomy as discussed in Section 5.2. Table 1 summarizes the used clustering method, frameworks, modalities, evaluation metrics, and the reported results.

6.1. Partition clustering

6.1.1. K-means clustering

K-means clustering, a fundamental Unsupervised method, is widely employed in various studies due to its simplicity and efficient computational time, especially for large datasets [71]. The use of K-means is prevalent, as indicated in Table 1, where it outperforms hierarchical methods in terms of computational speed. Notably, K-means yields compact clusters compared to hierarchical methods, contributing to its popularity [35]. In K-means, clusters are defined by their centroids, represented by the mean of points within the cluster [34]. The algorithm begins by specifying the number of clusters, denoted as k, followed by the selection of k centroid points. Subsequently, each data point is assigned to its closest centroid. The process iteratively updates centroids and re-assigns data points until convergence, where centroid points remain unchanged or exhibit minimal average change, often below a specified threshold.

Audio-visual modalities. Miao et al. [25] proposed an unsupervised multi-modal clustering method with cross-modal communication, optimizing visual clusters using audio information. Leveraging DL and metric learning on a multimedia feature extraction method named Deep Feature, the authors achieved improved clustering performance. They used raw audio features such as Mel frequency cepstral coefficients and fundamental frequency (F_0), and VGG-16 Net for visual features. Clustering was performed using K-means on the extracted audio-visual features, evaluated on datasets including The Big Bang Theory and raw

Table 1
Multi-modal data clustering techniques using DL.

Category	Clustering method	Deep Arch.	Ref.	Framework	Multi-modal					Evaluation metric	Evaluation results
					Audio	Visual	Text	Time series signals	Numeric		
Partition Clustering	K-means	CNN	[25]		✓	✓				ACC[a]	77.04%
		CNN	[9]	DMC	✓	✓					
		CNN	[60]	New York Melange		✓	✓				
		CNN	[61]	City Melange		✓	✓				
		AE	[23]	DMMC		✓	✓			ACC, NMI	ACC = 63.71, NMI = 75.42
		CNN	[20]	Sound	✓	✓				mAP[b]	44.1%
		CNN	[58]	MCN	✓	✓	✓			NMI, ARI, ACC, H[c], pmax	NMI = 65.5, ARI = 48.5, ACC = 57.6, H = 0.34, pmax = 83.8
		AE	[27]	DCUMC		✓	✓			ACC, NMI	ACC = 63.30, NMI = 76.63
		AE	[28]	DMCR		✓	✓			ACC, NMI, Purity	ACC = 66.68, NMI = 50.93, Purity = 61.70
		VAE[d]	[29]	DMGCR		✓	✓			ACC, NMI, Purity	ACC = 66.68, NMI = 50.93, Purity = 61.70
		VAE	[29]	DMLCR		✓	✓			ACC, NMI, Purity	ACC = 67.75, NMI = 52.84, Purity = 65.52
		AE	[62]	DMIM		✓	✓			ACC, NMI	ACC = 0.561, NMI = 0.371
		AE	[63]	AE^2-Nets		✓	✓			ACC, NMI, F1[e], RI	ACC = 77.75 ± 1.63, NMI = 78.61 ± 1.62, F1 = 70.96 ± 2.63, RI = 93.92 ± 0.58
		CAE	[31]	FDMMC		✓	✓			ACC, NMI, ARI	ACC = 0.684,NMI = 0.728, ARI = 0.692
		CNN	[13]	XDC	✓	✓				ACC	95.5%
		CNN	[32]	EFN	✓	✓				F_{co}	63
		CNN	[32]	LFN	✓	✓				F_{co}	64
		AE	[12]	TS2ACT		✓		✓		ACC, RC, F1	1-shot: ACC ≥ 0.7, 20-shot: ACC 0.82-0.92
		AE	[51]		✓	✓	✓			retrieval ACC, RC, training time	ACC 74.74
		AE	[52]		✓	✓				EER[f], minDCF[g], NMI, ACC, purity	EER 1.8%
		CNN	[53]		✓	✓				ACC	Increase 0.03% ACC
	Soft K-means	CNN	[30]	EAMC		✓	✓			ACC, NMI, Purity	ACC = 0.945, NMI = 0.937, Purity = 0.952
		GCN	[26]	GECMC		✓	✓			ACC, NMI	ACC around 0.5
	Kernel K-means	CNN/RNN	[55]	M&M TGM	✓	✓	✓			BLEU, METEOR, CIDEr	BLEU4 = 48.67, METEOR = 34.36, CIDEr = 80.45
	K-medoids	CNN	[22]			✓	✓			Purity, Entropy	Purity = 0.7709, Entropy = 0.2327
	Fuzzy Clustering	CNN	[21]	DFCN		✓	✓			NPR[h]	89.13%
		AE	[54]	F-HoFCM		✓			✓	SC[i] and ARI	5%–11% SC increase
	Temporal Clustering	AE (CNN+Bi-LSTM)	[56]					✓	✓	Sensitivity, Specificity, AUROC[j], AUPRC[k], SC	SC about 0.7
Hierarchical Clustering	Agglomerative	CNN	[33]	Deep-AD	✓	✓				Recall, precision, F1	R = 0.95, P = 0.81, F1 = 0.88
		CNN	[64]	MHCI		✓	✓			F-measure, NMI	F-measure = 0.998, NMI = 0.986
		CNN	[65]	NGW-CVT		✓	✓			BLEU-3 ratio[l]	0.079
		GCN + AE	[59]	TPIT-C		✓	✓			ACC, ARI, NMI, F-score, and purity	increase of 0.02–0.03 ACC
	BIRCH	CNN	[66]		✓	✓					
		AE	[67]			✓	✓			ROC[m] curve, AUC[n], AUPRC, F1, FP/FN[o]	increase of 0.074–0.01 AUC, and 0.169–0.054 AUPRC.

(*continued on next page*)

Table 1 (*continued*).

Category	Clustering method	Deep Arch.	Ref.	Framework	Multi-modal					Evaluation metric	Evaluation results
					Audio	Visual	Text	Time series signals	Numeric		
Graph-based Clustering	Spectral Clustering	CNN	[24]		✓	✓					–
		CNN	[68]			✓	✓			M_{iou}, F_{co}	M_{iou} = 0.51, F_{co} = 0.62
		GCN	[57]	MCGEC		✓		✓	✓	0.25 ACC, F1, NMI, and ARI	0.25 ACC, 0.31 F1, 0.08 NMI, and 0.11 ARI
Neural-Based Clustering	WTA hashing	MLP	[11]		✓	✓				ACC	ACC 50–56%
	OT-based [69]	MLP	[69]		✓	✓				NMI, ARI, ACC	ACC 32%, 6.6%, 43%, 55%
		AE + Faster R-CNN	[70]		✓	✓				mAP	comparable with supervised approaches

[a] Accuracy (ACC).
[b] Mean Average Precision (mAP).
[c] Mean Entropy Per Cluster (H).
[d] Variational Auto-Encoders (VAE).
[e] F1-score (F1).
[f] Equal Error Rate (EER).
[g] Minimum Detection Cost Function (minDCF).
[h] Naming Precision Rate (NPR).
[i] Silhouette Coefficient (SC).
[j] Area Under the Receiver Operating Characteristic (AUROC).
[k] Area Under the Precision-Recall Curve (AUPRC).
[l] The BLEU-3 ratio represents the percentage of times a model obtains values greater than 0 for BLEU-3.
[m] Receiver Operating Characteristic (ROC).
[n] Area Under the Curve (AUC).
[o] False Positive/False Negative (FP/FN).

videos, demonstrating the algorithm's effectiveness through accuracy comparisons with two other frameworks.

Owens et al. proposed the Sound approach [20] utilizes K-means clustering with audio and visual modalities, estimating statistical sound summaries, predicting sound textures from images using a CNN, and employing K-means to cluster sound textures. Trained on 360k videos from the Flickr dataset and tested on PASCAL VOC 2007 and SUN datasets, the model achieved a 15.8% successful sound label selection rate with 30 clusters.

Furthermore, Trojahn et al. [32] employed K-means clustering to cluster audio and CSIFT (visual color-based) features in the Early Fusion Network (EFN) and Late Fusion Network (LFN). Clusters were represented by selecting the medoid of each cluster. The architecture used CNN for visual feature extraction and proposed two RNN networks based on fusion type. The BBC Planet Earth dataset was used for evaluation with the F1-measure (F_{CO}) as the metric.

In [9], Hu et al. proposed MMDC of audio and visual data, focusing on tasks like sound localization and real-life sound event detection. Using 400K unlabeled videos from the SoundNetFlicker dataset, they extracted 5-second sound-image pairs. Visual and audio features were obtained using global pooling over VGG16 net's convolutional features and the VGG*issh* network, respectively. K-means clustering was applied to identify centers corresponding to specific objects or audio. Evaluations highlighted the method's efficacy in real-world sound detection.

In [13] Alwassel et al. proposed three self-supervised multi-modal frameworks for audio-visual modalities: MDC, CDC, and XDC. The authors employed K-means clustering on R(2+1)D visual features and ResNet audio features, achieving 95.5% accuracy on the UCF101 dataset.

Piergiovanni et al. [53] propose a representation learning model for video data. The model creates modal-specific feature representations and employs loss distillation to transfer knowledge to an RGB network, generating unified representations. An evolutionary search algorithm optimizes loss functions guided by a power law distribution. This self-supervised approach outperforms previous methods, except on large fully labeled datasets.

Cai et al. [52] introduce a self-supervised framework for video-based speaker recognition using audio-visual modalities. The method combines contrastive learning with clustering, training audio, and visual encoding networks iteratively. Pseudo-labels generated through clustering are fused using a cluster ensemble algorithm for improved performance.

Visual-text modalities. New York Melange [60], proposed by Zahálka et al. is a multi-modal city explorer using visual, text, and user data. CNNs extracted visual features and Latent Dirichlet Allocation processed text. K-means clustering was performed independently for images and text, with data collected from Foursquare, Picasa, and Flickr. With a similar approach and datasets, the authors extended this work in [61] by proposing the City of Melange; an interactive multi-modal city exploration system that recommends venues based on the user's preferences.

Zhang et al. in [63], the AE^2-Nets framework combines clustering and classification using K-means and k-nearest neighbors methods, respectively. It excels in view-specific and multi-view representation learning. Evaluation using five datasets and clustering measures: normalized mutual information (NMI), accuracy, Rand Index (RI), and F-score demonstrates AE^2-Nets' superior accuracy.

In [31], Yang et al. proposed a Fused Deep Multi-modal Clustering framework (FDMMC) utilizing K-means in the clustering layer. FDMMC employs a Convolutional Autoencoder (CAE) to encode text and enhance images, followed by clustering on the fused images. Evaluation on the Coco_cross Dataset includes metrics such as accuracy, NMI, and Adjusted Rand Index (ARI).

DMMC is an end-to-end model proposed by Zhang et al. [23]. DMMC employs two AEs, each handling one modality, and integrates clustering through a fusion layer. Using K-means, the framework achieves top accuracy on nine out of ten datasets for clustering tasks.

Zong et al. proposed Deep-Commonness and Uniqueness-Mining Clustering (DCUMC) [27], which consists of two steps: First, extracting shared and unique features per modality using an AE structure, facilitating cross-reconstruction of modality features. AEs, with encoder and decoder components, capture modality-common and modality-unique features, with cross-reconstruction linking these features across modalities. Second, shared consistent features are obtained by fusing modalities' features and clustering the data using K-means. DCUMC is evaluated on five benchmark datasets, showcasing superior performance in accuracy and NMI metrics on most datasets.

Zhang et al. [28] proposed Deep Multi-modal Clustering Cross-Reconstruction (DMCR) to mitigate distribution differences between modalities in feature space. DMCR comprises two phases: the first utilizes AEs and Variational Information Bottleneck for cross-reconstruction, while the second fuses common features for clustering using K-means. Evaluated on benchmark datasets, DMCR shows its efficacy through metrics like accuracy, NMI, and purity. Zhao et al. [29] extended this work with new cross-reconstruction AE algorithms: Deep Multi-modal Global Cross-Reconstruction (DMGCR) and Deep Multi-modal Local Cross-Reconstruction (DMLCR). Both employ the Variational Information Bottleneck (VIB) for mutual latent feature extraction, followed by K-means clustering. Evaluations of the algorithms demonstrate their effectiveness through accuracy, NMI, and purity metrics.

Mao et al. proposed Deep Mutual Information Maximin (DMIM) utilizing K-means clustering in a cross-modal setting [62]. This end-to-end method aligns latent feature distributions, maximizes shared data, and eliminates duplicated individual information through a multi-modal shared encoder, clustering, and over-clustering layers. Evaluated on four image and text datasets, DMIM outperforms ten baseline methods in accuracy and normalized mutual information metrics.

Audio-visual-text modalities. Multi-modal Clustering Network (MCN) [58] leverages audio, visual, and text modalities. Their K-means clustering utilizes a joint multi-modal space with semantically connected instances and a contrastive loss. Visual features are extracted from pre-trained ResNet-152 and ResNeXT-101 models, audio features are obtained through log-Mel spectrograms, and textual features use Word2 Vec embeddings. MCN is trained on the HowTo100M dataset and evaluated on YouCook2, MSR-VTT, CrossTask, and Mining Youtube datasets, assessing clustering performance with metrics including NMI, ARI, accuracy, Mean Entropy Per Cluster (H), and Mean Maximal Purity Per Cluster (Pmax).

Visual-signal modalities. Xia et al. [12] present TS2ACT; a cross-modal co-learning strategy designed for few-shot Human Activity Recognition (HAR) using embedded sensor data. TS2ACT employs an augmentation technique that utilizes semantic-rich label text to search for related web images and construct an augmented dataset containing partially-labeled time series and fully-labeled images. It utilizes contrastive learning, DC with a time series encoder, and a pre-trained image encoder to extract features from the time series data. TS2ACT achieves results comparable to fully supervised approaches.

Considering the increasing amount of web videos, e-commerce platforms process streaming multi-modal data for different purposes. Huang et al. [51] introduce a self-supervised multi-modal co-training that utilizes cross-modal pseudo-label consistency to facilitate the joint learning of representations across videos' different modalities, i.e. image, audio, video, motion, and text modalities. The authors clustered different modals and co-trained the model across modalities using credible samples only. They improve the training efficiency of their model to increase the scalability reducing the required time by 2.7 folds on a data set of 1.4 billion videos at the Alibaba website.

6.1.2. Soft k-means clustering

Visual-text modalities. Zhou and Shen [30] proposed the End-to-end Adversarial-attention Multi-modal Clustering (EAMC) method, incorporating modality-specific feature learning, modality fusion, and Soft k-means clustering for cluster assignment. Compared to nine baseline methods on five datasets, EAMC demonstrated superior performance in terms of accuracy, NMI, and purity, except on the MNIST dataset, where it outperformed accuracy and NMI.

Xia et al. [26] present the Graph Embedding Contrastive Multi-modal Clustering network (GECMC), a unified framework integrating representation learning and multi-modal clustering through deep NNs. GECMC utilizes contrastive loss with pseudo-labels to establish a common representation space for different modalities, followed by clustering. The soft k-means variant employs a Student t-distribution for direct prediction of soft clustering labels, eliminating the need for post-processing techniques to learn cluster centroids. The proposed method demonstrates superior intra-cluster representation similarities and inter-cluster dissimilarities, outperforming 14 competitive approaches.

6.1.3. Kernel k-means

Audio-visual-text modalities. The kernel K-means clustering algorithm enhances traditional K-means clustering. In [55], the authors introduced the M&M Topic-Guided Model (M&M TGM) for video captioning. M&M TGM incorporates topic mining, topic prediction, and topic-aware decoding. Topic mining involves extracting ground truth through stop-word removal and bag-of-audio-words conversion. A two-layer perceptron serves as the topic predictor based on multi-modal video features. Topic-aware decoding, using RNN-Long Short-Term Memory (LSTM), guides sentence generation with topic information. The end-to-end M&M TGM predicts latent video topics and generates topic-guided descriptions, utilizing Kernel K-means clustering to cluster video captions, where clusters represent latent topics. Evaluation on Youtube2Text and MSR-VT datasets utilized CIDEr, METEOR, BLEU, and ROUGE-L metrics, showcasing superiority over vanilla and single-task TGM methods.

6.1.4. K-medoids method

The K-medoid method, a partition clustering approach utilizing one cluster member as a centroid, calculates object similarities once.

Audio-visual-text modalities. In [22], an approach employing WordNet thesaurus for estimating textual semantic similarity in videos is presented. Visual features are extracted via a CNN, and web video search results are clustered, grouping videos of the same category. Utilizing visual and text modalities (video and its title, tags, and description), the authors compute similarity between visual features, resulting in a similarity matrix. K-medoid clustering is then applied. The evaluations used the TubeKit dataset, YouTube's open-source crawler, with 1580 video clips retrieved from 18 queries on the YouTube search engine.

6.1.5. Fuzzy clustering

The Fuzzy C-means clustering method [21] operates similarly to the K-means clustering method, as both fall under the category of partition methods. Fuzzy C-means aims to minimize the distance between samples (points) within a cluster. Specifically, it strives to minimize intra-cluster distances (distances between points within a cluster) while simultaneously maximizing inter-cluster distances (distances between clusters) [21].

Visual-text modalities. Tian et al. introduced Deep Cross-modal Face Naming (DCFN) [21], leveraging images (faces) and text (names) as dual modalities. DCFN employs Gravity Negative Fuzzy C-Means Clustering for Cross-modal Face-Name correlation learning. Fuzzy C-means reduce inter-modal correlations and enhance intra-modal correlations among name centers, while Negative Fuzzy C-means, aided by gravity, increase the distance between names and incorrect faces and reduce the distance between names and correct faces. Deep visual features are extracted using a CNN, and the Distributed Bag-of-Words model learns caption semantic features.

Visual-numeric modalities. Yu et al. [54] introduces F-HoFCM; an edge-cloud model to cluster multi-modal healthcare data, that mainly consists of numerical records and images. In their high-order multi-modal learning approach, The model leverages DL to extract features for each modality and fuse it using Tucker decomposition to reduce the dimensionality of the cross-modality feature tensor at the edge level. Fuzzy C-means clustering of the representations is conducted in a centralized manner in the cloud. The purpose of leveraging the edge computing architecture is to preserve the privacy of patients in healthcare centers.

6.1.6. Temporal clustering

Numeric-signal modalities. Ramazi et al. [56] classified mobility behavior of the elderly using wearable IoT devices, which continuously collect time series information of their activity patterns. The aim is to identify patients who are at high risk of acute events in long-term care facilities promptly. Static data, representing physical and cognitive condition information of the patients along with the time series signals are integrated into the system to improve prediction accuracy. First, time series representations are created through an AE that uses CNN and Bidirectional Long Short-Term Memory (Bi-LSTM) to learn meaningful representations. The authors in [72] introduced a novel method called deep temporal contrastive clustering (DTCC), which integrates the contrastive learning concept into deep time series clustering research. However, the incorporation of multiple techniques and loss functions may increase the complexity of the model, potentially requiring more computational resources and longer training times.

6.2. Hierarchical clustering

Another clustering technique is hierarchical clustering, which involves grouping unlabeled data in a dendrogram structure. Hierarchical clustering algorithms fall into two types: agglomerative clustering and divisive clustering. In the context of multi-modal clustering using DL, the existing literature predominantly employs agglomerative clustering.

Audio-visual modalities. Harwath et al. [66] introduced a matchmap NN for joint image and audio retrieval, extracting features from image pixels and audio waveforms. The proposed network utilized CNN for tasks such as searching for semantic image/spoken captions, speech-prompted object location, audio-visual clustering, and concept discovery. Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) was employed for clustering, starting with a target of 1000 clusters in the initial stage and reducing to a final target of 135 clusters. Hierarchical agglomerative clustering, known as the "bottom-up" approach, is applied in the Online Video Advertising system, Deep-AD, proposed by Tapu et al. [33]. Deep-AD utilizes CNNs for tasks such as face detection, boundary shot detection, voice activity detection, object detection, background information extraction, person/speaker re-identification, and thumbnail selection. Using agglomerative clustering, the system clusters videos into stories/scenes, considering both audio and video modalities. Scene Splitting involves recognizing indoor/outdoor locations. The authors introduce a thumbnail extraction technique based on visual quality and semantic representativeness for improved multimedia document access.

Visual-text modalities. A hierarchical clustering-based framework called N-Gram Weighted transfer by Clustering over Visual and Textual Information (NGW-CVT) is proposed by Gomez et al. in [65]. This framework utilizes image and text modalities to generate user comments for an image. Visual features are obtained using CNN. NGW-CVT consists of two steps: pin clustering and text transfer. In pin clustering, distinct clusters are formed using hierarchical clustering, creating a fine mapping between images and text. The visual centroids, representing average visual features, are computed along with the average number of words in the comments and an n-gram language model for each cluster. In text transfer, the nearest cluster for a test image is estimated based on Euclidean distance to visual centroids. The distances are used to weigh n-gram frequencies for generating comments.

Hierarchical Agglomerative Clustering is employed by Chaudhary et al. in the multi-modal Hierarchical Clustering for Image (MHCI) framework proposed by [64]. MHCI comprises four steps: pre-processing, leveraging bipartite graph structure, Hierarchical Agglomerative Clustering, and postprocessing. It clusters image and text features (tags, surrounding text, and query text). Visual features, including Scale-Invariant Feature Transform (SIFT), Hue, Saturation, Value, and Local Binary Pattern, are extracted and normalized using CNN, with SIFT features utilizing the k-means clustering method. Evaluation of three datasets showed MHCI's superior performance with CNN features, and MHCI demonstrated the best overall performance compared to other methods.

Guo et al. [59] addressed partial clustering, handling instances with missing modalities. They introduced the TPIT-C model, employing GCN and AEs for feature extraction from image data. Bidirectional mapping and adversarial learning align textual and visual instances in a unified space. K-means and hierarchical clustering were applied, with the latter showing superior performance. TPIT-C serves as a deep feature representation tool for preparing partial data for clustering.

Avellaneda et al. [67] focused on improving anomaly detection in human monitoring using images and deep captions. Their approach involves training a visual-text feature extractor through contrastive learning. The resulting feature vectors are combined and input to the BIRCH clustering algorithm to construct a clustering features tree. Anomaly detection relies on a predefined threshold based on the distance to the closest cluster in the tree.

6.3. Graph-based clustering

6.3.1. Spectral clustering

The spectral clustering algorithm relies on graph Laplacian matrices as a key tool. It operates without making assumptions about the cluster form and can efficiently handle large datasets when the similarity graph is sparse.

Audio-visual modalities. Baraldi et al. [24] presents a framework for the temporal and semantic segmentation of edited videos, focusing on storytelling structures. Their approach involves extracting perceptual features, including visual, temporal, speech quantity, and audio, to identify meaningful video segments. A Pooling Fully CNN is employed for high-level visual feature extraction. Texts are processed to handle duplications and synonyms, followed by clustering based on similarity. Spectral clustering is utilized to group mined terms, and the model is evaluated on BBC Planet Earth and Ally McBeal datasets, showcasing superior performance compared to other frameworks.

Visual-text modalities. Baraldi et al. [68] propose a spectral clustering-based framework for scene detection, integrating visual and text modalities of video shots and transcripts. The method employs a CNN's FC layer to merge both representations, incorporating weighting components to compute similarity scores and learn the final representation. Spectral clustering is applied to cluster adjacent shots based on the similarity matrix. The framework is evaluated on the BBC Planet Earth dataset, assessing scene detection through F-score, while the quality of detected scenes is evaluated using Intersection over Union (IoU).

Visual-numeric-signal modalities. Si et al. [57] employed a multi-modal approach for clustering traditional Chinese medicine patient data, utilizing images, clinical records, and pulse series. They introduced the Media Convergence and Graph Convolution Encoder Clustering (MCGEC), featuring a multi-modal graph convolution encoder to generate shared learning space graph representations. Modal-consensus representations are obtained through fusion in the convergence module. MCGEC incorporates spectral clustering and a self-supervision module, enhancing clustering by providing iterative feedback and optimizing loss functions. Experimental results on real-world clinical data demonstrate the superiority of MCGEC over baseline methods.

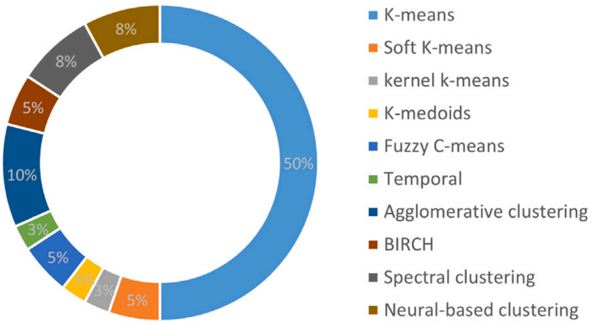


Fig. 6. Distribution of clustering methods in the literature.

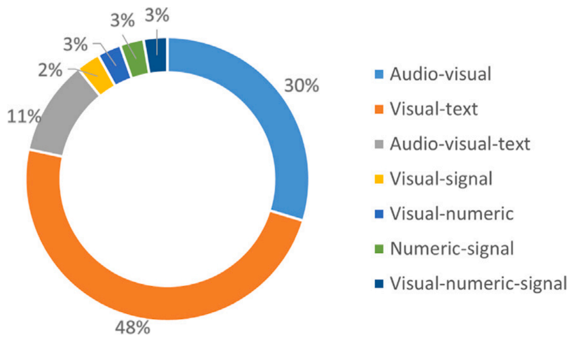


Fig. 7. Distribution of utilized modalities in the literature.

6.4. Neural-based clustering

Recent research has introduced specialized clustering algorithms for MMDC, diverging from traditional approaches. A distinct category, termed neural-based clustering, has emerged. In this category, DL is employed to automatically learn data representations, map similar instances closer in the learned space, and perform clustering simultaneously.

Diallo et al. [73] proposed an algorithm called deep embedding clustering, which is based on contractive autoencoder, to cluster documents. The deep embedding clustering algorithm was extended to provide multi-view clustering by utilizing different deep neural networks [74].

Audio-visual modalities. Asano et al. [69] proposed a multi-modal clustering approach using optimal transport to overcome assumptions of uniformly distributed data in clusters. The Sinkhorn–Knopp Algorithm is employed for solving the clustering problem, demonstrating semantically aligned clusters with human-provided labels. Afouras et al. [70] extended this method by incorporating sound source localization for object detection in videos. The model represents audio and image data, identifies sound sources, and clusters them, training a Faster Recurrent-CNN (R-CNN) for detecting silent objects. Cluster labels are matched with real labels using the Hungarian method [75].

Jia et al. [11] presented a semi-supervised solution for category discovery. Their framework employs contrastive learning, incorporating cross-modal discrimination at the instance and category levels for enriched audio-visual data representation. Modal-specific representations are fused into modal-consensus representations, processed by Multi-Layer Perceptron (MLP) with Winner-Take-All (WTA) to simultaneously transform and cluster data, generating pseudo-labels for unlabeled data. Evaluation of video and image datasets demonstrates superior performance compared to existing methods.

7. Discussions

As discussed in Section 6 and illustrated in Table 1, most frameworks employed partition methods, particularly K-means, for MMDC as can be seen in Fig. 6.

The rest of this section provides our concluding remarks on the utilized modalities, leveraged DL approaches, the involved mechanisms, and delves into the datasets utilized for model evaluations, as outlined in Section 6.

7.1. Utilized modalities

This section organizes literature according to the modalities used, as explained in Section 5.1. Fig. 7 displays the percentages of works using various combinations of modalities.

Audio-visual modalities. Around eleven works incorporated audio-visual modalities. These modalities are often integrated into videos, leading to the majority of video clustering methods relying on audio-visual modalities [13,24,25,33,69]. Some approaches address tasks like video scene segmentation [32], voice localization [9,70], speaker recognition [52], sound textures prediction from images [20], speech-prompted object location detection in images [66], and dynamic task prediction [53].

Visual-text modalities. Visual-text modalities represent one of the most frequently used combinations. Numerous studies proposed generic clustering of visual-text data [23,26–31,62–64,69]. Some works focused on clustering for venue recommendations [60,61], video broadcast segmentation [68], detecting new categories in videos [11], generating comments and descriptions for images [65], assisting search engines [22], face naming [21], HAR and behavior anomalies [67], and partial clustering [59].

Audio-visual-text modalities. Furthermore, some works employ the combination of all early-utilized modalities, namely, audio-visual-text modalities, for video clustering [51,55,58].

Visual-signal. Sensor data has recently become part of the literature, with one work utilizing images and time series information from sensors for human activity recognition (HAR) [12].

Visual-numeric. Recent studies combine numeric data, representing records, with images to support healthcare centers [54].

Numeric-signal. Similarly, recent research combines numeric data from clinical reports with time series signal data collected from sensors to detect behavioral movement events such as falls, urinary tract infections, and delirium [56].

Visual-numeric-signal. In addition to numeric and signal clinical data, images are utilized to cluster patient records [57].

7.2. DL approaches

Various deep learning approaches play a crucial role in clustering frameworks in the literature as explained in Section 5.3.

AE. Most frameworks leverage AEs, a widely adopted tool in MMDC. AEs consist of an encoder and a decoder, with the encoder extracting latent features and the decoder reconstructing the input data. Goals include extracting low-dimensional features for each modality, fusing multi-modality features, and reconstructing input data. Approaches that utilize AEs include [11,12,23,27,27,28,28,29,29,31,51,52,56,59,63,67,69,70].

Dual AEs are used [63], in addition to other AE variations. VAE is a specific type of AE capable of capturing the input distribution of features. VAE incorporates a VIB for feature extraction, distinguishing it from regular AEs [29]. On the other hand, CAE consist of a convolutional encoder and decoder, specializing in compressing data into low-dimensional feature vectors [31].

While VAE combines encoder and decoder functionality, some frameworks opt to use either the encoder or decoder independently. In [62], a multi-modal shared encoder with two fully connected layers aligns latent feature distributions across modalities. In contrast, [58] employed a two-layer encoder and decoder (not a full AE) for reconstructing features and computing the reconstruction loss. Additionally, in [13], an 18-layer ResNet variant and R(2+1)D encoders were used for audio and visual processing.

RNN. LSTMs are used for tasks beyond feature extraction. For instance, LSTM is employed to predict story boundaries in a video sequence [24] and for sentence generation with topic guidance [55]. Additionally, [32] introduced two RNN architectures based on the fusion strategy: LFN and EFN. Both network models incorporate a CNN architecture at their lower layers.

CNN. CNNs play a crucial role in feature extraction for visual modalities [9,13,24,25,32,55,58,60,61,64,65]. An RGB network is used to learn unified representations across audio and visual modalities of videos [53]. Cascaded CNNs are utilized for face detection and alignment [21]. Faster R-CNN is used to detect silent objects in videos [70].

Additionally, CNNs are used to predict statistical audio overviews from video frames [20], create a compact video representation to optimize processing [22], and for object recognition [68]. Moreover, the framework presented in [33] employs deep CNN for a range of tasks, including shot boundary detection, object identification, voice activity detection, speaker recognition, background information extraction, and thumbnail selection.

GCN. A recent approach aims to create modality-consensus representations through graph representations of data initially. GCNs are used to encode different data modalities [26,57]. Moreover, a GCN was combined with bottom-up attention used to extract global features from image data [59].

MLP. MLPs are utilized for various tasks, such as topic prediction [55] and speech clustering [76].

7.3. Mechanisms in deep multi-modal clustering

Various assumptions in the literature have led to distinct mechanisms in approaches for solving deep multi-modal clustering problems, as depicted in Fig. 5 and explained in Section 5.4. The following sections elaborate on these mechanisms.

Data representation. *Early fusion* Some studies adopt early fusion, merging modalities at the input level. A common approach is combining representations of modalities into unified representations before performing clustering [53,57,67]. On the other hand, modal-specific representations can be integrated before simultaneous mapping and clustering [11].

Late fusion Others perform late fusion, extracting features independently for each modality and combining them at a later stage [51,52,70]. Some methods were leveraged with late fusion leverage, such as utilizing pre-trained encoders for different modalities [12,69], and coordinated clustering [56].

Hybrid approaches Hybrid approaches, such as in [26], involve extracting modality-specific encodings before cross-modal fusion for subsequent clustering.

Feature learning. In the realm of feature learning, various strategies are employed to handle multi-modal data.

Shared space learning Some studies adopt shared space learning, aiming to map data from all modalities into a common feature space, promoting modality-invariant representations [26,57,67]. MLPs are utilized to learn a shared feature space for simultaneous mapping and pseudo-label creation [11]. Additionally, partial clustering is performed by aligning different modality instances into a unified space and filling missing modality data [59].

Modality-specific learning Other approaches employ modality-specific learning, training separate models for each modality. The separate features are then combined through learned representations during clustering [52], or they are clustered separately, and the separate clusters are later combined [51,56].

Hybrid approach In a hybrid approach, contrastive learning is leveraged to pair related embeddings of different modalities after modality-specific learning [12].

Architecture. In clustering frameworks, the architectural choices vary significantly, influencing how multi-modal data is processed.

Single NN Some approaches opt for a single NN to handle multi-modal data. Various types of NN architectures are employed, such as GCNs with shared weights across modalities [26], WTA hash-associated MLPs [11], and CNNs [53].

Multiple NNs Conversely, many works leverage multiple NNs, with each dedicated to processing a specific modality [51]. A common approach involves utilizing separate AEs [52] or GCN encoders [57] for different modalities. Some approaches synchronize the last layer across NNs [69].

Loss functions. In clustering frameworks, the choice of loss functions plays a crucial role in guiding the learning process.

Modality-specific losses. Some works employ loss functions tailored to individual modalities to guide the encoder learning [52].

Cross-modal losses. Others utilize cross-modal loss functions to encourage similarity or dissimilarity between instances of different modalities [51,56]. Commonly used loss functions include contrastive loss [26] and entropy loss [11].

Hybrid approach. A hybrid approach integrates both types of loss functions by combining them through a weighted sum [57]. For instance, in [53], multi-task losses are utilized, with each modality having its set of tasks and associated loss functions, and Distillation losses are used to combine the losses from different modalities.

Clustering algorithm. Clustering algorithms play a vital role in organizing data into meaningful groups within clustering frameworks.

Soft clustering. In soft clustering, instances are assigned to multiple clusters with probabilistic memberships, allowing for overlapping clusters. Approaches like Fuzzy C-means [21,54] and soft k-means [26,30] are commonly employed when clusters are considered non-mutually exclusive. In [26], a variant of soft k-means based on the Student t-distribution is introduced.

Hard clustering. On the other hand, literature employing k-means clustering, Kernel k-means [55], and K-medoids [22] assumes hard clustering, where instances are assigned to a single cluster based on the highest probability. Examples include [9,12,20,25,32,51–53]. In [67], a neural-based hard clustering approach is tailored.

Hybrid approach. A hybrid approach combining different clustering types for different modalities is utilized in [56], where soft temporal clustering is employed for time series data, and hard clustering is used for static numeric data.

Evaluation metrics. The evaluation of clustering performance in multi-modal clustering studies relies on various metrics tailored to different aspects of clustering quality.

Integrated metrics. Studies employing integrated metrics evaluate clustering performance by considering all modalities collectively. Common integrated evaluation metrics include Sensitivity, Specificity, AUROC, AUPRC, and SC [56]. Other metrics like mAP are also utilized [70], along with accuracy, RI, and training time considerations [51].

Modality-specific metrics. On the other hand, studies employing modality-specific metrics assess individual modality performance and overall modality clustering quality. Common modality-specific metrics include EER, minDCF, NMI, accuracy, purity, ARI, and F1-score [52,57,69].

Transfer learning. In multi-modal clustering studies, transfer learning techniques are frequently employed to leverage knowledge from pre-trained models or across modalities to enhance performance.

Pre-trained models. Many studies utilize pre-trained models on individual modalities for feature extraction, aiming to benefit from the learned representations [58].

Cross-modal knowledge transfer. Some works transfer knowledge learned from one modality to another during training. For example, in [53], a distillation of losses is employed across audio and visual modalities to transfer knowledge from modality-specific networks to an RGB network, facilitating the creation of unified representations for multiple tasks.

Hybrid approaches. Hybrid approaches, combining multiple strategies, are common in the literature. These approaches leverage both pre-trained models and cross-modal knowledge transfer techniques [12,52]. Co-learning methods are also utilized to transfer knowledge across modalities effectively [51].

Embedding techniques. Metric learning. In the realm of metric learning, several effective methods have gained prominence in the literature. Notable approaches encompass Siamese networks [68], triplet networks [24], contrastive learning [11,12,26,52,67], and margin-based loss functions [25,28,58]. These methodologies typically entail training models to minimize the distance between similar pairs of data points while simultaneously maximizing the distance between dissimilar pairs.

Moreover, CNNs are utilized in metric learning. For instance, in [33], a CNN model is employed to learn a metric, mapping audio-visual features to a space where Euclidean distances reflect the similarity between corresponding instances. Additionally, MLPs are utilized in [25] to identify individuals using visual and audio modalities, where CNNs are used for visual metric learning, and an MLP is used for audio metric learning.

Prototype learning. Several methodologies rely on prototype learning, involving the acquisition of representative prototypes for clusters, serving as pivotal reference points throughout the clustering process. A prominent exemplar of prototype learning in clustering is the K-means algorithm [12,27–29,58,62], alongside its variants such as soft K-means [26,30] and kernel K-means [55]. In these approaches, the prototype is determined as the centroid of the cluster, computed based on the mean of the points assigned to the cluster.

Fuzzy C-means employs a similar prototype learning strategy, but it assigns probabilities to the membership of a data point to several clusters [21,54]. Alternatively, K-medoids clustering [22] selects the medoid of the cluster instead of the centroid, designating an actual data instance as the prototype.

Furthermore, the BIRCH algorithm adopts an implicit prototype learning mechanism in hierarchical clustering by acquiring micro-clusters defined through the mean, radius, and the count of nearby objects [66,67].

Attention mechanisms. Attention mechanisms enable the model to selectively attend to specific features or regions, emphasizing important information during learning. Few approaches utilized attention mechanisms, as follows: In [59], global features are extracted from image data by GCN and bottom-up attention. In [30], an attention layer is utilized to quantify the importance of modalities, aiming to align latent feature distributions across modalities through the use of an adversarial regularizer. In [12], a multi-head self-attention mechanism is used in the time-series signal encoder to process the features from different positions in the signals for HAR.

7.4. Datasets

In this section, Table 2 presents the datasets for multi-modal data used to evaluate the performance of frameworks employing MMDC techniques, as summarized in Section 6 Table 1.

7.5. Research roadmap and gaps

Multimodal clustering has evolved with the prevailing trend involving deep encoding of individual modalities, followed by separate clustering and subsequent combination to derive final labels [12,52]. An emerging shift in this paradigm is towards representing features from all modalities in a shared space, enabling simultaneous clustering [53,57]. While this approach offers modality resilience, its implementation demands a robust feature extraction process.

Learning representations. Contrastive learning has gained prominence due to its inherent flexibility and efficiency in learning representations and capturing relationships between diverse modalities.

A common strategy involves leveraging limited labeled data from specific classes to cluster unlabeled data and discover new classes [11]. Recent trends emphasize the potential of few-shot learning for future exploration [12], mirroring the way humans discover new objects using prior knowledge about other objects. Consequently, research in semi-supervised and few-shot learning is highly encouraged.

Graph representations have recently gained traction for providing a unified space for different modalities [26,57,59]. This involves using GCNs to extract representations, a field that warrants further exploration and comparison with existing methods.

Exploration of modalities. While visual modalities have seen extensive exploration in conjunction with other types, there exists a noticeable gap in combining non-visual modalities, such as audio-text. Recent trends advocate for exploring combinations with numeric and signals modalities [12,54,56,57], especially in critical domains like healthcare services.

Despite the critical role that graph data, originally represented as graphs, play in various fields like network analysis, bioinformatics, and recommendation systems, there exists a research gap in the exploration of MMDC methods that integrate graph data with other modalities.

Clustering approaches. There is considerable room for investigating alternative clustering techniques, such as density-based clustering (e.g., DBSCAN and OPTICS), in MMDC. Additionally, Grid-based clustering techniques (e.g., STING, WaveCluster, and CLIQUE) could be explored in conjunction with DL tools for enhanced multi-modal data clustering.

Adaptability in big data. Partial clustering, addressing scenarios where data is missing in certain modalities, presents an intriguing avenue for exploration, particularly in the context of big data where missing attributes or modalities are prevalent [59]. Existing works predominantly focus on visual-text modalities.

Most clustering methods lack adaptability to the dynamic nature of big data. There is a need for research in adaptive clustering methods, including active learning and reinforcement learning.

This area of research has yet to address real-time multi-modal data clustering from a stream of multi-modal data. Novel representation techniques suitable for efficient clustering using DL tools, along with specialized indexing and storage techniques, should be studied to reduce clustering time for continuous multi-modal data streams.

Future directions. The exploration of soft clustering for anomaly detection stands as an untapped potential, offering more flexible detection for anomalies that do not neatly fit into a single class.

A significant gap exists in the exploration of deep multimodal clustering within federated learning scenarios.

Looking ahead, a possible future direction involves a comprehensive review of mixed-modal clustering, where data originate from different modalities without explicit pairings [45]. This scenario is distinct, as each data point exists exclusively within a single modality and there are no relations to explore between the different modalities.

Table 2
Datasets of multi-modal data.

Dataset	Size	Public	Modality	Ref.
The Big Bang Theory	60K visual, 10K audio	Yes	Audio-Visual	[25]
Sub-SoundNetFlickr	2786 videos		Audio-Visual	[9]
Flicker	261,494		Image-Text	[60,61]
Picasa	398,968		Image-Text	[60,61]
CUB-He[a]	2889	Yes	Image-Text	[23,27,63]
Flowers-He[b]	3235	Yes	Image-Text	[23]
CNN	2107	Yes	Image-Text	[23,28,29]
NPR	603	Request access	Image-Text	[23]
NUS-WIDE [77]	10K	Yes	Image-Text	[30,59]
SentencesNYUv2 (RGB-D)	1,449 images and texts		Image-Text	[30]
NUS	2000	Yes	Image-Text	[62]
APRTC12	7855		Image-Text	[62]
ESP-Game	11,032 images, 5 text/image		Image-Text	[62]
MIRFlickr	12,154		Image-Text	[62]
CCV [78]	9,317 videos	Yes	Image-Text	[30]
Coco_cross [79]	7429 data	Yes	Image-Text	[31]
UCF101 [80]	13K videos		Video-Audio	[13,53]
HMBD [81]	7K videos		Video-Audio	[13,53]
PASCAL VOC[c]	9,963	Yes	Image-Text	[20,30,82]
YouCook2	3.5K videos and texts	Yes	Video-Text	[58]
MSR-VTT	200K videos and texts		Video-Text	[58]
CrossTask [83]	2.7K videos	Yes	Video-Text	[58]
Youtube2Text [84]	1970 videos 40 caption/clip	Yes	Video-Text	[55]
MSR-VTT [85]	10k videos 20 caption/clip	Yes	Video-Text	[55]
TubeKit [86]	1580 videos	Yes	Video-Text	[22]
FAN [87]	194,046 faces244,725 names		Image-Text	[21]
MNDS_I [88],MNDS_II [89]	519,798 pairs,9,362 pairs	Yes	Image-Text	[21]
Video	30 movies (20-30 min)		Video-Audio	[33]
Clickture-1 [90]	497 queries, 136 images	Yes	Image-Text-#click	[64]
Clickture-2 [90]	22,342 queries, 2000 images	Yes	Image-Text-#click	[64]
Clickture-3 [90]	358,700 queries, 14k images	Yes	Image-Text-#click	[64]
ADE20k [91]	22,210 images	Yes	Image-Text	[66]
NUS-WIDE [77]	269,648 images,1000 & 5018 tags	Yes	Image-Text	[26,62,64]
SRC			Image-Text	[64]
Ally McBeal [92]	2660 shots-160 stories	Yes	Audio-Visual	[24]
BBC Planet Earth[d]	4900 shots-670 stories	Yes	Audio-Visual	[24,32,68]
Pinterest	70200 image-comment		Image-Text	[65]
UCI-HAR [93]	19-48 volunteers	Yes	Time series signals	[12]
HHAR [94]	9 volunteers	Yes	Time series signals	[12]
MotionSense [95]	24 volunteers	Yes	Time series signals	[12]
PAMAP2 [96]	9 volunteers	Yes	Time series signals	[12]
Clicical data [56]	23 volunteers	No	Clinical (numeric) and signals (timeseries)	[56]
MIMIC-IV [97]	40k patients	Yes	Electronic health records (numeric)	[56]
Alibaba	1.4 billion videos	No	Audio-visual-text	[51]
IAPR-TC12 [98]	20k images	Yes	Visual	[26]
ESP-Game [99]	20,770 images	Yes	Visual	[26]
MIRFlickr [100]	25k images, 1386 text tags	Yes	Visual-text	[26]
VGG-Sound [101]	183k videos (in 2021)	Yes	Audio-visual	[11,69,70]
AudioSet [102]	2,084,320 videos	Yes	Audio-Visual	[70]
OpenImages [103]	9.2M images	Yes	Visual	[70]
AVE dataset [104]	4k videos	Yes	Audio-visual	[69]
Kinetics-Sound [105]	22k videos		Audio-visual	[69]
Traditional Chinese medicine	1147 patients	No	Visual-numeric-Signal	[57]
Conceptual Captions [106]	3.3M images with captions	Yes	Visual-text	[67]
Human Activities Monitoring [107]	12K images with captions	No	Visual-text	[67]
Behavior Monitoring [108]	6.2K images	No	Visual	[67]
Wikipedia [109]	2866	Yes	Visual-text	[59]

[a] http://www.vision.caltech.edu/visipedia/CUB-200--2011.html.

[b] https://www.robots.ox.ac.uk/∼vgg/data/flowers/102/.

[c] https://deepai.org/dataset/pascal-voc.

[d] https://www.bbc.co.uk/programmes/b006mywy.

8. Conclusions and future directions

This survey systematically explores the landscape of MMDC techniques, providing valuable insights and advancing the field. Motivated by the increasing importance of MMDC in navigating the complexities of multi-modal data, this comprehensive review is driven by the imperative for researchers and practitioners to understand and categorize evolving techniques.

Through a detailed comparison of various MMDC methods, valuable insights into distinct approaches, datasets, assumptions, mechanisms, and limitations are revealed. The introduction of three novel taxonomies tailored for multi-modal data clustering provides a structured framework for classification.

Identification of the research roadmap and critical gaps in MMDC not only summarizes the current state but also inspires new avenues for future research. These include:

- Exploring real-time multi-modal data clustering, emphasizing novel representation techniques and specialized indexing for efficient processing of continuous multi-modal data streams.

- Researching adaptive clustering methods, incorporating active learning and reinforcement learning for enhanced adaptability to dynamic big data scenarios.
- Delving into few-shot learning and semi-supervised learning techniques, adopting a human-like approach for more effective exploration of data clusters.
- Investigating emerging representations, utilizing GCN and graph representations for unified and comprehensive multi-modal representations.
- Exploring less-explored modality combinations, such as audio-text, and investigating novel combinations with numeric and signal modalities.
- Performing MMDC on integrated graph data with other modalities.
- Investigating partial clustering, addressing scenarios with missing data in certain modalities, particularly relevant in big data with missing attributes or modalities.
- Exploring alternative clustering methods, including density-based clustering (e.g., DBSCAN and OPTICS) and Grid-based clustering techniques to broaden the spectrum of available techniques.
- Addressing the significant gap in the exploration of deep multimodal clustering within federated learning scenarios, ensuring a comprehensive understanding and application in distributed environments.
- Exploring soft clustering for anomaly detection to enhance flexibility in identifying anomalies that do not fit neatly into a single class.
- Conducting a comprehensive review of mixed-modal clustering, focusing on scenarios where data originates from different modalities without explicit pairings, considering the unique challenges posed when each data point exists exclusively within a single modality.

In conclusion, this survey paper is a foundational resource, contributing to the understanding and categorization of MMDC techniques amidst increasing multi-modal data complexities.

CRediT authorship contribution statement

Sura Raya: Writing – original draft, Resources, Methodology, Investigation, Formal analysis, Data curation. **Mariam Orabi:** Writing – review & editing, Resources, Methodology, Formal analysis. **Imad Afyouni:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis. **Zaher Al Aghbari:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] P. Michaud, Clustering techniques, Future Gener. Comput. Syst. 13 (2) (1997) 135–147, http://dx.doi.org/10.1016/S0167-739X(97)00017-4, Data Mining.

[2] H. Mittal, A.C. Pandey, M. Saraswat, S. Kumar, R. Pal, G. Modwel, A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets, Multimedia Tools Appl. (2021) 1–26.

[3] Y. Wang, Y. Zhao, T.M. Therneau, E.J. Atkinson, A.P. Tafti, N. Zhang, S. Amin, A.H. Limper, S. Khosla, H. Liu, Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records, J. Biomed. Inform. 102 (2020) 103364, http://dx.doi.org/10.1016/j.jbi.2019.103364.

[4] G. Marti, F. Nielsen, M. Bińkowski, P. Donnat, A review of two decades of correlations, hierarchies, networks and clustering in financial markets, in: Progress in Information Geometry: Theory and Applications, Springer, 2021, pp. 245–274.

[5] D. Jaiswal, V. Kaushal, P.K. Singh, A. Biswas, Green market segmentation and consumer profiling: a cluster approach to an emerging consumer market, Benchmark. Int. J. 28 (3) (2020) 792–812.

[6] Z. jiao Du, H. yang Luo, X. dong Lin, S. min Yu, A trust-similarity analysis-based clustering method for large-scale group decision-making under a social network, Inf. Fusion 63 (2020) 13–29, http://dx.doi.org/10.1016/j.inffus.2020.05.004.

[7] S. Amal, L. Safarnejad, J.A. Omiye, I. Ghanzouri, J.H. Cabot, E.G. Ross, Use of multi-modal data and machine learning to improve cardiovascular disease care, Front. Cardiovasc. Med. 9 (2022) http://dx.doi.org/10.3389/fcvm.2022.840262.

[8] D. Lahat, T. Adali, C. Jutten, Multi-modal data fusion: An overview of methods, challenges, and prospects, Proc. IEEE 103 (9) (2015) 1449–1477, http://dx.doi.org/10.1109/JPROC.2015.2460697.

[9] D. Hu, F. Nie, X. Li, Deep multi-modal clustering for unsupervised audiovisual learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[10] J. Gao, P. Li, Z. Chen, J. Zhang, A survey on deep learning for multi-modal data fusion, Neural Comput. 32 (5) (2020) 829–864, http://dx.doi.org/10.1162/neco_a_01273.

[11] X. Jia, K. Han, Y. Zhu, B. Green, Joint representation learning and novel category discovery on single- and multi-modal data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 610–619.

[12] K. Xia, W. Li, S. Gan, S. Lu, TS2ACT: Few-shot human activity sensing with cross-modal co-learning, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 7 (4) (2024) http://dx.doi.org/10.1145/3631445.

[13] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, D. Tran, Self-supervised learning by cross-modal audio-video clustering, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 9758–9770, URL https://proceedings.neurips.cc/paper/2020/file/6f2268bd1d3d3ebaabb04d6b5d099425-Paper.pdf.

[14] M. Caron, P. Bojanowski, J. Mairal, A. Joulin, Unsupervised pre-training of image features on non-curated data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2019.

[15] X. Ruhang, Efficient clustering for aggregate loads: An unsupervised pretraining based method, Energy 210 (2020) 118617, http://dx.doi.org/10.1016/j.energy.2020.118617.

[16] L. Yu, Y. Su, Y. Liu, X. Zeng, Review of unsupervised pretraining strategies for molecules representation, Brief. Funct. Genom. 20 (5) (2021) 323–332, http://dx.doi.org/10.1093/bfgp/elab036.

[17] D. Erhan, A. Courville, Y. Bengio, P. Vincent, Why does unsupervised pre-training help deep learning? in: Y.W. Teh, M. Titterington (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 9, PMLR, Chia Laguna Resort, Sardinia, Italy, 2010, pp. 201–208.

[18] M.A. Jamal, O. Mohareri, Multi-modal unsupervised pre-training for surgical operating room workflow analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 453–463.

[19] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multi-modal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.

[20] A. Owens, J. Wu, J.H. McDermott, W.T. Freeman, A. Torralba, Ambient sound provides supervision for visual learning, in: European Conference on Computer Vision, Springer, 2016, pp. 801–816.

[21] Y. Tian, L. Zhou, Y. Zhang, T. Zhang, W. Fan, Deep cross-modal face naming for people news retrieval, IEEE Trans. Knowl. Data Eng. 33 (5) (2021) 1891–1905, http://dx.doi.org/10.1109/TKDE.2019.2948875.

[22] P.Q. Nguyen, T. Do, A.-T. Nguyen-Thi, T.D. Ngo, D.-D. Le, T.-A.H. Nguyen, Clustering web video search results with convolutional neural networks, in: 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science, NICS, IEEE, 2016, pp. 135–140.

[23] X. Zhang, J. Mu, L. Zong, X. Yang, End-to-end deep multi-modal clustering, in: 2020 IEEE International Conference on Multimedia and Expo, ICME, 2020, pp. 1–6, http://dx.doi.org/10.1109/ICME46284.2020.9102921.

[24] L. Baraldi, C. Grana, R. Cucchiara, Recognizing and presenting the storytelling video structure with deep multi-modal networks, IEEE Trans. Multimed. 19 (5) (2017) 955–968, http://dx.doi.org/10.1109/TMM.2016.2644872.

[25] C. Miao, J. Feng, Y. Ding, Y. Yang, X. Chen, X. Ji, Unsupervised person clustering in videos with cross-modal communication, in: 2016 Visual Communications and Image Processing, VCIP, 2016, pp. 1–4, http://dx.doi.org/10.1109/VCIP.2016.7805581.

[26] W. Xia, T. Wang, Q. Gao, M. Yang, X. Gao, Graph embedding contrastive multi-modal representation learning for clustering, IEEE Trans. Image Process. 32 (2023) 1170–1183, http://dx.doi.org/10.1109/TIP.2023.3240863.

[27] L. Zong, F. Miao, X. Zhang, B. Xu, Multi-modal clustering via deep commonness and uniqueness mining, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 2357–2360.

[28] X. Zhang, X. Tang, L. Zong, X. Liu, J. Mu, Deep multi-modal clustering with cross reconstruction, Adv. Knowl. Discov. Data Min. 12084 (2020) 305.

[29] Q. Zhao, L. Zong, X. Zhang, Y. Li, X. Tang, A multi-modal clustering framework with cross reconstruction autoencoders, IEEE Access 8 (2020) 218433–218443.

[30] R. Zhou, Y.-D. Shen, End-to-end adversarial-attention network for multi-modal clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14619–14628.

[31] T.D. Do, K. Kim, H. Park, H.-J. Yang, Image and encoded text fusion for deep multi-modal clustering, in: The 9th International Conference on Smart Media and Applications, 2020, pp. 308–312.

[32] T.H. Trojahn, R.M. Kishi, R. Goularte, A new multi-modal deep-learning model to video scene segmentation, in: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, 2018, pp. 205–212.

[33] R. Tapu, B. Mocanu, T. Zaharia, DEEP-AD: A multi-modal temporal video segmentation framework for online video advertising, IEEE Access 8 (2020) 99582–99597.

[34] P. Rai, S. Singh, A survey of clustering techniques, Int. J. Comput. Appl. 7 (12) (2010) 1–5.

[35] J. Swarndeep Saket, D.S. Pandya, An overview of partitioning algorithms in clustering techniques, Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) 5 (6) (2016) 1943–1946.

[36] T. Sajana, C.S. Rani, K. Narayana, A survey on clustering techniques for big data mining, Ind. J. Sci. Technol. 9 (3) (2016) 1–12.

[37] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, J. Long, A survey of clustering with deep learning: From the perspective of network architecture, IEEE Access 6 (2018) 39501–39514, http://dx.doi.org/10.1109/ACCESS.2018.2855437.

[38] Y. Mehmood, N. Aziz, F. Riaz, H. Iqbal, W. Shahzad, PSO-based clustering techniques to solve multi-modal optimization problems: A survey, in: 2018 1st International Conference on Power, Energy and Smart Grid, ICPESG, 2018, pp. 1–6, http://dx.doi.org/10.1109/ICPESG.2018.8417315.

[39] D. Ramachandram, G.W. Taylor, Deep multi-modal learning: A survey on recent advances and trends, IEEE Signal Process. Mag. 34 (6) (2017) 96–108, http://dx.doi.org/10.1109/MSP.2017.2738401.

[40] K. Bayoudh, R. Knani, F. Hamdaoui, A. Mtibaa, A survey on deep multi-modal learning for computer vision: advances, trends, applications, and datasets, Vis. Comput. (2021) 1–32.

[41] W. Chen, W. Wang, L. Liu, M.S. Lew, New ideas and trends in deep multi-modal content understanding: a review, Neurocomputing 426 (2021) 195–215.

[42] S. Keele, et al., Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical report, Ver. 2.3 EBSE Technical Report, EBSE, 2007.

[43] X. Yan, S. Hu, Y. Mao, Y. Ye, H. Yu, Deep multi-view learning methods: A review, Neurocomputing 448 (2021) 106–129, http://dx.doi.org/10.1016/j.neucom.2021.03.090.

[44] X. Li, B. Liu, K. Zhang, H. Chen, W. Cao, W. Liu, D. Tao, Multi-view learning for hyperspectral image classification: An overview, Neurocomputing 500 (2022) 499–517, http://dx.doi.org/10.1016/j.neucom.2022.05.093.

[45] Y. Jiang, Q. Xu, Z. Yang, X. Cao, Q. Huang, DM2c: Deep mixed-modal clustering, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.

[46] Q. Yang, W. Jin, Q. Zhang, Y. Wei, Z. Guo, X. Li, Y. Yang, Q. Luo, H. Tian, T.-L. Ren, Mixed-modality speech recognition and interaction using a wearable artificial throat, Nat. Mach. Intell. 5 (2) (2023) 169–180.

[47] W.-N. Hsu, B. Shi, u-HuBERT: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, Vol. 35, Curran Associates, Inc., 2022, pp. 21157–21170.

[48] J. Kim, L. Billard, Dissimilarity measures and divisive clustering for symbolic multi-modal-valued data, Comput. Statist. Data Anal. 56 (9) (2012) 2795–2808.

[49] D.I. Ignatov, A. Semenov, D. Komissarova, D.V. Gnatyshak, Multi-modal clustering for community detection, in: Formal Concept Analysis of Social Networks, Springer, 2017, pp. 59–96.

[50] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, http://dx.doi.org/10.1038/nature14539.

[51] L. Huang, Y. Liu, X. Zhou, A. You, M. Li, B. Wang, Y. Zhang, P. Pan, X. Yinghui, Once and for all: Self-supervised multi-modal co-training on one-billion videos at alibaba, in: Proceedings of the 29th ACM International Conference on Multimedia, MM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1148–1156, http://dx.doi.org/10.1145/3474085.3481541.

[52] D. Cai, W. Wang, M. Li, Incorporating visual information in audio based self-supervised speaker recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2022) 1422–1435, http://dx.doi.org/10.1109/TASLP.2022.3162078.

[53] A. Piergiovanni, A. Angelova, M.S. Ryoo, Evolving losses for unsupervised video representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.

[54] H. Yu, Q. Zhang, L.T. Yang, An edge-cloud-aided private high-order fuzzy C-means clustering algorithm in smart healthcare, IEEE/ACM Trans. Comput. Biol. Bioinform. (2023) 1–10, http://dx.doi.org/10.1109/TCBB.2022.3233380.

[55] S. Chen, J. Chen, Q. Jin, A. Hauptmann, Video captioning with guidance of multi-modal latent topics, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1838–1846.

[56] R. Ramazi, M.E. Bowen, R. Beheshti, Predicting acute events using the movement patterns of older adults: an unsupervised clustering method, in: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '22, Association for Computing Machinery, New York, NY, USA, 2022, http://dx.doi.org/10.1145/3535508.3545561.

[57] J. Si, Z. Tian, D. Li, L. Zhang, L. Yao, W. Jiang, J. Liu, R. Zhang, X. Zhang, A multi-modal clustering method for traditonal Chinese medicine clinical data via media convergence, CAAI Trans. Intell. Technol. (2023).

[58] B. Chen, A. Rouditchenko, K. Duarte, H. Kuehne, S. Thomas, A. Boggust, R. Panda, B. Kingsbury, R. Feris, D. Harwath, J. Glass, M. Picheny, S.-F. Chang, Multi-modal clustering networks for self-supervised learning from unlabeled videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 8012–8021.

[59] D. Guo, X. Su, Y. Lian, L. Liu, H. Wang, Two-stage partial image-text clustering (TPIT-C), IET Comput. Vis. 16 (8) (2022) 694–708.

[60] J. Zahálka, S. Rudinac, M. Worring, New yorker melange: Interactive brew of personalized venue recommendations, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 205–208.

[61] J. Zahálka, S. Rudinac, M. Worring, Interactive multi-modal learning for venue recommendation, IEEE Trans. Multimed. 17 (12) (2015) 2235–2244, http://dx.doi.org/10.1109/TMM.2015.2480007.

[62] Y. Mao, X. Yan, Q. Guo, Y. Ye, Deep mutual information maximin for cross-modal clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 10, 2021, pp. 8893–8901.

[63] C. Zhang, Y. Liu, H. Fu, Ae2-nets: Autoencoder in autoencoder networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2577–2585.

[64] C. Chaudhary, P. Goyal, S. Tuli, S. Banthia, N. Goyal, Y.-P.P. Chen, A novel multi-modal clustering framework for images with diverse associated text, Multimedia Tools Appl. 78 (13) (2019) 17623–17652.

[65] J.C. Gomez, T. Tommasi, S. Zoghbi, M.F. Moens, What would they say? Predicting user's comments in Pinterest, IEEE Latin Am. Trans. 14 (4) (2016) 2013–2019.

[66] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, J. Glass, Jointly discovering visual objects and spoken words from raw sensory input, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 649–665.

[67] J.A. Avellaneda, T. Matsukawa, E. Suzuki, Cross-modal self-supervised feature extraction for anomaly detection in human monitoring, in: 2023 IEEE 19th International Conference on Automation Science and Engineering, CASE, 2023, pp. 1–8, http://dx.doi.org/10.1109/CASE56687.2023.10260493.

[68] L. Baraldi, C. Grana, R. Cucchiara, A deep siamese network for scene detection in broadcast videos, in: Proceedings of the 23rd ACM International Conference on Multimedia, 2015, pp. 1199–1202.

[69] Y. Asano, M. Patrick, C. Rupprecht, A. Vedaldi, Labelling unlabelled videos from scratch with multi-modal self-supervision, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 4660–4671.

[70] T. Afouras, Y.M. Asano, F. Fagan, A. Vedaldi, F. Metze, Self-supervised object detection from audio-visual correspondence, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 10575–10586.

[71] T.M. Kodinariya, P.R. Makwana, Review on determining number of cluster in K-means clustering, Int. J. 1 (6) (2013) 90–95.

[72] Y. Zhong, D. Huang, C.-D. Wang, Deep temporal contrastive clustering, Neural Process. Lett. 55 (6) (2023) 7869–7885.

[73] B. Diallo, J. Hu, T. Li, G.A. Khan, X. Liang, Y. Zhao, Deep embedding clustering based on contractive autoencoder, Neurocomputing 433 (2021) 96–107.

[74] B. Diallo, J. Hu, T. Li, G.A. Khan, X. Liang, H. Wang, Auto-attention mechanism for multi-view deep embedding clustering, Pattern Recognit. 143 (2023) 109764.

[75] H.W. Kuhn, The hungarian method for the assignment problem, Nav. Res. Logist. Q. 2 (1–2) (1955) 83–97.

[76] D. Hu, F. Nie, X. Li, Deep multimodal clustering for unsupervised audiovisual learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9248–9257.

[77] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: A real-world web image database from national university of Singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09, Association for Computing Machinery, New York, NY, USA, 2009, http://dx.doi.org/10.1145/1646396.1646452.

[78] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, A.C. Loui, Consumer video understanding: A benchmark database and an evaluation of human and machine performance, in: Proceedings of ACM International Conference on Multimedia Retrieval (ICMR), Oral Session, 2011.

[79] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[80] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, 2012.

[81] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: A large video database for human motion recognition, in: 2011 International Conference on Computer Vision, 2011, pp. 2556–2563, http://dx.doi.org/10.1109/ICCV.2011.6126543.

[82] R. He, M. Zhang, L. Wang, Y. Ji, Q. Yin, Cross-modal subspace learning via pairwise constraints, IEEE Trans. Image Process. 24 (12) (2015) 5543–5556.

[83] D. Zhukov, J.-B. Alayrac, R.G. Cinbis, D. Fouhey, I. Laptev, J. Sivic, Cross-task weakly supervised learning from instructional videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3537–3545.

[84] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, K. Saenko, YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 2712–2719, http://dx.doi.org/10.1109/ICCV.2013.337.

[85] J. Xu, T. Mei, T. Yao, Y. Rui, MSR-VTT: A large video description dataset for bridging video and language, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 5288–5296, http://dx.doi.org/10.1109/CVPR.2016.571.

[86] P.Q. Nguyen, A.-T. Nguyen-Thi, T.D. Ngo, T.-A.H. Nguyen, Using textual semantic similarity to improve clustering quality of web video search results, in: 2015 Seventh International Conference on Knowledge and Systems Engineering, KSE, 2015, pp. 156–161, http://dx.doi.org/10.1109/KSE.2015.47.

[87] M. Ozcan, J. Luo, V. Ferrari, B. Caputo, A large-scale database of images and captions for automatic face naming, in: Proceedings of the 22nd British Machine Vision Conference, (CONF) 2011.

[88] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? Metric learning approaches for face identification, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 498–505.

[89] Z. Chen, W. Zhang, B. Deng, H. Xie, X. Gu, Name-face association with web facial image supervision, Multimedia Syst. 25 (1) (2019) 1–20.

[90] X. Hua, L. Yang, M. Ye, K. Wang, Y. Rui, J. Li, Clickture: A large-scale real-world image dataset, in: Mocrosoft Research Technical Report MSR-TR-2013-75, 2013.

[91] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 633–641.

[92] P. Ercolessi, H. Bredin, C. Sénac, P. Joly, Segmenting TV series into scenes using speaker diarization, in: Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2011, Delft-Pays Bas, 2011, pp. 13–15.

[93] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, et al., A public domain dataset for human activity recognition using smartphones, in: Esann, Vol. 3, 2013, p. 3.

[94] A. Stisen, H. Blunck, S. Bhattacharya, T.S. Prentow, M.B. Kjærgaard, A. Dey, T. Sonne, M.M. Jensen, Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition, in: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, 2015, pp. 127–140.

[95] M. Malekzadeh, R.G. Clegg, A. Cavallaro, H. Haddadi, Protecting sensory data against sensitive inferences, in: Proceedings of the 1st Workshop on Privacy By Design in Distributed Systems, 2018, pp. 1–6.

[96] A. Reiss, D. Stricker, Introducing a new benchmarked dataset for activity monitoring, in: 2012 16th International Symposium on Wearable Computers, IEEE, 2012, pp. 108–109.

[97] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, R. Mark, Mimic-iv, 2020, PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/. (Accessed 23 August 2021).

[98] M. Grubinger, P. Clough, H. Müller, T. Deselaers, The iapr tc-12 benchmark: A new evaluation resource for visual information systems, in: International Workshop OntoImage, Vol. 2, 2006.

[99] L. Von Ahn, L. Dabbish, ESP: Labeling images with a computer game., in: AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors, Vol. 2, 2005, p. 1.

[100] M.J. Huiskes, M.S. Lew, The mir flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008, pp. 39–43.

[101] H. Chen, W. Xie, A. Vedaldi, A. Zisserman, Vggsound: A large-scale audio-visual dataset, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2020, pp. 721–725, http://dx.doi.org/10.1109/ICASSP40776.2020.9053174.

[102] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2017, pp. 776–780, http://dx.doi.org/10.1109/ICASSP.2017.7952261.

[103] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al., The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale, Int. J. Comput. Vis. 128 (7) (2020) 1956–1981.

[104] Y. Tian, J. Shi, B. Li, Z. Duan, C. Xu, Audio-visual event localization in unconstrained videos, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018.

[105] R. Arandjelovic, A. Zisserman, Look, listen and learn, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017.

[106] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2556–2565, http://dx.doi.org/10.18653/v1/P18-1238.

[107] M.F. Fadjrimiratno, Y. Hatae, T. Matsukawa, E. Suzuki, Detecting anomalies from human activities by an autonomous mobile robot based on" fast and slow" thinking, in: VISIGRAPP (5: VISAPP), 2021, pp. 943–953.

[108] K. Zhang, M.F. Fadjrimiratno, E. Suzuki, Context-based anomaly detection via spatial attributed graphs in human monitoring, in: Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part I 28, Springer, 2021, pp. 450–463.

[109] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 521–535, http://dx.doi.org/10.1109/TPAMI.2013.142.

Sura Raya received a B.Sc. degree in Mathematics in 2020 and is currently pursuing her M.Sc. in Data Science from the University of Sharjah, Sharjah, UAE. She worked as a researcher in the Big Data Mining research group, and currently she is a teaching assistant with the Department of Computer Science, College of Sciences, University of Sharjah. Her research interests include data mining and geo-spatial data.

Mariam Orabi received a B.Sc. degree in Mathematics, and an M.Sc. in Computer Science from the University of Sharjah, Sharjah, UAE, in 2019 and 2023, respectively. Currently, she is a Researcher in the Big Data Mining research group with the Department of Computer Science, College of Sciences, University of Sharjah. Her research interests include data mining, databases, big data, and social networks.

Imad Afyouni is an Assistant Professor in Computer Science at the University of Sharjah, UAE. He has previously worked as a Researcher and Principal Investigator at the Technology Innovation Center in Makkah, Saudi Arabia. He received a master's degree in computer science from Joseph Fourier University in 2009 (Grenoble, France), and a Ph.D. in Computer Science at the Naval Academy Research Institute in 2013, Brest, France. His research interests converge on data management and mining, exploring the intricate world of spatio-temporal databases, location-based services, and their dynamic applications. Besides, he is actively participating in research fields related to Multimodal GenAI, Machine Learning, and Natural Language Processing, in addition to exploring how video games and mixed reality can help with adaptive learning and healthcare.

Zaher Al Aghbari received a B.Sc. degree from the Florida Institute of Technology, Melbourne, USA, in 1987, and the M.Sc. and Ph.D. degrees in Computer Science from Kyushu University, Fukuoka, Japan, in 1998 and 2001, respectively. He was with the Department of Intelligent Systems, Kyushu University, from 2001 to 2003. Since 2003, he has been with the Department of Computer Science, University of Sharjah, United Arab Emirates. Currently, he is a Professor of Databases and Data Mining, coordinator of the M.Sc. in Data Science program, and the Vice Dean of the College of Computing and Informatics at University of Sharjah. His research interests include data mining, multimedia databases, spatio-temporal databases, big data, social networks, distributed computing, and data streams.