An analytical study of information extraction from unstructured and multidimensional big data

Kiran Adnan and Rehan Akbar*

*Correspondence:
rehan@utar.edu.my
Faculty of Information
and Communication
Technology, Universiti
Tunku Abdul Rahman,
31900 Kampar, Perak,
Malaysia

Abstract

Process of information extraction (IE) is used to extract useful information from unstructured or semi-structured data. Big data arise new challenges for IE techniques with the rapid growth of multifaceted also called as multidimensional unstructured data. Traditional IE systems are inefficient to deal with this huge deluge of unstructured big data. The volume and variety of big data demand to improve the computational capabilities of these IE systems. It is necessary to understand the competency and limitations of the existing IE techniques related to data pre-processing, data extraction and transformation, and representations for huge volumes of multidimensional unstructured data. Numerous studies have been conducted on IE, addressing the challenges and issues for different data types such as text, image, audio and video. Very limited consolidated research work have been conducted to investigate the task-dependent and task-independent limitations of IE covering all data types in a single study. This research work address this limitation and present a systematic literature review of state-of-the-art techniques for a variety of big data, consolidating all data types. Recent challenges of IE are also identified and summarized. Potential solutions are proposed giving future research directions in big data IE. The research is significant in terms of recent trends and challenges related to big data analytics. The outcome of the research and recommendations will help to improve the big data analytics by making it more productive.

Keywords: Big data, Information extraction (IE), Literature review, Learning-based techniques, Multimedia data, Unstructured data

Introduction

Information extraction (IE) process extracts useful structured information from the unstructured data in the form of entities, relations, objects, events and many other types. The extracted information from unstructured data is used to prepare data for analysis. Therefore, the efficient and accurate transformation of unstructured data in the IE process improves the data analysis. Numerous techniques have been introduced for different data types i.e. text, image, audio, and video.

The advancement in technology promoted the rapid growth of data volume in recent years. The volume, variety (structured, unstructured, and semi-structured data) and velocity of big data have also changed the paradigm of computational capabilities of the systems. IBM estimated that more than 2.5 quintillion bytes of data are generated every

day. Among these statistics, it was also predicted that unstructured data from diverse sources will grow up to 90% in few years. IDC estimated that unstructured data will be 95% of the global data in 2020 with estimated 65% annual growth rate [1]. The common characteristics of unstructured data are, (i) it comes in multiple formats [2–5] (text, images, audio, video, blogs, and websites, etc.) (ii) schema-less due to non-standardization [2–4] (iii) it comes from diverse sources (e.g. social media, clouds, sensors, etc.) [2–4, 6].

Due to the huge volume and complexity of unstructured data, it became a tedious task to extract useful information from different types of data. In this regard, systematic literature review have been conducted to identify state-of-the-art challenges. The primary contribution of this work is twofold. First, a systematic review of existing techniques for IE subtasks for each data type i.e. text, image, audio and video. The systematically extracted and synthesized knowledge can be leveraged by the researchers to understand the concept of IE, its subtasks for each data types and state-of-the-art techniques. Second, a taxonomy of IE research is designed to identify and classify the challenges of IE in big data environment. The main categories include task-related challenges and unstructured data-related challenges. Finally, the IE improvement model is designed to overcome the identified limitations of existing IE techniques for multidimensional unstructured big data.

The remaining document is organized as follows: research methodology with all phases and activities is presented in "Research methodology" section. "Information extraction from text" section presents detailed discussion on IE subtasks such as NER, RE, EE, their techniques and comparison of techniques for text data. In "IE from images" section, visual relationship detection, text recognition and face recognition techniques as IE subtask, recent work, and limitations have been described. "Audio IE" section presents the detailed discussion on IE from audio, its subtasks such as AED and ASR with state-of-the-art techniques and challenges. Text recognition and automatic video summarization are elaborated in "Video IE" section. Results and discussion on this systematic literature review are presented in "Results and discussion" section whereas "Conclusion" and "Future work" section present the conclusion and future work, respectively.

Research methodology

Systematic literature review (SLR) is a process to identify, select and critically analyzing the research to answer the identified research questions. Transparency, clarity, integration, focus, equality, accessibility and coverage are key principles of SLR. It is a comprehensive investigation of existing literature on the identified research question. Therefore, SLR has been selected for this review article on IE solutions for unstructured big data and followed the well-formed guidelines [7, 8]. SLR is more suitable for this study because it provides guidelines to conduct review and present findings in more systematic way. Generally, the process of SLR is divided into three main phases named as planning, conduct and reporting the review. These phases and their corresponding activities followed in this review are depicted in Fig. 1.

Planning the review

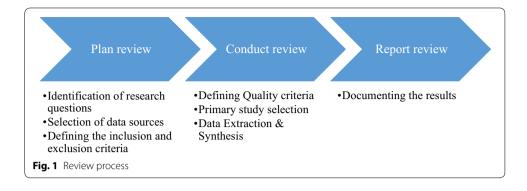The activities performed during the planning phase of the SLR are as follows:

Fig. 1 Review process

Table 1 Research questions and rationale

	Research question	Rationale
RQ1	What are the state-of-the-art approaches for IE from unstructured big data?	To explore the state-of-the-art approaches for IE in big data environment for text, images, audio, and video data
RQ2	What are the issues related to the unstructured big data IE for different types of data?	To investigate the impact of unstructured big data on IE techniques
RQ3	What are the common challenges of IE from a variety of big data?	To identify the common challenges for IE from the variety of unstructured data types i.e. text, images, audio, and video

A. Research questions

The research questions and their rationale have been given in Table 1.

B. Search string and data sources

The following search strings have been used to search the most relevant literature to address the research questions.

TITLE-ABS-KEY (("information extraction" OR "information extraction system" OR "visual relationship" OR "named entity" OR "relation extraction" OR "event extraction" OR "summarization" OR "speech recognition") AND ("big data" OR "large-scale data" OR "large data" OR "volume") AND ("unstructured data" OR "nonstructured data" OR "nonrelational data" OR "free text" OR "image" OR "audio" OR "video")).

ACM, IEEE Xplore, Springer, ScienceDirect, Scopus, and Wiley online library were selected as data sources for this review. The search was conducted in April 2019 using advanced search on the identified data sources. The details of searched and selected articles from each data source are presented in Table 2.
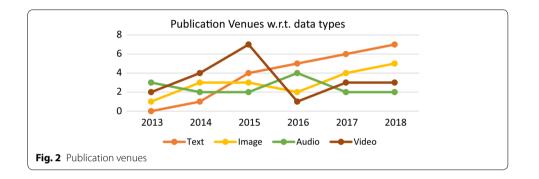
C. Inclusion conditions

The inclusion criteria have been defined to select the most relevant research studies according to the research questions. The inclusion criteria for this study are as follows:

    i.   Research work published between January 2013 and April 2019 inclusively.

   ii.   Studies conducted in the English language.

  iii.   Studies related to IE for text, images, audio and/or video.

  iv.   Research work on unstructured data.

   v.   Research work on data analytics.

  vi.   Research work related to the IE techniques for big data implicitly or explicitly.

Table 2 Data sources and publication for each step of phase 2 of SLR

Data sources	Publication count			
	Searched results	Selected based on title	Selected based on abstract	Selected based on full study + duplicate removal
Wiley Online Library	1012	531	24	3
Scopus	461	146	31	12
Springer	548	204	47	22
IEEE Xplore	203	124	68	36
ACM	633	183	42	10
ScienceDirect	281	122	36	8
Total	3138	1310	248	91

Fig. 2 Publication venues

D. Exclusion conditions

    i.   Studies that used other than the English language.

    ii.   Short papers, presentations, keynotes, and articles.

    iii.   Duplicate or redundant studies.

    iv.   Studies that are not relevant to the research questions.

    v.   Research work older than January 2013.

Conducting the review

After planning the review, studies were refined and selected based on the inclusion and exclusion criteria. The selected studies have been filtered based on the relevance to the study objectives. The selection process started with reading the "title" of the selected studies. Next, studies were filtered on the basis of "abstract" and "keywords" and finally selected on the basis of "full article reading". The publication count of each step to select the most relevant studies for this review is presented in Table 2.

Reporting the review

Figure 2 illustrates the publication venues for each data type from 2013 to 2018, and Fig. 3 illustrates the selected studies distribution over data sources.

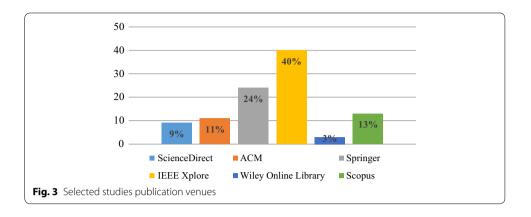Table 3 presents a summary of the categorization of selected studies according to each data type.

Fig. 3 Selected studies publication venues

Table 3 Distribution of selected studies w.r.t study type and data types

Category	Sub category	Selected studies	J	Ch	C	Total
Related to text IE	Named entity recognition	[9–16]	4	3	1	8
	Relation extraction	[17–23]	3	3	1	7
	Entity + relation extraction	[24–29]	4	1	1	6
	Event extraction	[30–35]	3	0	3	6
Total selected studies for text			14	7	6	27
Related to images IE	Visual relationship detection	[36–45]	3	0	7	10
	Text extraction from images	[46–57]	3	1	8	12
	Face recognition	[58–61]	2	2	0	4
Total selected studies for images			8	3	15	26
Related to audio IE	Acoustic event detection	[62–68]	4	0	3	7
	Automatic speech recognition	[69–79]	9	0	2	11
Total selected studies for audio data			13	0	5	18
Related to video IE	General information Extraction from video	[80–82]	0	0	3	3
	Text recognition	[83–92]	4	1	5	10
	Automatic video summarization	[93–99]	1	2	4	7
Total selected studies for video data			5	3	12	20
Total selected studies			40	13	38	91

J journal article, *Ch* chapter; *C* conference

A. Process validation

The key doubts about the SLR process validation depend upon "study selection", "inaccurate data extraction", "inaccurate classification" and "potential author bias". To ensure the process validation for this SLR, two authors were involved in the "selection" and "classification" of each study. Mutual understanding was developed for conflict resolution between authors.

Information extraction from text

The term NLP refers to the methods to interpret the data i.e. spoken or written by humans. In order to process human languages using NLP, several tasks like machine translation, question-answering system, information retrieval, information extraction and natural language understanding are considered high-level tasks. The process of

information extraction (IE) is one of the important tasks in data analysis, KDD and data mining [100] which extracts structured information from the unstructured data. IE is defined as "extract instances of predefined categories from unstructured data, building a structured and unambiguous representation of the entities and the relations between them" [101].

One of the intents of IE is to populate the knowledge bases to organize and access useful information. It takes collection of documents as input and generates different representations of relevant information satisfying different criteria. IE techniques efficiently analyze the text in free form by extracting most valuable and relevant information in a structured format. Hence, the ultimate goal of IE techniques is to identify the salient facts from the text to enrich the databases or knowledge bases. The following subsections discuss the literature selected in SLR process according to the IE subtasks for text data.

Named entity recognition (NER)

Named Entity Recognition is one of the important tasks of IE systems used to extract descriptive entities. It helps to identify the generic or domain-independent entities such as location, persons and organization, and domain-specific entities such as disease, drug, chemical, proteins, etc. In this process, entities are identified and semantically classified into pre-characterized classes [102]. Traditional NER systems were using Rule-Based Methods (RBM), Learning-Based Methods (LBM) or hybrid approaches [103]. IE together with NLP plays a significant role in language modeling and contextual IE using morphological, syntactic, phonetic, and semantic analysis of languages. Rich morphological languages like Russian and English make IE process easier. IE is difficult for morphologically poor languages because these languages need extra effort for morphological rules to extract noun due to non-availability of complete dictionary [104].

Question answering, machine translation, automatic text summarization, text mining, information retrieval, opinion mining and knowledgebase population are major applications of NER [105]. Hence, the higher efficiency and accuracy of these NER systems is very important but big data brings new challenges to these systems i.e. volume, variety and velocity. In this regard, this review investigates these challenges and explores the latest trends. Table 4 presents related work of NER using unstructured big data sets. It summarizes techniques, motivation behind research, domain analysis, dataset used in the research and evaluation of proposed solutions to identify the limitations of traditional techniques, impact of big data on NER systems and latest trends. Evaluation of proposed techniques for IE is performed using precision, recall and F1-score. Precision and recall are the measures for completeness and correctness, respectively. F1-score measures the accuracy of the system and harmonic combination of precision and recall [106, 107].

It has been identified that text ambiguity, lack of resources, complex nested entities, identification of contextual information, noise in the form of homonyms, language variability and missing data are important challenges in entity recognition from unstructured big data [11, 16, 105]. It is also found that the volume of unstructured big data changed the technological paradigm from traditional rule-based or learning-based techniques to

Table 4 Named entity recognition

	Technique	Purpose	Domain	Dataset	Results		
					P%	R%	F%
[9]	Self-training with CNN-LSTM-CRF	To improve the performance and accuracy of NER for large scale unlabeled clinical documents	Medical (clinical text)	19,378 patient data	84.2	85.5	84.4
[10]	CNN + Bi-LSTM and CRF	To improve the performance of NE extraction on OMD large size data and complex data structure without manual rules or features i.e. to deal with volume and variety	Online medical diagnosis text (EMR)	untagged corpus of 320,000 Q&A records of the online Q&A website	Trained with 1/3, 2/3 and all of the total data to compare the experimental performance which results 87.26%, 88.79% and 90.31% F-measure resp.		
[11]	Comparison of BioNLP task with 3 sequence labeling techniques: CRF, MEMM, SVM^{hmm} using one classifier SVM^{multiclass}	To evaluate the performance of ML methods and to identify best features for automatic extraction of habitat entities Features Used: orthographic, mor-phological, syntactic, semantic	Bacterial Biotope entities	BioNLP 2 datasets BB2013, BB2016	CRFs and SVM^{hmm} have comparable perfor-mance, but CRFs achieve higher precision whereas SVM^{hmm} has better recall CRFs and MEMM are shown to be more robust than SVM^{hmm} under poor feature conditions		
[12]	SML based pi-CASTLE: Crowd assisted IE system	To store text annotation in database and addresses the challenges of probabilistic data model, selection of uncertain entities, integration of human entities		For NER: CoNLL 2003 corpus, TwitterNLP dataset with 2400 unstructured tweets	pi-CASTLE achieves an optimal balance between cost, speedand accuracy for IE problems		
[13]	Hybrid method to automatically generate rule	To automatically extract and struc-tured patient related entities from large scale data	Diagnosis extraction	EHR clinical notes of 9.5M patient records	5 use cases applied to prove the modularity, extensibility, scalability, and flexibility		
[14]	Unsupervised ML (clustering)	Examine the impact of volume on three US ML (spectral, agglom-erative, and K-Means clustering)	Facebook posts	314,773 posts by companies and 1,427,178 posts by users for these companies	40.7　Spectral clustering performed better on larger datasets	83.5	56.3
[15]	Grammar rules + MapReduce	To handle large amount of data with parallelization Suitable for incomplete datasets	Free text	3 different text datasets with 1293, 689, 1654 sentences resp.	The results show better recall on 3 text datasets but low precision		

advanced techniques. Variations of deep learning techniques such as CNN are performing better for these NER systems [9, 10].

Relation extraction (RE)

Relation extraction (RE) is a subtask of IE that extracts substantial relationships between entities. Entities and relations are used to correctly annotate the data by analyzing the semantic and contextual properties of data. Supervised approaches use feature-based and kernel-based techniques for RE. DIPRE, Snowball, KnowItAll are some examples of semi-supervised RE [108]. Several supervised, weakly supervised and self-supervised approaches have been introduced to extract one to one and many to many relationships between entities. In the present study, various lexical, semantic, syntactic and morphological features have been extracted and then relationship between entities using learning-based techniques have been identified. Table 5 summarizes the work presented on relation extraction or entities relationship pairs.

Traditional learning-based or rule-based techniques are insufficient to handle the volume and dimensionality of unstructured big data [18]. The supervised LBM needs large annotated corpora and it is very laborious task to annotate large data sets manually. In order to reduce manual annotation effort, weakly supervised methods are more effective [20]. Semantic RE with appropriate features [17, 21] and semantic annotation [17, 20] are two critical challenges of RE.

Table 6 presents the research work related to extracted entities and their relationship from free text corpora. Most of the traditional RE techniques were extracting one to one relationship between entities due to limited text input. In this regard, many-to-many relations have been identified from the large scale datasets that reduce the time as well as increase the performance efficiency. Apache Hadoop provides a platform to adopt parallelization in many to many relation extraction tasks using MapReduce. The system was evaluated with 100 GB free text and many to many relationships were identified [24]. Traditional methods are ineffective to handle data sparsity and scalability [24]. Distant supervised learning, CNN and transfer learning have outperformed the existing traditional methods [23, 25, 26].

Event extraction (EE) and salient facts extraction

An event represents a trigger and arguments. A trigger is a verb or normalized verb that denotes the presence of an event whereas the arguments are usually entities which assign semantic roles to illustrate their influence towards event description [30]. The literature on event extraction and other salient fact extraction has been summarized in Table 7.

The present study identifies several challenges in IE from unstructured big data related to volume, variety and IE techniques. Unstructured big data comes with the heterogeneity of data types, different representations and complex semantic interpretation. These intrinsic problems of unstructured data generate challenges for big data analysis. In order to make unstructured data available in the form ready for analysis, it must be transformed into structured content and prepare for analysis. IE process must be efficient enough to improve the effectiveness of big data analysis. Heterogeneity, dimensionality and diversity of data are important to handle for IE using big data [32, 33]. However, volume of unstructured data is getting double every year [1], it is becoming

Table 5 Relation extraction

	Technique	Purpose	Domain	Dataset	Results		
					P%	R%	F%
[17]	CRF	To generate relationship knowledge base and annotation Lexical, POS and semantic features used	Chinese encyclopedia	52,975 web pages	Model trained for 9 attributes, accuracy of global training is higher than the local whereas recall rate was low		
[18]	Knowledge oriented CNN with clustering using word filters (WordNet)	To overcome the limitations of RBM and LBM and to reduce the dimensionality	Text	3 datasets were used: SemEval-2010 task 8 with 10,717 annotated samples, Causal-TimeBank dataset, Event StoryLine dataset	With max clustering achieved 91.34, 76.21, 81.84% macro averaged F1 on SemEval, Casual-TB, Event-SL resp., whereas, with average clustering, it achieved 91.20, 75.43, 81.96% F1 resp.		
[19]	Pattern-based method to build info network	To extract large-scale treatment drug-disease pairs and inducement drug-disease pairs	Medical literature for drug repurposing	27M abstracts and titles from PubMed	Algorithm has shown high precision but low recall		
[20]	Weakly supervised method without man-made annotation and SVM to train model	To reduce the manual annotation effort and expand the relation types using semantic and syntactic features	News text	Baidu encyclopedia, 50,000 entry pages of 10 GB size	83.61 Results proved that entity ambiguity, and poor universality affect the results	82.63	83.12
[21]	Multi-class SVM and syntactic model development	To detect semantic relation, model architecture with preprocessing phase to build feature vector using lexical, semantic and syntactic features, training phase and RE phase	News Text	ReACE	80.18	70.89	75.25

Table 6 Entity and relation pair extraction

	Technique	Purpose	Domain	Dataset	Results		
					P%	R%	F%
[25]	Transfer learning for domain dependent clustering	To adapt the world knowledge to the domain-dependent tasks by using semantic parsing and semantic filtering	News text	20 Newsgroups, RCV1	Case studies conducted to prove that conceptualization based semantic filter can produce more accurate indirect supervision		
[26]	Distant supervised learning (deep learning)	To overcome the limitations of text mining methods such as clustering or rule-based etc. in keyword and information extraction with technology dependency graph	Scientific literature	473,935 articles, labeled 38 relation instances from 20 articles and expanded to 573 instances by bootstrapping	Case study: Technology driven graph to analyze the technology architecture of DSSC		
[22]	MapReduce + semantic methods (attribute based, isA based and class based) + logistic regression	To overcome the long tail challenge using Sparse IE approach / To deal with scalability and effectiveness	Web pages	1.68 B web pages	Many entity pair identified and classified as good and bad pairs. Results of each entity pair with good and bad recall, precision abd F-measure is presented		
[27]	Supervised Kernel methods	To extract morpho-syntactic information from mined text / To deal with challenges of data prioritization and curation	Biomedical	EU-ADR	Proposed method using morpho syntactic and dependency information outperform to identify entity relationship		
[28]	Use of declarative rules in contextual exploration	Automatic detection and extraction of meaning from unstructured web using RDF WordNet, DBpedia, etc. / To bypass the limitation of lack of annotated data semantically and automatically usable using LOD	Free text	Large text corpuses provided by the labex OBVIL and the BNF (National Library of France)	EC3 software is implemented and shown considerable contribution in detecting real meaning of text		
[29]	CRF and dictionary for NER, word clustering through Unsupervised training	Chemistry aware NLP pipeline with tokenization, POS tagging, NRE and phrase parsing / To populate chemical databases with minimal time, effort and expense	Scientific documents	50 open access chemistry articles	89.1	86.6	87.8
[24]	Hadoop (MapReduce)	To identify many to many relationships with less training data	Free text	100 GB-sized corpus, baike.baidu.com: big encyclopedia having 700M entries	Proposed Snowball++ achieved higher positive pairs as compared to snowball and PROSPERA		
[23]	CNN (weakly supervised)	To obtain high-precision data and automatically generate annotated training sample set	Medical	Experiment selected seven medical sites, generate a total of 20,000 labeled samples at last and five categories of directional relations	91.87	91.58	89.08

Table 7 Event extraction and salient fact extraction

Task	Approach	Dataset	Results	Remarks
Term context understanding to deal with homonyms [31]	Semi-automated approach be combining automated content analysis and ANN	26,259 research articles from Web of science	Proposed solution evaluated with different sparsity parameters. Results showed different effects of different modeling terms on error rate	The proposed solution outperformed with manual classification in some instances that could not automatically be classified. Hence, improvement is required to automate the sifting process of homonyms context identification
IE from heterogeneous unstructured big data [32]	Unsupervised deep learning (multiple Kernel)	13 different datasets from UCI Machine Learning Repository	Performance of the proposed system was better in speed from other competitors and same in accuracy	Accuracy of heterogeneous data can be improved with unsupervised learning but advancement in approach is required to handle the dynamicity of such data
Deep semantic IE for big data mining from geoscience data [33]	convolutional neural networks (CNN) for classification and TF-IDF for word statistics	Multivariate and heterogeneous data of 16,098 PDN, 130 LAN's	classification accuracy of 99.9% and 99.8% at the sentence and paragraph levels, respectively	Insufficient comprehensiveness, poor correlation and inconsistent formats are problems of heterogeneous data
Open domain event extraction [34]	Schema discovery based on probabilistic generative models i.e. LinkLDA	Set of events generated and extracted from Twitter	The difference between proposed and related work is, it can handle complex queries and structured data browsing	The sparsity of unstructured big data can decrease the performance and scalability of solution. So, these are important factors to investigate the effectiveness of approach
Biomedical Event extraction [35]	Syntactic and semantic features to identify event trigger + Phrase Structure Tree	BioNLP-ST 2013	The solution was evaluated and shown 52.23% precision, 26.38% recall, and 35.06% F1-score	The proposed approach uses ML features that inherits the limitations of the ML feature based techniques

more critical to extract semantic information from such a huge deluge of unstructured data. Nevertheless, big data bring some challenges also for learning-based approaches which are dimensionality of data, scalability, distributed computing, adaptability and usability [109–111]. In this regard, advancements in learning-based approaches are trying their best to handle the complexity of big data.

State-of-the-art IE techniques

Two major categories of IE techniques are rule-based methods (RBM) and learning-based methods (LBM). It is difficult to identify which method is more popular and effective in IE. In this regard, two studies [112, 113] have shown totally different analysis. First, according to a systematic literature review on the popularity comparison of these two methods, it was concluded that more than 60% of the studies included in the review used pure rule-based IE systems. Whereas it was considered that rule-based IE techniques are obsoleted in academic research domain [112]. Another comparison has demonstrated totally different results by examining 177 research papers of four specific conferences on NLP. Among these 177 research papers, only 6 papers relied on pure rule-based IE approach [113]. It was also observed that the IE Systems by large vendors i.e. IBM, SAP and Microsoft are purely rule-based [113]. This review identifies that LBMs are more popular in academic research domain as compared to RBM but the importance of RBM could not be neglected. However, the debate on the comparison of these two approaches is subjective to various factors such as the cost, benefits and task specifications. Table 8 presents a comparison of these two approaches in general.

The comparative analysis explores different pros and cons of both approaches but the selection of approach for any task is highly dependent on the user needs and task at hand because IE is community-based process [100]. In general, learning-based approaches are divided into supervised, semi-supervised and unsupervised techniques. These techniques also have limitations to handle large scale big datasets and complexity of huge volume of unstructured data. Supervised techniques require manually labeled training data which is one of the major drawbacks of these techniques. Large scale labeled corpus construction is laborious and time consuming task [9]. These techniques are effective for domain-specific IE where specific information is required to be extracted. The efficiency

Table 8 Rule-based vs learning-based techniques

Rule-based approaches	Learning-based approaches
Interpretable and suitable for rapid development and domain transfer [114]	The performance of machine learning approaches is better in terms of precision and recall but appropriate feature selection is important [115]
Humans and machines can contribute to the same model. So it is easy to incorporate domain knowledge [114] Heavily rely on domain thesauri [11]	Generating training data is time consuming in learning-based approaches whereas rule-based approaches require pre-defined vocabularies [116]
Although rule-based systems require domain knowledge and are time consuming, results proved that these are more reliable and useful for automated processing [117]	No experts are required and system can be developed quickly with relatively low cost [118]
Declarative [119]	Adaptable [119]
Requires tiresome manual work [118]	Less manual effort [118]
Highly transparent and expressive	Higher portability than rule-based [9]

of these techniques also depends on the selected features like morphological, syntactic, semantic and lexical features. Whereas, unsupervised IE techniques do not need labeled data. These techniques extract entity mentions from the text, clusters the similar entities and identify relations [120]. In this case, intensive data preprocessing will be required for big data because unstructured big data sets have missing values, noise and other errors [16] that produce uninformative as well as incoherent extractions. Semi-supervised techniques use both labeled and unlabeled corpus with small degree of supervision [121]. For large scale data, distant supervised learning [26], deep learning (CNN, RNN, DNN) [9, 10, 18, 23, 31–33], transfer learning [25] techniques are more suitable for IE from free-text data.

Deep learning approaches show better results for large datasets despite its own limitations and challenges. It has the ability to generalize the learning and also has a unique characteristic to utilize unlabeled data during training. Deep learning has the ability to learn different features as it has multiple hidden layers. These techniques are more suitable for pattern recognition [122]. Unsupervised learning (deep) have large model capacity/complexity, high learning speed [32]. Feature learning-based systems are computationally expensive for large scale data [123]. For the selection of appropriate technique for large scale datasets, computational cost, scalability and accuracy are the key factors [124]. More advanced algorithms and techniques are required to achieve higher accuracy and efficiency [125]. Over-fitting can be resolved with self-training [18] and to overcome the limitation of large annotated dataset availability, reinforcement learning or distant supervision can be used because these techniques use small labeled dataset [26], [126]. Timeliness of distribution of data [126], balance of informativeness, representativeness, and diversity [127], data modeling performance for heterogeneous, dimensional, sparse and imbalance data [16] and structuring the unstructured data [10] are open challenges for IE using unstructured big data sets.

Unstructured big data barriers for IE

With huge volume and complexity of unstructured big data, natural language free text data implies various issues for the users to extract the most relevant and required information. Noisy and low-quality data is one of the major challenges in IE from big data [16, 31, 128, 129]. It causes difficulties in identifying semantic relatedness among entities and terms [130], improving the effectiveness and performance of IE systems [128], extracting contextually relevant information [31], data modeling [16] and structuring the data [10].

IE from text is also facing natural language barrier. Data diversity [124], ambiguities in text, nested entities [105], heterogeneity [131], automatic format identification [13], sparsity, dimensionality [16], homonym identification and removal [31] are some important challenges to IE from unstructured free text. The exponential growth of unstructured big data is making IE task more arduous. However, MapReduce has the capability to deal with large scale datasets by distributing the data into different clusters that increases the time efficiency [15, 22, 24]. Hence, the volume can be effectively handled using Apache Hadoop, whereas, the issues related to the variety of data needs to be focused. Unstructured big data is adding more challenges to IE from natural language text. Hence, advanced and adaptive preprocessing techniques are required to improve

the quality and usability of unstructured big data. After preprocessing the data, IE techniques i.e. RBM or LBM will be able to produce more effective and efficient results.

IE from images

The IE from images is a field with great opportunities and challenges such as extracting linguistic descriptions, semantic, visual and tag features, context understanding and face recognition. Content and context level IE from different types of images could improve image analytics, mining and processing. Following sections review the IE from images w.r.t. different subtasks.

Visual relationship detection

Visual relationship detection extracts interaction information of objects in images. These semantic representations of the relationship of objects are presented in the form of triples (Subject, Predicate, Object). The semantic triples extraction from images would benefit various real-world application such as content-based information retrieval [132], visual question answering [133], sentence to image retrieval [134] and fine-grained recognition [135]. Object classification and detection and context or interaction recognition are main tasks of visual relationship detection in image understanding.

In object detection and classification, objects are recognized based on appearance and its class labels have clear association. CNN based solutions in object classification are outperforming such as VGG [36] and ResNet [37].Whereas, Faster R-CNN and R-CNN achieved great success in deep learning [38, 39]. Unlike object detection, visual relationship detection extracts the interaction of objects. For example, "horse eating grass" and "person eating bread" are two visually dissimilar sentences but both are sharing the same interaction type "eating". Thus, subject, object and interaction are important in relationship detection as well as context of the interaction. The model of interaction and its context are treated as a single class where images are classified according to the interaction classes [40]. Single class modeling has poor generalization and scalability as it requires training images for each combination of interaction and context. Language priors [41] or structural language [38] are used to overcome the limitations of single class modeling. Intraclass variance, long-tail distribution, class overlapping are three major challenges of visual relationship detection [41]. Long-tail distribution challenges were addressed by introducing spatial vector for imbalance distribution of triples [41]. Long-tail distribution problem causes difficulties in collecting enough training images for all relationships. In this regard, incorporating linguistic knowledge to DNN can regularize the performance [42]. Several modified state of the art deep learning based techniques to extract context and interaction detection have been discussed Table 9.

It can be concluded that deep learning techniques are outperforming in IE from large scale unstructured images. CNN, RCNN and reinforcement learning achieved better recalls. Also, it has been observed that Faster-RCNN and R-CNN have achieved remarkable achievement in object detection [38, 39]. Whereas language prior and language structures are also improving performance of relationship detection [38]. CNN based VRD techniques extract features from subject and object union box before classification. The training samples contain various same predicate categories which can be used in different context with different entities. CNN based models have the limitation to learn

Table 9 Visual relationship detection

	Purpose	Technique	Dataset	Results	Limitations/benefits
[43]	To map the images with associated scene triples (subject, predicate, object)	Conditional multiway model with implicitly learned latent representation. Both semantic tensor model and object detection used RCNN model	Stanford visual relationship dataset	Model achieved better performance to predict unobserved triples	Results are comparable to Bayesian fusion model. Proposed approach has used implicit learned prior for semantic triples whereas BFM needs explicit
[38]	To improve global context cues extraction	Variation structured Reinforcement Learning (VRL): Directed semantic graph using language prior + variation structured traversal to construct action set + make sequential predictions using deep RL	VRD dataset with 5000 images and visual genome dataset with 87,398 images	The proposed approach outperforms baseline methods for attribute recall @100 and @50 i.e. 26.43 24.87 resp. other results also show pretty notable results compared to other methods	Although, the proposed approach outperformed but on comparison to VRL performance, the results are almost same as VRL with LSTM. But it was claimed that VRL with LSTM takes more training time
[39]	To identify unseen context interaction relationship	Context aware interaction classification i.e. Faster-RCNN + AP + C + CAT	VRD dataset and visual phrase dataset	The proposed approach has performed better than baseline methods using spatial and appearance features	Spatial feature representation produced better results than appearance based representation. Adding language prior to proposed approach does not bring benefit
[44]	1. To infuse semantic information and improve predicate detection 2. NMS was used to reduce redundancy and boost detection speed	To include spatial, classification and appearance information, feature extraction used + bidirectional RNN + paired non-maximum suppression (NMS)	Visual genome having 108,077 images, VRD having 5000 images	Results are compared with other existing methods for predicate, phrase and relationship detection for recall @ 50 and 100. Proposed solution gave better results for both datasets	Superfluous regions are filtered using NMS improves the performance
[41]	To overcome long tail distribution challenge To handle widely spread and imbalanced distribution of triples	Visual module using VGG16, Language module using softmax, the contribution was spatial vector using normalized relative location of object and intersection over union	VRD and VG dataset	The results showed that proposed vector improved the performance of 2% and 4% on Recall@50 as compared to other. The proposed solution capable to detect unseen visual relationship	The research only addressed the long tail distribution challenges

common features in same predicate category [45]. So, intraclass variance is a challenge for CNN based VRD. In order to overcome the limitation of CNN models in VRD, visual appearance gap between same predicates and visual relationship should be reduced. For this, Context and visual appearance features can be used to overcome the identified limitation [44, 45]. Further, modified deep learning techniques are required to overcome the challenges of visual relationship detection for large scale unstructured data. To the best of our knowledge, the impact of volume, variety and velocity of big data is not addressed well in visual relationship detection techniques.

Text recognition

A vast array of information can be extracted from the text content in images. Text within images and videos describes more about the useful information about the visual content and also improves the efficiency of keyword-based searching, indexing, information retrieval and automatic image captioning. Text information extraction (TIE) systems detect, localize and recognize the text in visual data like images and videos. The visual content can be categorized into perceptual content and semantic content. Perceptual content includes color features, shape, texture features, temporal attributes, whereas semantic content deals with the identification and recognition of objects, entities and events [136]. TIE systems follow detection, localization, tracking, extraction or enhancement and recognition phases in terms of detecting and identifying text in the visual data. Each subtask of TIE systems has different techniques, challenges and limitations. In TIE systems, text detection and localization tasks are used to identify different features such as color-based, edge based, texture-based and text-specific features [136, 137]. All these subtasks are important to extract useful information from visual data but only recognition task is more relevant to the identification of objects, entities and characters. Text recognition is a process to identify the character-forming meaningful words. So, recent literature have been discussed, in this section, to identify the potential challenges of text recognition task from images in information extraction.

Text recognition task is tightly coupled with the OCR (Optical Character Recognition) approach to recognize characters from images or scanned documents. Character recognition from the Tamil text in ancient documents and palm manuscripts to extract useful information from document images using OCR involved a segmentation technique which included different stages: image preprocessing, feature extraction, character recognition and digital text conversion. According to the experimental results, the accuracy of conversion for the Brahmi was 91.57% and 89.75% for the Vattezhuthu [46]. Whereas, character recognition using neural networks from handwritten text has shown different results. The Radial Basis Function (RBF) with one input layer and one output layer has been used to train RBF network. As compared to back propagation neural network, gradient feature extraction resulted in less accuracy with RBF using directional group values [47]. OCR systems perform better for scanned documents but different variation in images have shown inappropriate results [137]. The underlying reasons could be the geometric variation, complex background, variation of text layout and font, uneven illumination, multilingual content, low resolution and low quality [138].

Extracting text from the visual data, semantic features use learning-based approaches such as supervised and unsupervised. Supervised learning methods are used to learn

structure or concepts from the features such as Support Vector Machine (SVM) and Bayesian classifier. These classifiers are trained to learn the structure and are tested on the unlabeled regions. In this regard, distorted character recognition using Exempler SVM beat the existing state of the art by over 10% for English and 24% for Kannada on the benchmarked dataset Chars74k and ICDAR [48]. Similarly, CRF classifier was used in a framework to recognize characters with scores, spatial constraint and linguistic knowledge that performed 79.3% on ICDAR2003 and 82.79% on ICDAR2011 accuracy [49]. Another system, Stroklete, was designed to detect and recognize characters from the images using histogram features i.e. bag-of-Strokletes to learn the structure of the letters and train the system with Random Forest classifier. The system was trained and tested on English letters and Arabic numbers. It had shown 80% and 75% supporting results on ICDAR2003 and SVT respectively [50]. However, robustness to distortion and generality to variant language are challenging for these systems. To explore the advancement in TIE techniques, Table 10 summarizes the literature on the state of the art TIE techniques for high dimensional or large scale datasets.

Unlike traditional OCR techniques, CNN, RNN and LSTM are achieving high performance in text recognition in images. Deep learning techniques are showing prevalent results to date. CNN as feature extractor to detect, slice and recognize pipeline [57] and as encoder in attention mechanism outperformed others [56]. Although, these techniques are showing promising results, but diversity in data sources makes the system complex [55]. The effectiveness of these techniques for complex, diverse, high dimensional and heterogeneous datasets must be investigated. The huge volume of unstructured data is creating noisy and low-quality images such that multilingual text in images should be addressed to improve the IE from images [58, 59]. CNN based OCR have also shown pretty good results but the performance of technique on unstructured big datasets is still to be investigated. The attention mechanism is a new approach in text recognition [54, 56]. Initially, the results are satisfactory but there is a huge room for improvement in terms of unstructured and multidimensional big data. It is predicted that OCR with attention mechanism will be the emerging phenomenon in near future for text recognition [54]. In this regard, robust and adaptive techniques are required for unstructured big datasets for semantic understanding of text in images.

Face recognition

The task to recognize similar faces is a computational challenge. It is evident that humans have very strong face recognition abilities and these abilities are superior to known faces but ability to recognize the unfamiliar faces are error-prone [139]. This distinction of face recognition in human lead towards the finding that face recognition depends on different set of facial features for familiar and unfamiliar faces. These features are categorized into internal and external features respectively [140]. In this regard, [60] examined the role of high-PS and low-PS features in face recognition of familiar and unfamiliar faces and role of these critical features for DNN based face recognition. The review concluded that high-PS features are critical for human face recognition and are also used in DNN based trained on unconstrained faces.

Table 10 Text recognition from images

	Purpose	Technique	Dataset	Results	Limitations/benefits
[51]	To possess high learning capacity. To handle high dimensional data	CNN based OCR	Scanned Sanskrit document images (11,230)	Proposed approach outperform than existing. Accuracy was 93.32%	Training time as 1 h with GPU
[52]	To automatic recognition of hand-written text from images	CNN based OCR	MNIST	98.11% accuracy rate	DL should apply on large datasets
[53]	To compare the results of proposed DBN and CNN ECR	Unsupervised feature learning with DBN	HACDB dataset containing 6600 images	Experiments shown 3.64% and 14.71% for DBN and CNN resp.	DBN with unsupervised feature learning outperform CNN for high dimensional data
[54]	To develop end to end mechanism for Scene TR	FANet using resnet as encoder and seq2seq attention mechanism as decoder	5000 authentic seal dataset, 3660 real time train ticket dataset	Although, proposed approach could not achieve outperforming results but angular and horizontal TR was improved	Full attention mechanism was proposed to replace detect, slice, and recognize process with end to end recognition. Ineffective for long text recognition
[55]	To recognize text from handwritten and printed text images	TMIXT: tessetact for machine printed text recognition and LSTM for handwritten text recognition	IAM handwriting database	Achieved 80% average transcription accuracy	Heavy preprocessing is required for combined text recognition with proposed solution
[56]	To recognize text using attention mechanism	CAN (Convolutional Attention Network), 2D CNN as encoder and one dimensional CNN decoder	Street View text SVT, IIIT5K, and ICDAR 03, ICDAR 13 dataset	The proposed model performed better than others on SVT and ICDAR 03 datasets	Improvement in proposed method is required for promising results
[57]	Semantic based text recognition to extract useful information from images	CNN and bidirectional LSTM where convolutional part uses VGG and recurrent part uses bidirectional LSTM	Interior Design Dataset with 7708 images	Achieved 90% accuracy in word recognition	Generality improved but the text recognition from protest images is relatively an easy task. Evaluation of the system with complex and diverse datasets should be promising

In the domain of computer vision, face recognition is a holistic method that analyzes the face images. Various techniques have been proposed for face recognition for different datasets but these traditional techniques are inadequate to deal with large scale datasets efficiently. A comparative analysis shows that these traditional techniques have limitations to handle low-quality large scale image datasets whereas deep learning methods are producing better results for these datasets but with optimal architecture and hyper-parameters [58]. The face recognition in low quality i.e. blur and low-resolution images degrades its performance. Sparse representation and deep learning methods combined with handcrafted features outperformed in case of low-resolution images [59]. Face recognition techniques should be able to recognize faces with different face expressions and poses in different lighting conditions [58]. Various deep learning based solutions are proposed to address the limitations of traditional techniques. Deep CNN face recognition technique without extensive feature engineering reduces the effort of most appropriate feature selection. Deep CNN face recognition technique was evaluated on UJ face database of 50 images and the results have shown validation accuracy of 22% goes to 80% after 10 epochs and 100% after 80 iterations [52]. Certain limitations were also associated with the solution such as overfitting and very small dataset. To reduce overfitting, application of early stopping method will require extra effort. VGG-face architecture and modified VGG-face architecture with 5 convolutional layers, 3 pooling layers, 3 fully connected layers and softmax layer was evaluated using five different image datasets, i.e. ORL face database with 400 images, yale face database with 165 images, extended yale-B cropped face database with 2470 images, faces 94, Feret with 11,338 images and CVL face db. For all datasets, the proposed approach performed better as compared to traditional methods [58]. Although, the proposed technique outperformed five different datasets but the datasets were not complex and large-scale datasets. Deep learning based face recognition techniques such as deep convolutional network or VGG-face and lightened CNN have capability to handle huge amount of wild datasets [61].

Deep learning based face representations are more robust to handle misaligned images [61]. Deep CNN can perform better to recognize objects from partially observed data but image enhancement is important in deep CNN before the convolutional operation for low quality images [58]. Although deep learning techniques have capabilities to improve the performance of face recognition, certain challenges have also been associated with deep learning techniques that should be considered beforehand. Quality of images, missing data in images, noise should be handled because these factors degrade the performance of the deep learning based face recognition techniques [58, 59]. Face recognition with different face expressions, illuminations, using accessories causes partial occlusion [61]. This partial occlusion detection requires new optimal deep learning architecture and hyper-parameters to overcome these challenges. However, the selection of appropriate technique highly depends on the data size and quality. Further, more robust and optimal solutions are required for large scale datasets with high accuracy and low latency.

Audio IE

Companies like call centers and music files are the major sources which generate a huge volume of audio data. Different type of information can be extracted from this data to help predictive and descriptive analytics. The subtasks of IE from audio data are classified as acoustic event detection and automatic speech recognition.

Acoustic event detection

Sound event extraction or acoustic event extraction is an emerging field which aims to process the continuous acoustic signals, convert them into the symbolic description. The applications of automatic sound event detection are multimedia indexing and retrieval [141], pattern recognition [62], surveillance [142] and other monitoring applications. This symbolic representation of sound events is used in automatic tagging and segmentation [143]. These auditory sounds come from diverse sources and contain overlapping events and background noise [63, 64]. Moreover, parametric accuracy of training model on limited training data is also difficult to achieve [62].

As presented in Table 11, data scarcity and overfitting are common limitations of AED solutions. In this regard, modified data augmentation achieved better results due to modification in frequency characteristics with particular frequency band [65]. Context recognition is one of the solutions to overcome the overlapping issue and improve the accuracy of AED but identifying the specific context sound event is one of the critical challenges for AED. Adding language or knowledge prior can help to extract context sound events [64]. In recent work on AED, deep neural networks are outperforming traditional techniques. The capability to jointly learn feature representation is one of the major advantages of DNN. Whereas, supported by large amount of training data, DNN is well progressing in the field of computer vision. But non-availability of large scale datasets publically reduces the progress in this research area [64]. Creating large scale annotated data can be a time-consuming process. Therefore, weakly supervised or self-supervised data for training can perform better. In this context, CNN based weakly supervised technique was compared to the technique trained with fully-supervised data. On evaluating both techniques on UrbanSound and UrbanSound8k datasets, weakly supervised performed better for arbitrary duration without human labor for segmentation and cleaning [67]. On the implementation side of AED techniques on large scale, high computational power, efficient parallelism and support for training large models are important factors to consider [68]. The research on automatic AED is hindering by the complexity of overlapping sound events. Improved accuracy to handle overlapping sound events, efficient solutions to achieve labeled datasets, improved processing time with parallelism for large scale data are important dimensions for the development of optimal solutions for AED with unstructured big data.

Automatic speech recognition (ASR)

Automatic speech recognition (ASR) is a task to recognize and convert speech into any other medium such as text, that's why it is also known as speech to text (STT). Voice dialing, call routing, voice command and control, computer-aided language learning, spoken search and robotics are major applications of ASR [144]. In the process of speech

Table 11 Acoustic event detection

	Purpose	Technique	Dataset	Results	Limitations/benefits
[62]	Modeling data with exemplars To explicitly model the background events	Exemplar-based method with NMF	Office Live recordings from 1 to 3 min and office synthetic with bg noise	With time wrapping, Fscore improved from 50.2 to 65.2% in office live dataset whereas in office synthetic dataset, results were not promising	Proposed solution suffers from data scarcity, and overfitting
[65]	To overcome the overfitting limitation To improve the performance for large scale input	CNN to train AED end to end + data augmentation method to prevent overfitting	Acoustic event classification database	Achieved 16% improvement as compared to Bag of Audio Words (BoAW) and classical CNN	Results presented with and without data augmentation proved that augmentation improves the performance
[63]	To explore the impact of feature extraction in AED To explore the effectiveness of deep learning approaches	Multiple single resolution recognizers + selection of optimal set of events + merging or removing repeated labels	CHIL2007	CNN performed better with combination scheme of multi-resolution approach	DNN has the ability to model high dimensional data
[64]	To improve the detection accuracy by extracting context information	Context recognition phase: UBM to capture unknown events and sound event detection stage	Audio database consisting 103 recordings of 1133 min duration	Knowledge of context as context dependent event prior can be used to improve the accuracy	Context dependent event selection and accurate sound event modeling are two important factors for the improvement in AED
[66]	To improve the efficiency of acoustic scene classification and acoustic event detection	Gated recurrent neural networks (GRNN) + linear discriminant analysis (LDA)	DCASE2016 task 1	Achieved overall accuracy of 79.1% on DCASE2016 challenge. Relative improvement of 19.8% as compared to GMM	LDA minimizes inner class variance but not efficient for high dimensional data

recognition, sound waves of speaker's speech are converted into the electrical signal and then transformed into digital signals. These digital speech signals are then represented in discrete sequence of feature vectors [145]. The pipeline of speech recognition system consists of feature extraction, acoustic modeling, pronunciation modeling and decoder. Generally, these automatic speech recognition systems are divided into five categories according to classification methods such as Template-based approaches, Knowledge-based approaches, dynamic time warping (DTW), hidden Markov model (HMM) and artificial neural network (ANN) based approaches [146]. Recently, the exponential growth of unstructured big data and computational power, ASR is moving towards more advanced and challenging applications such as mobile interaction with voice, voice control in smart systems, communicative assistance [147]. For such large scale and real world applications, Table 12 presents the recent literature on ASR to discuss state-of-art classification approaches, its variants, evaluation results and remarks on the proposed solution.

ANN based approaches are followed in most of the research studies because these approaches can handle complex interactions and are easier to use as compared to statistical methods. ASR systems can be speaker-independent or speaker-dependent recognition systems. For speaker-dependent recognition systems, template-based methods are performing better due to individual reference template for each speaker which requires large training data from each individual [69]. Due to separate template for each individual, high accuracy can be achieved even in noise, but these methods are suitable for small scale data because, at large scale, it is ineffective to collect large training data from each individual. Rather than collecting large data for training, reinforcement learning can be adopted to make speaker identification automated. To implement speaker-dependent recognition systems at large scale, Apache Hadoop can be used to implement parallelism to make system computationally efficient. Whereas speaker-independent speech recognition systems are not achieving as high accuracy due to noise and overlapping in speech, and language used in speech. Rule-based approaches in speaker-independent recognition system require linguistic skills to implement rules that is a laborious task but rule-based approaches provide quality pronunciation dictionaries [70]. Rule-based methods have limitations of poor generalizability to implement multilingual recognition system or switching for different languages. HMM based speech recognition uses statistical method for data modeling [71]. These systems require large training data for huge number of parameters for HMM. In contrast, ANN-based methods are more flexible and nonlinear e.g. DNN [72, 73], CNN [69, 74], RNN [75]. ANN-based speech recognition systems are more generalize and have flexibility towards changing environments. ANN-based data models are informative and nonlinear. Several ANN-based solutions have been developed for different languages other than English such as Punjabi [76], Tunisian [70], Chhattisgarhi [77], Tamil [78], Amazigh [71] and Russian [79]. The evaluation of LSTM RNN based ASR have proved that word level acoustic models without language model are more efficient to improve accuracy [75]. The performance of ASR is sensitive to pooling size but insensitive to overlap between pooling units with CNN implementation [74]. Although ANN-based ASR systems achieved overall better performance, these systems also have some limitations. The quality of results is unpredictable due to its black box and empirical nature. To improve its computational power, cluster

Table 12 Automatic speech recognition

	Purpose	Approach	Technique	Dataset	Results/limitations
[68]	To improve computational power To enhance the training capability of larger models To ease the process	ANN	Mariana: GPU and CPU clusters for parallelism Three frameworks were developed: multi-GPU for DNN, multi-GPU for DCNN, CPU cluster for large scale DNN	With 6 GPUs, 4.6 times speedup over one GPU was achieved and character error rate was decreased by 10% as compared to existing techniques DNN framework with GPUs performed better for ASR
[72]	To investigate noise robustness on DNN based models	ANN	DNN-HMM: DNN based noise aware training	Aurora 4 w/o explicit noise compensation	7.5% relative improvement Dropout training in DNN with overlapping concern as compared to feature space and model space noise adaptive training
[73]	Bilingual ASR system for Frisian and Dutch languages	ANN	DNN with language dependent and language independent phone	FAME Speech database	Bilingual DNN trained on phones of both languages achieved best performance yielding CS-WER of 59.5% and WER of 38.8% Code switching ASR combines phones of two languages outperformed on WER whereas latency of switching is also an important factor for these systems
[75]	To improve performance for large vocabulary speech recognition	ANN	LSTM RNN	2800 utterance, each distorted once with held-out noise samples	On 25 k word vocabulary, 19.5% WER, and 14.5% in vocabulary WER Word level acoustic models without language model can achieve reasonable accuracy
[74]	To compare the performance of DNN-HMM with CNN-HMM	ANN	CNN with limited weight sharing scheme to model speech features	Small-scale phone recognition in TIMIT large vocabulary voice search task	CNN reduce error rate by 6% to 10% compared with DNN ASR performance is sensitive to pooling size but insensitive to overlap b/w pooling units The results were better for voice search experiment but not for phone recognition

Table 12 (continued)

	Purpose	Approach	Technique	Dataset	Results/limitations
[71]	To develop ASR for Amazigh language	HMM	GMM and tied states MFCC for feature extraction, phonetic dictionary, language model using CMU-Cambridge Statistical Language Modeling Toolkit, HMM based large vocabulary system	New corpus with 187 distinct isolated word speech recording by 50 speakers	Achieved reduced WER to 8, 20% The new corpus was collected. Results are not compared with existing state of the art techniques
[69]	LMS Adaptive filter are introduced to preprocess the speech signals and to identify speaker	Template based	Adaptive Filtering + feature extraction + dimensionality reduction + ensemble classification model using LSTM, ICNN, and SVM	IITG multi variability speaker recognition database	Achieved 95.69% accuracy for noisy data Follows sequential processing Require memory-bandwidth bound computation Required large amount of training data for each new speaker
[70]	ASR for Tunisian dialect	Rule based	G2P rules were defined to build pronunciation dictionaries	TARIC, 9.5 h speech for training and 43 min for testing	WER of 22.6% Validated on manually annotated dataset Improved quality pronunciation dictionaries can be build using expert knowledge but high linguistic skills are required

based solution was proposed with DNN framework that speeds up the process 4.6 times and reduces the error rate by 10% [68]. Overall, ANN based ASR systems are performing better than other classification approaches. Hence, modified ANN based ASR systems are required to improve the accuracy of these systems.

Video IE

The primary goal of IE from the video is to understand and extract relevant information from video content carried in videos. The applications of IE from video are semantic indexing [148], content-based analysis and retrieval, content-oriented video coding, Visually impaired people assistance and automation in supermarkets [149]. In the era of big data, social media and many other platforms are producing digital videos at very high speed. It is not only about size of data that matters, high computational power and speed are also essential to extract useful information from these digital videos. In this regard, Apache Hadoop has been used to implement an extensible distributed video processing framework in cloud environment [80]. FFmpeg and OpenCV for video coder and image processing respectively were implemented using MapReduce showing 75% scalability.

Generally, perceptual and semantic content can be extracted from videos. Semantic contents deal with the objects and their relationship [149]. The spatial and temporal association among objects and entities have been used to reduce the semantic gap between visual appearance and semantics with the help of fuzzy logic and RBM [81]. The proposed system achieved high precision but relatively low recall. Similarly, event extraction from audio-visual content consisting of CNN based audio-visual multimodal recognition was developed and incorporated knowledge from the website using HHMM was used to improve the efficiency. The proposed approach outperformed in terms of accuracy and concluded that CNN provides noise and occlusion robustness [82]. The following subsections extensively discuss the issues and state of the art techniques for subtasks of IE from video content.

Text recognition

The large volume of video data is produced and shared every day on social media. Text in videos plays an important role to extract rich information and provides semantic clues about the video content. Text extraction and analysis in video have shown considerable performance in image understanding. A wide variety of methods have been proposed in this regard. Caption text and scene text are two categories of text that can be extracted from videos [150]. Caption text provides high-level semantic information in captions, overlays and subtitles, whereas scene text is normally embedded in the images such as sign boards, trademarks, etc. Caption text or artificial text recognition is easier than scene text because caption text is added over the video to improve the understandability. Whereas, scene text recognition is complex due to low contrast, background complexity, different font size, orientation, type and language [83]. Besides, low-quality video frames, blur frames and high computation time are specific challenges related to video text extraction process [84].

The pipeline of text detection and extraction consists of text detection, text localization, text tracking, text binarization and text recognition stages. Focusing on IE techniques, this review presents only state of the art techniques for text recognition. Text

recognition system to extract semantic content from Arabic Tv channel using CNN with auto encoder was developed. The accuracy of character recognition was 94.6% [85]. Moreover, a similar system for Arabic News video was developed for video indexing using OCR engine ABBYY FineReader with linguistic analysis and achieved 80.52% F-measure [86]. Another text recognition system was developed for overlay text extraction and person information extraction using rule-based approach for NER to extract person, organization and location information. To extract text, ABBYY FineReader was used [148]. These text recognition systems deal with printed and artificial text only that is comparatively easy to extract. On the other hand, text binarization is important to segment natural scene text with filtering and iterative variance-based threshold calculation [87]. DNN has the ability to provide robust solution in end to end text recognition in videos. In this regard, Faster R-CNN [88], CNN [89, 90], LSTM based method [91] have shown comparatively better performance on scene text recognition. In general, temporal redundancy can be used in tracking for text detection and recognition from complex videos [92].

Traditional systems are not capable of managing and efficiently analyzing the complex big data. MapReduce based parallel processing system has been proposed to detect text in videos. The proposed system achieved high-speed performance on YouTube videos but the system only detects the text from videos using texture-based features [84]. Text recognition plays an important role in understanding multimedia data and multimedia retrieval, visually impaired people assistance, content-based multimedia analysis [151]. Multimedia big data is growing very fast in batch or streaming, more advanced and computationally powerful techniques are required for text recognition from multimedia big data. More robust algorithm to recognize variety of scene and artificial text from low-quality videos are required having the capability to address the space and speed performance in this area.

Automatic video summarization

Automatic tools are essential to analyze and understand visual content. People are generating huge volume of videos using mobile phones, wearable cameras and Google Glass, etc. Some examples of this explosive growth are: 144,000 h videos are uploaded daily on YouTube, lifeloggers generate Gigabytes videos using wearable cameras, 422,000 CCTV cameras are generating videos 24/7 in London [93]. The explosive growth of video data on daily basis highlighted the need to develop fast and efficient automatic video summarization algorithms. AVS has many applications in real life like surveillance, social media, monitoring, etc. [152]. It provides the summary of the video content in skim through video that presents the short video of semantic content of original long video, known as skimmed based summarization or dynamic video summarization. The second is key-frame based video summarization, a.k.a static video summarization, where frames and audio-visual features are extracted [94]. Selecting the most relevant or important frames or subshots from the video for video summarization is a critical task. Several supervised, unsupervised and other techniques are introduced in the literature of computer vision and multimedia. Selection and prioritization criteria for frames and skims is designed manually in unsupervised approach [95, 96] whereas supervised techniques leverage user-generated summaries for learning [94, 97, 98]. Each technique has

different properties for representativeness, diversity and interestingness [93]. Recently, supervised techniques are achieving promising results as compared to traditional unsupervised techniques [94]. Recent literature on user-generated videos have been presented in Table 13.

Poor quality e.g. erratic camera motion, variable illumination, etc. and content sparsity i.e. difficulty in finding representative frames, are two important challenges for AVS with user-generated videos [95]. Despite the limitations of unsupervised techniques, modifications such as incorporating prior information about category [95], selection of deep features rather than shallow features [96] have been presented. Unfortunately, the systems were unable to show promising improvement. Furthermore, it is difficult to define optimized joint criteria for frame selection due to the selection complexity of frame among large number of possible subsets. In contrast, supervised techniques require large annotated data that is one of its major limitations due to the shortage of large datasets [98]. Overall, supervised techniques are outperforming unsupervised techniques. However, more efficient and fast algorithms are required for AVS specially to deal with the variety and velocity of big data.

Results and discussion

This SLR distills the key insights from the comprehensive overview of IE techniques for a variety of data types and take a fresh look at older problems, which nevertheless are still highly relevant today. Big data brings a computational paradigm shift to IE techniques. In this regard, this SLR presents a comprehensive review of existing IE techniques for variety of data types. To the best of our knowledge, IE techniques from variety of unstructured big data at a single platform have not been addressed yet. In order to achieve this goal, SLR methodology has been followed to explore the advancements in IE techniques in recent years. To meet the objectives of the study, most relevant and up to date literature on IE techniques for text, images, audio and video data have been discussed. The selected studies have been classified according to IE subtasks for each data type and shown in Fig. 4.

Big data value chain defines high-level activities that are important to find useful information from big data where IE process is concerned with the data analysis. Therefore, the impact of inefficiencies of IE techniques will ultimately decrease the performance of big data analytics or decision making. In order to improve the big data analytics and decision making, this SLR was aimed to investigate the challenges of IE process in the age of big data for variety of data types. The objective of combining IE techniques for variety of data types at single platform was twofold. First, to identify the state of the art IE techniques for variety of big data and second, to investigate the major challenges of IE associated with unstructured big data. Further, the need for new consolidated IE systems is highlighted and some preconditions are also proposed to improve the IE process for the variety of data types in big data. This identified challenges of IE associated with unstructured big data have been discussed in the following subsection.

Unstructured big data challenges for IE

The challenges of IE from unstructured big data are categorized into task-dependent and task-independent categories. The task-dependent challenges have been discussed

Table 13 Automatic video summarization

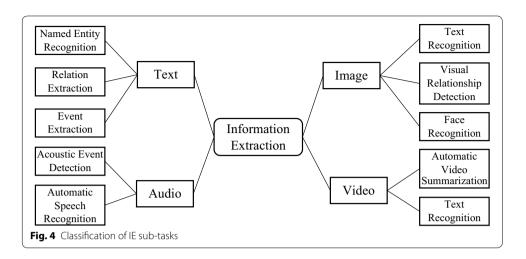|	Approach	Technique	Purpose	Dataset	Results/limitations
[97]	Supervised with prior segmentation	SVM based kernel video segmentation	Category specific video summarization	MED summaries training set 12,249 videos and testing set 60 videos	Higher quality video summaries can be produced with known categories than unsupervised approach
[95]	Unsupervised with web based prior information	Used four baseline algorithms: random and uniform sampling, k-means and spectral clustering followed by crowdsourcing	To deal with content sparsity and large scale evaluation	180 videos 25 for training and 155 for evaluation	Content sparsity and poor quality of user generated videos are major challenges Expert evaluation is not possible for large scale data, therefore crowdsourcing used Adding web images of category to incorporate knowledge is time consuming process especially for unknown categories
[98]	Supervised	Linear combination of Submodular maximization for each objective using structured learning	To implement interestingness, representativeness, and uniformity	Egocentric dataset and SumMe dataset	Shortage of large datasets for summarization
[94]	Supervised	vsLSTM to model variable range temporal dependency	To address the need for large amount of annotated data	SumMe and TVSum	Domain adaptation can improve learning and reduces discrepancies
[93]	Supervised	Sequential determinantal point process: supervised DPP coupled with NN representation	To incorporate human created summaries for selection of informative and diverse datasets	Open video project (50), YouTube (39), Kodak consumer video (18 videos)	Supervised approach with linear representation performed better
[99]	NA	MSR (Minimum Sparse Representation) based summarization	To utilize min number of keyframes To provide flexibility for practical applications	Open video project (50 videos), several genres dataset (50 videos)	Two variants were proposed for off-line and on-line applications Focused on selection of key frames
[96]	Unsupervised	General adversarial framework: summarizer (auto encoder LSTM) + discriminator LSTM (LSTM)	To regularize the summary length, diversity, and keyframes for	SumMe, TVSum, open video project, YouTube	Different performance on different datasets Deep features perform better than shallow features Frames with very slow motion and no scene change gave poor results

Fig. 4 Classification of IE sub-tasks

Table 14 Independent challenges identified from selected studies

Challenges of unstructured big data IE	Studies	Frequency
Data quality	[31, 58, 59, 63, 64, 84, 95, 138]	9
Data sparsity	[16, 22, 31, 34, 95]	6
Data volume	[10, 13, 15, 19, 65]	5
Data usability	[11–13, 27, 28]	5
Context understanding	[27, 31, 39, 64]	4
Computational requirements	[15, 68, 84, 124]	4
Data dimensionality	[16, 18, 66]	3
Heterogeneity	[33, 131]	2
Diversity	[55, 124]	2
Semantic understanding	[43, 44]	2
Data modeling	[16, 68]	2
Ambiguities in data	[31, 105]	2
Data scarcity	[62]	1
Balance among informativeness, representativeness, and diversity	[127]	1

in their corresponding sections with state of the art techniques in each area. Task-independent challenges are discussed in this section. Table 14 presents a summary of the challenges identified from the selected studies.

A. Quality of unstructured big data

Noise [31, 63, 64], missing data [59], incomplete data [15] and low quality data [58, 59, 84, 95, 138] are major quality issues of unstructured big data that degrades the performance of IE process. The quality issues of unstructured big data are huge barriers in extracting useful and most relevant information that makes IE process arduous. Quality improvement, early in the process, is the utmost requirement of IE from unstructured big data.

B. Data sparsity

The enormous growth of user-generated content increased the data sparsity (a.k.a. data sparseness and data paucity) issues where only small fraction of data contains interesting and useful information [16, 22, 31, 95]. Text analysis of social media data, summarization of visual data are directly associated with user-generated content. Due to the sparsity of content, it became difficult to find most relevant representative data to produce semantically rich results. There is a false assumption about large datasets that frequent extractions from large datasets can produce better results [22]. Extracting a small amount of evidence in the corpus to present useful information is a challenge for unstructured big datasets. Therefore, sparse IE for large scale and variety of big data for user-generated content have great opportunities along with the challenges to improve the IE process.

C. Volume of unstructured big data

People and machines are great producers of unstructured big data. The volume of data brings some opportunities as well as challenges for IE from the huge deluge of user and machine-generated content. Existing techniques should adopt new size and time requirements to deal with IE from big data [15, 84]. Automatic IE and structuring the unstructured big data requires the scaling of existing methods designed for very small data to process millions of data records [10, 13]. Therefore, distributed and parallel computing should be adopted for improved efficacy of IE from unstructured big data.

D. Dimensionality and heterogeneity

Unstructured big data comes with high dimensionality [16, 18, 66], diversity [55, 124], dynamicity [32] and heterogeneity [33, 131]. Dimensionality reduction [18] and semantic annotation [131] can further improve the IE performance of high dimensional and heterogeneous data respectively. The techniques with high representational power are appropriate for high dimensional data [66]. With the influx of data from increasingly diverse sources, big data IE and analytics require advanced techniques to handle more than data accessibility.

E. Data usability

Unstructured big data is a rich source of information but exploitation of relevant information is one of the major challenges [27, 28]. It is more relevant to the optimal data selection with balance of cost, speed and accuracy [12]. The main problem with unstructured big data is, huge deluge of data is available, but it is not usable. Usability of data is defined as the capacity of data to fulfill the requirements of user for a given purpose, area and epoch. According to the definition of data usability [153], "Usability is the degree to which each stakeholder is able to effectively access and use the data". Data usability helps to know more about data, its understanding and its usage. Therefore, usability varies due to the different interpretation of meaning of data values and different nature of tasks that relates IE process improvement to data usability improvement.

F. Context and semantic understanding

Identifying the context of interaction among entities and objects is a crucial task in IE [39, 64], especially with high dimensional, heterogeneous, complex and poor quality data. Data ambiguities add more challenges to contextual IE [31, 105]. Semantics are important to find relationship among entities and objects [44]. Entities and object

extraction from text and visual data could not provide accurate information unless the context and semantics of interaction are identified [43]. Efficient data prioritization and curation is important in this regard [27]. Therefore, semantic and context understanding is important as well as challenging for big data IE due to quality and usability issues.
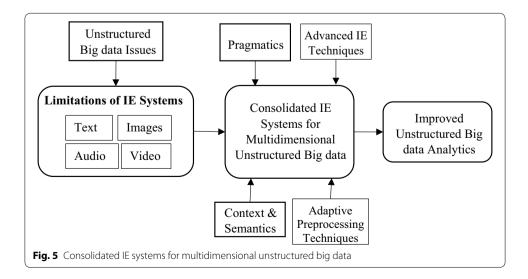
G. Data modeling

As discussed earlier, learning-based techniques are more popular for IE as it reduces manual intensive labor. Efficient data modeling is an important task in learning-based IE techniques. High dimensionality, heterogeneity and low quality of unstructured big data add complexities to data modeling process [16]. Efficient parallelism and computational power are required to support large data models [68].

Need for consolidated IE systems for multidimensional unstructured big data

The critical analysis of the existing literature selected in this SLR has identified various task-specific and data-specific challenges for big data IE. Based on the findings of this SLR, variety of big data is posing challenges to extract useful information. Every field is using IE systems for variety of data to perform mining and analysis. New consolidated systems to extract useful information from variety of data types can improve the efficiency of big data analytics by integrating the extracted information. For example, Healthcare systems are using variety of big data in different systems like decision support systems, disease identification, Pharmacovigilance and Healthcare analytics, etc. Consolidated IE systems would help to improve these systems by extracting useful information from variety of unstructured data. The analysis of existing IE techniques and limitations arises the need for the consolidation of IE techniques for variety of data types. The identified need has been depicted in Fig. 5.

As shown in Fig. 5, the identified task-specific and data-specific limitations of IE systems should be considered to design an IE system for more than one data type. Meanwhile, the proposed improvement preconditions should also be considered for the development of these systems. The identified challenges and proposed preconditions will help to extract relevant and useful information from variety of big data. Following

Fig. 5 Consolidated IE systems for multidimensional unstructured big data

are some improvement preconditions that have been proposed for these new consolidated IE systems for multidimensional unstructured big data.

Preconditions 1: Advanced preprocessing

Most of the challenges, identified in this SLR, are related to the quality and usability of unstructured big data. Data and process standardization, efficient data cleaning and quality improvement techniques are required for unstructured big data. Further, advanced and adaptive preprocessing techniques prior to IE are required to improve the effectiveness of big data analytics.

Precondition 2: Pragmatic IE

Pragmatics is a field of study that is related to the usefulness and usability of data [154]. It deals with the dimensions of data that are important to improve the usefulness and usability of data. As IE is a community based process, it depends on the user needs and available data source [100]. Therefore, IE equipped with pragmatics will help to improve unstructured data analysis as it will extract and select data according to the user needs. Pragmatic IE solutions are required to improve big data analytics and big data IE.

Precondition 3: Context and semantics are more important

Context and semantics play an important role in understanding relation among entities or objects. Extracting most relevant data is a difficult task for unstructured big data due to its complexity and quality. Therefore, contextually and semantically rich IE techniques will increase the robustness of big data IE.

Precondition 4: Selection of technique

Selection of appropriate techniques according to the data has strong impact on the results of IE process especially for unstructured big data due to its complexity and large size. Traditional IE techniques are inadequate to efficiently handle unstructured big data. It has been observed that selection of appropriate techniques highly depends on the data characteristics. Weakly supervised or distant supervised learning techniques are suitable for large scale and multi domain datasets as these techniques require small training samples [17]. Unsupervised techniques are suitable for heterogeneous data [32], whereas deep CNN have performed better on high dimensional data [36]. Therefore, understanding the data is an important factor for selection of IE technique.

Conclusion

The systematic literature review serves the purpose of exploring state-of-the-art techniques for IE from unstructured big data types such as text, image, audio and video investigating the limitations of these techniques. Besides, the challenges of IE in big data environment have also been identified. It is found that analysis and mining of data are getting more complex with massive growth of unstructured big data. Deep learning with its generalizability, adaptability and less human involvement capability is playing a key role in this regard. However, to process exponentially growing data, new flexible and scalable techniques are required to deal with the dynamicity and sparsity of unstructured data. Quality, usability and sparsity of unstructured big data are major obstruct

in deriving useful information. For improving the IE techniques, mining useful information and supporting versatility of unstructured data, it is required to introduce new techniques and make improvements and enhancements in existing techniques. Overall, the existing IE techniques are outperforming traditional techniques for comparatively larger datasets but inadequate to effectively deal with rapid growth of unstructured big data especially streaming data. Scalability, accuracy and latency are important factors in implementation of these IE techniques in big data platform. Apache MapReduce is also facing scalability issues in big data IE. To overcome these challenges, MapReduce based deep learning solutions are the future of big data IE systems. These systems will be helpful for healthCare analytics, surveillance, e-Government systems, social media analytics and business analytics. The outcome of the study shows that highly scalable and computationally efficient and consolidated IE techniques are required to deal with the dynamicity of unstructured big data. The study significantly contributes to the identification of the challenges to achieve more scalable and flexible IE systems. Quality, usability, sparsity, dimensionality, heterogeneity, context and semantics understanding, scarcity, modeling complexity and diversity of unstructured big data are major challenges in this field. Advanced data preparation techniques, prior to extracting information from unstructured data, semantically and contextually rich IE systems, the emergence of pragmatics and advanced IE techniques are essential for IE systems in unstructured big data environment. Hence, Scalable, computationally efficient and consolidated IE systems are required the can overcome the challenges of multidimensional unstructured big data.

Future work

The major focus of the review was to investigate the challenges of IE systems for multidimensional unstructured big data. The detailed discussion on IE techniques from variety of data types concluded that data preparation is equally important to the efficiency of IE systems. Advanced data improvement techniques will also increase the efficiency of IE systems. Therefore, the findings of the review will be used to develop a usability improvement model for unstructured big data to extract maximum useful information from these data.

References
1. Gantz J, Reinsel D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC iView IDC Analyze Future. 2012;2007(2012):1–16.
2. Wang Y, Kung LA, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change. 2018;126:3–13.
3. Lomotey RK, Deters R. Topics and terms mining in unstructured data stores. In: 2013 IEEE 16th international conference on computational science and engineering, 2013. p. 854–61.
4. Lomotey RK, Deters R. RSenter: terms mining tool from unstructured data sources. Int J Bus Process Integr Manag. 2013;6(4):298.
5. Scheffer T, Decomain C, Wrobel S. Mining the Web with active hidden Markov models. In: International conference on data mining. New York: IEEE; 2001; p. 645–6.
6. Lomotey RK, Jamal S, Deters R. SOPHRA: a mobile web services hosting infrastructure in mHealth. In: First international conference on mobile services. New York: IEEE; 2012; p. 88–95.
7. Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M. Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw. 2007;80(4):571–83.
8. Borrego M, Foster MJ, Froyd JE. Systematic literature reviews in engineering education and other developing interdisciplinary fields. J Eng Educ. 2014;103(1):45–76.
9. Che N, Chen D, Le J. Entity recognition approach of clinical documents based on self-training framework. In: Recent developments in intelligent computing, communication and devices. Singapore: Springer; 2019; p. 259–65.
10. Liu X, Zhou Y, Wang Z. Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network. J Vis Commun Image Represent. 2019;60:1–15.
11. Mao J, Cui H. Identifying bacterial biotope entities using sequence labeling: performance and feature analysis. J Assoc Inf Sci Technol. 2018;69(9):1134–47.
12. Goldberg S, Wang DZ, Grant C. A probabilistically integrated system for crowd-assisted text labeling and extraction. J Data Inf Qual. 2017;8(2):1–23.
13. Boytcheva S, Angelova G, Angelov Z, Tcharaktchiev D. Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybern Inf Technol. 2015;15(4):58–77.
14. Pogrebnyakov N. Unsupervised domain-agnostic identification of product names in social media posts. In: International conference on big data. New York: IEEE; 2018; p. 3711–6.
15. Napoli C, Tramontana E, Verga G. Extracting location names from unstructured italian texts using grammar rules and MapReduce. In: International conference on information and software technologies. Cham: Springer; 2016; p. 593–601.
16. Feldman K, Faust L, Wu X, Huang C, Chawla NV. Beyond volume: the impact of complex healthcare data on the machine learning pipeline. In: Towards integrative machine learning and knowledge extraction. Cham: Springer; 2017; p. 150–69.
17. Wang K, Shi Y. User information extraction in big data environment. In: 3rd IEEE international conference on computer and communications (ICCC). New York: IEEE; 2017; p. 2315–8.
18. Li P, Mao K. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. Expert Syst Appl. 2019;115:512–23.
19. Wang P, Hao T, Yan J, Jin L. Large-scale extraction of drug-disease pairs from the medical literature. J Assoc Inf Sci Technol. 2017;68(11):2649–61.
20. Guo X, He T. Leveraging Chinese encyclopedia for weakly supervised relation extraction. In: Joint international semantic technology conference. Cham: Springer; 2015; p. 127–40.
21. Torres JP, de Piñerez Reyes RG, Bucheli VA. Support vector machines for semantic relation extraction in Spanish language. In: Advances in computing. Cham: Springer; 2018; p. 326–37.
22. Li P, Wang H, Li H, Wu X. Employing semantic context for sparse information extraction assessment. ACM Trans Knowl Discov Data. 2018;12(5):1–36.

23. Liu Z, Tong J, Gu J, Liu K, Hu B. A Semi-automated entity relation extraction mechanism with weakly supervised learning for Chinese medical webpages. In: International conference on smart health. Cham: Springer; 2016; p. 44–56.

24. Li J, Cai Y, Wang Q, Hu S, Wang T, Min H. Entity relation mining in large-scale data. In: Database systems for advanced applications. Cham: Springer; 2015; p. 109–121.

25. Wang C, Song Y, Roth D, Zhang M, Han J. World knowledge as indirect supervision for document clustering. ACM Trans Knowl Discov Data. 2016;11(2):1–36.

26. Gao H, Gui L, Luo W. Scientific literature based big data analysis for technology insight. J Phys Conf Ser. 2019;1168(3):032007.

27. Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinform. 2015;16(1):55.

28. Fadili H, Jouis C. Towards an automatic analyze and standardization of unstructured data in the context of big and linked data. In: Proceedings of the 8th international conference on management of digital ecosystems—MEDES. New York: ACM Press; 2016; p. 223–30.

29. Swain MC, Cole JM. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. J Chem Inf Model. 2016;56(10):1894–904.

30. Miwa M, Thompson P, Korkontzelos Y, Ananiadou S. Comparable study of event extraction in newswire and bio-medical domains. In: 25th international conference on computational linguistics. 2014; p. 2270–9.

31. Roll U, Correia RA, Berger-Tal O. Using machine learning to disentangle homonyms in large text corpora. Conserv Biol. 2018;32(3):716–24.

32. Xiang L, Zhao G, Li Q, Hao W, Li F. TUMK-ELM: a fast unsupervised heterogeneous data learning approach. IEEE Access. 2018;6:35305–15.

33. Shi L, Jianping C, Jie X. Prospecting information extraction by text mining based on convolutional neural networks–a case study of the Lala copper deposit, China. IEEE Access. 2018;6:52286–97.

34. Mezhar A, Ramdani M, Elmzabi A. A novel approach for open domain event schema discovery from twitter. In: 2015 10th international conference on intelligent systems: theories and applications (SITA). New York: IEEE; 2015; p. 1–7.

35. Gong L, Zhang Z, Yang X, Huang D, Yang R, Yang G. A biomedical events extracted approach based on phrase structure tree. In: 2017 13th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD). New York: IEEE; 2017; p. 1984–88.

36. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.

37. KHe K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016; p. 770–8.

38. Liang X, Lee L, Xing EP. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2017; p. 4408–17.

39. Zhuang B, Liu L, Shen C, Reid I. Towards context-aware interaction recognition for visual relationship detection. In: Proceedings of the IEEE international conference on computer vision (ICCV). 2017; p. 589–98.

40. Ramanathan V et al. Learning semantic relationships for better action retrieval in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2015; p. 1100–9.

41. Jung J, Park J. Visual relationship detection with language prior and softmax. In: 2018 IEEE international conference on image processing, applications and systems (IPAS). 2018; p. 143–8.

42. Yu R, Li A, Morariu VI, Davis LS. Visual relationship detection with internal and external linguistic knowledge distilla-tion. In: Proceedings of the IEEE international conference on computer vision (ICCV). 2017; p. 1068–76.

43. Baier S, Ma Y, Tresp V. Improving information extraction from images with learned semantic models. arXiv preprint arXiv:1808.08941 2018.

44. Dai Y, Wang C, Dong J, Sun C. Visual relationship detection based on bidirectional recurrent neural network. Multi-media Tools and Appl. 2019. https://doi.org/10.1007/s11042-019-7732-z.

45. Han Y, Xu Y, Liu S, Gao S, Li S. Visual relationship detection based on local feature and context feature. In: 2018 International conference on network infrastructure and digital content (IC-NIDC). New York: IEEE; 2018; p. 420–4.

46. Vellingiriraj EK, Balamurugan M, Balasubramanie P. Information extraction and text mining of Ancient Vattezhuthu characters in historical documents using image zoning. In: 2016 international conference on Asian language processing (IALP). New York: IEEE; 2016; p. 37–40.

47. Singh D, Saini JP, Chauhan DS. Hindi character recognition using RBF neural network and directional group feature extraction technique. In: 2015 International conference on cognitive computing and information processing (CCIP). New York: IEEE; 2015; p. 1–4.

48. Sheshadri K, Divvala SK. Exemplar driven character recognition in the wild. In: Proceedings of the British Machine Vision Conference (BMVC). 2012; p. 13.1–13.10.

49. Shi Cun-Zhao, Wang Chun-Heng, Xiao Bai-Hua, Gao Song, Jin-Long Hu. Scene text recognition using structure-guided character detection and linguistic knowledge. IEEE Trans Circuits Syst Video Technol. 2014;24(7):1235–50.

50. Yao C, Bai X, Shi B, Liu W. Strokelets: a learned multi-scale representation for scene text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014; p. 4042–49.

51. Avadesh M, Goyal N. Optical character recognition for Sanskrit using convolution neural networks. In: 2018 13th IAPR international workshop on document analysis systems (DAS). New York: IEEE; 2018. p. 447–52.

52. Younis KS, Alkhateeb AA. A new implementation of deep neural networks for optical character recognition and face recognition. Jordan: Proc New Trends Inf Technol; 2017. p. 157–62.

53. Elleuch M, Tagougui N, Kherallah M. Towards unsupervised learning for Arabic handwritten recognition using deep architectures. In: International conference on neural information processing. Cham: Springer; 2015; p. 363–372.

54. Ding Z, Chen Z, Wang S. FANet: an end-to-end full attention mechanism model for multi-oriented scene text recognition. In: 2019 5th international conference on big data and information analytics (BigDIA). New York: IEEE; 2019; p. 97–102.

55. Medhat F et al. Theodoropoulos G, Obara B. TMIXT: a process flow for Transcribing MIXed handwritten and machine-printed text. In: 2018 IEEE international conference on big data (Big Data). 2018; p. 2986–94.

56. Xie H, Fang S, Zha Z-J, Yang Y, Li Y, Zhang Y. Convolutional attention networks for scene text recognition. ACM Trans Multimedia Comput Commun Appl. 2019;15(1s):1–17.

57. Zheng Y, Wang Q, Betke M. Deep neural network for semantic-based text recognition in images. Computer vision and pattern recognition. No. arXiv:1908.01403. 2019.

58. Wani MA, Bhat FA, Afzal S, Khan AI. Supervised deep learning in face recognition. Singapore: Springer; 2020. p. 95–110.

59. Heinsohn D, Villalobos E, Prieto L, Mery D. Face recognition in low-quality images using adaptive sparse representations. Image Vis Comput. 2019;85:46–58.

60. Abudarham N, Shkiller L, Yovel G. Critical features for face recognition. Cognition. 2019;182:73–83.

61. Prasad PS, Pathak R, Gunjan VK, Rao HR. Deep learning based representation for face recognition. In: ICCCE 2019. Springer: Singapore; 2019; p. 419–4.

62. Gemmeke JF, Vuegen L, Karsmakers P, Vanrumste B. An exemplar-based NMF approach to audio event detection. In: 2013 IEEE workshop on applications of signal processing to audio and acoustics. 2013; p. 1–4.

63. Espi M, Fujimoto M, Kinoshita K, Nakatani T. Exploiting spectro-temporal locality in deep learning based acoustic event detection. EURASIP J Audio Speech Music Process. 2015;2015(1):26.

64. Heittola T, Mesaros A, Eronen A, Virtanen T. Context-dependent sound event detection. EURASIP J Audio Speech Music Process. 2013;2013(1):1.

65. Takahashi N, Gygli M, Pfister B, Van Gool L. Deep convolutional neural networks and data augmentation for acoustic event detection. In: InterSpeech. arXiv:1604.07160. 2016.

66. Zöhrer M, Pernkopf F. Gated recurrent networks applied to acoustic scene classification and acoustic event detection. In: Proceedings of the detection and classification of acoustic scenes and events workshop (DCASE2016), Budapest, Hungary, 3 Sept 2016, p. 115–9.

67. Su TW, Liu JY, Yang YH. Weakly-supervised audio event detection using event-specific Gaussian filters and fully convolutional networks. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). 2017; p. 791–5.

68. Zou Y, Jin X, Li Y, Guo Z, Wang E, Xiao B. Mariana: tencent deep learning platform and its applications. Proc VLDB Endow. 2014;7(13):1772–7.

69. Devi KJ, Thongam K. Automatic speaker recognition with enhanced swallow swarm optimization and ensemble classification model from speech signals. J Ambient Intell Human Comput. 2019. https://doi.org/10.1007/s12652-019-01414-y.

70. Masmoudi A, Bougares F, Ellouze M, Estève Y, Belguith L. Automatic speech recognition system for Tunisian dialect. Lang Resour Eval. 2018;52(1):249–67.

71. El Ouahabi S, Atounti M, Bellouki M. Toward an automatic speech recognition system for amazigh-tarifit language. Int J Speech Technol. 2019;22(2):421–32.

72. Seltzer ML, Yu D, Wang Y. An investigation of deep neural networks for noise robust speech recognition. In: 2013 IEEE international conference on acoustics, speech and signal processing. 2013; p. 7398–402.

73. Yılmaz E, van den Heuvel H, van Leeuwen D. Investigating bilingual deep neural networks for automatic recognition of code-switching Frisian speech. Procedia Comput Sci. 2016;81:159–66.

74. Abdel-Hamid O, Mohamed A, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process. 2014;22(10):1533–45.

75. Sak H, Senior A, Rao K, Beaufays F. Fast and accurate recurrent neural network acoustic models for speech recognition. Computation and language. No. arXiv:1507.06947. 2015.

76. Kumar Y, Singh N. An automatic speech recognition system for spontaneous Punjabi speech corpus. Int J Speech Technol. 2017;20(2):297–303.

77. Londhe ND, Kshirsagar GB. Chhattisgarhi speech corpus for research and development in automatic speech recognition. Int J Speech Technol. 2018;21(2):193–210.

78. Lokesh S, Kumar PM, Devi MR, Parthasarathy P, Gokulnath C. An automatic Tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. Neural Comput Appl. 2019;31(5):1521–31.

79. Karpukhin IA. Contribution from the accuracy of phoneme recognition to the quality of automatic recognition of Russian speech. Moscow Univ Comput Math Cybern. 2016;40(2):89–95.

80. Ryu C, Lee D, Jang M, Kim C, Seo E. Extensible video processing framework in Apache Hadoop. In: 2013 IEEE 5th international conference on cloud computing technology and science. 2013; p. 305–310.

81. Manju A, Valarmathie P. Organizing multimedia big data using semantic based video content extraction technique. In: 2015 International conference on soft-computing and networks security (ICSNS). New York: IEEE; 2015; p. 1–4.

82. Kojima R, Sugiyama O, Nakadai K. Audio-visual scene understanding utilizing text information for a cooking support robot. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS). 2015; p. 4210–5.

83. Risnumawan A, Shivakumara P, Chan CS, Tan CL. A robust arbitrary text detection system for natural scene images. Expert Syst Appl. 2014;41(18):8027–48.

84. Ben Ayed A, Ben Halima M, Alimi AM. MapReduce based text detection in big data natural scene videos. Procedia Comput Sci. 2015;53:216–23.

85. Yousfi S, Berrani SA, Garcia C. Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos. In: 2015 13th international conference on document analysis and recognition (ICDAR) New York: IEEE; 2015; p. 1026–30.

86. Mansouri S, Charhad M, Rekik A, Zrigui M. A framework for semantic video content indexing using textual information. In: 2018 IEEE second international conference on data stream mining & processing (DSMP). 2018; p. 107–10.
87. Sudir P, Ravishankar M. An effective approach towards video text recognition. In: Advances in signal processing and intelligent recognition systems. Cham: Springer; 2014; p. 323–33.
88. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst. 2015;28:91–9.
89. Wang X et al. End-to-end scene text recognition in videos based on multi frame tracking. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR). New York: IEEE; 2017; p. 1255–60.
90. Ali A, Pickering M, Shafi K. Urdu natural scene character recognition using convolutional neural networks. In: 2018 IEEE 2nd international workshop on Arabic and derived script analysis and recognition (ASAR). 2018; p. 29–34.
91. Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans Pattern Anal Mach Intell. 2017;39(11):2298–304.
92. Tian S, Yin X-C, Su Y, Hao H-W. A unified framework for tracking based text detection and recognition from web videos. IEEE Trans Pattern Anal Mach Intell. 2018;40(3):542–54.
93. Gong B, Chao WL, Grauman K, Sha F. Diverse sequential subset selection for supervised video summarization. Adv Neural Inf Process Syst. 2014;27:2069–77.
94. Zhang K, Chao WL, Sha F, Grauman K. Video summarization with long short-term memory. In: European conference on computer vision 2016, Cham: Springer; 2016; p. 766–82.
95. Khosla A, Hamid R, Lin CJ, Sundaresan N. Large-scale video summarization using web-image priors. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2013. p. 2698–705.
96. Mahasseni B, Lam M, Todorovic S. Unsupervised video summarization with adversarial LSTM networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2017; p. 2982–91.
97. Potapov D, Douze M, Harchaoui Z, Schmid C. Category-specific video summarization. In: European conference on computer vision. Cham: Springer; 2014; p. 540–55.
98. M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3090–3098.
99. Mei S, Guan G, Wang Z, Wan S, He M, Feng DD. Video summarization via minimum sparse reconstruction. Pattern Recognit. 2015;48(2):522–33.
100. Lomotey RK, Deters R. Real-time effective framework for unstructured data mining. In: 2013 12th IEEE international conference on trust, security and privacy in computing and communications. 2013; p. 1081–8.
101. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investig. 2007;30(1):3–26.
102. Marrero M, Urbano J, Sánchez-Cuadrado S, Morato J, Gómez-Berbís JM. Named Entity recognition: fallacies, challenges and opportunities. Comput Stand Interfaces. 2013;35(5):482–9.
103. Abdallah ZS, Carman M, Haffari G. Multi-domain evaluation framework for named entity recognition tools. Comput Speech Lang. 2017;43:34–55.
104. Sazali SS, Rahman NA, Bakar ZA. Information extraction: Evaluating named entity recognition from classical Malay documents. In: 2016 third international conference on information retrieval and knowledge management (CAMP). 2016; p. 48–53.
105. Goyal A, Gupta V, Kumar M. Recent Named entity recognition and classification techniques: a systematic review. Comput Sci Rev. 2018;29:21–43.
106. Piskorski J, Yangarber R. Information extraction: Past, present and future. In: Multi-source, multilingual information extraction and summarization. Berlin: Springer; 2013; p. 23–49.
107. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: European conference on information retrieval. 2005; p. 345–59.
108. Konstantinova N. Review of relation extraction methods: What is new out there?. In: International conference on analysis of images, social networks and texts. Cham: Springer; 2014; p. 15–28.
109. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2(1):1.
110. Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: opportunities and challenges. Neurocomputing. 2017;237:350–61.
111. Wang W, et al. Deep learning at scale and at ease. ACM Trans Multimedia Comput Commun Appl. 2016;12(4s):1–25.
112. Wang Y, et al. Clinical information extraction applications: a literature review. J Biomed Inform. 2018;77:34–49.
113. Chiticariu L, Li Y, Reiss FR. Rule-based information extraction is dead! Long live rule-based information extraction systems! In: Proceedings of the 2013 conference on empirical methods in natural language processing 2013; p. 827–32.
114. Valenzuela-Escárcega MA, Hahn-Powell G, Surdeanu M, Hicks T. A domain-independent rule-based framework for event extraction. In: Proceedings of ACL-IJCNLP 2015 system demonstrations. 2015; p. 127–32.
115. Patel R, Tanwani S. Application of machine learning techniques in clinical information extraction. In: Smart techniques for a smarter planet. Cham: Springer; 2019; p. 145–65.
116. Topaz M, et al. Mining fall-related information in clinical notes: comparison of rule-based and novel word embedding-based machine learning approaches. J Biomed Inform. 2019;90:103103.
117. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. J Biomed Inform. 2009;42(5):923–36.
118. Gorinski PJ et al. Named entity recognition for electronic health records: a comparison of rule-based and machine learning approaches. Computation and language. 2019.
119. Atzmueller M, Kluegl P, Puppe F. Rule-based information extraction for structured data acquisition using Text-Marker. In: LWA. 2008; p. 1–7.
120. Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction. In: Proceedings of the conference on empirical methods in natural language processing. 2011; p. 1535–45.

121. Kanya N, Ravi T. Modelings and techniques in named entity recognition: an information extraction task. In: IET Chennai 3rd international conference on sustainable energy and intelligent systems (SEISCON 2012). 2012; p. 104–8.
122. Wani MA, Bhat FA, Afzal S, Khan AI. Introduction to deep learning. In: Advances in deep learning. Singapore: Springer; 2020; p. 1–11.
123. Coates A, Carpenter B, Case C, Satheesh S, Suresh B, Wang T, Wu DJ, Ng AY. Text detection and character recognition in scene images with unsupervised feature learning. In: ICDAR. 2011; p. 440–5.
124. Wang H, Nie F, Huang H. Large-scale cross-language web page classification via dual knowledge transfer using fast nonnegative matrix trifactorization. ACM Trans Knowl Discov Data. 2015;10(1):1–29.
125. Jan B et al. Deep learning in big data analytics: a comparative study. Comput Electr Eng. 2019;75:275–87.
126. Gheisari M, Wang G, Bhuiyan MZ. A survey on deep learning in big data. In: 2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC). 2017; p. 173–80.
127. Reyes O, Ventura S. Evolutionary strategy to perform batch-mode active learning on multi-label data. ACM Trans Intell Syst Technol. 2018;9(4):1–26.
128. Berndt DJ, McCart JA, Finch DK, Luther SL. A case study of data quality in text mining clinical progress notes. ACM Trans Manag Inf Syst. 2015;6(1):1–21.
129. Nuray-Turan R, Kalashnikov DV, Mehrotra S. Adaptive connection strength models for relationship-based entity resolution. J Data Inf Qual. 2013;4(2):1–22.
130. Zhang Z, Gao J, Ciravegna F. SemRe-rank: improving automatic term extraction by incorporating semantic relatedness with personalised pagerank. ACM Trans Knowl Discov Data. 2018;12(5):1–41.
131. Adrian WT, Leone N, Manna M, Marte C. Document layout analysis for semantic information extraction. In: Conference of the Italian association for artificial intelligence. 2017. Cham: Springer; 2017; p. 269–81.
132. C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual Relationship Detection with Language Priors," in *Computer Vision - ECCV 2016*, Springer, Cham, 2016, pp. 852–869.
133. Antol S et al. VQA: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. 2015; p. 2425–33.
134. Ma L, Lu Z, Shang L, Li H. Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE international conference on computer vision. 2015; p. 2623–31.
135. Yatskar M, Zettlemoyer L, Farhadi A. Situation recognition: visual semantic role labeling for image understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016; p. 5534–42.
136. Joan SF, Valli S. A survey on text information extraction from born-digital and scene text images. Proc Natl Acad Sci. 2019;89(1):77–101.
137. Jung K, Kim KI, Jain AK. Text information extraction in images and video: a survey. Pattern Recognit. 2004;37(5):977–97.
138. Zhang H, Zhao K, Song Y-Z, Guo J. Text extraction from natural scene image: a survey. Neurocomputing. 2013;122:310–23.
139. Young AW, Burton AM. Recognizing faces. Curr Direct Psychol Sci. 2017;26(3):212–7.
140. Young AW, Burton AM. Are we face experts? Trends Cognit Sci. 2018;22(2):100–10.
141. Peng YT, Lin CY, Sun MT, Tsai KC. Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models. In: 2009 IEEE International conference on multimedia and expo. 2009; p. 1218–21.
142. Harma A, McKinney MF, Skowronek J. Automatic surveillance of the acoustic activity in our living environment. In: 2005 IEEE international conference on multimedia and expo. 2005; p. 634–7.
143. Zhuang X, Zhou X, Hasegawa-Johnson MA, Huang TS. Real-world acoustic event detection. Pattern Recognit Lett. 2010;31(12):1543–51.
144. Li J, Deng L, Gong Y, Haeb-Umbach R. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans Audio Speech Lang Process. 2014;22(4):745–77.
145. Saini P, Kaur P. Automatic speech recognition: a review. Int J Eng Trends Technol. 2013;4(2):1–5.
146. Cutajar M, Gatt E, Grech I, Casha O, Micallef J. Comparative study of automatic speech recognition techniques. IET Signal Process. 2013;7(1):25–46.
147. He X, Deng L. Speech-centric information processing: an optimization-oriented approach. Proc IEEE. 2013;101(5):1116–35.
148. Lee S, Jo K. Automatic person information extraction using overlay text in television news interview videos. In: 2017 IEEE 15th international conference on industrial informatics (INDIN). 2017; p. 583–8.
149. Lu T, Palaiahnakote S, Tan CL, Liu W. Introduction to video text detection. In: Video text detection. London: Springer; 2014; p. 1–18.
150. Ye Q, Doermann D. Text detection and recognition in imagery: a survey. IEEE Trans Pattern Anal Mach Intell. 2015;37(7):1480–500.
151. Zhu Y, Yao C, Bai X. Scene text detection and recognition: recent advances and future trends. Front Comput Sci. 2016;10(1):19–36.
152. Rajpoot V, Girase S. A study on application scenario of video summarization. In: 2018 Second international conference on electronics, communication and aerospace technology (ICECA). New York: IEEE; 2018; p. 936–43.
153. Shanks G, Corbitt B. Understanding data quality: social and cultural aspects. In: Proceedings of the 10th Australasian conference on information systems. 1999; p. 785–96.
154. Price R, Shanks G. A semiotic information quality framework: development and comparative analysis. In: Enacting research methods in information systems. Cham: Springer; 2016; p. 219–50.

Publisher's Note