Heritage Science

RESEARCH

Open Access

CICHMKG: a large-scale and comprehensive Chinese intangible cultural heritage multimodal knowledge graph

Tao Fan[1,2], Hao Wang[1*] and Tobias Hodel[2]

Abstract

Intangible Cultural Heritage (ICH) witnesses human creativity and wisdom in long histories, composed of a variety of immaterial manifestations. The rapid development of digital technologies accelerates the record of ICH, generating a sheer number of heterogenous data but in a state of fragmentation. To resolve that, existing studies mainly adopt approaches of knowledge graphs (KGs) which can provide rich knowledge representation. However, most KGs are text-based and text-derived, and incapable to give related images and empower downstream multimodal tasks, which is also unbeneficial for the public to establish the visual perception and comprehend ICH completely especially when they do not have the related ICH knowledge. Hence, aimed at that, we propose to, taking the Chinese nation-level ICH list as an example, construct a large-scale and comprehensive Multimodal Knowledge Graph (CICHMKG) combining text and image entities from multiple data sources and give a practical construction framework. Additionally, in this paper, to select representative images for ICH entities, we propose a method composed of the denoising algorithm (CNIFA) and a series of criteria, utilizing global and local visual features of images and textual features of captions. Extensive empirical experiments demonstrate its effectiveness. Lastly, we construct the CICHMKG, consisting of 1,774,005 triples, and visualize it to facilitate the interactions and help the public dive into ICH deeply.

Keywords Digital humanities, Intangible cultural heritage, Multimodal knowledge graph

Introduction

Intangible Cultural Heritage (ICH), also known as the living culture, is formed from human activities over thousands of years and consists of various immaterial manifestations, e.g., skills, dancing, singing, as well as the witness of history, achievement of human creativity and reflection of cultural diversity [1–5]. It is crucial and meaningful to safeguard and take advantage of ICH.

In 2003, UNESCO passed the convention for the safeguarding of the ICH, which extremely facilitated the protection and record of ICH in the world [6–8]. Also, in China, influenced by the convention, a nation-level ICH list has been enacted by the ministry of culture and tourism, including 1557 ICH projects (composed of 3610 sub-projects) since 2006 where there are 42 world-level ICH projects,[1] which reflects the excellent representative traditional culture of Chinese nation, promotes the safeguarding and transmission of ICH, and enhances the nation confidence and identification. Besides, to accelerate the sustainability of ICH, great contributions have been made to record ICH in a form of texts, images, etc., producing a large amount of data with high value [1, 2, 9]. However, the invaluable ICH data is under-utilized,

*Correspondence:
Hao Wang
ywhaowang@nju.edu.cn
[1] School of Information Management, Nanjing University, Nanjing 210023, China
[2] Digital Humanities, University of Bern, 3012 Bern, Switzerland

[1] https://www.ihchina.cn/.

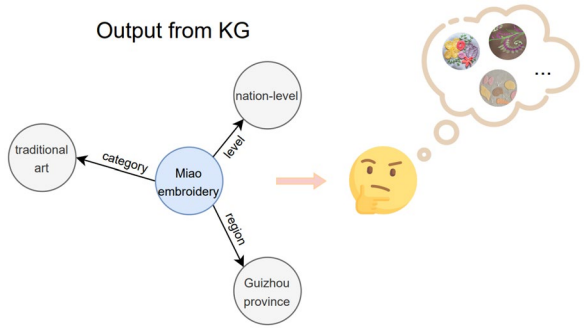

Fig. 1 An example of what text-based KG outputs partially with the ICH entity 苗绣 *(Miao embroidery)* as a query. Blue circle, gray circle, and edge represent the ICH entity, attribute values, and attributes respectively

Caption: The inheritor performed the casting skill of *Jingdezhen handmade porcelain craftsmanship.*

Fig. 2 An example of 景德镇手工制瓷工艺 *(Jingdezhen handmade porcelain craftsmanship)*

and stored in different databases and dispersed on official websites [10–12]. Consequently, concealed knowledge embedded in ICH data is not excavated and thus made accessible effectively, and in a state of fragmentation, which becomes a hindrance for the public and experts to perceive and analyse ICH in an interconnected manner.

Aimed at that problem, some researchers tried to draw into knowledge graph (KG), where ICH data is processed as nodes and they are linked based on relations and attributes [13–17]. For example, Carriero et al. [18] constructed the ArCo KG for Italian cultural heritage consisting of millions of triples. The KG was based on massive standard semi-structured textual data from multiple resources, and constructed under the guidance of the designed ontology through extending the eXtreme Design methodology. Also, Kalita & Deka [15] took the traditional dances in India as an example, and built the corresponding traditional dance KG based on the documents from museums and archives, depicting connections and interactions deeply between components of traditional dances, e.g., custom materials, acting roles, instruments. Though these KGs can provide rich text-based knowledge, they are not equipped with the capability to offer visual information. When the public tried to learn about an unfamiliar ICH through the text-based KG, the KG output is a bunch of related entities with the name of the ICH project (*Miao embroidery*) as an input, as shown in Fig. 1. But this does not address the question that how the ICH project is to be imagined.

When referring to the symbol *Car*, we can easily imagine what it represents because we have the related knowledge and experiences. But for the public without abundant ICH knowledge or experiences, the name of ICH is just an abstract symbol to them and the entities output from the KG are just textual descriptions of the symbol, which is unhelpful for the public to understand ICH and obtain a comprehensive overview. Besides, the text-based KG is not capable to empower downstream tasks, e.g., visual question-answering and cross-modal retrieval, which may make the multimodal ICH data under-utilized and under-cut its potential enormous value. Therefore, aimed at the gap, we propose to, taking the Chinese nation-level ICH list as an example, build a large-scale and comprehensive Multimodal Knowledge Graph (CICHMKG) integrating a great deal of text and image entities.

Great progress has been made in the construction of MKG, however, it is still a challenging task [19–21]. The main difficulties lie in how to select representative images for entities in KG from massive collections of images. Currently, there have been some explorations to tackle them. For example, Wang et al. [22] collected millions of images from several search engines and constructed the MKG *Richpedia* for city, sight, and celebrity entities in the world. In the process of noise filtration, they first employed VGG network to extract visual features as the representation of images. Next, a clustering-based method was utilized to remove noise images and select representative images. Similarly, Han et al. [23] built a tourist attraction MKG and developed a framework to remove noise images and select representative attraction images. In the proposed framework, geotagged information was incorporated to visual features, producing more distinguished image features. Then, the fused features, as the representation of images, were clustered and ranked to determine the representative images. We can find that the representation of images has a fundamental influence on the subsequent noise filtration and representative image selection. However, the utilized features in current research are mainly the global visual features and basic tagged information, which are not sufficient and complete to express the images especially in our ICH context. For instance, as exhibited in Fig. 2, it can be observed

that what the ICH image expresses focuses on the local regions with the blue box "inheritor" and orange box "the casting skill", which are the direct manifestation of ICH, and the caption also provides the description with key semantics for the ICH image. Nevertheless, these salient features are ignored in current studies about selecting representative images.

Based on that, in this paper, we first construct text-based KG for Chinese nation-level ICH entities, and propose a Clustering-based Noise Images Filtration Algorithm (CNIFA) to remove noise images in collected massive Chinese ICH images from multiple sources, integrating global and local features of images and textual features of attached captions. Thereafter, representative images for ICH entities are selected from denoised images with fused multimodal features based on the developed criteria. Lastly, we construct CICHMKG based on numerous text and image entities, consisting of 1,774,005 triples. Our contributions can be summarized as follows:

- We propose to adopt a way of from symbols to images to explore the construction of the large-scale and comprehensive multimodal knowledge graph in cultural heritage domain. Also, we establish a practical MKG construction framework.
- We propose to combine global and local visual features of images and attached textual features of captions as the representation of images, and develop a cluster-based method, composed of CNIFA and several criteria, to select representative images for ICH entities. Empirical experiments demonstrate that the proposed multimodal representation is satisfying and promising in selecting representative images.
- To help the public dive into ICH deeply, we also visualize CICHMKG and give examples of how to interact with CICHMKG.

The rest of the paper is organized as follows. "Related work" summarizes the studies about the KG construction in cultural heritage. "Methodology of CICHMKG construction" describes the proposed framework in the construction of CICHMKG, the denoising algorithm CNIFA, and the criteria to select representative images. "Evaluation of proposed methods" presents and discusses the experimental results. We give the statistics information and visualization of CICHMKG in "Statistics and visualization of CICHMKG" . Theoretical and practical implications are discussed in "Theoretical and practical implications" . The paper is summarized in "Conclusions" .

Related work

In this section, we will review the related work about the KG construction in culture heritage domain, not limited in the scope of ICH, and progress in MKG construction.

KG construction in cultural heritage

The rapid advancement of digital technologies promotes the record of cultural heritage, generating a sheer number of heterogeneous data which lays a solid foundation for the proceeding utilization and excavation, as well as provides general and strong models to help us to comprehend and dive into the culture [13]. However, these cultural heritage data are stored and presented in different databases and websites dispersedly, which results in these data being in a state of fragmentation [10, 24, 25].

With the emergence of KG, researchers seem to find out the method to resolve how to present the public more comprehensive and aggregated knowledge. Since KG possesses a strong capability to describe and link all things under a logical model, it can integrate all knowledge related to the cultural heritage and cope with the existing data fragmentation dilemma. Specifically, to empower the knowledge discovering and retrieval of Ireland's history, Debruyne et al. [26] first employed the CIDOC-CRM to develop the ontology. Then, they created massive metadata from the historical databases as the population for the ontology. Following the construction framework, the KG related with Ireland's history was built based on the triples processed from the metadata. To facilitate the ontology design on historical data and corresponding KG construction, *OntoME*[2] is developed based on CIDOC-CRM by Data for History Consortium [27]. Besides, cultural heritage data is not only disseminated in archives but also in digitalized texts, encyclopedias, etc. Integrating the knowledge from these channels is beneficial for the knowledge aggregation of cultural heritage. To tackle that, Buranasing and Lilakiataskun [28] proposed an effective model to extract entities and relations combining features of lexical features, multi-instance learning, etc. Furtherly, to discover the relations between historical persons in Finland, Hyvönen and Rantala [29] collected biographies, in the form of text, from museums, archives, libraries, etc., and employed the rule-based technique to clean data and produce triples. Finally, they constructed the KG for 13,000 historical persons in Finland, composed of 373,000 triples, which facilitated the application for seeking the relations of historical persons hidden in the data.

It can be discovered that existing research related to KG construction in the cultural heritage domain are

[2] https://ontome.net.

mainly textual data-based. Currently, there are some researchers turning their sights on the multimodal cultural heritage data [17, 25, 30–32]. For example, to promote data aggregation and alleviate publishing problems in Digital Humanities, Hyvönen [33] proposed the Sampo model to help users release linked data (compatible with multimodal data) under unified principles and create portals efficiently, which has been tested in a series of work [34, 35]. To retrieve and recommend European silk fabrics, Thomas et al. [36] developed a similarity calculation method combining rules formulated by expert knowledge and a knowledge graph built from the metadata of images, and siamese CNN. The proposed method was tested on several datasets and experimental results showed that it had a good performance in calculating the similarity of fabric images. Furtherly, to facilitate the preservation and understanding of European silk heritage, Puren and Vernus [37] established a conceptual model based on CIDOC CRM. The proposed conceptual model was employed to represent knowledge extracted from the metadata and data annotation results from free texts of silk including texts and images. Similarly, to give a complete description of the heritage object, Carboni and De Luca [38] developed an ontology integrating Visual and Iconographical Representations (VIR) and CIDOC CRM, which can be used to analyse the links between visual items in images.

Overall, their main aims do not focus on the construction of multimodal knowledge graph but the image similarity calculation, conceptual modelling, etc., although these studies utilized the multimodal data. These studies explored utilizing ontology to describe the multimodal data in cultural heritage [37, 38]. Moreover, they concentrated on the ontology construction and did not refer to how to generate large-scale triples automatically to populate the ontology. Additionally, it is time-consuming and laborious to annotate the entities occurring in images manually, which is not appropriate to create a large-scale MKG. In this paper, different from previous studies [10, 18, 24, 26, 39], we focus on the construction of MKG combining numerous text and image entities in cultural heritage domain. Based on that, we propose to, taking the Chinese nation-level ICH list as an example, construct a large-scale and comprehensive ICH multimodal knowledge graph (CICHMKG) combining massive text and image entities, and give the corresponding construction framework.

MKG construction

"Multimodal" is a buzzword in the Artificial Intelligence (AI) community, which is the combination of multiple unimodalities, e.g., texts, images, audios, and widely employed in several traditional AI tasks, sentiment analysis, named entity recognition, question-answering [40–44], etc. Due to the superior performance elicited from the empowerment of KG in the mentioned tasks [45–47], more and more researchers notice the potentiality of MKG in current times inundated with multimodal data, and make empirical explorations.

Currently, studies about the construction of MKG can be mainly divided into two forms, from symbols to images and from images to symbols [19–21, 48, 49]. The underlying idea of the former is to select representative images for the entities (symbols) in traditional KG and the latter, in an opposite way, is to label entities in images and generate triples based on visual semantic relations amid recognized entities. For example, to create KG for diagrams in deep learning papers, Roy et al. [50] proposed Dia2Graph model based on the pretrained CNN model to extract nodes, arrows, and texts. However, the model needs a large scale of annotated data with high quality to train the visual recognition model so that it is able to recognize the entities in new images. Also, it is restraint to the predefined types and can not recognize the entities outside the types, and hard to describe relations between entities showing in the images exactly. Given the scale of the processing data, it is extremely time-consuming and expensive. As aforementioned, there has been studies trying to employ the ontology to describe visual or iconographic items in images via labour in cultural heritage domain [37, 38]. However, it is not realistic to annotate visual or iconographic items in the large-scale images relying on labour. Additionally, if extracting these visual or iconographic items from images automatically, we first need to know what kinds of these items may exist in these images, and also requires a huge amount of well-annotated data by labour to train the recognition model. Moreover, it is of difficulty to give complete predefined types of visual or iconographic items and creating such annotated data needs huge resources. Essentially, we assume that the public are interested in ICH and the role of images is to help them comprehend ICH in MKG. Therefore, considering our task context, it is more appropriate and feasible to avail of the way of from symbols to images to construct the MKG oriented to ICH entities, which is also common-used in the construction of MKG.

However, in the strategy of from the symbols to images to construct MKG, the main challenge can be summarized as how to select representative images from a great deal of images for entities in traditional KG. In this study, M. Wang et al. [22] first selected city entities, sight entities and celebrity entities from Wikipedia as the core entities in the constructed MKG, and images related with entities are scraped from several search engines. To select representative images, they first utilized global VGG16 to extract visual features of images and then removed noise

based on clustering results. Lastly, representative images were selected based on the distance between images and the root of the corresponding cluster. To capture visual features of images, Y. Deng et al. [51] employed content and style features of paintings via VGG19 network as the representative features. Afterwards, they adopted the robust clustering method to select representative paintings relying on proposed constraint rules. To generate representative images for landmarks, Jiang et al. [52] first removed noise images from collected images, and clustered the images, using SIFT algorithm to represent global features of images. Lastly, representative images were determined through designed rules.

From the above research, we can discover that the representation of images is the key step for the selection of representative images. It can be observed that the global feature is mainly utilized as the representation of images, whereas, it is hard to capture salient features of images and not capable to express images completely. Local regions in images and attached captions can provide invaluable information for enhancing semantic understanding of images, supported by validations in several tasks, e.g., multimodal sentiment analysis and multimodal named entity recognition [44, 53]. Also, such clues are reflected in our context. As shown in Fig. 2, local visual features in the ICH image and captions can supplement extra fundamental semantic information to represent the ICH image. Inspired by that, different from previous studies [22, 23, 51, 52, 54], we propose to combine the global and local visual features, and the textual features of captions to represent the ICH images, and select representative ICH images utilizing the multimodal features.

Methodology of CICHMKG construction

In this section, we will give CICHMKG and problem definition respectively, and the construction framework of CICHMKG. In the following subsections, we will illustrate the construction pipeline with the framework in detail. Also, the algorithm of removing noise images and how to select representative images for ICH will be explained elaboratively.

CICHMKG definition

For conciseness, the traditional ICH KG only including textual entities is denoted as $\mathcal{G}_{\mathcal{KG}}$. Here, we denote CICHMKG as the set $\mathcal{G}_{\mathcal{MKG}} = \{\mathcal{E}, \mathcal{A}, \mathcal{V}, \mathcal{R}, \mathcal{T}\}$, where $\mathcal{E}$, $\mathcal{A}$, $\mathcal{V}$, $\mathcal{R}$, $\mathcal{T}$ represent ICH entities, attributes, literal attributes values, relations, triples respectively. $\mathcal{E} = \mathcal{E}_{\mathcal{ICH}} \cup \mathcal{E}_{\mathcal{IMG}}$ is the set of ICH entities (Chinese nation-level ICH projects) $\mathcal{E}_{\mathcal{ICH}}$ and image entities $\mathcal{E}_{\mathcal{IMG}}$, and $\mathcal{T} = \mathcal{T}_{\mathcal{A}} \cup \mathcal{T}_{\mathcal{R}}$ is the set of attribute triples $\mathcal{T}_{\mathcal{A}} = \mathcal{T}_{\mathcal{A}}^{\mathcal{ICH}} \cup \mathcal{T}_{\mathcal{A}}^{\mathcal{IMG}}$ and relation triples $\mathcal{T}_{\mathcal{R}} = \mathcal{E}_{\mathcal{ICH}} \times \mathcal{R} \times \mathcal{E}_{\mathcal{IMG}}$, where $\mathcal{T}_{\mathcal{A}}^{\mathcal{ICH}} = \mathcal{E}_{\mathcal{ICH}} \times \mathcal{A}_{\mathcal{ICH}} \times \mathcal{V}_{\mathcal{ICH}}$ and $\mathcal{T}_{\mathcal{A}}^{\mathcal{IMG}} = \mathcal{E}_{\mathcal{IMG}} \times \mathcal{A}_{\mathcal{IMG}} \times \mathcal{V}_{\mathcal{IMG}}$. For instance, a triple *(Beijing Opera, declaring area, Beijing)* $\in \mathcal{T}_{\mathcal{A}}^{\mathcal{ICH}}$ in CICHMKG with the form of *(s, p, o)* denotes the attribute *declaring area* of the ICH entity *Beijing Opera* is the attribute value *Beijing*, and similarly a triple *(IMG, ImageOf, Beijing Opera)* $\in \mathcal{T}_{\mathcal{R}}$ denotes that the image entity *IMG* and the ICH entity *Beijing Opera* have the relation *ImageOf*.

Problem definition

What consist of $\mathcal{G}_{\mathcal{MKG}}$ are numerous triples defined in CICHMKG. Hence, the main problem is how to produce such triples. Following the framework in our previous work [39], we can construct $\mathcal{G}_{\mathcal{KG}}$, which is composed of triples $\mathcal{T}_{\mathcal{A}}^{\mathcal{ICH}}$. Also, the attribute information of ICH images can be extracted through deep pretrained visual models, producing attribute triples $\mathcal{T}_{\mathcal{A}}^{\mathcal{IMG}}$. Then, the difficulty lies in how to select representative images for ICH entities to construct triples $\mathcal{T}_{\mathcal{R}}$ like the form *(IMG, ImageOf, ICH entities)*. Though the collected images are of high quality, there are still noise images. Here, we denote noise images as the images unrelated to the corresponding ICH project. We slash the gordian knot into two parts according to our task context. The first is to remove noise images from collected massive images $\mathcal{I}$, and the second is to select representative images $\mathcal{E}_{\mathcal{IMG}}$ from denoised images $\mathcal{D}$, constructing the multimodal triples.

Overview of the framework

The construction framework for CICHMKG is shown in Fig. 3, which is mainly composed of data acquisition, the selection of representative images, and CICHMKG generation. Basically, the main aim of the proposed framework is to serve the MKG construction in cultural heritage domain, expected to be capable to guide the MKG construction in different nations and communities. In the following, we will demonstrate construction steps and details with the framework.

Ontology for CICHMKG

Here, we introduce the utilized ontology. Normally, before starting to construct the ontology, we first need to consider whether it is possible to reuse existed ontology because rich expert knowledge is required to build a complete ontology. Through searching for related literatures, we discover that the ontology used for *Richpedia* and *IMGpedia* [22, 54] is suitable for our context. Although the ontology is lightweight, it is still capable to provide descriptions for our ICH multimodal data. Basically, we reuse the ontology to describing resources with the Resource Description Framework (RDF), and make

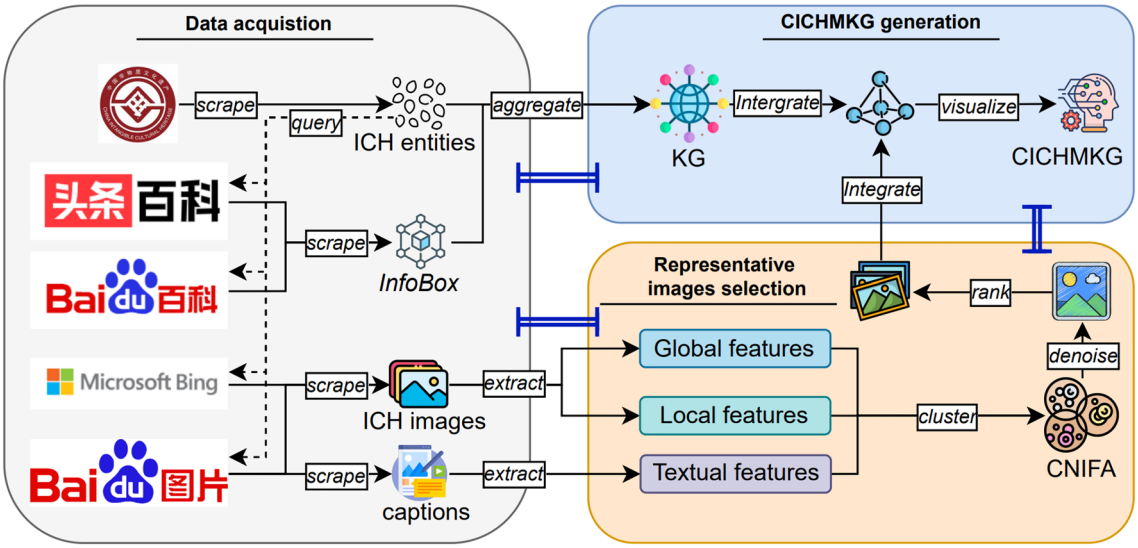

Fig. 3 The construction framework of CICHMKG

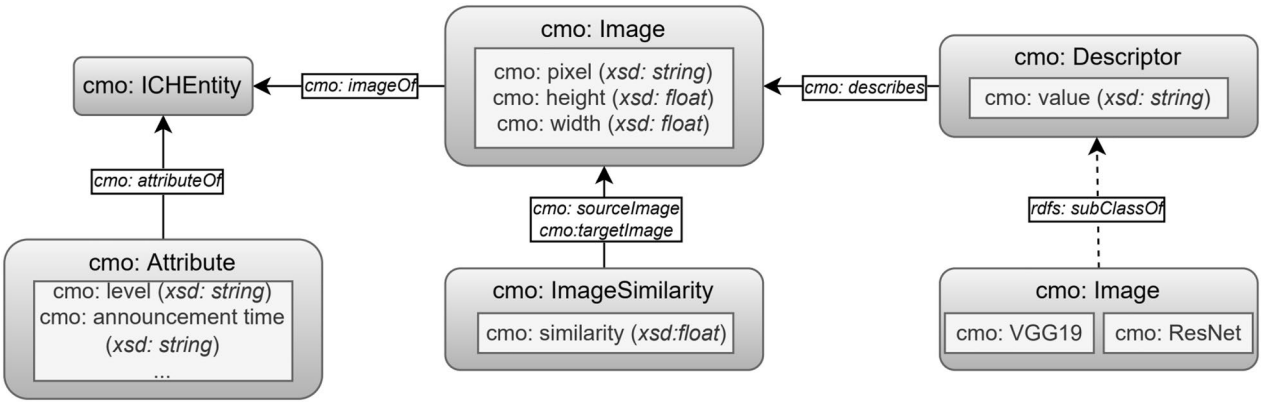

Fig. 4 Ontology for CICHMKG

some modifications to adapt our context. Specifically, compared with them, we add the attribute information of ICH entities and replace visual descriptors. Here, we describe the modified ontology briefly, shown in Fig. 4.

The cmo: ICHEntity denotes an ICH project in the range of Chinese nation-level ICH list, which can have several textual attributes and attribute values, including cmo: level, cmo: announcement time, etc., through the cmo: attributeOf. Also, A cmo: ICHEntity can be linked by several ICH image entities cmo: Image via the relation cmo: ImageOf. Furtherly, cmo: pixel with data type *xsd: string*, cmo: height and cmo: width with data type *xsd: float* are employed to describe the basic properties of cmo: Image. Differently, original low-level features, e.g., HOG, GLCM, in the visual descriptor cmo:Descriptor are replaced as high-level visual features cmo: VGG19 and cmo: ResNet, output from pretrained vision models with the deep structure, which can be employed directly and easily to calculate *cosine* similarity cmo: ImageSimilarity between images. In CICHMKG, triples related to image similarity are not included. Besides, the solid line denotes the relation or attribute between instances produced from classes and the dotted line represents the relation inside the class.

Data acquisition for CICHMKG

Following the construction framework, we separate the data acquisition process into two parts, which are Chinese nation-level ICH entities collection and corresponding image collection. The main target of the former is to obtain all ICH entities, related *InfoBox* information, and description texts, and the latter is to scrape images and attached captions for all ICH entities.

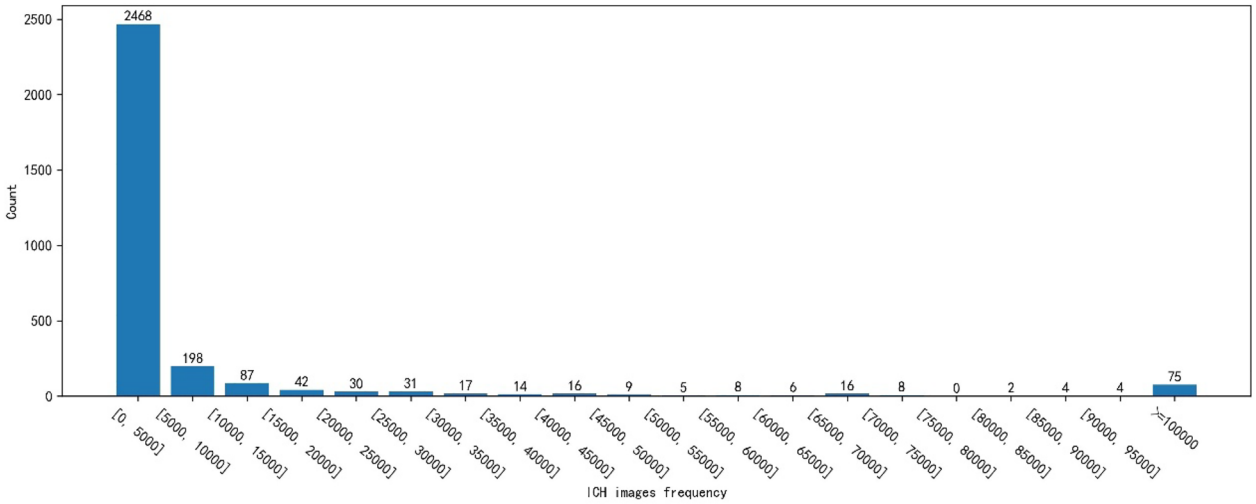

Fig. 5 The frequency distribution of ICH images

ICH entities collection

To safeguard ICH, the ministry of culture and tourism of China has established the official website for documenting the information of nation-level ICH projects, which is an authoritative data source for crawling ICH entities. Moreover, to obtain more comprehensive knowledge, *Baike* and *Toutiao* are utilized as the data sources to extend KG. They are the biggest encyclopedias in Chinese communities, and with high-quality and -accuracy in created contents, which are widely employed as data sources for the construction of KG [10, 14, 22, 39, 55]. As a result, we obtain 3040 nation-level ICH entities after removing deduplication from the official website. Then, the names of ICH entities are employed to compile queries to scrape the *InfoBox* information, which can be transformed into triples easily, and descriptions from the official website, *Baike*, and *Toutiao*. Following our previous work [39], we perform the same filtration and construction process but extend the data sources furtherly, adding the data from *Toutiao* and generating 37,527 triples for $\mathcal{G}_{\mathcal{KG}}$.

Images collection

In this paper, we select *Baidu Image* and *Bing Image* as image sources, which are common-utilized data resources for the construction of MKG in several studies [21, 22, 54], and are reliable and can meet our needs in the preliminary investigation. The former is the biggest image search engine in Chinese community and the latter is employed as the supplementation image resource. If there are no related ICH images in *Baidu Image*, we will seek images from *Bing Image*. Names of ICH entities are employed as the keywords to compile queries. Next, we write a crawler to scrape all images and corresponding captions from chosen image sources, and all scraped images and captions are saved to the server, in a form of. *jpg* and.*txt* respectively. In the crawling process, images of 12 ICH projects are not retrieved in *Baidu Image* and we scrape missing ICH images and captions from *Bing Image*.

Consequently, 37,765,371 ICH images and captions are obtained respectively, and next we conduct a simple statistical analysis on the scraped images. The frequency distribution of ICH images is shown in Fig. 5. It can be observed that the distribution complies with the long-tailed distribution, and the image number of most ICH projects is in the range of [0, 5000].

Chinese ICH is divided into 10 categories listed in Fig. 6, which are formulated by the Chinese related official department, and based on the opinions from experts and scholars in ICH domain [56]. When observing the image number distribution from the view of the ICH category, we can find that the percentage of 民俗 *(Folk-custom)* images in all categories ranks first, and the least is 传统音乐 *(Traditional music)*. It hints that the category occupying a huge percentage of images has higher visibility and popularity, which is also reflected in the image number distribution of different ICH projects. According to the image amount of different ICH projects, the top 3 are 春节 *(Chinese New Year),* 中秋节 *(Mid-Autumn Festival), and* 端午节 *(Dragon Boat Festival)*, which are all the important festivals for Chinese and belong to the category *Folk-custom*, and have an extremely high influence on the society. The last 3 are from the west of China, which are not performed by a large group.

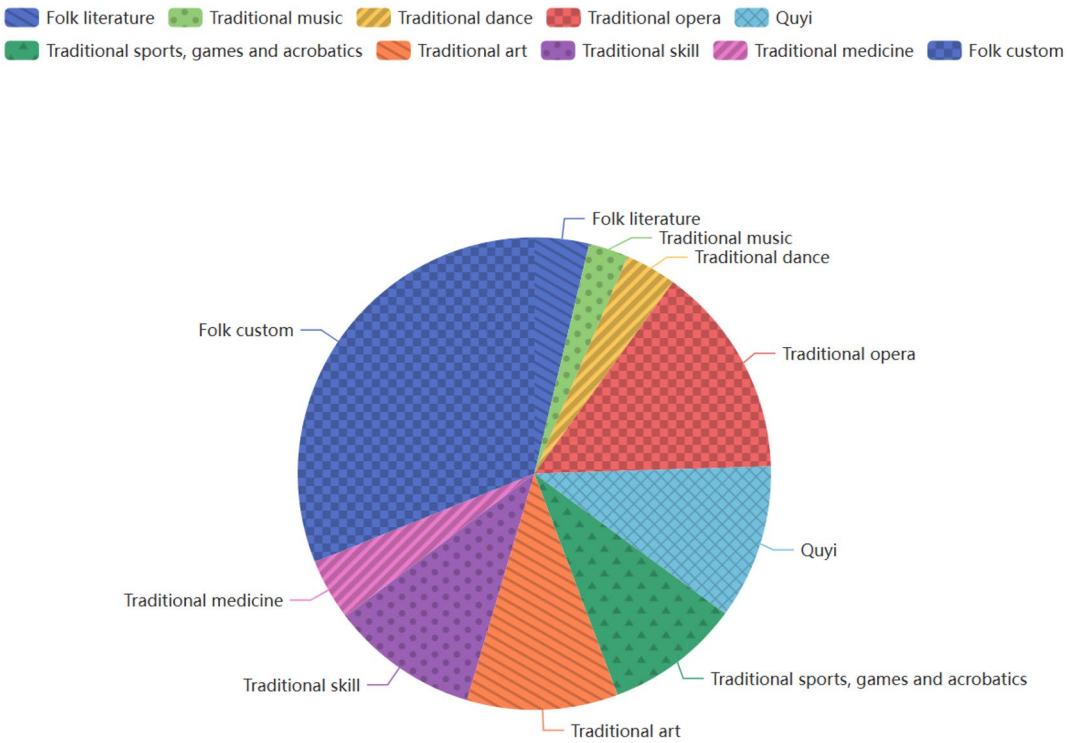

Fig. 6 The image number distribution of different ICH categories

Noise images filtration

Although we have obtained massive images highly related to ICH entities, there are still noise images among them. Previous research about filtering noise images mainly adopted a clustering-based method with the global visual features or tagged information as the input [23, 57, 58], ignoring the local features and attached captions. Moreover, it is not sufficient to use those features to represent ICH images, which are not capable to describe the salient semantic information embedded in ICH images. Two extra instances are shown in Table 1, and we can find that the most distinguished features are concentrated on regions of red boxes, which are the direct manifestation of ICH. In addition, captions attached to images can enhance the semantic understanding of images effectively shown in Fig. 2, and improve the quality of the image representation.

Based on that, different from previous studies [23, 51, 52], we propose to integrate local and global visual features of ICH images and textual features of captions into the multimodal features, as the representation of ICH images, and establish a Clustering-based Noise Images Filtration Algorithm (CNIFA) for ICH images.

Table 1 Examples of ICH images

Name of ICH	Global features	Local features
淮剧 *(Huai Opera)*		
少林功夫 *(Shaolin Kung Fu)*		

Feature representation

Here, we adopt the VGG19 model [59] pretrained on ImageNet [60] to extract visual features from ICH images. To make extracted features contain more spatial and aggregated visual information, the output from the last pooling layer is utilized. To locate the regions of interest in ICH images, the state-of-the-art segmentation YOLOv5 model [61] pretrained on COCO dataset [62] is employed. Through applying the model, we can obtain several locations of the regions, which is outputted in the form of a *.json* file including the coordinates of segmented regions. Through utilizing these locations, we can gain the local features of ICH images via the VGG19 model. For captions, the pretrained BERT model [63] is employed to extract textual semantic features, which has been proven to be effective in several NLP tasks and is usually adopted as the representation of texts.

Formally, given the ICH image $\boldsymbol{V}$ and corresponding caption $\boldsymbol{T}$, the extracted global visual feature after average-pooling and textual feature through VGG19 and the caption after preprocessing can be represented via BERT as $\boldsymbol{V}_{global} \in \mathbb{R}^{512}$ and $\boldsymbol{T}_{bert} \in \mathbb{R}^{768}$ respectively, where 512 and 768 denote the number of the dimension. Let $\boldsymbol{V}_{local} = \{\boldsymbol{V}^1_{lcoal}, \boldsymbol{V}^2_{lcoal}, \ldots, \boldsymbol{V}^j_{lcoal}\}$ be the set of detected local regions in $\boldsymbol{V}$, where j is the number of recognized local regions. Then, local features can be represented as $\boldsymbol{V}_{local} = AVG(\boldsymbol{V}^1_{lcoal}, \boldsymbol{V}^2_{lcoal}, \ldots, \boldsymbol{V}^j_{lcoal})$, where $\boldsymbol{V}^1_{lcoal} \in \mathbb{R}^{512}$ is the extracted region feature and AVG denotes the average-pooling. Lastly, the multimodal representation of ICH image $\boldsymbol{V}$ is $\boldsymbol{V}_{mm} = \boldsymbol{V}_{global} \oplus \boldsymbol{V}_{local} \oplus \boldsymbol{T}_{bert}$, where $\oplus$ is the concatenation operation.

Determine the optimal K value

In this paper, what we use is a clustering-based method to remove ICH noise images. Through clustering, all ICH images will be divided into several clusters based on the represented feature $\boldsymbol{V}_{mm}$. It is assumed that if an ICH image belongs to some cluster, it should be close to the cluster rather than a discrete point in the two-dimension visualization map. Otherwise, it will be seen as the noise.

Here, the classical clustering algorithm K-means [64] is utilized to cluster ICH images. However, an important issue occurs that how to determine the optimal K value because it needs to be set in advance. K value is an import parameter in K-means which has a huge impact on the clustering quality and subsequent noise removal. Normally, it is determined by empirical tuning experiments. Given that we have numerous ICH images, if all images are clustered, it will be very coarse-grained, bring huge computational resource consumption, produce the cluster that we do not want, and unbeneficial to the selection of representative images. Also, topics of ICH images are different from category to category and from project to project.

A natural idea is that we search the optimal K-value for each ICH project in the Chinse nation-level ICH list. However, it is not practical because we have thousands of ICH projects and in the tuning process for each ICH project dozens of experiments are performed which needs huge computational resources and a large amount of running time. Hence, we adopt a more practical strategy. The reason why ICH projects are put into a category is that they have common points from the perspective of expression, manifestation, form, etc. Based on that, we decide that each ICH project in the same category is assigned to an identical K-value. Specifically, 20 ICH projects from each category are picked randomly. The average value of the optimal K-value of each randomly selected 20 ICH projects from the same category is utilized as the optimal K-value, which means that the average optimal K-value is employed for all ICH projects in the same category.

In the process of determining the optimal K-value, we employ a common-used and effective method *Elbow* [65], to determine the optimal K-value. We introduce the main idea of *Elbow* briefly. The core metric of *Elbow* is the Sum of the Squared Errors (*SSE*):

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{1}$$

where p, C_i, and m_i denote image points, the ith cluster, and the centroid in the ith cluster. In the process of K-means clustering, we utilize Principal Component Analysis (PCA) technique to decrease the dimension of $\boldsymbol{V}_{mm}$, preserving more valuable information and increasing the computation efficiency, and $\boldsymbol{V}_{mm}$ processed by PCA is employed to represent each ICH image. When K value is rising, samples will be divided finely and become more aggregated with *SSE* decreasing sharply. When K value is approaching and exceeding the optimal value, the *SSE* will decrease slowly, and a curve of elbow shape will be formed based on the coordinates composed of K value and *score* in the plot. Then, we can find the optimal K-value from the curve. Here, we give 3 examples of how to select the optimal K-value for 3 ICH projects, presented in Fig. 7. The range of K-value is set from 2 to 30, and the tuning step is 1. The optimal K-value is found in the inflection point on the elbow curve.

Lastly, through extensive empirical experiments, the optimal K-value for each ICH category is obtained, shown in Table 2. Next, we will cluster images of different ICH projects based on the optimal K-value of the corresponding category.

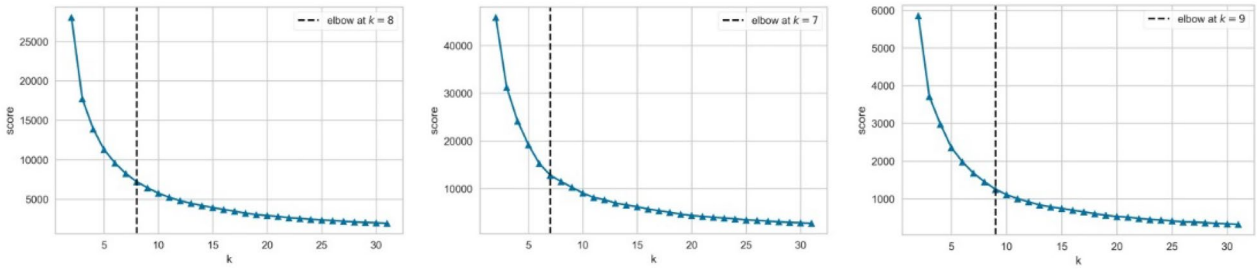

Fig. 7 Examples of selecting the optimal *K*-value through *Elbow* method, which are *Jingdezhen handmade porcelain craftsmanship*, 淮剧 *(Huai opera)*, and 锡伯族贝伦舞 (*Sibe Belem dance*) respectively

Table 2 The optimal *K*-value for each ICH category

ICH category	Optimal *K* value
Traditional sports, games and acrobatics	8
Traditional medicine	8
Traditional opera	10
Traditional skill	11
Traditional art	10
Traditional dance	9
Traditional music	10
Quyi	8
Folk-custom	8
Folk literature	9

Besides, we give instances of the visualization on the clustering results of 3 ICH projects with the optimal *K*-value of the belonged category, shown in Fig. 8. It can be seen that ICH image points are divided into several clusters and there are still some discrete points far away from the clusters. Basically, if an image is related to the belonging cluster, it is supposed to be close to the cluster in visualization. If not, it means that the image point is far away from the cluster and has a long distance from the centroid point of the cluster, and also indicates the image point is not related to the topic of the cluster and is of high possibility to be the noise image. The specific filtration process is introduced in the following.

Removing noise images

To this end, we obtain the clustering results for each ICH project. Next, we need to remove noise images from clustering results. Referring to the noise image removal, a usual idea is that we can train a binary classifier to recognize the noise. Whereas, in our context, if we are desired to construct such a classifier, it means that massive labeled ICH images are needed to support the training of the classifier. But we have millions of images, it is extremely time and labor-consuming to do the annotation. Additionally, noise images only account for a small percentage in all images and it also incurs the problem of unbalanced data and restrains the performance of the model. Therefore, we adopt a clustering-based method, in an unsupervised manner, which is widely used in noise removal, outlier detection etc., and propose the algorithm CNIFA illustrated in Table 3 to remove noise images in each ICH project.

CNIFA is established on the classic and robust *Cluster-Based Local Outlier Factor* (*CBLOF*) method [66], which is applied to help us to calculate the score of each image in the ICH project. Basically, the higher the score is, the more possible the image is noise. The main idea of *CBLOF* can be concluded as following 3 steps: the first

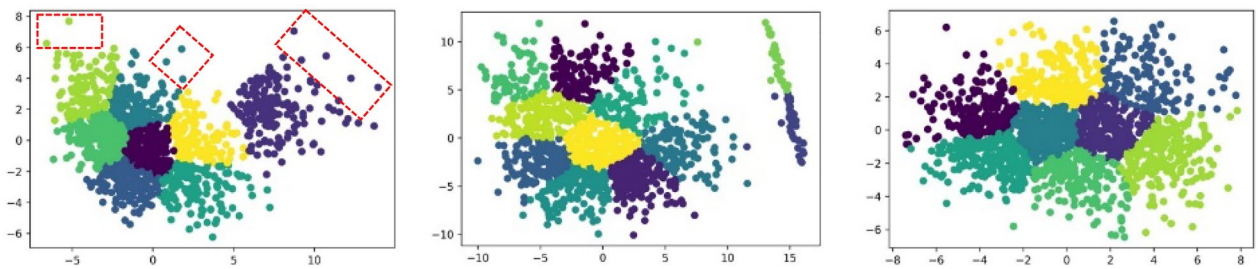

Fig. 8 Visualization of clustering results of 3 ICH projects, 抖空竹 *(Diabolo)*, 陶瓷 *(Ceramic)*, and 丝弦 *(Silk string)*. An example of partial discrete points is shown in the first image, which are inside the red dotted box

Table 3 The algorithm of CNIFA

Algorithm: Noise images filtration algorithm

Input: A dataset $D = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$. *K*: the optimal *K* value after *K-means* clustering, *n*: the number of data, *N*: the number of CICH noise images
Output**:** CICH noise image point set

Extracted representation features $E = \{e_1, e_2, \ldots, e_i, \ldots, e_n\} = Extractor(D)$
Centroid point set $C = \{c_1, c_2, \ldots, c_j, \ldots, c_K\} = Kmeans(E, K)$
For each cluster Q_j in $Q = \{Q_1, Q_2, \ldots, Q_j, \ldots, Q_K\}$ do
    $Radius_{Q_j} = radius(Q_j)$
End for
If the number of points in $Q_j > N$:
    For each point $e_i \in Q_j$ do
        If $distance\ (e_i, c_j) > Radius_{Q_j}$ then
            Add e_i to *S*
        Else
            Continue
    End for
Else
    For each point $e_i \in Q_j$ do
        Add e_i to *S*
    End for
For each point e_i in *S* do
    Calculate *CBLOF* (e_i)
End for
The top-*N* points are the noise images based on the *CBLOF* (e_i) value ranking results

step is to cluster all data points in the dataset through clustering algorithm, e.g., *K*-means; the second step is to divide clusters into *large* and *small clusters* (*LC* and *SC*) based on the number of data points in the cluster; the last step is to calculate the *CBLOF* score for each data point *t*:

$$CBLOF(t) = \begin{cases} |C_i| * min\ (distance\ (t, C_j), where\ t \in C_i, C_i \in SC\ and\ C_j \in LC\ for\ j = 1\ to\ b) \\ |C_i| * (distance\ (t, C_i), where\ t \in C_i\ and\ C_i \in LC) \end{cases} \quad (2)$$

where $|C_i|$, C_j, and b denote the number of data points in the i th cluster, the centroid point of the j th cluster, and the boundary of *LC* and *SC* respectively. Though *CBLOF* algorithm has been proved reliable and strong, it needs huge time if we perform the algorithm on all ICH images because it needs to compute the score of every image in the dataset. Nevertheless, as aforementioned, the scale of noise images is small and the normal images occupy most of images. Hence, it is unnecessary to figure up the score of every image. Based on the idea, CNIFA first executes a preprocessing step to eliminate some normal images to decrease the computation burden. In the process of *K*-means clustering, *K* centroid points will first be determined and then all points will be clustered centered on the centroid points through extensive iterations. Basically, if a data point belongs to a cluster, it should be as close as the corresponding centroid point. Otherwise, it will be classified as the possible noise set. Specifically, we employ *Euclidean Distance* to calculate the distance $dis_{i,j}$ between each point e_i in the cluster Q_j and its centroid point c_j:

$$dis_{i,j} = \sqrt{(x_{c_j} - x_{e_i})^2 + (y_{c_j} - y_{e_i})^2} \quad (3)$$

Then, we can get the average distance $Radius_{Q_j}$ of all distances between data points and centroid points. If the distances between points and centroid points are less than the average distance, we identify them as the normal data and exclude these data points, which are not involved in the next processing, and the left data is added to set *S*. Lastly, we compute the *CBLOF* score of each point in *S*, and obtain the top-*N* ICH noise images, which are discarded. The performance of CNIFA is evaluated in "Evaluation of proposed methods".

Representative images selection for CICHMKG

After obtaining the denoised ICH images through CNIFA, according to the framework in Fig. 3, we need to select representative images for ICH from the clustering results. In the clustering process, several clusters are produced, which indicates the diversity of ICH images. The diversity can be different from angles, a variety of work, or styles of acting. But do they all represent ICH well? Also, if we pick out representative images only from one or two clusters, it may make the presentation of ICH incomplete and is incapable to give a visual overview for the public. Therefore, to resolve that, we split the problem into two parts. The first is to rank the clusters and the second is to select representative ICH images from the ranked clusters.

Guarantee the diversity

The main reasons why we rank the clusters are that on one hand we can guarantee the diversity of selected representative ICH images and on the other hand it can make the selected ICH images more representative if they belong to the top-ranked clusters. Hence, clusters are ranked to decide which one of them is more representative. Top-ranked clusters are reserved to serve as the candidate clusters and low-ranked clusters are thrown away. Here, we utilize the following metrics [21, 23, 67] to rank the clusters:

Number of images It is an intuitive metric that if a cluster include a lot of images, it reveals that what the cluster expresses and presents is an important topic in the ICH project, and can also be regarded as the integral part of the ICH project in vision.

Inter- and intra-cluster distance From the perspective of the space, assuming a scenario where several clusters are overlapped partially, it indicates that what clusters expressing in semantics and topics have a certain of similarity. Hence, we want that the different clusters are far away and the inter-cluster distance is large. Besides, if the points in the cluster are sparse, it also reflects that the cluster is not aggregated well and the topic of the cluster is not consistent. In other words, the shorter the average distance (intra-cluster distance) from different points to the centroid point is, the more representative the cluster is. Concerning the above aspects, the ratio of inter-cluster distance to intra-cluster cluster is utilized as a metric to evaluate the representativeness of clusters.

Then, we normalize the scores with *L1-norm* method. The average score of two normalized values is utilized to rank the clusters. The higher the score is, the more representative the cluster is. Following their studies [22, 52, 67], we select representative images for each ICH project from top-5 clusters based on ranking results.

Select representative images from ranked clusters

Here, we rank each image in each top-ranked cluster to find out how well the image can represent the cluster. Based on the ranking results of images and clusters, representative ICH images are determined.

Intra-cluster distance Each image in the cluster is aggregated with the centroid point as the centre. If the image is far away from the centroid point, it indicates that the image may not related with the topic of the cluster. Hence, the intra-cluster distance is a good reflection for measuring the representativeness of images.

Length of image caption Every image of an ICH project has an attached caption, which is a description for the ICH. Presuming that an image has a longer caption, it reveals that what the image conveys and includes is richer and more specific. Hence, the image with longer caption is counted as an evaluation criterion.

Image quality Additionally, for the benefit of the subsequent CICHMKG visualization and the downstream tasks, selecting images with high quality is reasonable and appropriate, which also can provide the public with a good sense of visual experience. Here, we adopt the BRISQUE [68] as the metric to assess the image quality, which is an efficient algorithm regarding the facets of image resolution, colour, brightness, etc. Basically, the higher the metric value is, the better the image quality is.

Lastly, every metric score is normalized with *L1-norm*. The average score of the sum of the *length of image caption score* and *image quality* minus the *intra-cluster distance* for each ICH image is employed as the final score for ranking. The higher the value is, the more representative the image is. Additionally, in this process, if repeated captions or images are shown in the selected representative images, only one image is reserved through comprehensive crosscheck. As a result, the top 20 images ranked by scores in the top 5 clusters of each ICH project are selected as the representative images. Note that if the number of images in the selected cluster is less than 20, all images are chosen.

Evaluation of proposed methods

To this end, following our proposed construction framework, we finish the construction of CICHMKG. In this process, we propose a noise image filtration algorithm CNIFA and apply several criteria to select representative images for ICH, where the proposed multimodal features composed of the global features (GF) and local features (LF) of images, and textual features of captions are the

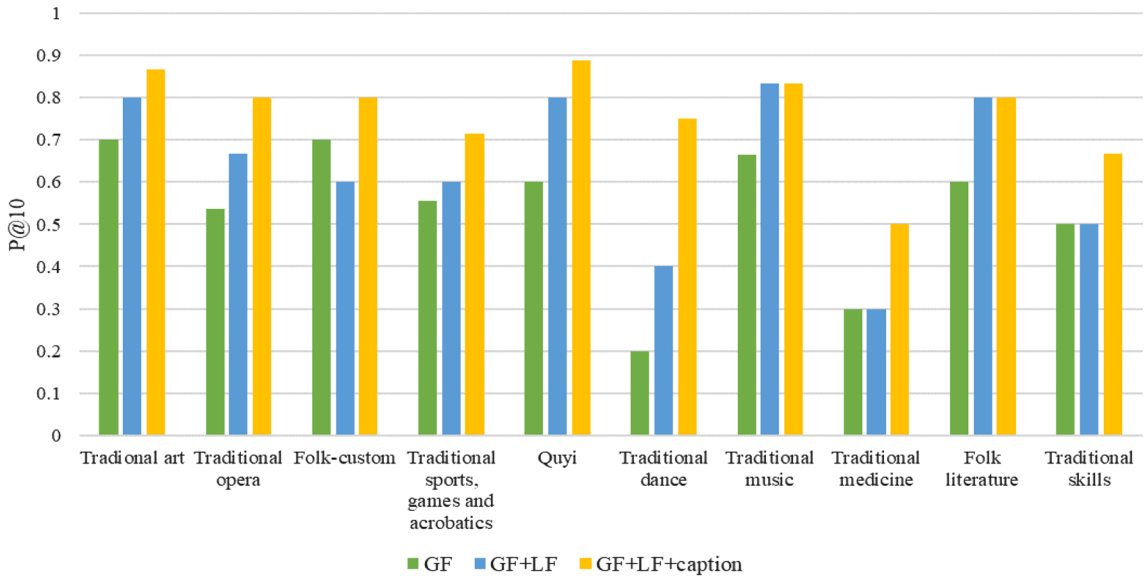

Fig. 9 Experimental results of removing noise images. (GF and LF denote global features and local features of images respectively. Caption denotes the textual features of the caption)

foundation. Nevertheless, the effectiveness of them is not tested. Hence, in this section, we will conduct extensive experiments to assess the performance of the proposed methods and report detailed experimental results.

Results of noise image filtration

To test the performance of CNIFA completely, we conduct experiments on each ICH category. We select one ICH project randomly from each ICH category and the performance of CNIFA on the randomly selected ICH project is taken as the performance of the category. Specifically, following their methods [51, 57, 69], we annotate 200 normal images and 20 noise images for each selected ICH project. These images are annotated by a group consisting of 3 graduates with the knowledge of cultural heritage to ensure the quality of annotation and the annotation result is highly consistent. We adopt the *P@10* as the evaluation metric, which is often utilized in noise filtration, outlier detection, etc. [67, 69].

The fusion of global and local features of images, and textual features of captions is the most important component of CNIFA. It is expected to give a more representative multimodal feature for the ICH image, helping the model to distinguish the noise. To examine the advantage, we apply GF and GF+LF as the representation of the ICH image respectively to remove noise images for comparison. Experimental results are shown in Fig. 9. It can be discovered that CNIFA performs best in almost categories combining deep visual features of GF and LF, and textual features of captions, which demonstrates the superiority of multimodal feature and valid our initial idea. Additionally, we can observe that the integration of GF and LF is superior to GF on the whole, which furtherly indicates the role of salient local regions in ICH images for filtering noise images.

Results of assessing the representativeness

Here, following their study [23, 52, 67], we employ a way of questionary survey to evaluate the performance of our proposed method in selecting representative images for ICH, and select one ICH project from each category randomly and pick up 10 images from the set of representative images in each selected ICH project for evaluation. We reutilize the questions in the papers [23, 67] and modify them to conform to our ICH context. The simplified evaluation questions are listed as follows, concerning the four aspects: (1) *Representative (0–10 scale).* How many images can be considered as the representative for the ICH project? We take the 花灯戏 *(Huadeng Opera)* as an example. The set question is "How many images of the above 10 exhibited images can be seen as the representative images of *Huadeng Opera*?" (2) *Unique (0–10 scale).* How many images of the ICH project are unique without repetition? (3) *Comprehensive (0–5 scale).* Do these images provide a comprehensive view for the ICH project? We take the 京绣 *(Beijing Embroidery)* as an instance. The set question is "Do you think the exhibited images can provide a comprehensive view of *Beijing Embroidery*". (4) *Satisfying (0–5 scale).* Are you satisfied with the images of the ICH project? To facilitate

Table 4 Results of assessing the performance of different methods in selecting representative ICH images

Questions	GF	GF + LF	GF + LF + caption
Representative	8.0	8.3	9.2*
Unique	7.7	8.6	8.6*
Comprehensive	3.4	3.4	3.6*
Satisfying	3.1	3.6	3.8*

Note that * denotes the results of GF + LF + caption are statistically significant over that of GF with *p-value* < 0.05

Table 5 Statistical information of CICHMKG

CICHMKG	Number
ICH entities	3040
ICH images	289,413
Triples	1,774,005

understanding and perception, we also add corresponding description texts to introduce the brief history of each ICH project in the questionary. The presented score for each evaluation question is the corresponding average score based on the results of the questionary survey.

To test the role of multimodal features in selecting representative images, we also select representative images for the selected ICH projects based on GF and GF + LF features respectively, and make the questionaries with the same questions for comparison. The questionary tool *Wenjuanxing* is applied to make the above three kinds of questionaries. 100 voluntary students, including undergraduates and graduates with the knowledge background of cultural heritage, are recruited to fill out the questionaries through the *Wenjuanxing* link of questionaries. In all, we receive 100, 98, and 98 effective questionaries of three kinds (GF, GF + LF, and GF + LF + caption).

Results are presented in Table 4. In term of *representative*, our proposed method performs well and achieves the highest score, which hints that the combination of GF, LF, and captions is more capable to capture salient features of ICH projects, and provides a good summarization for images of ICH projects as a result. Regarding *unique*, though our method obtains the same score as GF + LF, it still shows the ability to avoid the images with similar contents and is better than the single GF. It also indicates that the import of LF and captions can supply more distinguished ICH features to remove similar ICH images. Besides, in the facets of *comprehensive* and *satisfying*, our method provides better results, which also means that not only can the selected images by our method satisfy the requirement of representativeness but also contain diverse images to express ICH comprehensively. It demonstrates the merit of the integration of GF, LF, and caption in providing representative ICH images from massive images furtherly.

Statistics and visualization of CICHMKG

In this section, we report the statistical information of the constructed CICHMKG, illustrated in Table 5. CICHMKG is oriented to the Chinese nation-level ICH list, including 3040 ICH entities after deduplication. Through scraping and organizing textual attribute information in *Baike*, *Toutiao*, and the official website, 37,527 triples are obtained containing abundant textual knowledge. Subsequently, we utilize proposed method to remove noise images and select representative images for ICH entities. As a result, 289,413 ICH representative images are determined. Additionally, to empower downstream image retrieval tasks, we also provide the deep visual features for the images extracted from the pretrained VGG19 and ResNet as the attribute information of images, and basic property information (pixel, height, and width). Finally, in all, 1,774,005 triples are generated and compose CICHMKG.

To help the public dive into ICH deeply and facilitate interactions between the public and CICHMKG, we utilize the *InteractiveGraph* framework [70] to visualize CICHMKG, compared with the *Neo4j* database, which is capable to give a multimodal exhibition. The outline of CICHMKG is presented in Fig. 10(a) where the center is the mark of Chinese ICH, linked blue points are the categories, and points surrounding blue points represent ICH projects.

When we zoom on CICHMKG and click one node, taking *Jingdezhen handmade porcelain craftsmanship* as an example, the linked nodes are shown in Fig. 10(b). When the public are desired to know what the ICH project is in the vision, they can click the node with the image and the corresponding image will be exhibited automatically. Besides, if the public only desire to learn the history of the ICH project and its related textual knowledge, they can click the cmo:ImageOf in the control panel on the upright of Fig. 10(b) to hide images by the interaction. Correspondingly, the textual knowledge of the ICH is shown in Fig. 10(c). Especially, when clicking the ICH project node, the corresponding description text is shown in the light brown frame, introducing the history of the ICH project.

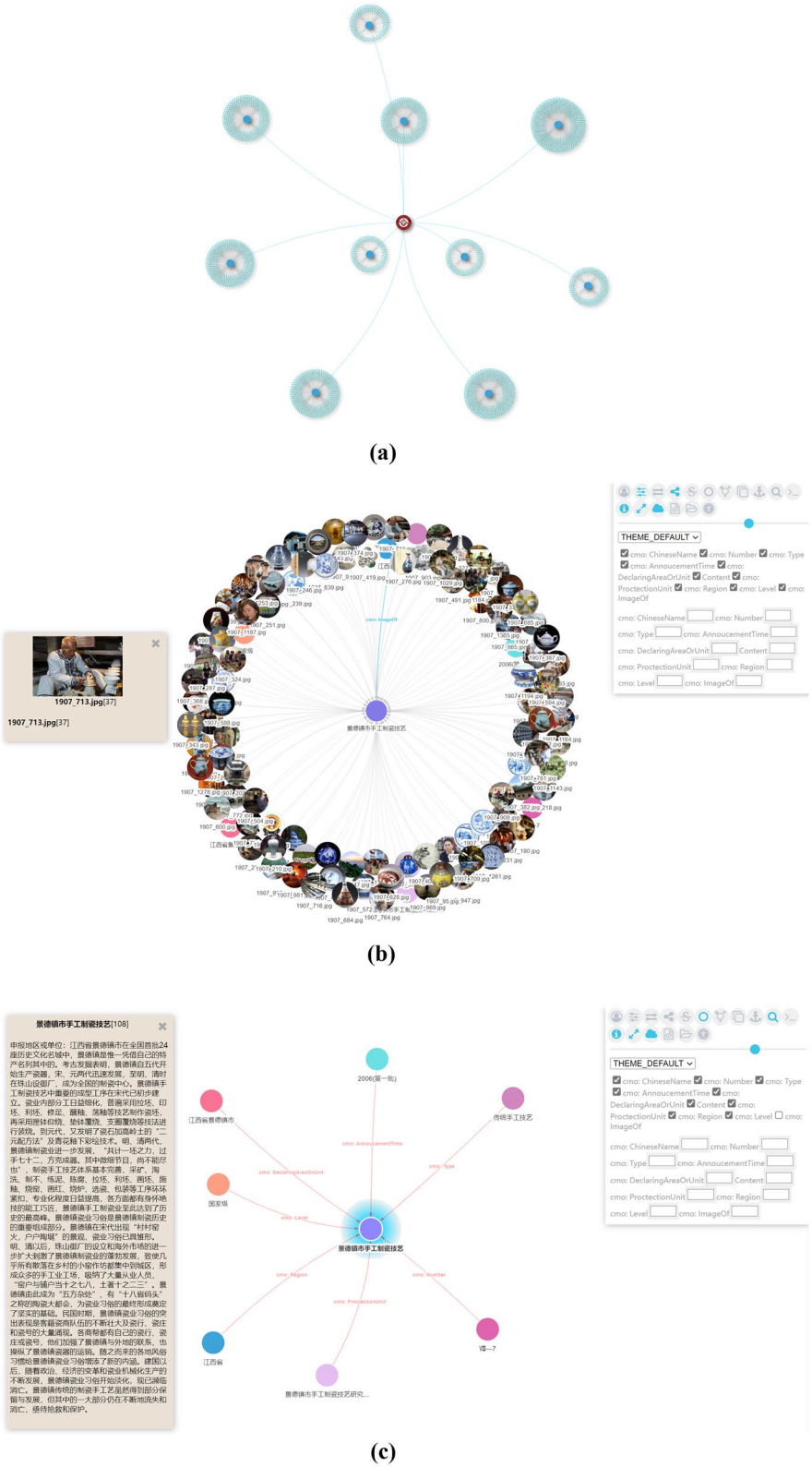

Fig. 10 **a**. The outline of CICHMKG. **b**. The overall exhibition of *Jingdezhen handmade porcelain craftsmanship*. Note that the attribute information of images is not given in the visualization process. **c**. The exhibition of *Jingdezhen handmade porcelain craftsmanship* without images

Theoretical and practical implications

In this section, we will discuss the potential implications of the paper from the aspect of theory and practice respectively.

Theoretical implications

Currently, KG in the cultural heritage domain is mainly text entities-based [15, 26, 39]. However, such traditional KG all suffer to give the public a visual overview due to the lack of image entities and not beneficial for professionals to analyse ICH. Essentially, the reason why we introduce KG is to reduce the data fragmentation and aggregate the related ICH knowledge, supporting the public to obtain complete knowledge and bridging the distance between the public and ICH. But when traditional KG is not capable to establish the connection between the visual perception and abstract symbol, it becomes the stumbling stone rather than the stepping stone for the public to dive into ICH deeper.

Therefore, different from previous studies [10, 13, 18, 24, 26], we take the Chinese nation-level ICH list as an example, and integrate ICH text and image data to construct CICHMKG in a way of from symbols to images. Adopting this way to construct a large-scale and comprehensive MKG in the cultural heritage domain has not been well explored. Also, we provide a framework to construct MKG with the potentiality to guide the MKG construction in different communities and nations.

Practical implications

In practice, we contribute CICHMKG consisting of 1,774,005 triples, providing ample text and image knowledge of Chinese nation-level ICH projects, and explore a viable way about how to utilize and link the multimodal data in cultural heritage domain. Compared with traditional KG, CICHMKG can empower several multimodal tasks effectively, e.g., visual question-answering, multimodal named entity recognition, and multimodal link prediction, releasing the potential of KG furtherly. Apart from that, we visualize CICHMKG and exhibit the interactions between CICHMKG and users, aiming to give users a complete and deep understanding of ICH.

In this paper, we adopt the route from symbols to images to construct CICHMKG given our task context. The most difficult part lies in how to select representative images for ICH entities. Usually, previous studies utilized global features of images or tagged information as the representation but they are insufficient to express images, ignoring salient and complementary global visual features and textual features of captions. Hence, different from previous studies [23, 51, 58], we propose to fuse global and local visual features of images, and textual features of captions into multimodal features as the representation of ICH images, which are the foundation of proposed denoising algorithm CNIFA and subsequent representative image selection. Our extensive empirical experiments in removing noise images and the evaluation of representativeness demonstrate the superiority.

Conclusions

In this paper, aimed at aggerating the disseminated knowledge of ICH and giving the public a visual perception, we propose to, taking the Chinese nation-level ICH list as an example, build the large-scale and comprehensive multimodal knowledge graph CICHMKG integrating enormous text and image entities, and give a practical construction framework. Specifically, we first acquire ICH entities from the official website, which then are employed as the seeds to obtain the textual knowledge from *Baike*, *Toutiao*, etc. Thereafter, massive images are scraped from *Baidu* and *Bing Image*. We propose the CNIFA combining global and local visual features and textual features of captions to remove noise images, and select representative images for ICH entities from denoised images depending on the defined ranking criteria. To test the performance of CNIFA and the representativeness of selected images, empirical experiments demonstrate that our proposed method is promising and effective. Finally, we create CICHMKG consisting of 1,774,005 triples and visualize CICHMKG to promote the interactions between the public and CICHMKG.

Though we make progress in utilizing MKG to dive into ICH, there are still rooms to be improved. The general ability of CNIFA needs to be enhanced furtherly and tested in the dataset from different cultural communities. Also, the visual features in ICH images need to be excavated furtherly, and connections between ICH images and attached captions are deserved to be explored deeply.

Acknowledgements

The authors are appreciated for anonymous reviewers' insightful comments.

Author contributions

TF: conceptualization, methodology, experiments, prepare original manuscript, revision, funding acquisition; HW: supervision, methodology, revision, funding acquisition; TH: supervision, methodology, revision. All authors read and approved the final manuscript.

Funding

This paper is supported by the National Natural Science Foundation of China (No. 72074108), Special Project of Nanjing University Liberal Arts Youth Interdisciplinary Team (010814370113), Jiangsu Young Social Science Talents, "Tang Scholar" of Nanjing University, and China Scholarship Council (202206190149).

Availability of data and materials

Data was available on request from the authors.

Declarations

Competing interests
The authors declare no competing interests.

Received: 7 January 2023 Accepted: 4 April 2023
Published online: 23 May 2023

References

1. Giglitto D, Ciolfi L, Bosswick W. Building a bridge: opportunities and challenges for intangible cultural heritage at the intersection of institutions, civic society, and migrant communities. Int J Herit Stud. 2022;28:74–91. https://doi.org/10.1080/13527258.2021.1922934.
2. Hou Y, Kenderdine S, Picca D, Egloff M, Adamou A. Digitizing intangible cultural heritage embodied: state of the art. J Comput Cult Herit. 2022. https://doi.org/10.1145/3494837.
3. Lenzerini F. Intangible cultural heritage: the living culture of peoples. Eur J Int Law. 2011;22:101–20. https://doi.org/10.1093/ejil/chr006.
4. Lu Z, Annett M, Fan M, Wigdor D. "I feel it is my responsibility to stream": Streaming and Engaging with Intangible Cultural Heritage through Livestreaming. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: Association for Computing Machinery; 2019, p. 1–14. https://doi.org/10.1145/3290605.3300459.
5. Vecco M. A definition of cultural heritage: From the tangible to the intangible. J Cult Herit. 2010;11:321–4. https://doi.org/10.1016/j.culher.2010.01.006.
6. Carrillo Yap SL. The role of the UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage (ICH Convention) in the protection of traditional forest-related knowledge (TFRK) of Amazonian indigenous peoples. Int J Human Rights. 2021;25:853–69. https://doi.org/10.1080/13642987.2020.1859481.
7. Lázaro Ortiz S, Jiménez de Madariaga C. The UNESCO convention for the safeguarding of the intangible cultural heritage: a critical analysis. Int J Cult Policy. 2022;28:327–41. https://doi.org/10.1080/10286632.2021.1941914.
8. Melis C, Chambers D. The construction of intangible cultural heritage: a Foucauldian critique. Ann Tourism Res. 2021;89:103206. https://doi.org/10.1016/j.annals.2021.103206.
9. Do T-N, Pham T-P, Pham N-K, Nguyen H-H, Tabia K, Benferhat S. Stacking of SVMs for Classifying Intangible Cultural Heritage Images. In: Le Thi HA, Le HM, Pham Dinh T, Nguyen NT, editors. Advanced Computational Methods for Knowledge Engineering. Cham: Springer International Publishing; 2020. p. 186–96. https://doi.org/10.1007/978-3-030-38364-0_17.
10. Dou J, Qin J, Jin Z, Li Z. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage. J Vis Lang Comput. 2018;48:19–28. https://doi.org/10.1016/j.jvlc.2018.06.005.
11. Skublewska-Paszkowska M, Milosz M, Powroznik P, Lukasik E. 3D technologies for intangible cultural heritage preservation—literature review for selected databases. Heritage Sci. 2022;10:3. https://doi.org/10.1186/s40494-021-00633-x.
12. Zhao H. The database construction of intangible cultural heritage based on artificial intelligence. Mathemat Prob Eng. 2022;2022:e8576002. https://doi.org/10.1155/2022/8576002.
13. Castellano G, Digeno V, Sansaro G, Vessio G. Leveraging knowledge graphs and deep learning for automatic art analysis. Knowledge-Based Syst. 2022;248:108859. https://doi.org/10.1016/j.knosys.2022.108859.
14. Chen X, Xie H, Li Z, Cheng G. Topic analysis and development in knowledge graph research: a bibliometric review on three decades. Neurocomputing. 2021;461:497–515. https://doi.org/10.1016/j.neucom.2021.02.098.
15. Kalita D, Deka D. Ontology for preserving the knowledge base of traditional dances (OTD). Electron Libr. 2020;38:785–803. https://doi.org/10.1108/EL-11-2019-0258.
16. Liu S, Tan N, Yang H, Lukač N. An intelligent question answering system of the liao dynasty based on knowledge graph. Int J Comput Intell Syst. 2021;14:170. https://doi.org/10.1007/s44196-021-00010-3.
17. Wang X, Chang W, Tan X. Representing and linking dunhuang cultural heritage information resources using knowledge graph. KO. 2020;47:604–15. https://doi.org/10.5771/0943-7444-2020-7-604.
18. Carriero VA, Gangemi A, Mancinelli ML, Nuzzolese AG, Presutti V, Veninata C. Pattern-based design applied to cultural heritage knowledge graphs. Semantic Web. 2021;12:313–57. https://doi.org/10.3233/SW-200422.
19. Kannan AV, Fradkin D, Akrotirianakis I, Kulahcioglu T, Canedo A, Roy A, et al. Multimodal Knowledge Graph for Deep Learning Papers and Code. Proceedings of the 29th ACM International Conference on Information & Knowledge Management, New York, NY, USA: Association for Computing Machinery; 2020, p. 3417–20. https://doi.org/10.1145/3340531.3417439.
20. Li N, Shen Q, Song R, Chi Y, Xu H. MEduKG: a deep-learning-based approach for multi-modal educational knowledge graph construction. Information. 2022;13:91. https://doi.org/10.3390/info13020091.
21. Zhu X, Li Z, Wang X, Jiang X, Sun P, Wang X, et al. Multi-Modal knowledge graph construction and application: a survey. arXiv. 2022. https://doi.org/10.1109/TKDE.2022.3224228.
22. Wang M, Wang H, Qi G, Zheng Q. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. Big Data Res. 2020. https://doi.org/10.1016/j.bdr.2020.100159.
23. Han S, Ren F, Du Q, Gui D. Extracting representative images of tourist attractions from flickr by combining an improved cluster method and multiple deep learning models. ISPRS Int J Geo Inf. 2020;9:81. https://doi.org/10.3390/ijgi9020081.
24. Faralli S, Lenzi A, Velardi P. A Large Interlinked Knowledge Graph of the Italian Cultural Heritage. Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France: European Language Resources Association; 2022, p. 6280–9.
25. Fan T, Wang H. Multimodal sentiment analysis of intangible cultural heritage songs with strengthened audio features-guided attention. J Inform Sci. 2022. https://doi.org/10.1177/01655515221114454.
26. Debruyne C, Munnelly G, Kilgallon L, O'Sullivan D, Crooks P. Creating a knowledge graph for ireland's lost history: knowledge engineering and curation in the beyond 2022 project. J Comput Cult Herit. 2022;1:25–25. https://doi.org/10.1145/3474829.
27. Beretta F. A challenge for historical research: making data FAIR using a collaborative ontology management environment (OntoME). Semantic Web. 2021;12:279–94. https://doi.org/10.3233/SW-200416.
28. Buranasing W, Lilakiataskun W. Semantic relation extraction from cultural heritage archives. J Web Eng. 2022. https://doi.org/10.1305/jwe1540-9589.2145.
29. Hyvönen E, Rantala H. Knowledge-based Relation Discovery in Cultural Heritage Knowledge Graphs: Digital Humanities in the Nordic Countries. Digital Humanities in Nordic Countries 2019:230–9.
30. Dimitropoulos K, Tsalakanidou F, Nikolopoulos S, Kompatsiaris I, Grammalidis N, Manitsaris S, et al. A Multimodal approach for the safeguarding and transmission of intangible cultural heritage: the case of i-treasures. IEEE Intell Syst. 2018;33:3–16. https://doi.org/10.1109/MIS.2018.111144858.
31. Montalvo M, Calle-Ortiz E, Chica J. A Multimodal Robot Based Model for the Preservation of Intangible Cultural Heritage. Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA: Association for Computing Machinery; 2017, p. 213–4. https://doi.org/10.1145/3029798.3038315.
32. Qiu Q, Zhang M. Using content analysis to probe the cognitive image of intangible cultural heritage tourism: an exploration of Chinese social media. ISPRS Int J Geo Inf. 2021;10:240. https://doi.org/10.3390/ijgi10040240.
33. Hyvönen E. Digital humanities on the semantic web: sampo model and portal series. Semantic Web. 2022. https://doi.org/10.3233/SW-223034.
34. Hyvönen E, Leskinen P, Heino E, Tuominen J, Sirola L. Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web. In: Gracia J, Bond F, McCrae JP, Buitelaar P, Chiarcos C, Hellmann S, editors. Language, Data, and Knowledge. Cham: Springer International Publishing; 2017. p. 113–9. https://doi.org/10.1007/978-3-319-59888-8_9.
35. Hitzler P, Janowicz K, Hyvönen E. Using the Semantic Web in digital humanities: shift from data publishing to data-analysis and serendipitous knowledge discovery. Semant Web. 2020;11:187–93. https://doi.org/10.3233/SW-190386.

36. Schleider T, Troncy R, Ehrhart T, Dorozynski M, Rottensteiner F, Sebastián Lozano J, et al. Searching Silk Fabrics by Images Leveraging on Knowledge Graph and Domain Expert Rules. Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents, New York, NY, USA: Association for Computing Machinery; 2021, p. 41–9. https://doi.org/10.1145/3475720.3484445.
37. Puren M, Vernus P. Conceptual Modelling of the European Silk Heritage with the SILKNOW Data Model and Extension 2022.
38. Carboni N, de Luca L. Towards a semantic documentation of heritage objects through visual and iconographical representations. Int Inform Library Rev. 2017;49:207–17. https://doi.org/10.1080/10572317.2017.1353374.
39. Fan T, Wang H. Research of Chinese intangible cultural heritage knowledge graph construction and attribute value extraction with graph attention network. Inform Proc Manage. 2022;59:102753. https://doi.org/10.1016/j.ipm.2021.102753.
40. Ghorbanali A, Sohrabi MK, Yaghmaee F. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. Inform Proc Manage. 2022;59:102929. https://doi.org/10.1016/j.ipm.2022.102929.
41. Jain DK, Boyapati P, Venkatesh J, Prakash M. An intelligent cognitive-inspired computing with big data analytics framework for sentiment analysis and classification. Inform Proc Manage. 2022;59:102758. https://doi.org/10.1016/j.ipm.2021.102758.
42. Mai C, Liu J, Qiu M, Luo K, Peng Z, Yuan C, et al. Pronounce differently, mean differently: a multi-tagging-scheme learning method for Chinese NER integrated with lexicon and phonetic features. Inf Process Manage. 2022;59:103041. https://doi.org/10.1016/j.ipm.2022.103041.
43. Yang L, Na J-C, Yu J. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. Inform Proc Manage. 2022;59:103038. https://doi.org/10.1016/j.ipm.2022.103038.
44. Yu J, Jiang J, Yang L, Xia R. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics; 2020, p. 3342–52. https://doi.org/10.18653/v1/2020.acl-main.306.
45. Ma J, Li D, Zhu H, Li C, Zhang Q, Qiao Y. GAFM: a knowledge graph completion method based on graph attention faded mechanism. Inform Proc Manage. 2022;59:103004. https://doi.org/10.1016/j.ipm.2022.103004.
46. Nguyen H-V, Gelli F, Poria S. DOZEN: Cross-Domain Zero Shot Named Entity Recognition with Knowledge Graph. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: Association for Computing Machinery; 2021, p. 1642–6. https://doi.org/10.1145/3404835.3463113.
47. Zhao A, Yu Y. Knowledge-enabled BERT for aspect-based sentiment analysis. Knowledge-Based Syst. 2021;227:107220. https://doi.org/10.1016/j.knosys.2021.107220.
48. Min W, Liu C, Xu L, Jiang S. Applications of knowledge graphs for food science and industry. Patterns. 2022;3:100484. https://doi.org/10.1016/j.patter.2022.100484.
49. Wilcke WX, Bloem P, de Boer V, van Veer RH, van Harmelen FAH. End-to-End entity classification on multimodal knowledge graphs. arXiv. 2020. https://doi.org/10.48550/arXiv.2003.12383.
50. Roy A, Akrotirianakis I, Kannan AV, Fradkin D, Canedo A, Koneripalli K, et al. Diag2graph: Representing Deep Learning Diagrams In Research Papers As Knowledge Graphs. 2020 IEEE International Conference on Image Processing (ICIP), 2020, p. 2581–5. https://doi.org/10.1109/ICIP40778.2020.9191234
51. Deng Y, Tang F, Dong W, Wu F, Deussen O, Xu C. Selective clustering for representative paintings selection. Multimed Tools Appl. 2019;78:19305–23. https://doi.org/10.1007/s11042-019-7271-7.
52. Jiang S, Qian X, Xue Y, Li F, Hou X. Generating representative images for landmark by discovering high frequency shooting locations from community-contributed photos. 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2013, p. 1–6. https://doi.org/10.1109/ICMEW.2013.6618374
53. Mai S, Zeng Y, Zheng S, Hu H. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Trans Affective Comp. 2022. https://doi.org/10.1109/TAFFC.2022.3172360.
54. Ferrada S, Bustos B, Hogan A, et al. IMGpedia: A Linked Dataset with Content-Based Analysis of Wikimedia Images. In: d'Amato C, Fernandez M, Tamma V, Lecue F, Cudré-Mauroux P, Sequeda J, et al., editors. The Semantic Web – ISWC 2017. Cham: Springer International Publishing; 2017. p. 84–93. https://doi.org/10.1007/978-3-319-68204-4_8.
55. Liu S, Yang H, Li J, Kolmanič S. Preliminary study on the knowledge graph construction of Chinese ancient history and culture. Information. 2020;11:186. https://doi.org/10.3390/info11040186.
56. Tan N, Anwar S, Jiang W. Intangible cultural heritage listing and tourism growth in China. J Tourism Cult Change. 2022. https://doi.org/10.1080/14766825.2022.2068373.
57. Lei D, Zhu Q, Chen J, Lin H, Yang P. Automatic K-Means Clustering Algorithm for Outlier Detection. In: Zhu R, Ma Y, editors. Information Engineering and Applications. London: Springer; 2012.
58. Zeng J, Wang J, Guo L, Fan G, Zhang K, Gui G. Cell scene division and visualization based on autoencoder and K-Means algorithm. IEEE Access. 2019;7:165217–25. https://doi.org/10.1109/ACCESS.2019.2953184.
59. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ICLR 2015.
60. Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, p. 248–55. https://doi.org/10.1109/CVPR.2009.5206848.
61. Jiang P, Ergu D, Liu F, Cai Y, Ma B. A review of Yolo Algorithm developments. Procedia Computer Sci. 2022;199:1066–73. https://doi.org/10.1016/j.procs.2022.01.135.
62. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors., et al., Computer Vision – ECCV 2014. Cham: Springer International Publishing; 2014. p. 740–55. https://doi.org/10.1007/978-3-319-10602-1_48.
63. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics; 2019, p. 4171–86. https://doi.org/10.18653/v1/N19-1423.
64. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. J Roy Stat Soc: Ser C. 1979;28:100–8. https://doi.org/10.2307/2346830.
65. Liu F, Deng Y. Determine the number of unknown targets in open world based on elbow method. IEEE Trans Fuzzy Syst. 2021;29:986–95. https://doi.org/10.1109/TFUZZ.2020.2966182.
66. He Z, Xu X, Deng S. Discovering cluster-based local outliers. Pattern Recogn Lett. 2003;24:1641–50. https://doi.org/10.1016/S0167-8655(03)00003-5.
67. Kennedy LS, Naaman M. Generating diverse and representative image search results for landmarks. Proceedings of the 17th international conference on World Wide Web, New York, NY, USA: Association for Computing Machinery; 2008, p. 297–306. https://doi.org/10.1145/1367497.1367539.
68. Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain. IEEE Trans Image Process. 2012;21:4695–708. https://doi.org/10.1109/TIP.2012.2214050.
69. Pamula R, Deka JK, Nandi S. An Outlier Detection Method Based on Clustering. 2011 Second International Conference on Emerging Applications of Information Technology, 2011, p. 253–6. https://doi.org/10.1109/EAIT.2011.25.
70. Zhao Z, Shen Z. An interactive analysis framework for multivariate heterogeneous graph data management system. Data Anal Knowledge Discovery. 2019;3:37–46. https://doi.org/10.11925/infotech.2096-3467.2019.0252.

Publisher's Note