IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 November 2022, accepted 28 November 2022, date of publication 1 December 2022,
date of current version 8 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3226070

RESEARCH ARTICLE

CulturAI: Semantic Enrichment of Cultural Data Leveraging Artificial Intelligence

SALVATORE M. CARTA[1], SERGIO CONSOLI[2], ALESSANDRO GIULIANI[1], ALESSANDRO SEBASTIAN PODDA[1], AND DIEGO REFORGIATO RECUPERO[1]
[1]Department of Mathematics and Computer Science, University of Cagliari, 09121 Cagliari, Italy
[2]European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Corresponding author: Sergio Consoli (sergio.consoli@ec.europa.eu)

ABSTRACT In this paper, we propose an innovative tool able to enrich cultural and creative spots (*gems*, hereinafter) extracted from the European Commission *Cultural Gems* portal, by suggesting relevant keywords (*tags*) and YouTube videos (represented with proper *thumbnails*). On the one hand, the system queries the YouTube search portal, selects the videos most related to the given *gem*, and extracts a set of meaningful thumbnails for each video. On the other hand, each tag is selected by identifying semantically related popular search queries (i.e., trends). In particular, trends are retrieved by querying the Google Trends platform. A further novelty is that our system suggests contents in a dynamic way. Indeed, as for both YouTube and Google Trends platforms the results of a given query include the most popular videos/trends, such that a *gem* may constantly be updated with trendy content by periodically running the tool. The system has been tested on a set of *gems* and evaluated with the support of human annotators. The results highlighted the effectiveness of our proposal.

INDEX TERMS Computer science in cultural heritage, heterogeneous data analysis, modeling, interlinking, and browsing, semantic-aware representation of cultural data, machine learning, social media.

I. INTRODUCTION

Cultural heritage is an expression of the ways of living developed by a community and passed on from generation to generation, including customs, practices, places, objects, artistic expressions, and values. As a source of identity, heritage is a valuable factor for empowering local communities and enabling vulnerable groups to fully participate in social and cultural life. It can also provide time-tested solutions for conflict prevention and reconciliation.

Cultural Gems (CG)[1] is a free and open source web platform, crowdsourced, conceived by the European Commission's Joint Research Centre (DG JRC)[2] to map cultural and creative venues in European cities. The main purpose is to capture diversity in culture and creativity among European cities and towns, creating community-led maps and a common repository for European cultural and creative places. CG includes data on selected cultural venues from OpenStreetMap,[3] and information provided by European cities, universities, and other public and private organizations, allowing users to share and visualize information on city maps. The application is tailored for local authorities and for people working (or interested) in the cultural and creative sectors, providing a digital tool and references to promote culture and creativity in their cities [17].

How many stories can describe culture in a European city? Which places were the stage for hidden secrets or well-known events? Each corner in a city can tell something about the local culture. CG is aimed at addressing this kind of questions, supporting users in collecting and visualizing European stories on culture and intangible heritage [1]. Local authorities, universities, and people interested in the topic can

The associate editor coordinating the review of this manuscript and approving it for publication was Varun Gupta.

[1] Cultural Gems application: https://culturalgems.jrc.ec.europa.eu/

[2] https://joint-research-centre.ec.europa.eu/index_en

[3] https://www.openstreetmap.org/

contribute and visualize information on cultural and creative places in their cities.

CG was launched in December 2018 as a legacy of the European Year of Cultural Heritage.[4] Currently, the application contains information on more than 130,000 cultural and creative places in over 300 European cities and towns. Since its launch, CG has been continuously evolving to meet users' needs. In 2022, the platform is changing data infrastructure to improve data interoperability and information accessibility. Visualization of CG maps will help users to better explore data and guide them to find city stories. A technology upgrade for large-scale implementation of cultural heritage data is currently ongoing to allow a proper management of cultural properties, events, stories, geo-locations (e.g., points, lines, or polygons), time (one-time, recurrent), intangibles, etc.

In particular, we discuss here our currently ongoing developments and future directions, focusing the attention on those improvements aimed at enhancing the interoperability of the application by designing an innovative tool able to enrich a given *gem* by linking it to relevant social media information. For such a purpose, we leverage the enormous potential of Artificial Intelligence for cultural heritage (see, for example, the recent works by Fiorucci et al. [10] and Díaz-Rodríguez and Pisoni [8]), and in particular, in this work, we aim to better capture user attention by associating relevant keywords (*tags*) and YouTube videos (represented with proper *thumbnails*) to the *gem*. In particular, the system is able to select the videos that are most related to the given *gem* from the YouTube portal and extracts a set of meaningful thumbnails for each video. Subsequently, the tags are selected by identifying semantically related popular search queries (i.e., trends). A significant novelty is that our system suggests contents in a "dynamic" way. Indeed, as the system retrieves the most popular videos/trends, a *gem* may constantly be "refreshed" with trendy content by periodically running the tool.

Our main motivation consists in improving the attractiveness of CG to users by increasing the interoperability of the application and including additional social media content dynamically. Given that the application is crowdsourced, the enrichment with further appealing material brings benefits in terms of users interest and interaction experience [15]. Accordingly, a *gem* should be able to attract and captivate the users (intended here as the visitors of the portal) from different perspectives. In particular, the users play a crucial role, as they are the final beneficiaries of the information the *gem* provides. With a proper *gem* enhancement, the final user may be involved in tuning the locals' stories, discovering hidden cultural treasures, or picking up cultural vibes. The problem of how to properly enhance a given *gem* is an open challenge. For this purpose, we devised an innovative tool that helps the application overcome the aforementioned *gem* enhancement problem.

We named the proposed model as *CulturAI*. Let us clarify that it partially exploits, and widely extends, a small set of functionalities from a previous work of ours, i.e., *VSTAR* (Visual Semantic Thumbnails and tAgs Revitalization), devised for generating suitable thumbnails and tags [4].

Therefore, the main innovations of our paper are the following:

- we contribute to the body of knowledge by providing a digital tool to the European Commission to map cultural and creative venues in European cities with the goal of capturing diversity in culture and creativity among European cities and towns;
- we leverage Artificial Intelligence technologies by associating relevant tags and YouTube videos and thumbnails to the gem;
- we carry out an evaluation study by exploiting Amazon Mechanical Turk where we assigned tags and thumbnails produced for a set of videos to different workers who labeled them with a proper relevance score;
- we present athe use case related to a historical building complex in Vienna where we show the information that has been extracted (description, figures, videos, tags) and indicate step by step how our pipeline of modules is processed;
- our proposed system is dynamic in the sense that it retrieves the most popular videos/trends and, therefore, a gem can be constantly refreshed with trendy content by periodically running the tool.

The remainder of this paper is organized as follows. Section II discusses the related work. Section III illustrates the contributions of our work. Section IV describes the CG platform and the methodology we have followed. Section V describes the architecture of the system and how it extends *VSTAR*. In Section VI, we included an example of the execution of the system, including all the intermediate outcomes of each module. Section VII presents the evaluation process we have carried out, whereas Section VIII contains the results we have obtained along with a discussion on them. A use case of our approach showing the *Lower Belvedere* gem has been included in Section IX. Finally, Section X ends the paper with concluding remarks and future directions where we are headed.

II. RELATED WORK

The scientific literature shows numerous studies in the area of content enrichment for entertainment and tourism, and, among them, there are several whose focus is on data concerning cultural heritage. Of these, the prominent focus in scholarly research is to automatically generate information that can enhance, or sometimes customize, the user experience with artistic and cultural items in the context of exhibitions, museums, guided tours, but also, more recently, digital resources such as web portals or mobile applications.

Where pioneering work in the area of semantic enrichment of cultural content, such as those by Mäkelä et al. [13] and by Oh and Moon [18], has emphasized the importance and

[4] https://culture.ec.europa.eu/cultural-heritage/eu-policy-for-cultural-heritage/european-year-of-cultural-heritage-2018

type of features to be considered in the development of effective interfaces to the transfer of generated information, the contemporary studies, driven by the emergence of new Artificial Intelligence methods, particularly those based on (Deep) Machine Learning and Natural Language Processing techniques, have turned toward proposing innovative approaches for generating (or enriching) the content that would populate the aforementioned interfaces.

More in general, other authors have analyzed how the cultural heritage domain has changed due to the effects of technology (i.e., Internet, Virtual and Augmented Reality, Semantic Web, Content Digitization, Machine Learning, Big Data, Social Media) [21]. In particular, they have examined how culture is accessed, presented, recorded, and spread to a broader audience. Furthermore, they have depicted some scenarios where organizations surrounding the cultural heritage domain should adapt their processes.

Mudge et al. [17] have exploited the technology to represent digital twins of cultural objects. There are several advantages: first of all, the digital representation of the real world gets rid of physical obstacles to scholarly and public access, and encourages knowledge and entertainment of cultural archives. Moreover, the conservation of digital contents is perpetual and can be accessed by multiple users without deterioration.

As far as specific technologies are concerned, Fiorucci et al. [10] provides a survey of machine learning techniques applied within the cultural heritage domain. The authors describe papers about classical classification and regression techniques having useful applications in conservation efforts, such as historical building integrity prediction or in recognition of iconographic elements in artworks. They also mention how Deep Learning models have been employed for transfer learning; these approaches are applied in the presence of small amounts of labelled data, a common issue in cultural heritage, and usually applied for digital artwork classification.

One more survey about human-centered artificial intelligence for designing accessible cultural heritage is represented by [20]. This paper discusses state-of-the-art technologies to design and deliver accessible museum and cultural heritage sites experiences. The focus is on the heterogeneity of the users with different areas of expertise and, therefore, need artificial intelligence approaches to improve their accessibility. Furthermore, the authors design a conceptual framework with an interaction of key elements forming museum and cultural heritage online experiences.

Remaining within the artificial intelligence area, authors in [8] discuss some challenges and research questions to be addressed by the latest explainable artificial intelligence (XAI) models. Fairness, accountability, and transparency in machine learning are the first topics which specific challenges defined by the authors revolve around. For example, one challenge targets the ways artificial intelligence can help accessibility to audiovisuals.

Furthermore, Egarter Vigl et al. [9] have examined a user-friendly artificial intelligence-based approach to infer visual-sensory landscape values from Flickr. They show that, with a mixture of text mining and computer vision techniques, and by exploiting the semantic content of about 640,000 artificially generated tags of photos taken in the UNESCO world heritage site: "The Dolomites" (Italy), it is possible to relate photographers' preferences in capturing landscape elements to a set of cultural ecosystem services with high accuracy. Moreover, the authors demonstrate that geographic information in the data can be used to i) link preferences to different natural and human variables and ii) leverage to forecast cultural ecosystem services patterns.

Differently from the past works in literature, we exploit Semantic Web, Artificial Intelligence, and Social Media techniques to propose an innovative tool with the goal of enriching a given *gem* by suggesting relevant keywords and YouTube videos. Videos are related to the given *gem*, while the associated keywords are also trendy, that is, retrieved by querying the Google Trends platform. Moreover, our system is dynamic in the sense that it can be periodically deployed to constantly refresh a *gem* with trendy content.

For the sake of clarity, the aforementioned related works are summarized in Table 1.

III. RESEARCH AIM

The main goals of the research work described in this contribution are listed in the following:

- Increase user interoperability and attractiveness of CG by data augmentation;
- Automatic selection of relevant social media content for cultural data;
- Identification of the most related YouTube videos for each cultural *gem*;
- Optimized extraction of the set of summarizing and meaningful *thumbnails* for each identified video;
- Semantic keywords retrieval from Google Trends for cultural data, and association of the most popular *tags* to the *gems*;
- Enrichment of the CG data with the identified relevant *tags* and *thumbnails*;
- Dynamic maintenance and update of the most popular videos and trends for each of the *gems* in the application.

IV. MATERIALS AND METHODS

A. METHODOLOGY

As described in Section I, our model, named *CulturAI*, extends a small set of our previous work (VSTAR), which has been devised for generating suitable thumbnails and tags [4]. As the original VSTAR model cannot be highly effective in the specific cultural heritage domain, *CulturAI* has been devised to provide high effectiveness in the specific domain. In so doing, *CulturAI* identifies relevant YouTube videos and combines visual and semantic information for suggesting,

TABLE 1. Related work summary.

Reference	Main Topic
[13]	Development of an expressive exhibition generation interface with multiple visualizations based on combining narrative query patterns for forming exhibitions with the concept of domain-centric view-based search.
[18]	Proposal of 3-D integrated design principles of modeling, design, and system for the design of cultural user interface generation reflecting the users potential culture models.
[21]	Study on the role of technology, especially for big data, with a discussion on the advances that led to paradigm shifts in the research area of cultural informatics.
[17]	Study and discussion on emerging digital technologies for cultural heritage.
[10]	Survey of machine learning techniques applied within the cultural heritage domain, focusing on classical classification and regression techniques and Deep Learning models.
[20]	Survey on technologies aimed at designing and delivering accessible museum and cultural heritage sites experiences.
[8]	Discussion on challenges and research questions to be addressed by the latest explainable artificial intelligence (XAI) models for cultural heritage.
[9]	Investigation on a user-friendly artificial intelligence (AI)-based approach for inferring visual-sensory landscape values from Flickr data, combining computer vision with text mining.

at the same time, both thumbnails and tags. It exploits image captioning to extract semantic information from visual features, which permits the identification of relevant frames and related popular topics (trends), the former being suggested as thumbnails, whereas the latter being selected as new tags.

The proposed model presents several advantages in real-world applications:

- (i) *CulturAI* can suggest thumbnails and tags simultaneously, basing on a unique captioning algorithm and, consequently, reducing the computational resources. To the best of our knowledge, no state-of-the-art models suggest both tags and thumbnails relying on a sole algorithm;
- (ii) as the tags are suggested considering the latest relevant popular cultural heritage topics, unlike most of the state-of-the-art tools (e.g., the proposal by Jin et al. [11]), our proposal can constantly revitalize the associated tags;
- (iii) the option, for the final user, of setting a trade-off between quality and quantity of suggested items. Unlike the current tools (e.g., the work of Patwardhan et . [19]), the final user can decide if it would be worth enriching a *gem* with many tags/thumbnails (with the risk of assigning less relevant items) or preferring to select only relevant tags/thumbnails having a confidence score higher than a certain threshold, assuming the risk of not retrieving any items (if there are no items with the confidence score higher than the fixed threshold).

B. ALGORITHMS

In the following, we report the main algorithms and methods adopted for processing the gem and the video metadata.

1) WORD EMBEDDING

As the system aims to select semantically relevant trends and frames, all textual elements are projected in a proper vector space to represent the semantics of the videos and the given *gem*. In doing so, to project the textual data into a suitable embedding space, we adopt the Sentence Transformer [23], a framework based on BERT (Bidirectional Encoder Representations from Transformers). BERT is a well-known Natural Language Processing (NLP) model developed by Google [7]. The choice of this model is motivated by several key points: first, BERT is highly suitable for task-specific models. Indeed, it has been pre-trained on a large corpus, making it easier for smaller, more defined tasks. Metrics can be fine-tuned and used immediately. Moreover, it is frequently updated and available in more than 100 languages, which is a significant strength in cultural and tourist-oriented services like CG. Let us point out that BERT generates *contextual embeddings*, meaning that the input of the embedding model should be a sentence rather than a single word. In our model, although most textual data is represented by entire sentences, it also relies on identifying relevant trends, which usually are represented as single tokens (single terms or short n-grams); in this case, a context-independent model (e.g., Word2Vec [14]) could seem more appropriate. Nevertheless, the main weakness is that such models usually do not address out-of-vocabulary (OOV) words, meaning that they can compute embedding vectors only for words included in the training vocabulary. Conversely, a trend might have a high probability of not being included in the usual model vocabularies, with the consequent loss of meaningful information. BERT is a more suitable choice, as it is not limited to the vocabulary space [25]. Indeed, it supports OOV words, generating a vector representation for any arbitrary word. Each step regarding textual data is performed in the same BERT-based embedded space.

2) COSINE SIMILARITY

In a word embedding scenario, the cosine similarity gives a valuable measure of how similar two embedding vectors are likely to be. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. In more detail, let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two embedding vectors, with $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$, m being the number of embedding space dimensions. When plotted on a multi-dimensional space, the cosine similarity captures the orientation (the angle) of the data objects and not the magnitude. Smaller the angle, the

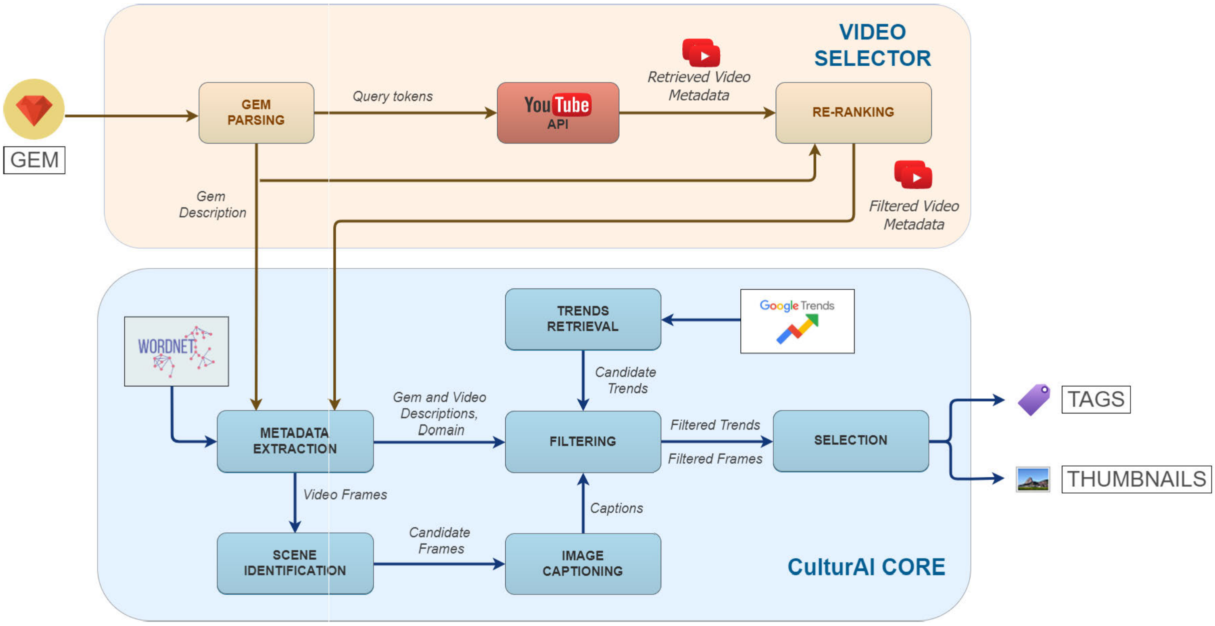


FIGURE 1. **Schematic diagram of *CulturAI*.**

higher the similarity. The cosine similarity (*sim*) between $\boldsymbol{x}$ and $\boldsymbol{y}$ is, in formula:

$$sim(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|} = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} x_i^2}\sqrt{\sum_{i=1}^{m} y_i^2}}. \tag{1}$$

When plotted on a multi-dimensional space, the cosine similarity captures the orientation (the angle) of the data objects and not the magnitude. Smaller the angle, the higher the similarity.

V. SYSTEM OVERVIEW

The approach is schematized in Figure 1. The process is accomplished in two main stages: first, a set of relevant videos is retrieved from the YouTube portal (*Video Selector*), and then relevant tags and thumbnails are generated for each selected video (*CulturAI Core*). The *gem* represents the input from which the system extracts (*Gem Parsing* module) the *gem* description and suitable textual tokens, the latter being sent as a query to the YouTube API, which retrieves relevant videos. Subsequently, the retrieved videos are filtered and re-ranked to keep the videos most related to the *gem* content (*Re-ranking* module). Then, the entire frames and metadata are extracted for each filtered video. On the one hand, the module *Metadata Extraction* extracts and aggregates meaningful textual metadata, i.e., (i) the description and the title of the video (*video description*), (ii) the textual description of the *gem* (*gem description*), and (iii) the description of the specific scenario (*domain*), e.g., cultural heritage semantic information obtained from an external lexical resource; subsequently, a set of suitable trends (module *Trends retrieval*) is retrieved and filtered. On the other hand, the *Scene Identification* module identifies the scene changes in the video at hand by analyzing the entire frameset, splits it into its scenes, and selects the most relevant frame for each scene, each frame representing a candidate frame. For each candidate frame, an *Image captioning* technique is applied to generate a proper textual caption. The generated captions are used to filter both frames and trends (module *Filtering*). After this step, the *Selection* module is aimed at identifying the most suitable trends and frames; the selected trends are suggested as new tags, and the selected frames will be suggested as new thumbnails.

Let us remark that, although the proposed model is based on the previous work (VSTAR), only the *Scene Identification*, *Image Captioning*, and *Selection* modules provide the same functionalities as the original VSTAR model.

In the following, we describe each module in depth.

VIDEO SELECTOR

A. GEM PARSING

The *gem* is represented with a proper document containing all its info and metadata (e.g., a JSON document). The module is aimed at extracting the textual information from the input document. It returns two outputs:

- The query tokens, which are sent to YouTube API to identify the relevant videos to the *gem*. In particular, the *gem* title, the location, and the city are extracted and aggregated. We have chosen to include the geoinformation to avoid ambiguities (e.g., a museum named ‘‘National Museum’’ can be found in numerous cities around the world);
- The token containing the textual *gem* description; such textual data is cleaned by removing noisy elements and sent directly to the *CulturAI* core. We denote the resulting text as *gem description*.

B. RE-RANKING

The YouTube API returns a ranked list of videos related to the given query. The returned ranking is affected by two main issues: (i) it may not be optimal, with the risk of proposing in first positions videos less related to the *gem*; (ii) the videos may contain textual data in different languages. At this stage, we are interested in selecting text in a specific language (i.e., English). To this end, the module first filters the list by removing videos having non-English text (title and description), both projected in a proper embedding space; then, the module re-ranks the remaining items according to a similarity score σ being, for a given video i, the cosine similarity between the *gem* description g and the textual data of the video (description and title) v_i:

$$\sigma_i = \text{sim}(g, v_i) \tag{2}$$

The module returns a number of videos (N_K) corresponding to the first K positions of the new ranking (K may be chosen by the final user). Each video is represented by the entire set of video frames and its title and description.

CulturAI CORE

Let us note that, in this stage, each selected video returned by the *Re-ranking* module is processed as described in the following.

C. METADATA EXTRACTION

The task aims to process the textual information from the various sources of the systems. To this end, the module integrates proper packages, i.e., Youtube-dl,[5] for accessing and extracting video metadata and OpenCV[6] for extracting visual features from frames. In detail, the following elements are extracted and grouped:

(i) The video info: the textual description of each video and its title are extracted and aggregated after removing noisy elements (e.g., links or emojis). For the sake of clarity, we denote the resulting textual element as *video description*.

(ii) The description of the *gem*, extracted by the *gem parsing* module.

(iii) Cultural heritage semantic information: to semantically enrich the textual data, we rely on a third-party lexical resource, i.e., WordNet [16], a widespread lexical network that has been extensively adopted in several fields of text mining and categorization. WordNet (WN hereinafter) groups words basing on their meanings, the direct relation among words being synonymy. Synonyms (i.e., words or phrases having a meaning that is the same as, or very similar to, other words or phrases) are grouped into unordered sets (named *synsets*). Furthermore, all synsets are organized in a proper taxonomy. To provide semantically relevant textual data, we manually collected a set of synsets related to the cultural heritage field. More details on the selected synsets are given in Section VII-E. Let us note that the optimal selection of synsets and of a different semantic resource is a current challenge. The selected set of synsets is denoted as *domain*.

(iv) The video frames. The entire frameset is extracted from the YouTube portal for each selected video. The final thumbnails will be selected from this set.

Although we have employed YouTube as a video platform in our instance, our approach can be applied to any streaming video-sharing platform.

D. SCENE IDENTIFICATION

Since a selected video comprises a high number of frames, mainly redundant and strongly similar to each other, analyzing the full frameset may be worthless and entail an expensive effort in terms of execution time and computational resources. A suitable choice for overcoming this issue has been relying on a Scene Identification tool, which aims to split the video into different scenes and select one representative frame for each scene, to remove frames too similar. In particular, we rely on PySceneDetect,[7] an efficient Python-based framework for scene detection. In detail, as numerous frames might be blurred or muddied, the module identifies the sharpest frame according to the percentage of "blurriness" (picking the least blurred frame) for each scene. The identified scene frames will compose the candidate frames set $\mathcal{CF}$.

E. IMAGE CAPTIONING

As already pointed out, the visual content of a given frame would be better represented by a proper sentence, which could lead to more accurate outcomes than using a simple object detection approach. As Deep Learning is a flourishing strategy in image and video captioning [24], we exploited a deep neural network-based Image Captioning framework to label each candidate frame $f_C \in \mathcal{CF}$ with an appropriate caption. In detail, *CulturAI* embeds an Image Captioning framework[8] compliant with the architecture proposed by Xu et al. [27]. The framework adopts a convolutional neural network to extract from each candidate frame its visual features, which are subsequently decoded into human-friendly sentences by an LSTM recurrent neural network.

F. TRENDS RETRIEVAL

A tag should be an attractive word (or, in general, an *n-gram*) popular amongst users. In the context of social media, which is remarkably "dynamic", a user is often interested in seasonal/hot events or resources (e.g., a company, a tool, a person). In other words, a user may often be interested in current trends rather than a more generic topic. To this end, the Google Trends engine[9] has been chosen as the optimal tool. Briefly, Google Trends provides several functionalities that

[5] https://youtube-dl.org/
[6] https://opencv.org/
[7] https://pyscenedetect.readthedocs.io/en/latest/
[8] https://github.com/DeepRNN/image_captioning
[9] https://trends.google.it/

permit one to infer information about searches performed on Google platforms over time. Also, Google Trends allows filtering results across targeted platforms, e.g., Google Search, Image Search, News Search, Google Shopping, and YouTube Search. For obvious reasons, we focus only on YouTube Search queries. We aim to collect the most popular searches related to the cultural heritage domain in a given period. Let us point out that Google Trends organizes the search information into a taxonomy of about 1400 nodes. Nevertheless, there are no specific nodes associated with the cultural heritage domain. To this end, we selected a set of nodes strictly related to our domain, as reported in Section VII-E. Each retrieved trend is a candidate for being a new tag. We denote the set of candidate trends with $\mathcal{CT}$.

G. FILTERING

TRENDS FILTERING

There may still be several irrelevant or unrelated results among all retrieved trends. To filter them (and reduce the computational effort), all trends belonging to $\mathcal{CT}$ are ranked by their cosine similarity, computed against the description vector d_v:

$$\psi_i = \mathrm{sim}(t_j, d_v) \tag{3}$$

All trends with a similarity score ψ_i greater than a given threshold τ_T are considered related to video content and taken as candidate tags.

Captions Filtering. Although the adopted captioning framework is a reliable tool, some generated captions might still not be accurate. Therefore, we rank captions of each candidate frame $f_i \in \mathcal{CF}$ according to their cosine similarity Υ_i, computed between a caption c_i and the domain s_{CH}:

$$\Upsilon_i = \mathrm{sim}(c_i, s_{CH}) \tag{4}$$

The system filters frames by selecting only the highest-ranked captions (we denote this number, hereinafter, with N_C). Identifying an appropriate threshold for captions is more challenging than trend filtering. Indeed, whereas a trend is related to a definite concept (i.e., the corresponding Google Trends category), a caption is strictly related to an uneven context (i.e., the frame content). Further study on how to select an appropriate threshold is currently under investigation.

H. SELECTION

According to the following phases, we use captions to generate tags and thumbnails: starting from the highest-ranked caption, we select the associated frame as a final thumbnail. Subsequently, we pick the most similar filtered trend to the given caption by computing their cosine similarity σ. The goal is to provide only tag-thumbnail pairs in which both elements are relevant and related to the video content and, consequently, related to the given *gem*. Subsequently, we iterate through the captions' ranking in the same way. Let us remark that a trend is associated with one caption only. Hence, for a given caption, if the most similar trend has already been assigned to a higher-ranked caption, we will select the second most similar, and so on.

This process returns, for each video, M tag-thumbnail pairs, with M specified by the final user u. Let us point out that u may also select to output different numbers of tags and thumbnails. In particular, u can specify to retrieve (i) M_{th} thumbnails, corresponding to the frames associated with the first M_{th} ranked pairs, and (ii) M_{ta} tags, corresponding to the trends associated with the M_{ta} high ranked pairs. Furthermore, the user may choose a threshold θ, intending to discard all pairs with a similarity $\sigma \leq \theta$ and, therefore, retrieve only trends and thumbnails highly related to video content.

Finally, let us remark that the described tag and thumbnail generation process is made for each video selected by the *Re-ranking module*. To this end, as the final output is the aggregation of tags and thumbnails of all selected videos, the *Selection* module aggregates and outputs the generated items of each video, removing the possible duplicated tags.

VI. A PRACTICAL EXAMPLE

To better clarify the whole process, Figure 2 illustrates a practical toy example of the deployment of the system, including all intermediate outcomes of each module described in the system's architecture (Section V). In the example, the input is the *gem* associated with the famous "Louvre Museum" in Paris. Furthermore, let us suppose that the museum hosts a temporary exhibition dedicated to Leonardo da Vinci when the system is deployed.

For the sake of clarity, since the final user may set the number of re-ranked videos, we have chosen to select only one video in the *Re-ranking* phase (module (3) in Figure 2), in order to let the reader better understand all the intermediate outcomes.

Let us now describe the process and the outcomes of our system step by step for the given example, depicted in Figure 2. The *Gem Parsing* module (1) analyzes the input *gem* and extracts the query tokens and the *gem* description. The tokens are sent to the *YouTube API* (2), which returns a ranked list of related videos, which is sent to the *Re-ranking* module (3). Videos having non-English text (e.g., video #1) or that are unrelated to the *gem* (e.g., video #3, which is a music video) are removed, and the remaining videos are re-ranked using the cosine similarity score between the *gem* description, and the description and title of each video. Since we have assumed that the user selects one video only for this example, after this phase only one video is retained by the algorithm. Then, the *Metadata Extractor* (4) returns two outputs for the selected video. On the one hand, it outputs the aggregation of the processed textual input items, i.e., the cultural heritage *domain* (represented by relevant synsets extracted by WordNet), the *video description* (extracted by the video metadata), and the *gem description* (extracted by the *Gem Parsing* module). On the other hand, the *Metadata Extractor* returns the entire frame set of the selected video, from which the *Scene Identification* module (5) identifies the

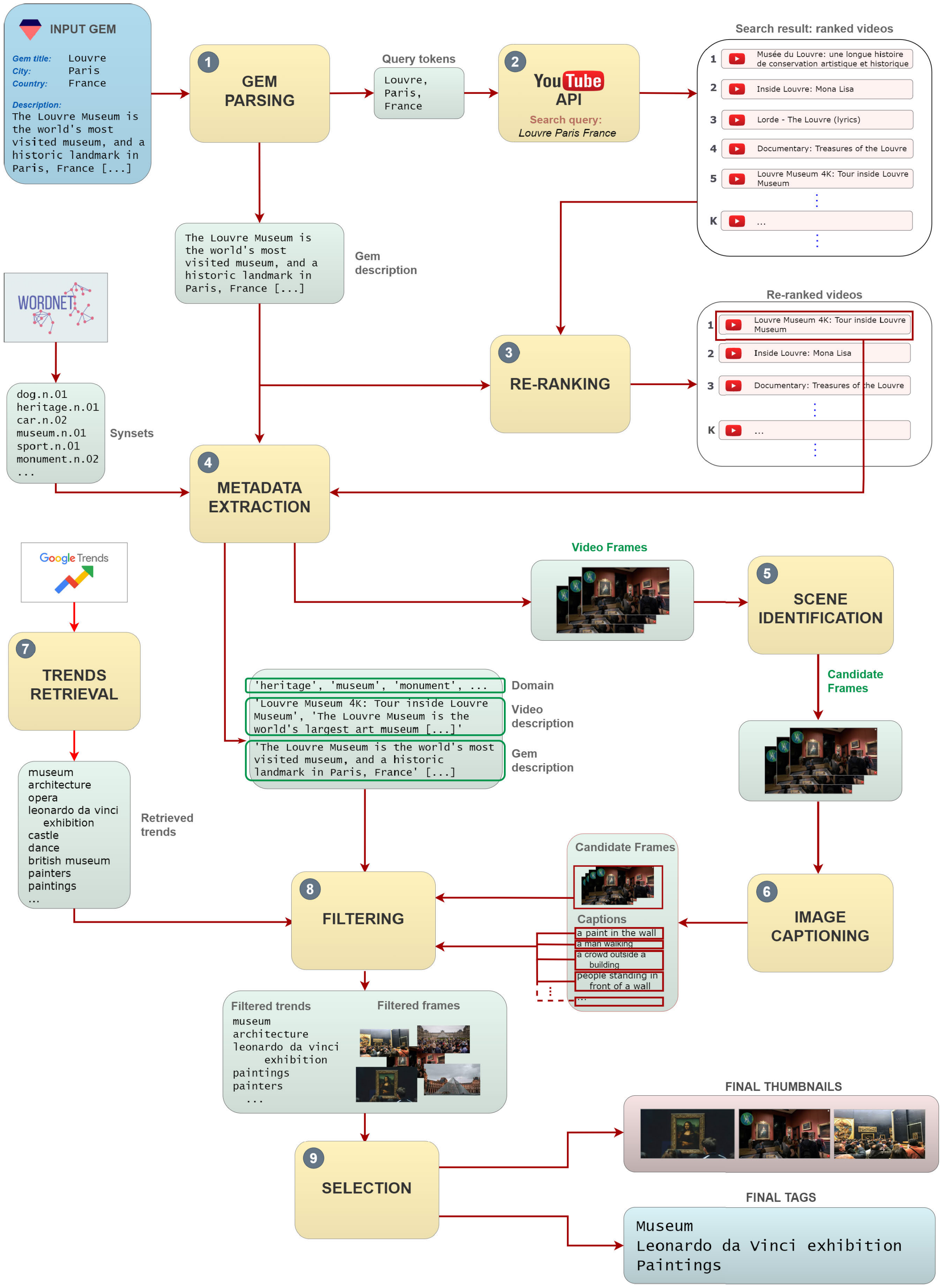

INPUT GEM
Gem title: Louvre
City: Paris
Country: France
Description:
The Louvre Museum is the world's most visited museum, and a historic landmark in Paris, France [...]
1 GEM PARSING
Query tokens
Louvre, Paris, France
2 YouTube API
Search query: Louvre Paris France
Search result: ranked videos
1 Musée du Louvre: une longue histoire de conservation artistique et historique
2 Inside Louvre: Mona Lisa
3 Lorde - The Louvre (lyrics)
4 Documentary: Treasures of the Louvre
5 Louvre Museum 4K: Tour inside Louvre Museum
K ...
The Louvre Museum is the world's most visited museum, and a historic landmark in Paris, France [...]
Gem description
WORDNET
dog.n.01
heritage.n.01
car.n.02
museum.n.01
sport.n.01
monument.n.02
...
Synsets
3 RE-RANKING
Re-ranked videos
1 Louvre Museum 4K: Tour inside Louvre Museum
2 Inside Louvre: Mona Lisa
3 Documentary: Treasures of the Louvre
K ...
4 METADATA EXTRACTION
Google Trends
Video Frames
5 SCENE IDENTIFICATION
Candidate Frames
7 TRENDS RETRIEVAL
'heritage', 'museum', 'monument', ...
Domain
'Louvre Museum 4K: Tour inside Louvre Museum', 'The Louvre Museum is the world's largest art museum [...]'
Video description
'The Louvre Museum is the world's most visited museum, and a historic landmark in Paris, France' [...]
Gem description
museum
architecture
opera
leonardo da vinci exhibition
castle
dance
british museum
painters
paintings
...
Retrieved trends
Candidate Frames
Captions
a paint in the wall
a man walking
a crowd outside a building
people standing in front of a wall
...
8 FILTERING
6 IMAGE CAPTIONING
Filtered trends
museum
architecture
leonardo da vinci exhibition
paintings
painters
...
Filtered frames
FINAL THUMBNAILS
9 SELECTION
FINAL TAGS
Museum
Leonardo da Vinci exhibition
Paintings

FIGURE 2. Example of tag and thumbnail generation for the "Louvre Museum" *gem*.

scene changes and extracts a set of candidate frames. Each candidate frame is then annotated with a proper caption generated by the *Image Captioning* module (6). Subsequently, the set of captions of the corresponding frames, the aggregated text extracted by (4), and the list of the most recent trends extracted by the *Trends Retrieval* module (7) are sent to the *Filtering* module (8), which identifies the most appropriate trends and frames (together with the corresponding captions). The *Selection* module (9) ranks the filtered items, returning the most relevant trends and frames according to the cosine similarity between trends and captions. The highest-ranked items represent the final tags and thumbnails for the selected video. In this example, we selected three tags and three thumbnails. Finally, let us remark that we decided to consider only one video in this example to give more clarity to the reader. In choosing more videos, the output would be the aggregation of generated tags and thumbnails.

As already remarked, one of the goals of the algorithm is to suggest hot popular topics as relevant tags for the input *gem*. In this case, the expected behavior is to suggest also tags related to the abovementioned exhibition, e.g., as reported in Figure 2, the tag "Leonardo da Vinci exhibition".

VII. EVALUATION

This section concerns the evaluation process of the described system. First, we illustrate the underlying scenario and the goal of our evaluation. Then, the evaluation methodology and the metrics are described.

A. UNDERLYING SCENARIO

According to the following phases, we use captions to generate tags and thumbnails for each video. First, the frame associated with the highest-ranked caption is selected as a final thumbnail. Then, the filtered trend most similar to such caption (according to their cosine similarity σ) is picked as a final tag. The goal is to provide only tag-thumbnail pairs in which both elements are relevant and related to the video content and, consequently, related to the given *gem*. Similarly, the following thumbnails and tags are suggested by iterating through the ranked captions. As the system assigns a trend to one caption only, if the most similar trend has already been associated with a higher-ranked caption, the system selects the second most similar, and so on. Therefore, M tag-thumbnail pairs will be generated for each video, with M specified by the final user u. It is worth noticing that u may also select to suggest different numbers of tags and thumbnails. i.e., u could specify to generate (i) M_{th} thumbnails, being the frames associated with the first M_{th} ranked pairs, and (ii) M_{ta} tags, being the trends associated with the M_{ta} high ranked pairs. Finally, let us remark that the described tag and thumbnail generation process is made for each video selected by the *Re-ranking module*. To this end, as the final output is the aggregation of tags and thumbnails of all selected videos, the *Selection* module aggregates and outputs the generated items of each video, removing the possible duplicated tags.

To this end, we estimate the performance in enhancing the original metadata with items highly related to the given *gem* by assessing the generated tags and thumbnails, i.e., the final output of the system. This preliminary study aims to obtain assessments comparable with the optimum.

B. EVALUATION METHODOLOGY

The focus of the evaluation is to estimate the relevance of each suggested tag or thumbnail regarding the given *gem*. It is not easy to define an automated evaluation. Indeed, the relevance of an item is generally a result of subjective factors, as human evaluation may vary along with personal tastes, interests, preferences, or emotions. To this end, as recent trends leverage collective human intelligence for such tasks, we opted to perform a manual assessment. However, the assessments may often be laborious and require extensive human effort. In this scenario, a widespread approach to reducing human effort relies on crowdsourcing services. Crowdsourcing in machine learning entails facilitating recruiting of data annotators at a large scale. In detail, we rely on a popular tool for crowdsourcing and human validation, i.e., Amazon Mechanical Turk (AMT hereinafter) provided by the Amazon Sagemaker service.[10] We assigned each tag or thumbnail to five workers, who labeled an item with a proper relevance score.

1) RELEVANCE SCORE

The relevance of each suggested tag/thumbnail is determined by relying on a similar approach presented by Konjengbam et al. [12]. We adopt a three-point relevance scale. Human assessors assign a relevance score to each suggested item (thumbnail or tag). The score can be any of the following values:

- **Non-relevant (score 1):** The tag/thumbnail has no association with the *gem* at hand.
- **Somewhat relevant (score 2):** Although the tag or the thumbnail might not fit with the specific *gem*, it may belong to a more generic or somewhat related concept (e.g., an event or a place located in the same city).
- **Relevant (score 3):** The tag/thumbnail concept is fairly related to the *gem*.

2) BINARY SCORE

With the aim of using classical evaluation metrics, as described in the following section, we need to label each suggested element with a binary score. To this end, we differentiate between *positive* and *negative* suggested elements (either tag or thumbnail). For each item x, labeled by an assessor a with a relevance score $r_a(x)$, we perform the following steps:

- As *relevant* and *somewhat relevant* items are considered potentially meaningful for a given user, we first convert

[10] https://aws.amazon.com/sagemaker/

$r_a(x)$ in an appropriate binary score $b_a(x)$ according to the following formula:

$$b_a(x) = \begin{cases} 1 & \text{if} \quad r_a(x) = [2, 3], \\ 0 & \text{if} \quad r_a(x) = 1. \end{cases} \tag{5}$$

- As each item x is assessed by N workers, x will be annotated with N binary relevance scores. To compute the final score, $\sigma(x)$, we adopt a majority vote strategy for each item according to the following formula:

$$\sigma(x) = \begin{cases} 1 & \text{if} \quad \sum_{a=1}^{N} b_a(x) > \theta, \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

θ being a threshold dependent on N. Each item has been assessed by five workers, meaning that the sum in the previous formula may vary between 0 and 5. We decide to set $\theta = 2$. In doing so, an item annotated with a score $\sigma = 1$ is considered *positive*, otherwise is *negative*.

C. METRICS

1) INTER-ANNOTATOR AGREEMENT

As described in the previous section, the assessment relies on manual evaluations performed by workers of the AMT service. In such a task, the risk is dealing with a high grade of disagreement among annotators, which may lead to an unreliable evaluation. Usually, the reliability of the agreement among annotators is estimated through Cohen's Kappa coefficient [6], a statistical metric that compares the agreement between two annotators when they label some assigned items into mutually exclusive categories. It calculates the degree of agreement while excluding the probability of consistency expected by chance. In our case, let us point out that AMT assigns tasks (HITS) to human workers, who are compensated for finishing the tasks. A HIT should cover a small piece of work paid with a small reward. The significant difference between a classical human assessment and the AMT evaluation is that most assessors do not label all items but only a small part; furthermore, each assessor evaluates a variable number of items. This problem makes the classical Cohen's Kappa ineffective. To address this issue, we rely on the so-called Free Marginal Kappa (k_m) [22], [26], an alternative algorithm that can be adopted when annotators are not forced to label a fixed number of elements. It is suitable for any number of annotators. As a rule of thumb, a k_m value above 0.6, in the [0, 1] range, represents a satisfactory annotator agreement.

2) EVALUATION MEASURES

As the problem can be seen as a typical information retrieval task, the evaluation has been carried out using a classical state-of-the-art metric, i.e., the precision score, being, in this field, the percentage of relevant items (i.e., the number of positive suggestions) among the total retrieved items (i.e., the total number of suggested items). In detail, either for tags or thumbnail suggestions, the precision $\mathcal{P}$ is given by:

$$\mathcal{P} = \frac{M_P}{M}, \tag{7}$$

where M_P is the number of positives, and M is the total number of suggestions. Let us note that the meaning and usage of *precision* in this field differs from the definition of accuracy and precision within other branches of machine learning.

As our system ranks the suggested items, another helpful metric is the *precision at K ($\mathcal{P}@K$)*, corresponding to the percent of *positive* elements among the top K suggestions, averaged over all suggestions. It is also useful for better investigating the optimal number of suggestions. For example, let us suppose that the precision at 10 in a top-10 recommendation problem is 80%. This means that 80% of the recommendations are relevant to the user. This metric helps to better understand the behavior of our system, but it cannot be adopted for other comparisons. Indeed, the main drawback in our experiments is that the optimum does not provide an actual ranking of items.

D. DATASET

Let us remark that we rely on manual assessment, for which an extensive evaluation would require considerable effort in terms of investments. Therefore, we have focused on identifying a suitable real-world dataset to assess our model. The adopted dataset has been manually built with 38 *gems* directly extracted from the Cultural Gems portal. Each *gem* is represented by its name, description, and location (city and country). The *gems* belong to different points of interest in several European cities, covering most European countries.

Furthermore, to compare with the ideal optimum, we select *gems* for which all retrieved videos contain enough original tags.

E. PARAMETERS SETTING

The main parameters of the system are:

- N_v – the number of videos retrieved for each *gem*;
- M_{ta} – the number of generated tags;
- M_{th} – the number of selected thumbnails.

For each *gem*, we have chosen $N_v = 3$. As we intend to compare with the optimum, let us remark that each YouTube video is associated with a variable set of tags and three different thumbnails. In particular, as no explicit ranking is given for the optimum, we have decided to extract the first M_{ta} tags from the list of original tags. We have set, for both our system and the optimum, $M_{ta} = 5$ (which we empirically estimated as an appropriate number of tags) and $M_{th} = 3$ (as only three thumbnails are associated with each uploaded video in the YouTube portal). Therefore, as we adopt a dataset built with 38 *gems*, the system generates 570 total tags (38 *gems* × 3 videos × 5 tags) and 342 total thumbnails (38 *gems* × 3 videos × 3 thumbnails). Let us remark that such amount of generated tags and thumbnails fulfill the trade-off

TABLE 2. Cultural heritage domain: selected WordNet synsets.

cultural_movement.n.01	heritage.n.01	museum.n.01
memorial.n.03	artwork.n.01	monument.n.01
monument.n.02	theatre.n.01	library.n.03
church.n.02	monastery.n.01	tourism.n.01
art_gallery.n.01		

TABLE 3. Google Trends: selected categories.

Historical Sites & Buildings	Dance	Opera
Libraries & Museums	Architecture	History
Myth & folklore		

between the need for an adequate number of annotations and the limited budget reserved for the AMT service.

Cultural Heritage Domain – As mentioned in previous sections, we adopt WordNet in order to semantically enrich the textual data and to represent the underlying cultural domain. In detail, we manually select the synsets more related to the cultural heritage domain. Each synset is represented by the textual description extracted from WN by the *Metadata Extraction* module. Each synset is aggregated with the other textual elements (i.e., the *gem* and the video descriptions), which are used to filter the most relevant trends and captions, as detailed in Section V-G. Let us remark that the selected synsets aim to semantically enrich the textual data and, consequently, to improve the relevance of the final generated tags/thumbnails.

To this end, we selected the synsets reported in Table 2 and their hyponyms to represent the cultural heritage domain.

Google Trends Domain – As already reported, Google Trends taxonomy does not provide a specific category for cultural heritage. To address this issue, we select a list of relevant categories from which to extract potentially relevant trends. The list of categories is reported in Table 3.

Trends Retrieval – We retrieve trends belonging to the ''last 30 days'' from the time of the query, as smaller periods imply smaller numbers of candidate trends.

Caption Filtering – We set the similarity threshold τ_T to 0.7. Trends having a similarity $\sigma_i \geq 0.7$ are considered similar and taken as candidates. For the experiments, if less than N_C trends exceed the threshold, θ_T is iteratively lowered by 0.1 until at least N_C trends are retrieved. Let us point out that the initial value of τ_T has been empirically chosen with several preliminary experiments, as reported in our previous work [4].

VIII. COMPUTATIONAL RESULTS

This section reports the obtained experimental results and some discussion around them. Let us point out that we devised a prototype entirely implemented in Python to assess the model.

A. ANNOTATOR AGREEMENT

Table 4 reports the inter-annotator agreement in tags and thumbnails suggestions for both our system and the optimum.

TABLE 4. Inter-annotator agreements for tags ($\mathcal{K}_{TA}$) and thumbnails ($\mathcal{K}_{TH}$).

	$\mathcal{K}_{TA}$	$\mathcal{K}_{TH}$
CulturAI	0.59	0.63
Optimum	0.62	0.65

TABLE 5. Precision scores for suggesting tags ($\mathcal{P}_{TA}$) and thumbnails ($\mathcal{P}_{TH}$).

	$\mathcal{P}_{TA}$	$\mathcal{P}_{TH}$
CulturAI	0.776	0.742
Optimum	0.899	0.787

The annotation agreement shows satisfactory values. The score is around 0.6 for the proposed model and the optimum. AMT workers might not always be experts in specific cultural heritage domains and, therefore, might not always agree.

B. PRECISION

The comparative results, in terms of the precision metric, are reported in Table 5. In detail, we report the total precision in both tag and thumbnail suggestions for our system and for the ideal optimum.

The obtained results by our system reported in Table 5 appear to be quite satisfactory. In particular, tags and thumbnail suggestions have similar performances, abundantly above a 0.7 precision score, representing a solid indicator of the effectiveness of the algorithm in suggesting both types of items. Moreover, the performances of tag and thumbnail suggestions appear to be quite comparable with that of the optimum, which represents a theoretical ideal. These results confirm the effectiveness of our methodology for proposing relevant tag and thumbnail suggestions, highlighting the reliability of the system for both tasks. Furthermore, another indicator of the model effectiveness is that for all selected videos (114 in total), the system has generated at least one positive tag, meaning that each analyzed video, and, consequently, each *gem*, has been enhanced with relevant content.

Furthermore, as already pointed out, the optimum does not always guarantee to be the best choice for the suggestion of the relevant tags and thumbnails. Indeed, let us remark that we built an ad-hoc dataset in which all retrieved videos contain an adequate number of tags and appropriate thumbnails. Such items are usually derived from several optimization processes (mainly manual tasks), and they may seem the optimal choice for obvious reasons. Actually, many YouTube videos are not processed in such a way, and they may not have enough or even have no tags, or may have randomly chosen thumbnails. To confirm this point, during the task of dataset building, we estimated that about 20% of videos contained 1 or 0 tags. Furthermore, let us remark that the retrieved tags are related to trendy popular topics and usually differ from the original tags. We estimated that less than 3% of the tags suggested by the system occur in the list of original tags. Finally, it is worth noticing that the original tags of a YouTube video are

TABLE 6. Precision @K for the suggested tags.

P@1	P@2	P@3	P@4	P@5
0.804	0.797	0.785	0.779	0.776

not often "refreshed". Our approach, instead, as it relies on popular trends, may be constantly refreshed with trendy tags, making a given *gem* more attractive, and further confirming the superiority of our approach.

Our system is able to enhance a given video and, in particular, the input *gems*, with popular and attractive tags and thumbnails, which can be adopted to augment original metadata rather than substituting it.

C. PRECISION @K

For a given video, the system outputs a ranked list of tags. In order to emphasize the effectiveness of the returned tag ranking, we also calculated the *Precision@K*, i.e., the averaged percent of positive elements among the top K suggested items. Table 6 reports the results in terms of *Precision@K*, (with K varying from 1 to 5), averaged over all 114 selected videos. The results reported in the table indicate the effectiveness of our approach for all values of K. Furthermore, the table conveys that increasing K leads to a precision decrease, which is the expected behavior of effective ranking algorithms. Indeed, the ideal scenario is that the most relevant items should be placed in the first positions. Hence, the more one goes down along the ranking, finding less relevant items, the more the total precision decreases. This is a strong indicator of the capability of generating an effective tag ranking and, hence, suggesting the most relevant tags.

D. DISCUSSION AND FUTURE IMPROVEMENTS

The previous experiments confirm that the system is able to effectively enrich a given *gem*. Although the study is still in its preliminary stage, the reported assessments encourage future improvements to address several crucial points that currently represent open challenges. In particular, the following issues are currently under investigation:

- The retrieved trends have presented a high "bias" towards UK locations, monuments, or events, as we have set the query to retrieve only popular queries in the UK. This choice is motivated by the need to address some limitations of the Google Trends platform. Indeed, Google Trends does not permit to specify a preferred language for the retrieved trends. Furthermore, each query may return a limited number of trends (up to only 25 trends may be retrieved for each query). We performed explorative experiments querying Google Trends for retrieving popular searches without setting the geolocalization, but the platform always returned mainly irrelevant trends. We estimated that more than 90% of the retrieved trends were non-English terms. Moreover, we have noticed a low variance in retrieving trends, meaning that, also varying the period of time to consider, the set of trends is always the same in each run. To address these problems, we plan to investigate more complex algorithms for retrieving trends, e.g., combining different queries varying the geolocalization or the period of time, or using more sophisticated NLP techniques to define specific tokens to query the portal.
- The given domain has been defined by manually choosing proper WordNet synsets. A future improvement is to define a more effective method to characterize the cultural heritage domain, for example, by enhancing the manually chosen synsets with further related synsets or concepts, or by using different lexical resources or ontologies, such as, e.g., Europeana[11] [5] and ArCo[12] [2], [3], increasing in this way the interoperability of the system.
- As the captions represent the key elements of the entire process, as both thumbnail and tag filtering and selection are based on the semantic information associated with the generated captions, exploring and analyzing different image captioning models may guide to a better choice and, consequently, to improve the effectiveness of our approach.
- Several algorithm steps are based on the cosine similarity metric calculated among textual items. We deem that adopting or devising more complex semantic methods may improve the matching between textual items, resulting in better final performances.

IX. USE CASE

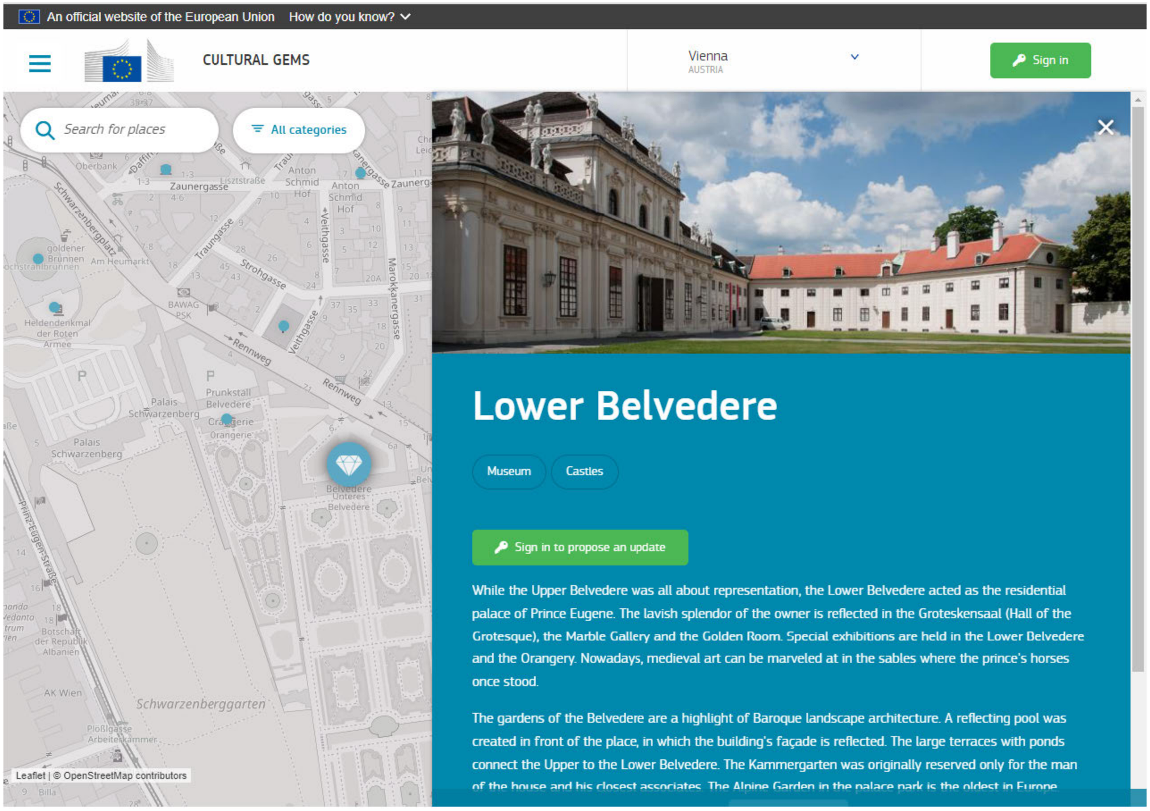
To give a better understanding of the effectiveness of our proposal, let us describe an appropriate use case. Let us consider the *"Lower Belvedere" gem*, a Baroque palace belonging to the historic building complex *Belvedere* in Vienna, Austria. The buildings are set in a Baroque park landscape, which houses the Belvedere museum.

Figure 3 depicts a screenshot of the *gem*, taken from the Cultural Gems web portal. For the sake of completeness, the *gem* description, partially shown in the figure, is the following:

While the Upper Belvedere was all about representation, the Lower Belvedere acted as the residential palace of Prince Eugene. The lavish splendor of the owner is reflected in the Groteskensaal (Hall of the Grotesque), the Marble Gallery, and the Golden Room. Special exhibitions are held in the Lower Belvedere and the Orangery. Nowadays, medieval art can be marveled at in the sables where the prince's horses once stood. The gardens of the Belvedere are a highlight of Baroque landscape architecture. A reflecting pool was created in front of the place in which the building's faïçade is reflected. The large terraces with ponds connect the Upper to the Lower Belvedere. The Kammergarten was originally reserved only for the man of the house and his closest

[11] Euopeana: https://www.europeana.eu/en

[12] ArCo - Knowledge Graph of the Italian Cultural Heritage: http://wit.istc.cnr.it/arco/index.php?lang=en



Lower Belvedere

While the Upper Belvedere was all about representation, the Lower Belvedere acted as the residential palace of Prince Eugene. The lavish splendor of the owner is reflected in the Groteskensaal (Hall of the Grotesque), the Marble Gallery and the Golden Room. Special exhibitions are held in the Lower Belvedere and the Orangery. Nowadays, medieval art can be marveled at in the sables where the prince's horses once stood.

The gardens of the Belvedere are a highlight of Baroque landscape architecture. A reflecting pool was created in front of the place, in which the building's façade is reflected. The large terraces with ponds connect the Upper to the Lower Belvedere. The Kammergarten was originally reserved only for the man of the house and his closest associates. The Alpine Garden in the palace park is the oldest in Europe.

FIGURE 3. **Screenshot of the selected "Lower Belvedere" *gem* taken from the Cultural Gems portal.**

associates. The Alpine Garden in the palace park is the oldest in Europe.

We choose to select three relevant videos and generate three relevant thumbnails and five related tags for each video, according to the experiments reported in Section VIII. According to the architecture depicted in Figure 1, let us describe the process and the outcomes of our system step by step. First, the *Gem Parser* extracts the textual information from the *gem* metadata and queries the YouTube portal with the following tokens: "Lower Belvedere, Vienna, Austria". The *Re-Ranking* module returns a ranked list of relevant videos. In our experiments, the module returns the videos summarized in Table 7, which reports the links, titles, and description snippets for each video.

The *Metadata extraction* module, for each video summarized in Table 7, extracts all meaningful information, i.e., the textual information, represented by the title and the description reported in the table, and the entire frameset of the video, the former being sent directly to the *Filtering* module. On the one hand, a set of candidate frames are extracted by the *Scene identification* module and annotated in the proper captions with the *Image captioning* stage. On the other hand, the *Trends retrieval* extracts the most popular searches related to the cultural heritage domain in the last 30 days. An excerpt of the retrieved trends list is reported in Table 8.

Afterward, the *Filtering* module processes, for each video of Table 7, the textual video information, the retrieved trends, the domain (represented by the synsets reported in Table 2), and the set of candidate frames and the associated captions. The module outputs two sets, one containing the most relevant trends and one containing the most relevant frames. Among them, the *Selection* module identifies, for each video, the most proper thumbnails and tags. Finally, the module outputs the set of selected videos (three in our example), each represented with the generated thumbnails (three per video) and annotated with the associated YouTube link, and the set of tags, obtained by gathering all tags generated for the selected videos (five tags per video) and removing the duplicates.

To give the reader a better understanding of the system effectiveness, we report both generated thumbnails and tags and the corresponding comparisons with the optimum (i.e., the original items) of the videos summarized in Table 7.

On the one hand, Figures 4-6 report the comparisons between the original thumbnails and the generated ones.

The figures confirm the insights about the potential of our proposal. Suggesting thumbnails highly related to the *gem* is one of the main strengths, confirmed especially in Figure 6. Indeed, the main content of all generated thumbnails is the "Belvedere Palace", whereas all the original thumbnails depict a close-up of a woman. Furthermore, the system generates clearer and brighter thumbnails, as highlighted, for example, in Figure 5. From the final user perspective, the insight is that providing a colorful and bright image may be more attractive than a dark and colorless thumbnail.

TABLE 7. Retrieved YouTube videos for the "Lower Belvedere" use case.

	YouTube ID	Title	Description
1	V-2TdpZObo8	VIENNA - Belvedere Palace in 4K	The Belvedere consist of two Baroque palaces (the Upper and Lower Belvedere), the Orangery, ...
2	0fsAc_ommik	Vienna Austria Travel Guide: Belvedere Palace 4K	Vienna Travel Guide : Belvedere Palace 4K. Top Tourist destination in Vienna...
3	7V5izRLTK74	Vienna's Belvedere Palace & Art Museum - Klimt's "The Kiss"	The Belvedere Palace complex in Vienna used to be a summer residence for royalty, hosting balls...

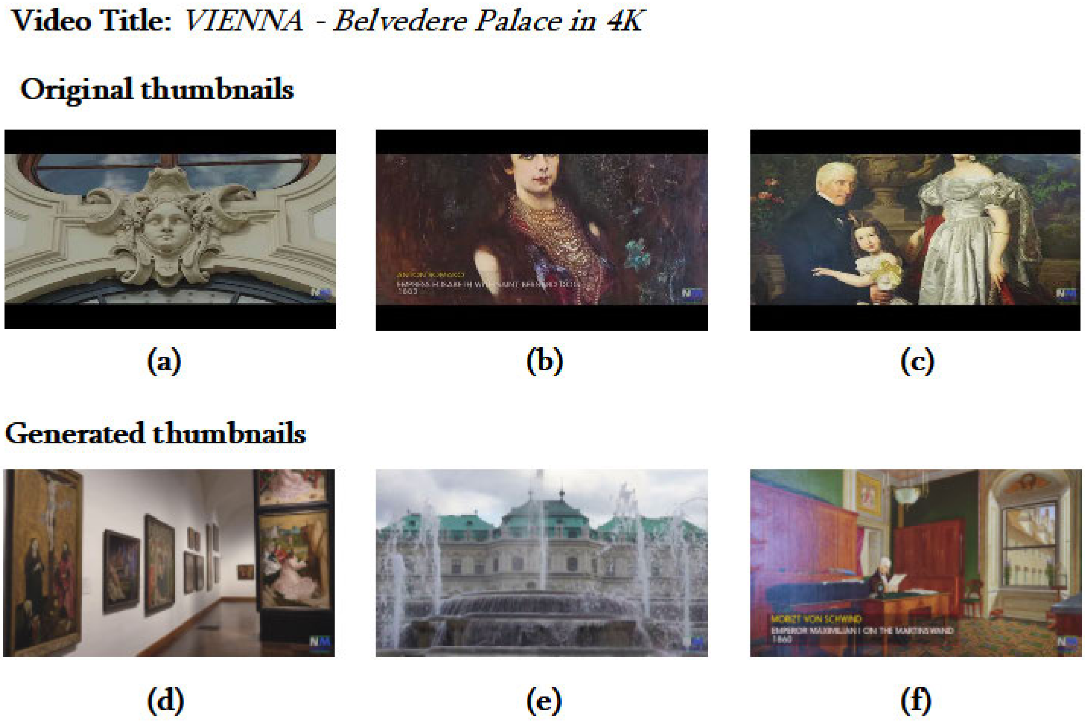


FIGURE 4. Use case - original thumbnails (a,b,c) and generated thumbnails (d,e,f) for video 1.

TABLE 8. Retrieved trends for the "Lower Belvedere" use case.

tower bridge	the shard london	london museum
chateau	history museum	museum
theatre	wall	house
...	...	...

TABLE 9. Use case: original and generated tags.

	Original tags	Generated tags
Video 1	vienna vienna austria viena viena austria wien	house history museum library chateau london museum
Video 2	Vienna Travel Guide Belvedere Palace 4K vienna austria vienna austria	theatre iconic london history museum trade center uk museum
Video 3	belvedere palace belvedere palace vienna belvedere palace austria art galleries in vienna art galleries in austria	exhibition history museum museum london museum iconic london

On the other hand, Table 9 reports, for each selected video, the original and the generated tags. The table highlights that the generated tags set has no common elements with the original set, meaning that, as already pointed out in the previous section, the system may propose different tags and, therefore, potentially *enrich* the original tags set with relevant items. In particular, several tags highly related to the given *gem* that do not occur in the original tag set are suggested, e.g., "museum", "exhibition", or "history museum".

The system returns the aggregated generated tags filtered by removing duplicates – i.e., "house", "theatre", "exhibition", etc. Conversely, some tags are related to specific locations in the UK, like "london museum". As already pointed out in Section VIII-D, this is due to the limitation of the Google Trends platform, which has led to retrieve

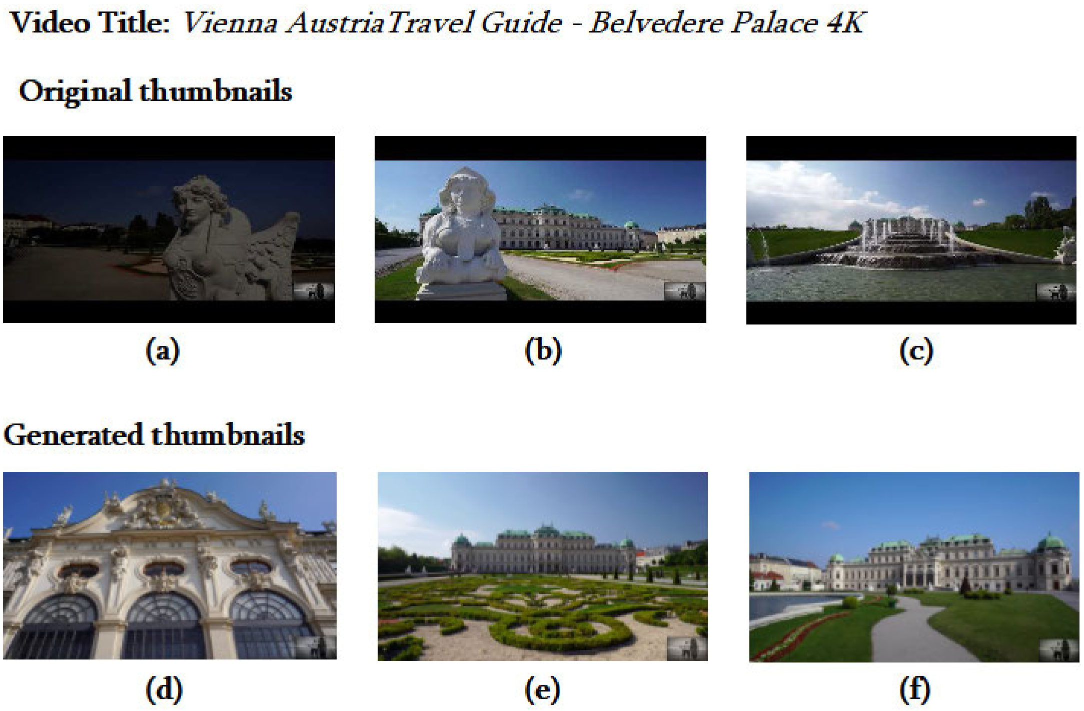

Video Title: *Vienna Austria Travel Guide - Belvedere Palace 4K*
Original thumbnails
(a)
(b)
(c)
Generated thumbnails
(d)
(e)
(f)

FIGURE 5. Use case - original thumbnails (a,b,c) and generated thumbnails (d,e,f) for video 2.

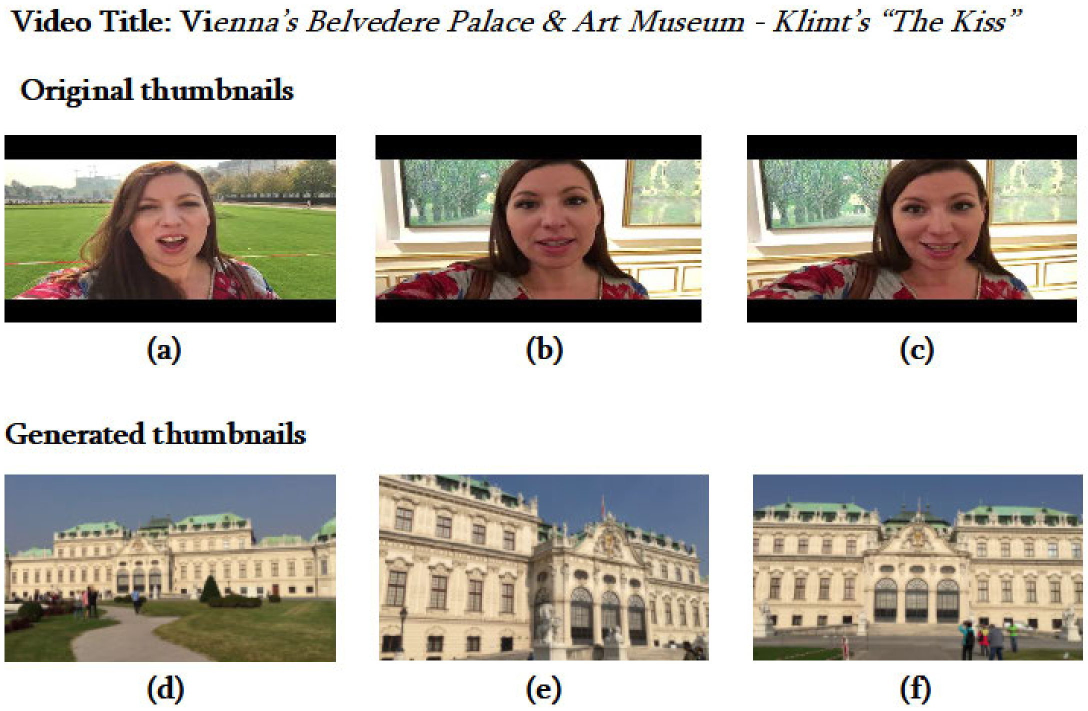

Video Title: Vie*nna's Belvedere Palace & Art Museum - Klimt's "The Kiss"*
Original thumbnails
(a)
(b)
(c)
Generated thumbnails
(d)
(e)
(f)

FIGURE 6. Use case - original thumbnails (a,b,c) and generated thumbnails (d,e,f) for video 3.

only trends popular in the UK territory for language reasons. We expect to limit and possibly solve this issue in the future.

X. CONCLUSION

In this manuscript, we proposed a novel system aimed at enriching the *gems*, i.e., cultural and creative venues of many European cities mapped in a free and open source web platform (namely, *Cultural Gems*). In this scenario, the effectively addressed challenge has been to efficiently increase user interoperability and attractiveness of Cultural Gems by data augmentation. In particular, the proposed model suggests appropriate social media content strongly related to each *gem*.

In detail, the system outputs relevant tags and videos (represented by suitable thumbnails), the former being selected by identifying semantically related popular search queries from the Google Trends platform, and the latter extracted by querying the YouTube portal. A further strength of our proposal is denoted by dynamic maintenance and update of the suggested contents for each *gem* in the application. Indeed, as the system retrieves the most popular videos/trends, a *gem* may constantly be ''refreshed'' with trendy content by periodically running the tool.

To assess the effectiveness of our proposal, we developed and tested an implementation of the whole system. To the best of our knowledge, no similar models have been proposed in the related literature. Therefore, the approach has been compared with an ideal optimum. In particular, the suggested tags and thumbnails have been compared with the corresponding elements originally associated with the given YouTube videos. Results confirmed the potential of the proposal and encourage future investigations and improvements. In particular, the experiments highlighted that tags and thumbnail suggestions have similar performances, representing a solid indicator of the effectiveness in suggesting both types of items, which are comparable with the optimum, highlighting the reliability of both tag and thumbnail generation. Moreover, the system generated at least one positive tag for each selected video, and nearly the totality of generated tags was not included in the original video tags, emphasizing the effectiveness of enriching the original tag set of a video with new tags related to the given *gem*. However, some limitations persist: the verticality of the domain under consideration, as well as its strong connotation as a humanistic field, in which elements of, e.g., subjective perception and artistic feeling prevail, make it complex to identify content capable of comprehensively enriching the cultural *gems*. Indeed, the existing artificial intelligence approaches, as they mostly make use of learning techniques that exploit the syntactic and semantic context of a (narrow) source description of a *gem*, are not yet perfectly capable of also capturing the emotional features described above. As for future work, we aim to deeper investigate and address the main weaknesses of our proposal, with the goal of improving the model. First, more complex algorithms should be devised for retrieving more relevant trends, as the Google Trends platform currently presents limited functionalities (e.g., up to only 25 trends per query may be retrieved). Furthermore, we deem to develop a more efficient method to represent the cultural heritage domain, e.g., by using further third-party lexical resources and ontologies. We also plan to study more sophisticated algorithms for capturing and comparing the semantic information associated with textual elements.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of the Joint Research Centre of the European Commission. The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the European Commission.

REFERENCES

[1] M. Artese and I. Gagliardi, ''Cataloging intangible cultural heritage on the web,'' in *Proc. Euro-Medit. Conf.*, in Lecture Notes in Computer Science, vol. 7616, 2012, pp. 676–683.

[2] V. Carriero, A. Gangemi, M. Mancinelli, L. Marinucci, A. Nuzzolese, V. Presutti, and C. Veninata, ''ArCo: The Italian cultural heritage knowledge graph,'' in *Proc. Int. Semantic Web Conf.*, in Lecture Notes in Computer Science, vol. 11779, 2019, pp. 36–52.

[3] V. A. Carriero, A. Gangemi, M. L. Mancinelli, A. G. Nuzzolese, V. Presutti, and C. Veninata, ''Pattern-based design applied to cultural heritage knowledge graphs,'' *Semantic Web*, vol. 12, no. 2, pp. 313–357, Jan. 2021.

[4] S. Carta, A. Giuliani, L. Piano, A. S. Podda, and D. R. Recupero, ''VSTAR: Visual semantic thumbnails and tAgs revitalization,'' *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116375.

[5] P. Clough, T. Hill, M. L. Paramita, and P. Goodale, ''Europeana: What users search for and why,'' in *Proc. Int. Conf. Theory Pract. Digit. Libraries*, in Lecture Notes in Computer Science, vol. 10450, 2017, pp. 207–219.

[6] J. Cohen, ''A coefficient of agreement for nominal scales,'' *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[8] N. Díaz-Rodríguez and G. Pisoni, ''Accessible cultural heritage through explainable artificial intelligence,'' in *Proc. 28th ACM Conf. User Modeling, Adaptation Personalization*, New York, NY, USA, Jul. 2020, pp. 317–324.

[9] L. E. Vigl, T. Marsoner, V. Giombini, C. Pecher, H. Simion, E. Stemle, E. Tasser, and D. Depellegrin, ''Harnessing artificial intelligence technology and social media data to support cultural ecosystem service assessments,'' *People Nature*, vol. 3, no. 3, pp. 673–685, Jun. 2021.

[10] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. D. Bue, and S. James, ''Machine learning for cultural heritage: A survey,'' *Pattern Recognit. Lett.*, vol. 133, pp. 102–108, May 2020.

[11] D. Jin, Z. Qi, Y. Luo, and Y. Shan, ''TransFusion: Multi-modal fusion for video tag inference via translation-based knowledge embedding,'' in *Proc. 29th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2021, pp. 1093–1101, doi: 10.1145/3474085.3481535.

[12] A. Konjengbam, N. Kumar, and M. Singh, ''Unsupervised tag recommendation for popular and cold products,'' *J. Intell. Inf. Syst.*, vol. 54, no. 3, pp. 545–566, Jun. 2020.

[13] E. Mäkelä, O. Suominen, and E. Hyvönen, ''Automatic exhibition generation based on semantic cultural content,'' in *Proc. Cultural Heritage Semantic Web Workshop*, 2007, pp. 1–11.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ''Efficient estimation of word representations in vector space,'' 2013, *arXiv:1301.3781.*

[15] G.-Z. Miliopoulou, ''Brand communities, fans or publics? How social media interests and brand management practices define the rules of engagement,'' *Eur. J. Marketing*, vol. 55, no. 12, pp. 3129–3161, 2021.

[16] G. A. Miller, ''WordNet: A lexical database for English,'' *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[17] M. Mudge, M. Ashley, and C. Schroer, ''A digital future for cultural heritage,'' *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.-ISPRS Arch.*, vol. 36, no. 5, pp. 1–6, 2007.

[18] J.-M. Oh and N. Moon, ''Towards a cultural user interface generation principles,'' *Multimedia Tools Appl.*, vol. 63, no. 1, pp. 195–216, Mar. 2013.

[19] A. A. Patwardhan, S. Das, S. Varshney, M. S. Desarkar, and D. P. Dogra, ''ViTag: Automatic video tagging using segmentation and conceptual inference,'' in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 271–276.

[20] G. Pisoni, N. Díaz-Rodríguez, H. Gijlers, and L. Tonolli, ''Human-centered artificial intelligence for designing accessible cultural heritage,'' *Appl. Sci.*, vol. 11, no. 2, p. 870, Jan. 2021.

[21] V. Poulopoulos and M. Wallace, ''Digital technologies and the role of data in cultural heritage: The past, the present, and the future,'' *Big Data Cognit. Comput.*, vol. 6, no. 3, p. 73, Jul. 2022.

[22] J. J. Randolph, ''Free-marginal multirater Kappa (multirater K[free]): An alternative to Fleiss' fixed-marginal multirater Kappa,'' ERIC, Tech. Rep., ED490661, pp. 1–20, 2005. [Online]. Available: https://eric.ed.gov/?id=ED490661

[23] N. Reimers and I. Gurevych, ''Sentence-BERT: Sentence embeddings using Siamese BERT-networks,'' 2019, *arXiv:1908.10084*.

[24] R. Shetty, H. R. Tavakoli, and J. Laaksonen, ''Image and video captioning with augmented neural architectures,'' *IEEE Multimedia Mag.*, vol. 25, no. 2, pp. 34–46, Apr./Jun. 2018.

[25] S. T. Kokab, S. Asghar, and S. Naz, ''Transformer-based deep learning models for the sentiment analysis of social media data,'' *Array*, vol. 14, Jul. 2022, Art. no. 100157. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005622000224

[26] M. J. Warrens, ''Inequalities between multi-rater Kappa,'' *Adv. Data Anal. Classification*, vol. 4, no. 4, pp. 271–286, Dec. 2010.

[27] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, ''Show, attend and tell: Neural image caption generation with visual attention,'' in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 2048–2057.

SALVATORE M. CARTA received the Ph.D. degree in electronics and computer science from the University of Cagliari, in 2003. In 2005, he joined the Department of Mathematics and Computer Science, University of Cagliari, as an Assistant Professor. From 2006 to 2007, he joined the Swiss Federal Institute of Technology as an Invited Researcher. He is currently a Full Professor with the Department of Mathematics and Computer Science, University of Cagliari. He is the author of more than 130 conference and journal papers in the research fields of artificial intelligence, recommendation and computer vision, with more than 2000 citations. He is a member of the ACM. He founded three hi-tech companies, spin-offs of the University of Cagliari, and currently leading one of them.

SERGIO CONSOLI received the Ph.D. degree. He is currently a Scientific Project Leader with the European Commission, Joint Research Centre (DG JRC), Ispra, Italy, and working with the Competence Centre on Composite Indicators and Scoreboards, and formerly within the Centre for Advanced Studies on the project: Big data and forecasting of economic developments. Previously, he was a Senior Scientist at the Data Science Department, Philips Research, Eindhoven (NL), focusing on advancing automated analytical methods used to extract new knowledge from data for HealthTech applications. Other former experiences include the Italian Presidency of the Council of Ministers and the National Research Council of Italy. He also provided ICT consultancy services to Isab, the largest oil refinery in the Mediterranean area. His education and scientific experience fall in the areas of data science, operational research, artificial intelligence, knowledge engineering, semantic reasoning, and machine learning. He is the author of several research publications in peer-reviewed international journals, granted EPO and WIPO patents, edited books, and leading conferences in the fields of his work. He is also co-edited two books. He is an Associate Editor for IEEE Access and a member of the Editorial Board of *PLOS ONE* and the *Journal of Big Data*.

ALESSANDRO GIULIANI received the Ph.D. degree in electronic and computer engineering from the University of Cagliari, Italy, in 2012, supported by a grant from RAS (Autonomous Region of Sardinia). In 2011, he was a Visiting Ph.D. student, he joined the Yahoo! Research Laboratory in Barcelona (Spain). From 2012 to 2020, he was a Research Associate at the Department of Electrical and Electronic Engineering of the University of Cagliari. From 2016 to 2017, as a Visiting Postdoctoral, he joined the ''HeiderLaboratory'' at the University of Marburg (Germany), directed by Prof. Dominik Heider. He is currently a Postdoctoral Researcher with the Department of Mathematics and Information Technology, University of Cagliari. His main research interests include ranges over several domains, such as artificial intelligence, machine learning, deep learning, recommender and advertising systems, text categorization, data mining, and information retrieval. Currently, he is a member of the Artificial Intelligence and Big Data Laboratory at the University of Cagliari, in which he has been involved in several research projects.

ALESSANDRO SEBASTIAN PODDA received the Ph.D. degree in mathematics and computer science, in 2018. He is currently an Assistant Professor with the Department of Mathematics and Computer Science, University of Cagliari. He is also a Research Unit Coordinator (AI for eHealth and Smart Cities) with the Artificial Intelligence and Big Data Laboratory, as well as a member of the Blockchain Laboratory, University of Cagliari. He is also the Former Technical Director and a Solution Architect of the Doutdes and Sardioin projects and participates in numerous research projects including AlmostAnOracle, Safespotter, and Mister. To date, he has been the coauthor of more than 25 journal articles and conference papers.

DIEGO REFORGIATO RECUPERO received the Ph.D. degree in computer science from the University of Naples Federico II, Italy, in 2004. From 2005 to 2008, he was a Postdoctoral Researcher at the University of Maryland, College Park, USA. He has been a Full Professor with the Department of Mathematics and Computer Science, University of Cagliari, Italy, since February 2022. His current research interests include sentiment analysis, semantic web, natural language processing, human–robot interaction, financial technology, and smart grid. He is the author of more than 190 conference and journal papers in these research fields, with more than 2400 citations. He won different awards in his career (such as Marie Curie International Reintegration Grant, Marie Curie Innovative Training Network, and Best Researcher Award from the University of Catania, Computer World Horizon Award, Telecom Working Capital, Startup Weekend, Best Paper Award). He has co-founded six companies within the ICT sector and is actively involved in European projects and research (with one of his companies he won more than 40 FP7 and H2020 projects).