


# Extracting semantic knowledge from web context for multimedia IR: a taxonomy, survey and challenges

Teresa Bracamonte<sup>1</sup> · Benjamin Bustos<sup>1</sup>  ·  
Barbara Poblete<sup>1</sup> · Tobias Schreck<sup>2</sup>

Received: 4 May 2016 / Revised: 29 June 2017 / Accepted: 3 July 2017 /  
Published online: 25 July 2017  
© Springer Science+Business Media, LLC 2017

**Abstract** Since its invention, the Web has evolved into the largest multimedia repository that has ever existed. This evolution is a direct result of the explosion of user-generated content, explained by the wide adoption of social network platforms. The vast amount of multimedia content requires effective management and retrieval techniques. Nevertheless, Web multimedia retrieval is a complex task because users commonly express their information needs in semantic terms, but expect multimedia content in return. This dissociation between semantics and content of multimedia is known as the *semantic gap*. To solve this, researchers are looking beyond content-based or text-based approaches, integrating novel data sources. New data sources can consist of any type of data extracted from the *context* of multimedia documents, defined as the data that is not part of the raw content of a multimedia file. The Web is an extraordinary source of context data, which can be found in explicit or implicit relation to multimedia objects, such as surrounding text, tags, hyperlinks, and even in relevance-feedback. Recent advances in Web multimedia retrieval have shown that context data has great potential to bridge the semantic gap. In this article, we present the first comprehensive survey of *context-based* approaches for *multimedia information retrieval on the Web*. We introduce a data-driven taxonomy, which we then use in our literature review of the most emblematic and important approaches that use context-based data. In addition, we

---

✉ Teresa Bracamonte  
tbracamo@dcc.uchile.cl

Benjamin Bustos  
bebustos@dcc.uchile.cl

Barbara Poblete  
bpoblete@dcc.uchile.cl

Tobias Schreck  
tobias.schreck@cgv.tugraz.at

<sup>1</sup> Department of Computer Science, University of Chile, Santiago, Chile

<sup>2</sup> Computer Graphics and Knowledge Visualization, Graz University of Technology, Graz, Austria

identify important challenges and opportunities, which had not been previously addressed in this area.

**Keywords** Semantic knowledge extraction · Multimedia retrieval · Web retrieval · Context data · Big data

## 1 Introduction

The wide adoption of online social media platforms for publishing and sharing content has produced an explosion of multimedia data on the Web. This, along with the massification of digital cameras and smartphones, has transformed the way users interact with each other and express their ideas to the world. User exchanges have experienced an important shift, from being mostly text-based, towards becoming rich in multimedia data. Online social platforms have had a great influence in this paradigm shift with the rise of websites such as Facebook,<sup>1</sup> Instagram,<sup>2</sup> YouTube<sup>3</sup> and Flickr,<sup>4</sup> to name a few. These websites promote the publication and dissemination of user-generated content, and provide tools customized for the publication of multimedia (i.e., mostly of videos, images and text). Within this data overload scenario on the Web, efficient multimedia search has become a very relevant information retrieval (IR) topic.

Since its beginnings, *Multimedia Information Retrieval* (MIR) research has been based mostly on using the content of multimedia objects as features for retrieval, or so called *content-based* [44] features. In particular, an important drawback of content-based MIR systems is the difficulty of adapting them for information retrieval based on humans' information needs. These needs, which are usually expressed in semantic terms (i.e., natural language text), depend on different user factors, such as the user's particular circumstances and experience. This problem is called the *semantic gap*, which Smeulders et al. [62] define as "... the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation"; and is without a doubt the driving challenge of user-designed MIR systems, in particular for those created for retrieval on the Web. In view of this problem, Web MIR systems have adapted not only to include multimedia content-based features, but also include *context-based* features [72]. Context, as described by Westerveld [72], is information about the content of an image that stems from sources other than the image itself (i.e., not from visual properties of the images). In particular, context is defined (for images) as the textual information that comes with an image. Context can consist of annotations that were manually added for describing the images (keywords, descriptions), or collateral text that is more or less *loosely coupled* with an image (e.g., captions, subtitles, surrounding text). For instance, Fig. 1 shows an example of an image's surrounding text in a Web document. The main idea of introducing context-based features in MIR is that they can be used to enhance the semantic information about a multimedia object, or in addition, the information that we have about its relevance.

The vast amount of user-generated multimedia content published on a daily basis on the Web, has produced a wide variety of context information. This has gradually influenced

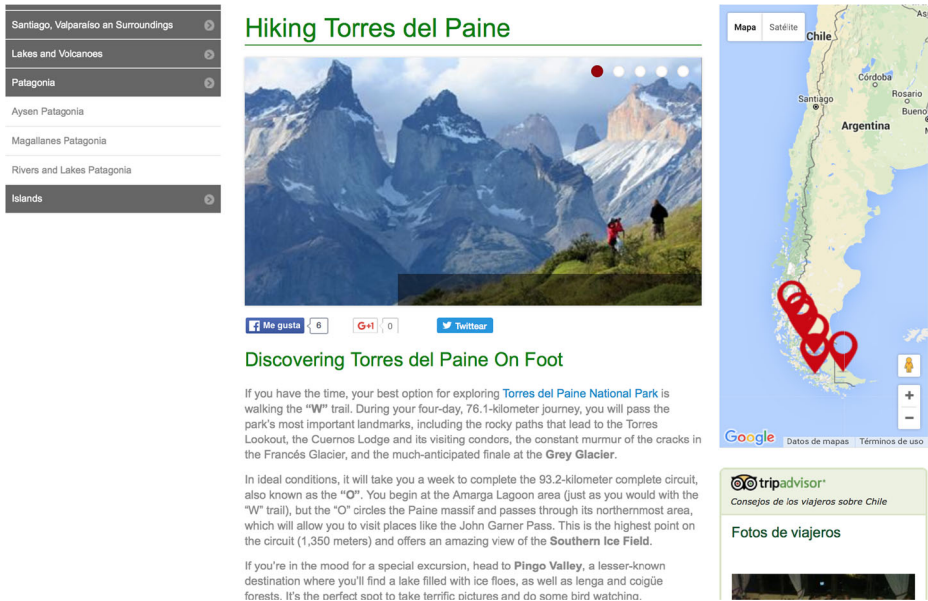
---

<sup>1</sup><http://www.facebook.com>

<sup>2</sup><http://instagram.com>

<sup>3</sup><http://www.youtube.com>

<sup>4</sup><http://flickr.com>



**Fig. 1** Image embedded in a Web page

the Web MIR research community, which has started to incorporate more sophisticated approaches that use different types of context-based features, and also combinations of context- and content-based approaches. Recent context-based features are extracted from sources which go beyond those proposed by Westerveld [72]. For example, Fig. 2 shows an image on Flickr that includes context data from the image's title, user-generated tags, community comments, and number of favorites. Figure 3 shows another example of context data for a video published on YouTube, consisting of comments, likes and shares.

## 1.1 Our contribution

Our research on the use of context data for Web MIR has shown this to be an emerging area, with promising results in terms of bridging the semantic gap on the Web. Nevertheless, although there is a growing body of work on this topic, we have found that there are currently no studies that analyze existing approaches in a comprehensive manner. Indeed, we have found a vast amount of work related to traditional content-based MIR, such as the surveys by Datta et al. [12], Hu et al. [29] and Typke et al. [67]. In addition, there are recent surveys that partially cover certain subtopics of this survey, such as the use of certain context information, like tags for image retrieval in Li et al. [46]. Also, there have been works that survey approaches that exploit context information focused on a specific multimedia type, such as music [40], or the task of reranking [50], or with more emphasis on user intent [41].

Our survey differs from the rest in its scope, with its use of context data in any of its forms, for the purpose of improving Web Multimedia IR and its related tasks. Given the importance of context information for obtaining semantic and relevance information for Web MIR, we believe there is need for a survey, such as this one, that provides a comprehensive view of this area. We focus on MIR for the Web because the Web is by far the largest multimedia repository available and is immensely rich in context information. Specifically,

**David Min** + Follow

Torres Del Paine, Chile

21,330 Views 318 faves 337 comments Taken on February 21, 2013

© All rights reserved

Getty | 500px | Facebook Fan Page | Twitter | Google + | Blogger | Tumblr | Formspring

Torres del Paine is a national park in the Extreme South region of Patagonian Chile. It is located in the southern tiers of the Andes and features mountains, lakes and glaciers. The Torres del Paine (Spanish for "Towers of Paine" and "Paine" is the old indigenous name for the colour blue), three immense rock towers give the park its name.

Torres del Paine, Magallanes and Antartica Chilena Region, Chile

This photo is in 3 galleries [Add to gallery](#)

- [Great Landscapes](#) 49 photos
- [Landscapes 231\(2180\)](#) 16 photos
- [Lugares Fantasticos 20](#) 14 photos and 1 video

Tags **BETA**  [Add tags](#) [People in photo](#) [Add people](#)

- Andes Argentina Chile
- Cuernos Grey
- Grey Glacier Lake
- Mountain Nordenskjöld
- Paine Park Patagonia
- Pehoe Puerto Natales

☆ [Viewminder, Michele Calabretta](#) and **316 more people** faved this

[View 200 previous comments](#)

- [Kev Hill](#) **PRO** 3y Beautiful shot David
- [Francisco Curbelo](#) 3y Fantástica composición. Nice shot. Ciao.
- [alphawolf\\_2013](#) **PRO** 3y Beautiful shot and wonderful gallery!
- [Dan Sobkowik](#) **PRO** 3y Once again excellent shot, your photo stream a pleasure to view, very nice work.
- [CameliaTWU](#) **PRO** 2y Excellent shot!

**Fig. 2** Image posted in Flickr

we **contribute a data-driven taxonomy for existing Web MIR approaches**. This taxonomy is based on the three main data sources used for feature extraction: *content* data, *explicit context* data and *implicit context* data. We propose an extension to the notion of *context data* defined by Westerveld [72], by subdividing it into *explicit* and *implicit* context. The first, explicit context, considers context as text data that *directly* describe a multimedia object (corresponding to Westerveld's original definition of context, such as: tags, caption, surrounding text, file name, hyperlink anchor text that references the multimedia object). The second, implicit context, which is novel in this work, refers to other context from which the description or relevance of a multimedia object can be *derived*. Such context can be the result of user interaction with the platform where the multimedia object is published or with



**Fig. 3** Video posted in YouTube

the Web MIR system itself. Examples of such implicit context are: social interactions, e.g., comments, likes, shares, favorites, relevance feedback, and queries. The effective extraction of context information can be challenging. There are several issues related to dealing with massive volumes of data (i.e., Big Data), such as noise, varying data modalities and uncertainty. Our second contribution is to **provide a compact, yet extensive literature review of approaches that use context-based data** for Web MIR, including those which use combinations with content information. In addition, we survey benchmarks and datasets used in the MIR. We notice an important lack in the use of standards in Web MIR context-based research community. This, along with very few public datasets, makes reproducibility and comparison difficult, which is necessary for advancing the field.

## 1.2 Our findings

Our work reveals that most research using context data has been focused on explicit context; specifically using text that has been published along with the multimedia object (e.g.: annotations and descriptions). But recently, systems have started to incorporate implicit context generated from user social interactions. These interactions differ from traditional text annotations and include things such as *favorites*, *shares*, *comments*, etc. Furthermore, there is a significant amount of research that incorporates implicit context data, usually in combination with other types of data, producing very interesting and promising results. This indicates that implicit context contains new and valuable information for improving Web MIR that it is widely available in many Web platforms. Context provides an opportunity to enhance semantic information about multimedia objects.

In particular, the large amount of data produced by users on the Web contributes important context information, indicating that social media has the potential to improve MIR systems in an unprecedented manner. Particularly, social media and crowdsourcing are excellent ways to automatically label media and for adding users' experiences into multimedia semantic information.

In addition, the main challenges that arise within the Web MIR research field are the lack of proper benchmarks and standard datasets. These issues make repeatability and comparison of results extremely difficult.

## 1.3 Organization of the survey

The remainder of this survey is organized as follows: Section 2 describes the traditional Information Retrieval system architecture as well as an enhanced version of this architecture for the Web. Section 3 lays out the taxonomy that we use to classify existing work in the field of Web MIR. Section 4 contains a literature review of Web MIR approaches that use context information. In Section 5 we describe a set of public datasets and benchmarking efforts in the area of Web MIR. In Section 6 we discuss main challenges identified for this emerging area, including an in depth analysis of existing benchmarks. Section 7 presents our main conclusions.

## 2 Multimedia information retrieval architecture

As the volume of a repository increases in size, it becomes more and more difficult to manually search for a specific document. To solve this problem information retrieval systems were created, allowing users to search by formulating a query that describes the information that they need. Multimedia objects, in particular, require multimedia-specific information retrieval systems, called MIR systems. Originally, MIR systems were designed to retrieve information based on two types of queries: queries-by-example and queries-by-keyword. Query-by-example uses a multimedia object as an example to query the system in order to retrieve similar objects based on content similarity. Query-by-keyword uses keywords to query the system in order to retrieve multimedia objects whose metadata text matches the query.

The exponential growth of the Web has made users very familiar with query-by-keyword search engines for finding information within Web documents. This has influenced how users search for multimedia, creating a high demand for query-by-keyword MIR systems on the Web as well. MIR systems found on the Web cover a diverse variety of multimedia



content, such as images and video. Query-by-keyword systems match queries to multimedia metadata, like, anchor text, titles, file name, tags. MIR systems that rely on keyword queries do not usually search by content-based similarity. On the other hand, query by example MIR systems in general only use content-based features to retrieve relevant multimedia objects.

In this section we give an overview of the main characteristics of MIR systems and how they have evolved to incorporate the different types of data sources available on the Web. In Section 2.1 we explain the original architecture for basic MIR systems, and, in Section 2.2 we introduce an extended architecture representing requirements of current Web MIR systems.

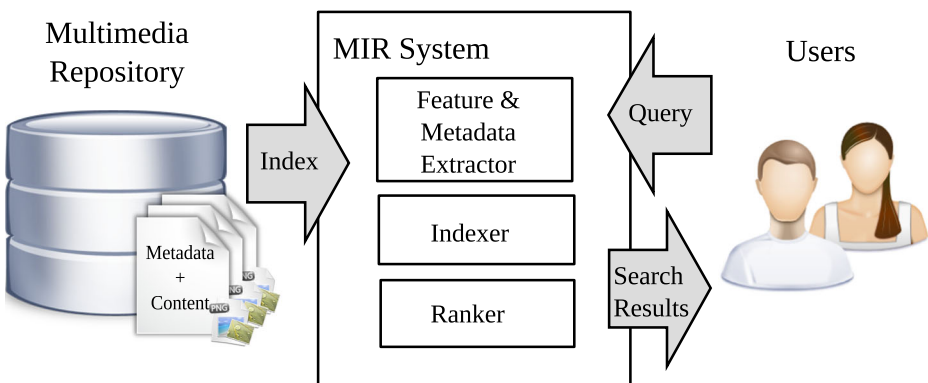
## 2.1 Traditional MIR architecture

Figure 4 describes the architecture of a traditional MIR system. This architecture is based on the model described in Blanken et al. [1], which is centered on the tasks of archiving and retrieving multimedia objects. The archiving process relies on indexing multimedia by their metadata descriptions. The retrieval process is centered on the user interaction with the MIR system in a process of querying and browsing search results. In general, we consider the following to be the main components involved in a traditional MIR process, as depicted in Fig. 4:

**Multimedia repository** Archive of multimedia documents that comprises a single modality (e.g., image, audio) or a specific format of multimedia document (e.g., MPEG-4 videos).

**Multimedia document** A multimedia object that is included in a multimedia repository. It comprises two types of data:

1. **Content**, which represents the raw multimedia content, such as an image, video, or audio clip.
2. **Metadata**, the data that is provided manually or automatically in order to enrich the descriptions of multimedia content. For example, images can often be found stored along with metadata that describes the location where the image was taken (if it is a picture), the camera that was used, the date that was taken, etc. Additionally, metadata



**Fig. 4** Architecture of a traditional MIR system which can incorporate features from multimedia metadata and content

can include human descriptions of the multimedia objects, commonly known as *tags*, which commonly describe objects and concepts represented in the data. This information is usually found as structured text. We emphasize that not all multimedia objects have metadata.

**Feature/metadata extractor** This process is in charge of computing the content-based descriptors and extracting the metadata for each multimedia object in the repository, and for each query object.

**Indexer** This process is responsible for generating a structure, called index, that efficiently retrieves relevant multimedia objects at query time. There are two types of indexes which allow users to search the system using different types of queries:

1. **Index based on content:** This type of index organizes objects according to their content-based similarity, such as color distribution or audio pitch. Search trees or other spatial index structures are often employed to index vector-oriented descriptors.
2. **Index based on metadata:** Indexing by metadata is similar to indexing Web text documents, as it only uses text in multimedia metadata to index multimedia documents. Index structures include B-Trees, Hash tables, or Tries.

**User** It is the entity with an information need related to multimedia content. There are many models of the information seeking process with varying task and user roles (such as the expert user, exploratory searcher and surfer, for more details refer to White et al. [73]).

**Query** It is the representation of an information need. The query may have the same format as the multimedia objects included in the archive or be expressed as text. Examples of types of queries are query-by-keyword and query-by-example (including query-by-sketch, query-by-humming, etc., for more details refer to [15] and [21]).

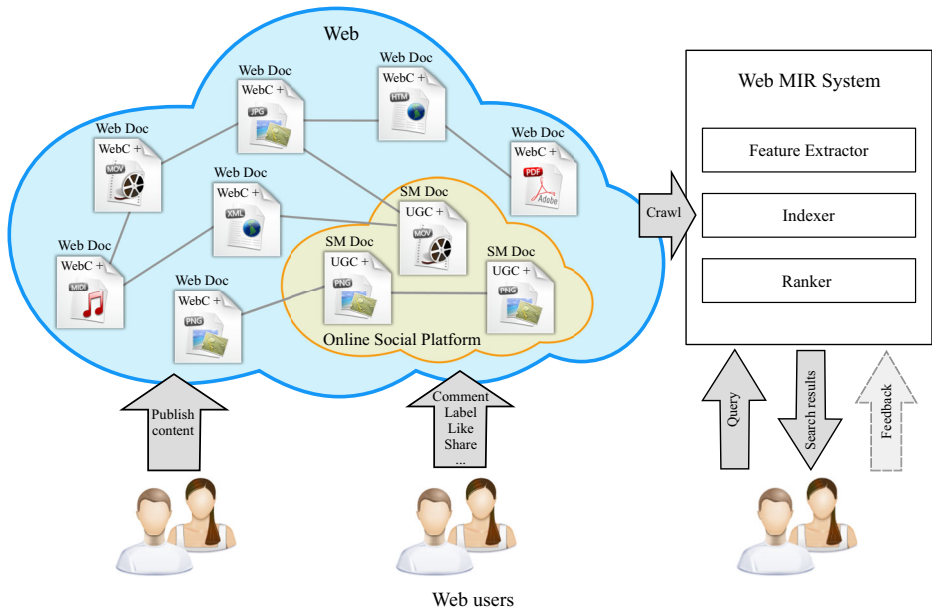
**Ranker** This process is in charge of sorting candidate multimedia objects based on their relevance with respect to the query. In some MIR systems, the ranker module can include user feedback to improve the search results. This means that, based on user actions upon an initial search results list, multimedia objects are re-ranked.

**Search results** List of multimedia documents ranked according to their relevance with respect to the query.

## 2.2 Extended MIR architecture for the web

Traditional MIR systems are typically driven by the data that they *consume*, being the content and text of the documents in their collection. Kherfi et al. [38] survey the main functions of initial Web image retrieval systems. In their survey they describe the usage of text data obtained from resources apart from multimedia content. Lately, with the incorporation of the Web and social media, users have become actively engaged with the retrieval systems themselves as well as with social media platforms. From their interactions, potentially valuable information may be extracted about multimedia content and then added to the functionality of the Web MIR systems. This has brought on a complete new set of components that can potentially be incorporated into MIR systems on the Web. Next, we describe these sources of information related to Web MIR architectures, depicted in Fig. 5.





**Fig. 5** Web Multimedia Information Retrieval Architecture model includes the social/user sphere with edit and share/label operations, among others. Lines between documents represent hyperlinks (other relationships can be mined and exploited as implicit context). Web Documents (*Web Doc*) include the Web Content (*WebC*) and the Multimedia Document (*metatada + multimedia content*). Social Media Documents (*SM Doc*) include User-Generated Content (*UGC*) and the Multimedia Document

**Web** As opposed to the traditional MIR system described in Section 2.1, the multimedia repository in the case of Web MIR systems is the Web. The Web is a multimodal repository, i.e., it contains documents with different multimedia types of content, such as images and videos. Multimedia content is usually embedded within a *Web document* which can provide additional information about the multimedia object itself, called context data (described in Section 3). In addition, the Web is characterized by having a network structure, because Web documents are interconnected to other Web documents with *hyperlinks*. These hyperlinks between documents include *anchor text* which provides information about the document being pointed to.

**Web documents** Documents on the Web can also have structure, for example, the title of a document can be considered more relevant than the content of the document. Also, the caption or surrounding text of multimedia data embedded in Web documents is thought of as being descriptive of the actual (semantic) content of the multimedia object itself.

**Online social platform** These are user-centric platforms that have flourished in the past decade, in which users have become editors, publishers and consumers alike of the content they generate. On a daily basis, online social platforms grow with multimedia data generated by users in real time at extremely high rates. Social data includes different types of information, and is very rich in many ways. For example, user preferences for certain content can be extracted from social data. Also user-generated comments for multimedia content can be exploited to have a better understanding of the information in the multimedia content.

**Feature extractor** Within Web MIR systems, features can be extracted from all of the data available for multimedia objects. This can include the raw content of the multimedia or metadata that is derived in some way from the context provided by the Web. In Section 3 we discuss in detail the different types of data exploited for feature extraction in this scenario.

**Indexer & ranker** These are extensions of those mentioned in Section 2.1, with the difference that the indexing and ranking processes take into account the different types of information found on the Web. This information can be much richer than that found in traditional MIR repositories and in many cases data analysis techniques are applied to enhance the system's performance.

**Web Users:** The Web MIR system interacts with Web users. Web users may have different roles, such as publishers of multimedia content and of providing feedback about the content by their interactions with the Web MIR system or with the platform where the content is published. The role of users is fundamental to Web MIR systems in general.

The main difference with traditional MIR systems and Web MIR systems, are the additional data sources available. Web MIR systems increasingly apply techniques from Machine Learning and Data Mining [13] to extract and process features from the Web structure, content and interactions (see Sections 3 and 4).

While a survey focused on data mining methods used in Web MIR systems is beyond the scope of this work, we mention that in general two methodologies can be applied. Unsupervised methods typically rely on hand-crafted features and similarity functions to compare data objects, based on experimentation and domain experience. Supervised methods rely on training data which is used to automatically determine the features and/or similarity function parameters to compute a ranking. Recently, Deep Learning approaches have shown superior performance in classification of multimedia content with respect to other algorithms, such as Decision Trees or Support Vector Machines (SVM). Caffe [33] is one recent software framework based on deep learning for classification of large image content which could be used to automatically determine candidate visual and context features to classify content.

### 3 A taxonomy for web MIR approaches

In this section, we introduce a taxonomy for Web MIR approaches based on the different types of data that they use. The Web is composed of all sorts of multimedia content such as: images, videos, audio and text. Furthermore, most of the documents found in the Web correspond to combinations of different modalities, which in composition create a single document. For example, an HTML document may consist mostly of semi-structured text, but also include images and video content as well. This multimodal richness that is found on the Web allows us to obtain new information for describing the multimedia content that has been published. As mentioned in Section 1, multimedia raw content is difficult to understand from a semantic point of view (i.e., the semantic gap problem). This limits greatly query-by-keyword MIR systems, which have high demand from users when searching for content on the Web. Therefore, alternative information sources for describing multimedia content on the Web constitute immense possibilities for MIR.

In general, MIR systems use two main data sources for feature extraction, *content* and *context* [72]. The first data source, *content* data, refers directly to the digital content of multimedia objects posted on the Web (i.e., the different raw types of data). Content data feature extraction requires ad-hoc processing for each particular type of raw multimedia content.

This process is usually computationally expensive for non-textual documents. In addition, features extracted from this process depict multimedia objects from a purely computational perspective (low-level features), lacking any relation to possible higher-level human abstractions (semantic features). MIR approaches based on content data have been widely studied, details can be found in the surveys of Lew et al. [44], Data et al. [12], Fu et al. [18] and Hu et al. [29].

The second, less studied type of data source used by MIR systems is *context* data, which is the data found in direct or indirect relation to a multimedia object, but is not part of the raw content itself. Context data is believed to contain information that can be used to describe the contents of the multimedia object that they are related to. The advantage of using context data for feature extraction is that it is usually found in text format and often has been written by humans. Therefore, it can provide descriptive semantic information about multimedia objects as interpreted by a human observer. Context-based data has been originally defined by Westerveld [72] as surrounding text, image caption, image metadata, tags, and other data directly related to the multimedia object (file name, alt text, etc). This type of context has been the basis<sup>5</sup> of many large search engines on the Web's query-by-keyword approach, such as Google Images,<sup>6</sup> Flickr Search,<sup>7</sup> Yahoo! Image Search.<sup>8</sup>

In our survey, we have found that research has extended beyond the context sources described by Westerveld [72]. Therefore, we create a sub-division of the context data available on the Web:

1. **Explicit context data:** This type of data is a generalization of the original definition of context by Westerveld [72] for images. Thus, we define explicit context data as data that can be used to *directly* describe a multimedia object. Examples of explicit context data are: tags and annotations<sup>9</sup> (automatically generated and user generated), captions, surrounding text, file name, hyperlink anchor text that links to the multimedia object, among others. All of this data is found in direct relation to the multimedia object (i.e., tags, caption, hyperlink anchor text) or in the vicinity of the object (i.e., surrounding text). It should be noted that in most cases (except for surrounding text) explicit context data *has been created with the purpose of describing the multimedia object*.
2. **Implicit context data:** This type of data is any type of context (that is not explicit context) from which information about a multimedia object can be inferred. In general, we can say that implicit context data is data from which we can *derive* a description or indication of relevance of a multimedia object. Implicit context data can be the result of user interactions with the platform where the multimedia object is published (i.e., social interactions such likes, comments, shares, favorites) or with the Web MIR system (i.e., relevance feedback, queries, clickthrough data, dwell time). It should be noted that in most cases implicit context data has been provided by users unaware that it will be used as input to describe a multimedia object at some point. In addition, unlike explicit context, implicit context is usually aggregated with other data and intensively processed before being used as input feature extraction.

---

<sup>5</sup>It is clear that these search engines relied heavily on context data in their beginnings, although their technology is proprietary and it is very likely that they are currently using content data too.

<sup>6</sup><http://images.google.com>

<sup>7</sup><https://www.flickr.com/search/>

<sup>8</sup><http://images.yahoo.com/>

<sup>9</sup>We refer to tags, as text associated to a multimedia resource, and annotations as the text associated to only part of the multimedia object (i.e., sub-area of an image, fragment of an audio).

**Taxonomy definition** We create a taxonomy for Web MIR approaches based on the type of data that they use for feature extraction. The three types of data that we have identified are *content data*, *explicit context data* and *implicit context data*. Figure 6 shows all of the possible categories in the taxonomy: *content-based approaches (C)*, *implicit context-based approaches (I)*, *explicit context-based approaches (E)*, *explicit + implicit context-based approaches (E+I)*, *implicit context + content-based approaches (I+C)*, *explicit context + content-based approaches (E+C)*, *(explicit + implicit) context + content based approaches (E+I+C)*. Areas in color indicate the topics that are covered in our survey.

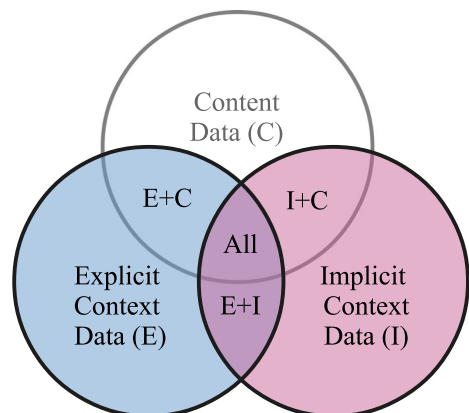
In the following section, Section 4, we classify the most relevant research that makes use of implicit and/or explicit context information. First, we provide an introduction to the approaches for each main category: (I), (E) and (E+I). Inside each category, we classify the most representative methods based on the task that they address: ranking (or re-ranking), classification (or clustering), and labeling (or indexing). For each method, we provide a brief description of its important aspects. We focus our literature review only on context-based approaches, i.e., all the categories of the taxonomy, except for approaches that use content data exclusively (without combining it with context data, category (C)).

#### 4 Survey of web MIR approaches that use context data

We review a number of advanced multimodal Web MIR systems which have recently been proposed. For our survey, we systematically searched for articles related with Web MIR that dealt with multimedia retrieval, multimodality, and explicit/implicit context data. Table 1 presents an overview summary of the full literature review. We organize the summary according to the use of multimedia (abbreviated “MM”) content data: (a) with MM content data (b) without MM content data; the type of context data employed: (a) explicit context data (b) implicit context data (c) explicit and implicit context data; and the type of task performed: (a) ranking/re-ranking (b) classification/clustering (c) indexing/labeling.

To explain the different approaches that we surveyed, we divide the literature into three main sections: approaches based on *explicit context data (E)*, approaches based on *implicit context data (I)*, and approaches based on *explicit and implicit context data (E+I)*. We note that these sections can include approaches which combine context with content data, and cover all different tasks mentioned before. In each section, we provide an overview of the

**Fig. 6** Venn diagram of the proposed taxonomy for organizing existing Web MIR approaches. The taxonomy categories are based on the three groups of data that Web MIR approaches use: content data (C), implicit context data (I), explicit context data (E) and all of the possible combinations between them, i.e., E+I, I+C, E+C, All(E+I+C). Colored areas indicate those that will be covered in this survey



**Table 1** Literature review organized according to the taxonomy proposed in Fig. 6

	Without MM Content Data	With MM Content Data
Explicit Context Data	<i>Task: Ranking/ Re-ranking<sup>a</sup></i>	
	Shen et al. [61] (2000)	La Cascia et al. [5] (1998)
	Yang et al. [76] (2005)	Zhao et al. [79] (2002)
	Popescu and Grefenstette [57] (2011)	Blei and Jordan [2] (2003)
		Wang et al. [70] (2004)
		Gui et al. [23] (2009)
		Chen et al. [8] (2010)
		Xu et al. [75] (2011)
		Tan and Ngo [64] (2011)
		Wang et al. [69] (2011)
	Chen et al. [9] (2013)	
	Gao et al. [20](2013)	
	<i>Task: Classification / Clustering</i>	
		Gao et al. [19] (2005)
		van Leuken et al. [42] (2009)
		Liu and Huet [47] (2013)
	<i>Task: Indexing / Labeling<sup>b</sup></i>	
		Wu et al. [74] (2009)
		Tan et al. [65] (2011)
		Dmitri et al. [54] (2012)
		Li et al. [45] (2012)
		Kannan et al. [37] (2013)
Implicit Context Data	<i>Task: Ranking / Re-ranking</i>	
	Craswell et al. [11] (2007)	Poblete et al. [56] (2010)
	Nie et al. [53] (2012)	Yu et al. [78] (2015)
	Morrison et al. [51] (2013)	
	<i>Task: Classification / Clustering</i>	
	Not found	Not found
	<i>Task: Indexing / Labeling</i>	
	Tsikrika et al. [66] (2011)	
	Leung et al. [43] (2012)	
	Eickhoff et al. [16] (2013)	
Exp. & Imp. Context Data	<i>Task: Ranking / Re-ranking</i>	
	Hanjalic et al. [24] (2012)	Chen et al. [7] (2001)
	Kaminskas et al. [36] (2013)	He et al. [27] (2006)
	Bota et al. [3] (2014)	Jain and Varma [32] (2011)
		Jiang et al. [34](2015)
	<i>Task: Classification / Clustering</i>	
	Hauff and Houben [26] (2012)	
	He et al. [28](2014)	
	<i>Task: Indexing / Labeling</i>	
	Feng and Wang [17] (2012)	Wang et al. [71] (2013)

<sup>a</sup> For more literature on this topic refer to [50]<sup>b</sup> For more literature on this topic refer to [46]

commonalities of different approaches, and we present relevant literature for each case. We group systems (and the corresponding references) according to the input and output data involved in the approach, and give a short description. For each group, we explicitly indicate if they use (E), (I), or both (E+I), and in addition, if content data (C) is used as well.

#### 4.1 Approaches based on explicit context data

Figure 7 highlights approaches which use explicit context data; i.e., surrounding text [5, 70, 79], tags [2, 64] or external resources [57]; to retrieve multimedia objects. Table 1 shows that most context-based approaches use explicit context data. Also, we notice that for tasks, such as classification and indexing content data is employed to improve accuracy on state-of-the-art approaches. This trend is most likely the result of the wide availability and easy access to explicit context data. The Web is full of pages that contain images or videos surrounded by text. In addition, the adoption of multimedia sharing platforms has increased the use of tags and annotations as a tool for organizing user-generated multimedia content. In particular, one of the earliest works in this category is that of Shen et al. [61] proposed to retrieve multimedia documents, specifically images on the Web, by indexing them using the text from HTMLfields in Web pages. Their approach is directed towards answering text-based user queries.

Next, we detail the most relevant approaches in this group:

##### 4.1.1 Ranking-related approaches using explicit context data

**La Cascia et al. [5], Zhao et al. [79], Wang et al. [71] (E+C)**

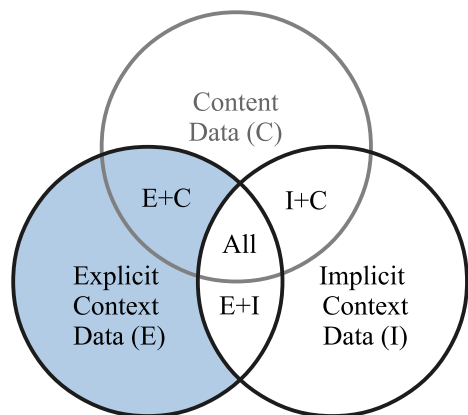
Input: Web docs.

Output: Images ranking

These systems rely on HTML tags to determine which parts of Web pages contain relevant content for the multimedia resources (i.e. images) embedded in Web pages. La Cascia et al. [5] propose to model Web pages as a vector of text extracted from HTML tags using Latent Semantic Indexing (LSI), and then combine this vector with the visual descriptor of the embedded images.

Similarly, Zhao et al. [79] employ LSI to determine the concepts related to keywords in Web pages (specifically, title and body) with images. They propose the *structure anglogram*, which describes the content of a Web document using textual descriptors from its text and

**Fig. 7** Venn Diagram of the proposed taxonomy highlighting the section for proposals that employ explicit context data





visual descriptors from its images. The authors show that the retrieval performance increases when the search system considers the underlying structure of the Web document.

Wang et al. [70] also exploit the hierarchy of HTML tags to determine the relevance of text in a Web page with respect to an image and build a visual hierarchy. Different from the proposal in [5], relevant terms are assigned to specific key image regions. Also, while La Cascia et al. [5] apply the image ranking in the area of content-based image retrieval, the hierarchy built as an outcome of Wang et al. [70] is applied to perform query expansion.

### **Shen et al. [61] (E)**

*Input:* Query, Web doc.

*Output:* Image ranking

Based on empirical observations, Shen et al. [61] propose the use of HTML tags: *page title*, *image title*, *ALT description*, and *image caption* to represent image embedded in Web documents. They also propose a similarity function which takes into account the overlap between the search queries (set of words) and a set of HTML fields read from Web pages. Within the caption field, a differentiated similarity function based on overlap of query terms with (1) a best matching single sentence, (2) an aggregate of two interleaved sentences, and (3) all sentences in the caption is proposed. The similarity computation encodes a prioritization of these fields and the degree of matching within them.

### **Blei and Jordan [2], Tan and Ngo [64], Chen et al. [6], Gao et al. [20] (E+C)**

*Input:* Images/MM doc., tags

*Output:* Image/MM ranking

Blei and Jordan [2] propose adapting Latent Dirichlet Allocation (LDA) to discover correspondences between annotations and images. To achieve this, an image is split into regions and each region is represented with a set of content-based features extracted from it, and a set of annotations previously assigned to the image. Their model, Corr-LDA (Correspondence Latent Dirichlet Allocation), is able to find conditional relationships between sets of words and regions inside an image.

In the same manner, Chen et al. [9] propose a framework that allows for multimedia concept retrieval by means of a multimodal scheme. Their scheme integrates the results obtained using visual features with results using manual annotations. An important feature of their model is that it employs Multiple Correspondence Analysis (MCA) to remove noisy tags, which are not closely related to the concept represented in a multimedia object.

Aside statistical models employed in [2, 9], Tang et al. [64] propose extending PageRank [4] to work on multiple types of heterogeneous resources. The key idea is to build a similarity graph among the documents for each modality (e.g., visual and textual features); and apply PageRank to produce a ranking for each modality, which is later fused in a unique ranking function.

Gao et al. [20] also use a graph representation to combine visual content and tags. The approach is based on a hypergraph model, where images are represented by nodes. Two sets of hyperedges connect images that either share a tag, or a visual feature (from a bag of visual words representation), respectively. From this hypergraph, images are retrieved by a probabilistic learning scheme including hypergraph edge weight estimation.

### **Yang et al. [76] (E)**

*Input:* MM doc., hyperlinks

*Output:* MM ranking

This method takes advantage of the structure of the Web to improve multimedia retrieval performance. They use hyperlink analysis and propose a modified PageRank [4] algorithm to apply to multimedia objects in the Web. In their PageRank version, they replace Web pages with multimedia objects embedded in them as the source and target of hyperlinks.

Since their model does not perform any content-based analysis of multimedia documents, it can be applied to different types of multimedia data.

**Gui et al. [23] (E+C)**

*Input:* Image, tag, Web doc.

*Output:* Image ranking

The system performs a content-based image retrieval using query-by-example. It then learns semantics from the retrieved images, using the tags associated with those images and calculating a score from text-based image retrieval. Finally, it uses a co-search that linearly combines the similarity scores from the content-based and text-based retrieval results.

**Chen et al. [8], Xu et al. [75] (E+C)**

*Input:* Query, image, Web doc.

*Output:* Image ranking

Chen et al. [8] propose a method that combines image features with textual features from surrounding text, using the hypothesis that “items [that] share the same keyword(s) may also share some consistency in selected visual features”. Their system, iLike, first performs a query-by-keyword, and then weighs visual features according to tags associated with the images, giving more weight to visual features that are significant for the keyword. These weights are used for building a new query specification that combines both visual and textual features.

Likewise, Xu et al. [75] consider queries as starting point for their system. Their method aims to enhance image retrieval over unannotated resources in a system based on query-by-keyword requests. The basic idea is to retrieve candidate Web pages for a given keyword image query relying on a given image search engine (e.g., Google image search). From the number of initially retrieved candidates, a number of most relevant images are identified by aligning the query text with context text from the Web pages retrieved. Then, the system performs content-based retrieval over the most appropriate candidate images.

**Popescu and Grefenstette [57] (E)**

*Input:* Query, Wikipedia

*Output:* Image ranking

This method proposes a query expansion model for keywords-based image retrieval, showing improved retrieval performance over the use of non-expanded queries. The key idea is to do query term expansion using Wikipedia, where a weighting scheme is applied for computing conceptual distances between expanded keywords. The methodology was tested on Flickr image data and ImageCLEF benchmarks, showing good performance. Though the retrieval results are images, the approach is inherently driven by text search.

**Wang et al. [70] (E+C)**

*Input:* Query, color map<sup>10</sup>

*Output:* Image ranking

This method proposes using colors to enhance query-by-keyword searches. They introduce *color maps*, which are grids that users can fill with different colors. Each cell of the grid can store a single color, and the spatial distribution of colors in the grid represents the image a user desires. By using color maps, it is possible to specify the spatial distribution of dominant colors and then re-rank search results using this information. The authors state that despite the fact that their method adds a simple representation of color in the search process, it can effectively handle the user intention behind the textual query.

---

<sup>10</sup>It refers to a discrete color distribution manually specified by the user.

#### 4.1.2 Classification-related approaches using explicit context data

##### **Gao et al. [20] (E+C)**

*Input:* Images, Web doc.

*Output:* Image clustering

This method addresses the problem of Web image clustering by combining low-level features and surrounding text. To do this, they proposed the “consistent bipartite graph co-partitioning”, which combines the clustering obtained by visual features and textual features.

##### **van Leuken et al. [42] (E+C)**

*Input:* Query, image

*Output:* Image clustering

This method focuses on clustering images from search result lists in order to diversify results for queries with multiple meanings. They use clustering and dynamically combine visual features for performing the diversification analysis, while obtaining the original image results through a query-by-keyword search.

##### **Liu and Hue [47] (E+C)**

*Input:* Images, tags

*Output:* Event classification

This method addresses the detection of *events*, such as concerts or breaking news, linked to images provided by users in social media channels. The authors train a multimodal classifier and input both user-generated text features (timestamps, location and descriptive tags) as well as image content features.

#### 4.1.3 Labeling-related approaches using explicit context data

##### **Wu et al. [74], Li et al. [45] (E+C)**

*Input:* Images, tags

*Output:* Labeled images

Using labeled MM documents to enriched unlabeled ones allows multimedia systems to improve their answers with respect to user information needs expressed by query-by-keyword. Xu et al. [74] propose using labeled set of images and content-based descriptors for generating a bag-of-words approach for image annotation that is semantic preserving. That is, each *codeword* in the *codebook* tries to minimize the semantic gap, defined as the distance between “semantically identical features”. This allows the system to produce meaningful annotations that can later be used for image retrieval.

Aside bag-of-words models, Li et al. [45] introduces the notion of *bi-concepts*, which corresponds to images that concurrently depict two different concepts. For example, the bi-concept composed of the words “cat” and “flower” stands for images showing both of them. For the concept detection in images, they proposed a multimodal approach that uses tagged images from social networks like Flickr and content-based visual descriptors that estimate the co-occurrence of the two different concepts. They propose to directly detect the bi-concept, as detecting two concepts independently and then combining both results was shown to be ineffective.

Additional, more in depth analysis on the exclusive use of *tags* for labeling multimedia content can be found in the survey by Li et al. [46].

##### **Tan et al. [65] (E+C)**

*Input:* Video, metadata, Wikipedia

*Output:* Visual snippets

This method focuses on linking heterogeneous media types by similarity for creating *visual snippets*. They define visual snippets as cross-linking structures between heterogeneous

media types for browsing and retrieval. To build visual snippets, video shots are clustered by content to form the view of interest. Clusters metadata is also correlated with concepts from Wikipedia entries to group them by topic. Then, the set of clusters is synchronized on a single timeline for exploration which is shown in a network visualization.

**Dmitri et al. [54] (E+C)**

*Input:* Images, Web doc.

*Output:* Labeled face image

This method addresses the problem of labeling face images in Web pages with names extracted from the surrounding Web text. To achieve this goal, the user first manually searches a small number of samples showing portraits of the queried persons. Second, a face detector segments the face images. Third, a multilabel classifier is trained for the annotated images and entity names extracted from the text. The method can robustly label face images in Web pages and be used for querying-by-keyword person images.

**Kannan et al. [37] (E+C)**

*Input:* Images, Web doc.

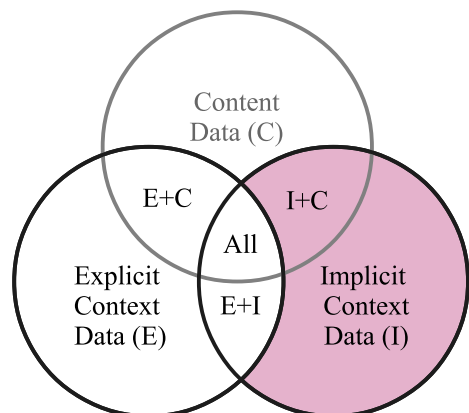
*Output:* Text snippets

This method produces text snippets for images. First, the method finds all Web pages that contain the same or near duplicates of a given image. For this task, the system uses global and local visual descriptors. Second, the method produces a set of candidate text snippets from the text of the Web pages that contain the image (or a near duplicate). Finally, the method performs a top-k snippet selection posed as an optimization problem. The final results are snippets that are both relevant to the images and show diversity.

## 4.2 Approaches based on implicit context data

Approaches in the areas of our taxonomy highlighted in Fig. 8 use implicit context data; i.e., search engine logs [11, 32, 51, 56], or user-generated comments [16]; to retrieve multimedia items. From Table 1 we notice that only few context-based approaches employ implicit context data. This happens because most implicit context data is restricted to proprietors, such as query logs collected by commercial search engines. Also, the richness of implicit context data is for most multimedia content proportionally inverse to its freshness on the Web. An illustrative example of a Web MIR system that uses implicit context data is the approach by Eickhoff et al. [16]. They propose to mine user comments for inferring tags and to index terms of audio-visual data based on the idea of bursts in the comments' timeline.

**Fig. 8** Venn Diagram of the proposed taxonomy highlighting the section for proposals that employ implicit context data



Next, we detail the most relevant approaches in this group:

#### 4.2.1 Approaches for ranking-related tasks that use implicit context data

##### **Craswell et al. [11], Morrison et al. [51] (I)**

Input: Search engine log                      Output: Image ranking

Clickthrough is a valuable source of implicit context data since it represents the behavior of users during searches. Craswell et al. [11] use the information of the click log to produce a ranking of images for a given query. The main idea is to use the query log information to build a click-graph, by linking an image with a query if the user clicked the image after performing that query. The ranking of images is obtained by performing a Markov random walk model over the click-graph. Similarly, Morrison et al. [51] use clickthrough data for modeling hidden topics in the queries, which can be used to improve the retrieval performance of the image search engine.

##### **Poblete et al. [56] (I+C)**

Input: Images, search engine log                      Output: Image ranking

Combining visual descriptors from images and clickthrough data collected from Web search engines are effective at improving image ranking. Poblete et al. [56] introduce the *Visual-Semantic Graph* that is a weighted undirected graph representing the relation between content-based information of images and textual information from their corresponding clickthrough data. Images and queries are represented as nodes in the graph, while the content-based similarity and clickthrough data are represented as edges. The re-ranking process on the visual-semantic graph is based on a random walk.

##### **Nie et al. [53] (I)**

Input: Query                      Output: Image ranking

This work proposes a scheme for Web image reranking from complex text queries. First, the search system obtains an initial image ranking. Second, the system detects the main visual concepts of the complex text query, performing individual searches for each detected concept using different Web search engines and retrieving additional collections of images (one per visual concept). Finally, the method uses a heterogeneous probabilistic network to estimate the relevance of each image on the original ranking, based on the search results from the visual concepts. The heterogeneous network considers semantic, visual, and cross-modality relatedness estimation. The computed relevance scores are used to obtain the final ranking of images. Using a self-collected dataset of complex queries and images from Google Image, the authors show that the proposed reranking approach outperforms random walk re-ranking and pseudo-relevance feedback.

##### **Yu et al. [78] (I+C)**

Input: Query                      Output: Image ranking

This is a supervised method that uses user clicks and visual features for learning to rank, which is used for Web image retrieval. First, the method learns a ranking model using the clickthrough data stored in the search engine log. It then enhances the performance of the model by reducing the noise from the clickthrough data using data from visual features like color, texture, and edge histograms. The main idea of this method is to obtain a ranking model of images that is relevant regarding the clickthrough data and that is visually consistent. An experimental evaluation using a dataset collected from the Microsoft Bing search engine shows that the proposed learning to rank model is both robust and accurate.

#### 4.2.2 Approaches for labeling-related tasks that use implicit context data

##### **Tsikrika et al. [66] (I)**

*Input:* Search engine log

*Output:* Image labels

Clickthrough data might be sparse and noisy, however it can be used to produce good automatic descriptions for multimedia content. Tsikrika et al. [66] propose using clickthrough data for labeling images. The method considers a positive sample collection (images clicked while searching for a specific concept) and a negative sample collection (randomly selected). Given an image, a *surrogate textual description* is built based on all concepts for which the image was clicked. This description is later used with a probabilistic approach for relaxing the matching criteria, so images clicked using different but related concepts can be grouped together. Since the approach is purely text-based it can be generalized to any type of multimedia data.

##### **Leung et al. [43] (I)**

*Input:* Query, search engine log

*Output:* MM doc. index

This method proposes a framework of a search engine that adapts its response based on usage information. In order to achieve the gradual adaptation, Leung et al. [43] apply an evolutionary self-organizing approach, using genetic algorithms. Thus, multimedia objects get linked to search terms representing deeper semantics. They claim this approach is effective for searching multimedia that cannot be directly associated with a specific concept due to the semantic gap. The outcome of their framework includes an index that is updated based on user querying behavior and feedback. The major advantage of this approach with respect to others is its flexibility when introducing new concepts, and the addition of more relevant objects in the repository.

##### **Eickhoff et al. [16] (I)**

*Input:* Videos, comments

*Output:* Labeled videos

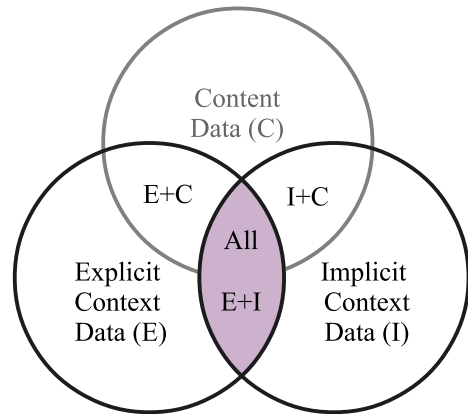
This approach identifies tags by analyzing the comments from users of videos in YouTube. It does not rely on direct annotations of the videos. The method detects first “bursts” in comments, which are short peaks of activity. These selected *bursty* comments are then used to infer the tags. Their model consists of three main stages: (1) using language models to predict relevant tags, (2) representing conversations in comments using time series and (3) reducing noise from predicted tags using an external source of information, such as Wikipedia. They show that comment streams are a suitable source of information to deduce meaningful tags without exploiting other metadata or content *per se*.

### 4.3 Approaches based on explicit and implicit context data

Web MIR systems belonging to this category (area highlighted in Fig. 9) combine both types of context data for the retrieval tasks. The main characteristic of approaches in this category is that they focus on modeling Web multimedia documents using all of the available context, therefore maximizing semantic information extraction. Despite that approaches in this category (E+I) are those that leverage the most amount of context information, and therefore, could stand a better chance towards reducing the semantic gap, there are not many proposals in this category: Chen et al. [7] present a Web image retrieval system that encompasses all of the different data sources in a relevance-feedback oriented architecture. Their proposal combines visual descriptors from the images with text from tags, captions, and other explicit context data. They collect implicit context information from the search engine log for improving image ranking.



**Fig. 9** Venn Diagram of the proposed taxonomy highlighting the section for proposals that employ explicit and implicit context data.



#### 4.3.1 Approaches for ranking-related tasks that use explicit and implicit context data

##### **Chen et al. [9] (E+I+C)**

*Input:* Images, tags, Web doc., search engine log

*Output:* Image ranking

This system crawls the Web gathering images, extracting low-level visual features and high-level semantic features in parallel. Semantic features are obtained from file names, ALT descriptions, and surrounding HTML. High-level features are used to retrieve relevant images when the user submits a text-based query into the system. Relevance feedback is incorporated by allowing the user to provide input on irrelevant images in the search results. Low-level image feature representations are used to find visually similar images to results marked as relevant by the user. The system monitors all user feedback and updates a log of correspondences between query terms, images and user relevance judgments.

##### **He et al. [27] (E+I+C)**

*Input:* Images, query, tags

*Output:* Image ranking

This method focuses on solving the problem of Web image retrieval without adding the user in the loop. Their main idea is to combine text keywords and visual features using query-by-keyword and query-by-example. Their approach uses association rules to link visual features to keywords and keywords to visual features. The ranking is generated using a combined similarity score based on the computed “hybrid” query.

##### **Jain and Varma [32] (E+I+C)**

*Input:* Image, Web doc., search engine log

*Output:* Image ranking

The hypothesis behind Jain and Varma [32] proposal is that highly relevant images to a textual query are also images with a large number of clicks from previous queries. Indeed, the authors propose the method *Click Boosting*, which improves image retrieval effectiveness by re-ranking images based on the number of clicks. However, clicking non-relevant images could add noise to the answer, and relevant unclicked images would never be promoted in the ranking. To overcome these issues, the authors proposed a framework that joins query-independent static features, textual features, and visual features. PCA is applied to reduce the feature vector dimension. A regression model based operating on the projected

feature vectors, based on a Gaussian Process that takes into account the log-normalized clicks, propagates the relevance of the images.

### **Hanjalic et al. [24] (E+I)**

*Input:* Speech transcript, video metadata, SM content

*Output:* Video ranking (intent of a video)

This method detects the “intent” of a video for improving Web video retrieval systems. The intent tries to capture why a user is making a specific video search, e.g. for learning purposes, for searching information, and so on. The main idea is that depending on the user intent, different results should be given for better satisfying the needs of the user. The authors propose a social-Web mining method for discovering classes of “intent”. Also, they propose a learning method based on explicit context data from the video from the video, like the speech transcript and associated metadata, for detecting the intent of a video. The authors test the learning approach by using it for video retrieval, showing that an intent-based ranking of the videos can improve the user satisfaction with respect to their needs.

### **Kaminskas et al. [36] (E+I)**

*Input:* Music, location, tags, DBpedia

*Output:* Music ranking/recommendation

The goal of this approach is to recommend music pieces to users appropriate for a given location. Both music and locations are described by emotion- and physical-related tags obtained from an initial user survey. The manually obtained music tag set is enlarged by a classifier using low-level music features and provided labels as input. For a given location, music pieces are recommended based on tag set similarity and semantic relationships using a rank aggregation function. The semantic relationships are obtained from DBpedia queries relating locations to music, e.g., by composer birthplace or coinciding time frames. A user study shows that the approach is effective and is an improvement over genre-based recommendations. It is also more effective than using only tag sets or semantic relationships alone.

### **Bota et al. [3] (E+I)**

*Input:* Query, Web doc.

*Output:* Composite ranking

The authors address the issue of combining heterogeneous resources to build answers for user needs. They propose to solve this problem by building *bundles* (semantically cohesive groups that include several types of multimedia content). Each bundle is built based on entity relationships between Web documents and their multimedia content. Besides employing the document content, authors also propose to use query intent to increase diversity in search results.

### **Jiang et al. [34] (E+I+C)**

*Input:* Query, videos

*Output:* Video ranking

The authors propose E-Lamp, a semantic video search system. The query is a textual description of semantic aspects of the desired video (e.g., name, definition, and textual description of visual/audio evidence). Content-based audio-visual features are used for high-level concept learning for indexing the videos into four different modalities: speech, text, audio, and visual. The video search system consists of three main steps. The first step translates the original textual query into the four different modalities. The second step performs a multimodal search using the indexed features. The last step uses re-ranking to improve the retrieval performance. An experimental evaluation of the system using the TRECVID Multimedia Event Detection (MED) Test shows that E-Lamp obtains a statistically significant improvement on MAP compared with the best methods from the MED14Test.

### 4.3.2 Approaches for classification-related tasks that use explicit and implicit context data

#### Hauff and Houben [26] (E+I)

*Input:* Images, tags, comments

*Output:* Image location class

This method classifies the geolocation of user-provided image data for cases in which users do not provide appropriate location metadata to their content. Specifically, the authors propose to build a Bayes classifier based on text features obtained from the uploaded image description, and also from text related to this content in social media channels such as Twitter.

#### He et al. [28] (E+I)

*Input:* Audio, comments

*Output:* Audio clustering

This method proposes clustering audio data using text-based content data, as well as user-generated comment data from Last.fm and Yelp on the respective items. A multi-view clustering method is applied on both modalities and applied to a test collection of music items. It is shown that while the independent clustering of two modalities might differ in quality (clustering precision), a non-negative matrix factorization approach can balance for these differences. The merging of both modalities improves the obtainable results in terms of clustering precision, as measured on a groundtruth.

### 4.3.3 Approaches for labeling-related tasks that use explicit and implicit context data

#### Feng and Wang [17] (E+I)

*Input:* MM doc., tags, comments, Wikipedia

*Output:* Labeled MM doc.

This is an approach for recommending appropriate tags to users for annotating multimedia items. The approach supports efficient and effective annotations by users, in turn, helping with heterogeneous retrieval and reducing the semantic gap. The problem is modeled as a graph set up by users, tags, and items. They enrich the explicit relationship between tags and multimedia objects by mining implicit relationships between users who use those tags and the co-occurrence of tags. A graph search strategy based on random walks with restarts is applied. As a result, new tags are recommended to users.

#### Wang et al. [69] (E+I+C)

*Input:* Face img, tags, SM content

*Output:* Labeled face image

This method proposes a multimodal approach for tagging large amounts of faces in the Web. Given a face image as query, they propose a content-based search using several visual descriptors (colors, textures, etc.) for getting a set of similar faces available from a “Web facial image dataset”. This dataset contains facial images and annotations, which are extracted from the Web from sites like Flickr or Facebook. The annotations may be noisy, incomplete or even incorrect, thus they are regarded as *weak labels*. These labels, together with the content-based descriptors, are used to learn an optimal distance metric that refines the annotations and get the final ranked name list.

## 5 Reproducibility in web MIR research

Being able to reproduce and compare existing research on multimodal features for Web MIR is essential for advancing in the field. Important comparison dimensions include the

efficiency (the system resources needed to operate the MIR system), and the effectiveness (the quality of the rankings generated with respect to the user information needs and relevance concepts). While efficiency is rather straightforward to measure empirically, measuring effectiveness is more difficult as it requires models that represent user preferences or similarity concepts.

In many MIR communities, benchmark tests are defined and shared, allowing the comparative analysis of effectiveness of methods. Conferences such as The Text Retrieval Conference (TREC<sup>11</sup>) and The Conference and Labs of the Evaluation Forum (CLEF Initiative<sup>12</sup>) are a good opportunity to interact with huge amounts of non-synthetic data. For example, TREC organizes a Video-specialized Retrieval Conference referred to as TRECVID. Currently, TRECVID datasets are the *de facto* datasets to assess video-oriented retrieval systems. However, only very few video context-based approaches have used these datasets. It is reasonable to expect this to change in upcoming years, as current research trends indicate.

The CLEF initiative has campaigns to evaluate and compare retrieval tasks in multimedia documents. CLEF organizes a track specialized in images, referred to as ImageCLEF. ImageCLEF<sup>13</sup> datasets are of great interest for the Web MIR community. For example, Popescu and Grefenstette [57] assess the performance of their proposal using the ImageCLEF collection. We believe that ImageCLEF datasets are some of the most suitable datasets to extract and exploit Web explicit and implicit context data. However, the use of ImageCLEF datasets has not been consistent through the years, and in the last years it has noticeably decreased. A possible explanation is that larger datasets have been published by other organization, and the access to these new datasets is less restrictive than the ones ImageCLEF provides (usually only to participants of the campaigns). Similarly, the MediaEval initiative<sup>14</sup> has brought together various research groups interested in addressing multimedia retrieval using multimodal approaches that involve user-generated data. The main difference between MediaEval and the rest of the previous initiatives is that MediaEval emphasizes the use of multimodal data, and encourages the use of context data, specially centered on the social and human aspects of multimedia retrieval. The tasks defined by MediaEval are usually different from the ones proposed by other initiatives.

In the latest built and published datasets, we notice a big change over recent years with the influence of social media platforms as a source of data. Currently, there are some datasets obtained from online social networks such as Flickr and last.fm which are rich in explicit and implicit context data. Another source of public multimedia datasets is provided by ACM Multimedia Systems Conference (ACM MMSys) and its dataset archive.<sup>15</sup> This archive contains links to datasets employed and related to articles published between 2011 and 2015 in ACM MMSys. In 3D Object Retrieval, the International Shape Retrieval Contest<sup>16</sup> has provided numerous benchmarks on which 3D similarity search methods have been compared. Table 2 contains descriptions to the main TREC, CLEF and MediaEval initiative datasets, and also to other independent efforts.

---

<sup>11</sup><http://trec.nist.gov>

<sup>12</sup><http://www.clef-initiative.eu>








<sup>13</sup><http://imageclef.org/2016>

<sup>14</sup><http://www.multimediaeval.org>


<sup>15</sup><http://traces.cs.umass.edu/index.php/Mmsys/Mmsys>

<sup>16</sup><http://www.aimatshape.net/event/SHREC>

**Table 2** Datasets for benchmarking - Part 1

Dataset	Size	Content	Data available	
			Implicit Context	Explicit context
<b>TRECVID</b>				
IACC <sup>a</sup>	 7,300 videos	Video files, shot reference		Title, keywords, description
BBC EastEnders	 244 videos	Video files, shot segmentation, face recognition		Metadata embedded in video
BBC Archive	 6,000 hours	Video files, shot segmentation, face recognition		Subtitles, descriptions, UK celebrities
HAVIC	 9,300 hours	Video files		Events related to videos
<b>MediaEval</b>				
Flickr Places [10]	 5M images  25K videos	Image and video files, visual features	Geographic data (lat, lon), text with geo location	
Flickr Images [30]	 90K images	Images files	300 location queries	Related Wikipedia pages
Flickr Events [55]	 500K images	Image files		Title, description, tags, geo coordinates
MusicBrainz music [59]	 13K songs	Audio features	Artist, title, last.fm tags, genre, mood	
Blip 10000 [60]	 14K videos	Video files, shot boundaries and key frames	25,000 tweets and 8,800 Twitter users	Title, description, duration, tags, transcripts
<b>Other resources</b>				
TREC Twitter Dataset <sup>b</sup>	 240M tweets			Tweets posted from Feb. 2013 to Mar. 2013
CLEF Images from Search Engines [68], [22]	 500K images	Image files and visual features	Web pages: (word, source, rank), (word, score)	
MSD <sup>c</sup>	 1M songs	Audio features	Hotness	Title, year, artist, album
NUS-WIDE <sup>d</sup>	 269K images	Visual features	Image URL, tags	

**Table 2** (continued)

Dataset	Size	Content	Data available	
			Implicit Context	Explicit context
MIRFlickr <sup>e</sup>	 1M images	Image files and thumbnails, and visual features	Creator, license, image URL, title, tags, exif	

<sup>a</sup> Internet Archive videos under Creative Commons license

<sup>b</sup> <https://github.com/lintool/twitter-tools/wiki>

<sup>c</sup> <https://aws.amazon.com/es/datasets/million-song-dataset>

<sup>d</sup> <http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm>

<sup>e</sup> <http://press.liacs.nl/mirflickr>

In addition to the initiatives of the multimedia community, major commercial Web search engines (i.e., Google Research,<sup>17</sup> Microsoft Research<sup>18</sup> and Yahoo! Lab<sup>19</sup>) have an area dedicated to research and development of new technologies that lead to improvements of user experience. For instance, Microsoft Research has made various datasets available that combine information from the Web with crowdsourced context data. The main challenges for gathering crowdsourced data are defining human intelligence tasks (a.k.a. *HITS*), and ensuring the quality of data gathered on this process. For Google Research, we notice a bigger effort in building video-oriented datasets. This makes sense in that YouTube is a service owned by Google and has been widely adopted in the Web community. Finally, regarding the variety of services provided by Yahoo!, the most diverse datasets are published by Yahoo! Labs. For example, Yahoo! Labs provides many datasets based on user opinions, specifically ratings. Also, given that Yahoo! owns Flickr, they have built one of the largest image datasets which can be used for many different tasks such as classification, geo-location, automatic tagging, among others. Table 3 summarizes the main datasets provided by commercial Web search engines.

Research in the multimedia community would greatly benefit from the use of benchmarks with a well-defined ground truth as well as standard evaluation metrics, for example the benchmark could include an evaluation software which computes the performance metrics. In this manner, the amount of reproducible research would increase and the comparison between methods would be easier. Beside the ground truth and metrics, it is important to build benchmarks based on public datasets like the ones described before, or publishing all data employed in current Web MIR research in standard formats. In addition, the datasets should be representative of real-scenario problems, in the sense that they should have a large enough size to require automatic approaches, and diverse enough to represent the problem well. Unfortunately, most research in MIR is done on hand-crafted datasets, which are not always available to other research groups. This is sometimes due to the lack of specific information required to boost the methods proposed. Another common reason is that research














<sup>17</sup> <https://research.google.com/research-outreach.html#/research-outreach/research-datasets>

<sup>18</sup> <http://research.microsoft.com/en-US/projects/data-science-initiative/datasets.aspx>

<sup>19</sup> <https://webscope.sandbox.yahoo.com/#datasets>



**Table 3** Datasets for benchmarking - Part 2

Dataset	Size	Content	Data available		
			Implicit Context	Explicit context	
<b>Google Research</b>					
YouTube Video	 120K	Audio-visual features			Title, tags, comments
Games	videos				
YouTube	 1,111				Speaker ID, video URL
Speakers	videos				
Youtube What's Cookin'	 365K videos				Video URL, start/end timestamps, event name, tags
<b>Microsoft Research</b>					
Flickr Visual Annotations [77]	 500 images				Image URL; 100,000 labels for objects in images
YouTube Video Description [6]	 2,000 videos				85,000 descriptions about actions in videos
<b>Yahoo! Lab</b>					
YFCC100M <sup>a</sup>	  100M media objects	Audio-visual features	Comments and favorites can be obtained using Flickr API		Id, user, URL, camera, timestamp, location, title, description, tags
Flickr European Cities (EC1M)	 910K images	Visual features			Image URL, image relevance for 25 queries
Y! Musical Artist Ratings	 10M ratings		10M artist ratings		
Y! Song Ratings	 717M ratings		717M ratings of 136,000 songs		Song, artist, album, genre
Y! Internet Radio Playlists	 4,000 stations		track play, local/system		Radio station, time of play
Y! Movie Ratings	 220K ratings		220,000 ratings of 14,000 movies		Cast, crew, awards, synopsis, genre, avg. rating
TVSum50 [63]	 50 videos	Video files	Shot-level importance scores		Video URL

<sup>a</sup> A subset of 200,000 images from YFCC100M with labels for 10 classes is also available

is conducted within private organizations, therefore in-house data cannot be distributed to people outside these organizations. With this situation, it is difficult if not impossible to properly compare different methods with the state-of-the-art.

## 6 Discussion and research challenges

In this section, we discuss the advantages and difficulties of using context data for Web MIR. We present our perspectives for this area based on the direction of the research we notice in our literature review. In addition, we outline the research challenges that we identify for the area in Section 6.3.

### 6.1 Taxonomy & literature review discussion

**Explicit context data** The use of explicit context data has been very useful for improving the semantic understanding of multimedia content, therefore helping improve Web MIR considerably. Explicit context data is the most widely used type of context data, due to the fact that most of it is publicly available on the Web (e.g., Web document text, hyperlinks and tags). Hence, its access is not limited under proprietary policies. In addition, explicit context data, though it is not always of good quality, is usually directly related to the content of the multimedia object. In this sense, explicit context data is not commonly subjected to any type of quality control, which is very time-consuming and unfeasible at Web scale. Thus, explicit context data is often more useful when combined with content data.

**Implicit context data** Implicit context data is mostly generated by users who do not realize they are producing this type of data. This makes it less susceptible to malicious manipulations (such as spam), and to the user it is effortless to produce, since this process is almost transparent to him/her. Implicit context data, being less prone to manipulation, can be used without necessarily introducing context data. Nevertheless, implicit context data is usually under proprietary repositories and it is not usually accessible to outsiders. Such is the case of query logs and clickthrough data, which are not freely available and mostly owned by large companies. In addition, implicit context data is not always in text form, and in order to be used, some transformations and assumptions must be made (e.g., number of clicks on a link, number of favorites on a picture are assumed to indicate user preference). These direct assumptions are not always reliable and require learning processes for correctly tuning certain parameters.

**Explicit and Implicit context data** The combination of implicit and explicit context data provides a full view of the context of a multimedia object, including how it is perceived by users in terms of relevance and possibly semantics. State-of-the-art work demonstrates that through the combination of contextual data and multimedia content data, the semantic gap can be effectively addressed. With this, we can achieve better semantic understanding of the similarity between multimedia documents, when both types of context data are combined with content data. However, the combination of both types of context data is subject to the availability of each data source, specially to that of implicit context data. In addition, multimedia documents that have been exposed in the Web for a longer time will be associated with more contextual information in general. This will produce biases that might affect Web MIR algorithm outcomes. This is also the case of multimedia objects that are deemed as relevant by Web MIR engines, and which will end up having much more relevance feedback data in the long term than other less relevant objects.

## 6.2 Perspectives

The main perspective that we foresee for Web MIR context-based systems has to do with the importance of leveraging social media data. We detail our observations:

**Emergence of social media as a new modality** The enrichment of Web-based multimedia documents using social media and crowd services has led to a shift in Web MIR system requirements. Naaman [52] analyzes the challenges and new research opportunities related to multimedia content distributed along the Web on social sharing platforms. Previously, the main task was to retrieve objects based on their similarity to a query. Currently, more implicit forms of retrieval, such as recommendation, come into focus. At the same time, the wealth of data generated by users on the social Web adds to the explicit and implicit context that is available, which should be reflected properly in retrieval models.

**Improvement of retrieval with social web data** The availability of social Web data as a new multimedia modality will lead to improved MIR services. For example, Hauff and Houben [26] have shown that taking into account user comments for images can improve the accuracy of automatic geo-location of image visual content, which is a difficult task. Another example is the approach proposed by Liu and Huet [47], where multimodal image analysis provides better event classification results. Both these approaches are examples of how social media data is taken into account to enhance existing content-based analysis techniques. Such approaches are also useful for generating explicit semantic labels (or improving existing labels), therefore leading to a more complete or accurate content retrieval.

**Social web data applications** New applications come into sight with the rise of social Web data. Traditionally, multimedia retrieval tasks searched for similar content on multimedia objects. Based on social annotations, we can now also search for communities of users who share similar preferences or opinions. For example, Kamath and Caverlee [35] apply a community detection algorithm in a stream of text messages, based on the similarity of the topics that users posted. Another social media application is recommendation to new user domains, as shown in Low et al. [48], where they used a cross-domain recommendation approach. The novelty of their recommendation system is that it can transfer recommendations from a given data domain to another one which was previously unknown.

**Reducing the semantic gap** The semantic gap is probably the ultimate problem in multimedia retrieval. Besides content-based media characteristics, similarity and relevance depend on context information, such as time, user and the task. Semantic annotations (based on Ontologies) are seen as a solution for bridging the semantic gap; yet due to the sheer growth of the media volume and the emergence of new multimodal media, it may never be possible to annotate all content in a controlled manner. Haslhofer et al. [25] introduce the Open Annotation Model aimed at supporting user annotations to further allow multimedia document exchange via Web standard protocols. Similarly, Mallik et al. [49] propose Multimedia Web Ontology Language (MOWL) along with a perceptual modeling technique that supports reasoning with multimedia content. Social media information may be the best choice for scaling label generation for multimedia. Furthermore, social data carries *user context*. By combining data about user access to multimedia objects, content of the objects

and user profile, it may eventually become feasible to bridge the semantic gap, or at least, make it smaller.

**Confluence of multimedia retrieval and data analysis** It is now widely accepted that MIR systems increasingly rely on advanced data analysis techniques. Clustering, classification and association algorithms play an important role in preprocessing and mining the available media data for implicit relationships among modalities, users, and preferences. With the use of advanced data mining methods, comes a need to embed these methods within the multimedia architecture, as well as to have tools to store the obtained implicit context information as part of the multimedia database. Dupplaw et al. [14] propose a platform that integrates different state-of-the-art analysis techniques, e.g., fact extraction, opinion mining, sentiment analysis, among others. The main features of their platform is that it is open-source and that the included techniques combine content and context information.

### 6.3 Research challenges

The main challenges that we identify are related to obtaining good quality user-generated data to reliably improve Web MIR systems. We detail these challenges next:

**Assuring the availability of implicit context data** The potential for improvement of MIR systems using implicit context data depends directly on the quality with which this information can be extracted. This is, if implicit context data is mined from social media and user interactions, certainly the quality of user information is critical. For example, user comments on social platforms may be short, ambiguous, or sparse. How to assess the quality and completeness of user data is a challenging problem. There is work that focuses on computing measures such as credibility and trustworthiness of comments in social media [39]. However, this type of assessment is not easy to do, as it often depends on the availability of training data (required for supervised methods), or on effective filtering heuristics. Gamification or other incentive schemes may help provide better user feedback data and more of it, however, these also incur costs for MIR system providers.

**Ownership of social information** The quality of improvements in multimedia retrieval depends on the availability of user-generated data. The question that arises is: Who is this data owned by? And how can researchers outside of large IR companies access it? In the case of user-generated content from social platforms, this data can be obtained publicly with certain bandwidth and sampling limitations (e.g., Twitter, Flickr, etc.). But in the case of user relevance feedback within the MIR system, this data is proprietary to the company.

**Homogeneous label space** This point reflects the need for a unique dictionary of labels for multimedia content. How many different visual object categories exist in the real world? If it were possible to answer that question, we might be able to exhaustively annotate all visual content simply by training an object detector for all these classes (or mining user-generated content to detect all possible classes). However, if the types of visual objects are different per domain, it may be complicated to obtain a globally consistent labeling scheme, which could be interchanged between MIR repositories.

**Evaluation and reproducibility** Benchmarking of solutions for multimodal retrieval is difficult. We observe from the literature that most approaches are assessed using handcrafted

datasets. Ideally, generally accepted benchmark datasets would provide for objective, reproducible results. However, due to the diverse nature of the modalities and their combinations, such consistent benchmarking is hard to do. Development of a taxonomy for the different types of multimodal objects and typical retrieval scenarios may be helpful in guiding the creation of multimodal benchmarks (see also Section 5). Ionescu et al. [31] propose a framework for benchmarking diversification on search results. This framework has been designed based on the tasks proposed by the MediaEval initiative. Similarly, Popescu et al. [58] also address this issue focusing mainly on the potential of the Yahoo Flickr Creative Commons (YFCC) dataset.

We note that the evaluation of MIR systems does not only span retrieval tasks evaluation, but potentially includes criteria such as usability or user experience. Such user-oriented evaluation is also expensive to do. A particular challenge in the UI design area here is, how to design search interfaces which can span complex, multimodal query formulations, but at the same time, are easy and efficient for users. Some studies show that users prefer simple interfaces, maybe as simple as a text query field. How to map the complexities of multimodal retrieval to a minimalist interface remains a challenge.

**Explorative and visual search systems** Finally, we consider the need for explorative and analytic MIR systems. In particular for expert users, it may not suffice to just produce a ranking of multimedia objects to inspect. But users may need to see why objects are retrieved. For example, it may be necessary to know how many users commented positively on a given search result, what their profile characteristics are, and from which parts of the social network they come. An example for this kind of requirement are scientists in search of a data repository for reliable experiment or observation data for synthesizing a new theory, or for validating previous research results. Such aspects may then call for enhanced user interfaces that go beyond result presentation, but include analytical functionality for information experts. Methods from Information Visualization and Visual Analytics may prove useful for designing respective systems.

## 7 Conclusions

Multimedia information retrieval systems are evolving from single-modal, content-based retrieval systems to multimodal, context-based systems which can take into account explicit and implicit context data, provided e.g., from the social media domain, to improve retrieval tasks. We have introduced the topic of context-based approaches for improving Web MIR. We have extended the traditional MIR architecture to explicitly take into account context information, social Web platforms, and analysis algorithms, which are increasingly employed in advanced MIR system. We defined a taxonomy for different types of data used for Web MIR, including implicit and explicit context data, digital content data, and combinations thereof. Along the lines of this taxonomy, we surveyed works focused on Web MIR that use context data in their processes. We found that context data obtained from social media is increasingly being used for improving existing retrieval services. Although there are new data sources that are becoming available, our taxonomy will be able to accommodate new approaches. Besides, social media arises as a valuable source of information, which could help bridge the semantic gap. From the survey, we get an insight on how the Web MIR community is using context data to leverage multimedia systems, as well as the perspectives and challenges this community needs to address in the near future.

Furthermore, future Web MIR approaches are expected to be scalable and quality-aware, and provide effective user interfaces.

**Acknowledgements** This work was partially supported by the Millennium Nucleus Center for Semantic Web Research, Grant No. NC120004. In addition, B. Poblete was also partially supported by Project Enlace-Fondecyt ENL011/16 and Project Fondef ID16—10222. T. Bracamonte was also supported by PhD Scholarship Program of Conicyt, Chile (CONICYT-PCHA/Doctorado Nacional/2013-63130260).

## References

1. Blanken HM, de Vries AP, Blok HE, Feng L (eds) (2007) *Multimedia Retrieval*. Springer, Berlin
2. Blei DM, Jordan MI (2003) *Modeling annotated data*. ACM, New York
3. Bota H, Zhou K, Jose JM, Lalmas M (2014) Composite retrieval of heterogeneous web search. ACM, New York
4. Brin S, Page L (2012) Reprint of: the anatomy of a large-scale hypertextual web search engine. *Comput Netw* 56(18):3825–3833. doi:[10.1016/j.comnet.2012.10.007](https://doi.org/10.1016/j.comnet.2012.10.007)
5. Cascia ML, Sethi S, Sclaroff S (1998) Combining textual and visual cues for content-based image retrieval on the World Wide Web. In: *Proceedings of the IEEE workshop on content-based access of image and video libraries, CBAIVL '98*. IEEE, Washington, p 24
6. Chen DL, Dolan WB (2011) Collecting highly parallel data for paraphrase evaluation. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, HLT '11, vol 1*. Association for Computational Linguistics, Stroudsburg, pp 190–200
7. Chen Z, Wenyin L, Zhang F, Li M, Zhang H (2001) Web mining for web image retrieval. *J Am Soc Inf Sci Tec* 52(10):831–839
8. Chen Y, Yu N, Luo B, Chen X (2010) iLike: integrating visual and textual features for vertical search. In: *Proceedings of the 18th international conference on multimedia, MM '10*. ACM, New York, pp 221–230
9. Chen C, Zhu Q, Lin L, Shyu ML (2013) Web media semantic concept retrieval via tag removal and model fusion. *ACM Trans Intell Syst Technol* 4:61:1–61:22
10. Choi J, Thomee B, Friedland G, Cao L, Ni K, Borth D, Elizalde B, Gottlieb L, Carrano C, Pearce R, Poland D (2014) The placing task: a large-scale geo-estimation challenge for social-media videos and images. In: *Proceedings of the 3rd ACM multimedia workshop on geotagging and its applications in multimedia, geoMM '14*. ACM, New York, pp 27–31. doi:[10.1145/2661118.2661125](https://doi.org/10.1145/2661118.2661125)
11. Craswell N, Szummer M (2007) Random walks on the click graph. In: *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '07*. ACM, New York, pp 239–246
12. Datta R, Joshi D, Li J, Wang J (2008) Image retrieval: ideas, influences, and trends of the new age. *ACM Comput Surv* 40(2):1–60
13. Duda R, Hart P, Stork D (2001) *Pattern classification*. 2nd edn. Wiley
14. Dupplaw DP, Matthews M, Johansson R, Boato G, Costanzo A, Fontani M, Minack E, Demidova E, Blanco R, Griffiths T, Lewis P, Hare J, Moschitti A (2014) Information extraction from multimedia web documents: an open-source platform and testbed. *Int J Multimed Inf Retr* 3(2):97–111. doi:[10.1007/s13735-014-0051-2](https://doi.org/10.1007/s13735-014-0051-2)
15. Egenhofer MJ (1997) Query processing in spatial-query-by-sketch. *J Vis Lang Comput* 8(4):403–424. doi:[10.1006/jvlc.1997.0054](https://doi.org/10.1006/jvlc.1997.0054)
16. Eickhoff C, Li W, de Vries A (2013) Exploiting user comments for audio-visual content indexing and retrieval. In: *Proceedings of the 35th european conference on advances in information retrieval, ECIR '13*. Springer, Berlin, pp 38–49
17. Feng W, Wang J (2012) Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In: *Proceedings of the 18th international conference on knowledge discovery and data mining, KDD '12*. ACM, New York, pp 1276–1284
18. Fu Z, Lu G, Ting KM, Zhang D (2011) A survey of audio-based music classification and annotation. *IEEE Trans Multimedia* 13(2):303–319. doi:[10.1109/TMM.2010.2098858](https://doi.org/10.1109/TMM.2010.2098858)
19. Gao B, Liu TY, Qin T, Zheng X, Cheng QS, Ma WY (2005) Web image clustering by consistent utilization of visual features and surrounding texts. In: *Proceedings 13th annual ACM international conference on multimedia, MM '05*. ACM, New York, pp 112–121
20. Gao Y, Wang M, Zha ZJ, Shen J, Li X, Wu X (2013) Visual-textual joint relevance learning for tag-based social image search. *IEEE Trans Image Process* 22(1):363–376. doi:[10.1109/TIP.2012.2202676](https://doi.org/10.1109/TIP.2012.2202676)

21. Ghias A, Logan J, Chamberlin D, Smith BC (1995) Query by humming: musical information retrieval in an audio database. In: Proceedings of the 3rd international conference on multimedia, MULTIMEDIA '95. ACM, New York, pp 231–236. doi:[10.1145/217279.215273](https://doi.org/10.1145/217279.215273)
22. Gilbert A, Piras L, Wang J, Yan F, Dellandrea E, Gaizauskas R, Villegas M, Mikolajczyk K (2015) Overview of the imageclef 2015 scalable image annotation, localization and sentence generation task. In: CLEF (Online working notes/labs/workshop)
23. Gui C, Liu J, Xu C, Lu H (2009) Web image retrieval via learning semantics of query image. In: Proceedings of the IEEE international conference on multimedia and expo, ICME '09. IEEE, pp 1476–1479
24. Hanjalic A, Kofler C, Larson M (2012) Intent and its discontents: The user at the wheel of the online video search engine. In: Proceedings of the 20th ACM international conference on multimedia, MM '12. ACM, New York, pp 1239–1248. doi:[10.1145/2393347.2396424](https://doi.org/10.1145/2393347.2396424)
25. Haslhofer B, Sanderson R, Simon R, van de Sompel H (2014) Open annotations on multimedia web resources. *Multimed Tool Appl* 70(2):847–867. doi:[10.1007/s11042-012-1098-9](https://doi.org/10.1007/s11042-012-1098-9)
26. Hauff C, Houben GJ (2012) Placing images on the world map: a microblog-based enrichment approach. In: Proceedings of the 35th international conference on research and development in information retrieval, SIGIR '12. ACM, New York, pp 691–700
27. He R, Jin H, Tao W, Sun A (2006) Unifying keywords and visual features within one-step search for web image retrieval. In: Advances in multimedia information processing, PCM '06. Springer, pp 527–536
28. He X, Kan MY, Xie P, Chen X (2014) Comment-based multi-view clustering of web 2.0 items. In: Proceedings of the 23rd international conference on World Wide Web, WWW '14. ACM, New York, pp 771–782
29. Hu W, Xie N, Li L, Zeng X, Maybank S (2011) A survey on visual content-based video indexing and retrieval. *IEEE Trans Syst Man Cybern Part C Appl Rev* 41(6):797–819
30. Ionescu B, Popescu A, Lupu M, Gînsca AL, Müller H (2014) Retrieving diverse social images at mediaeval 2014: challenge, dataset and evaluation. In: Mediaeval 2014 workshop
31. Ionescu B, Popescu A, Radu AL, Müller H (2016) Result diversification in social image retrieval: a benchmarking framework. *Multimed Tool Appl* 75(2):1301–1331. doi:[10.1007/s11042-014-2369-4](https://doi.org/10.1007/s11042-014-2369-4)
32. Jain V, Varma M (2011) Learning to re-rank: query-dependent image re-ranking using click data. In: Proceedings of the 20th international conference on World Wide Web, WWW '11. ACM, New York, pp 277–286
33. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia, MM '14. ACM, New York, pp 675–678. doi:[10.1145/2647868.2654889](https://doi.org/10.1145/2647868.2654889)
34. Jiang L, Yu SI, Meng D, Mitamura T, Hauptmann AG (2015) Bridging the ultimate semantic gap: a semantic search engine for internet videos. In: Proceedings of the 5th ACM on international conference on multimedia retrieval, ICMR '15. ACM, New York, pp 27–34. doi:[10.1145/2671188.2749399](https://doi.org/10.1145/2671188.2749399)
35. Kamath KY, Caverlee J (2012) Content-based crowd retrieval on the real-time web. In: Proceedings of the 21st international conference on information and knowledge management, CIKM '12. ACM, New York, pp 195–204
36. Kaminskas M, Ricci F, Schedl M (2013) Location-aware music recommendation using auto-tagging and hybrid matching. In: Proceedings of the 7th ACM conference on recommender systems, recsys '13. ACM, New York, pp 17–24. doi:[10.1145/2507157.2507180](https://doi.org/10.1145/2507157.2507180)
37. Kannan A, Baker S, Ramnath K, Fiss J, Lin D, Vanderwende L, Ansary R, Kapoor A, Ke Q, Uyttendaele M, Wang XJ, Zhang L (2014) Mining text snippets for images on the web. In: Proceedings of the 20th international conference on knowledge discovery and data mining, KDD '14. ACM, New York, pp 1534–1543
38. Kherfi ML, Ziou D, Bernardi A (2004) Image retrieval from the World Wide Web: issues, techniques, and systems. *ACM Comput Surv* 36(1):35–67. doi:[10.1145/1013208.1013210](https://doi.org/10.1145/1013208.1013210)
39. Kim YA, Ahmad MA (2013) Trust, distrust and lack of confidence of users in online social media-sharing communities. *Knowl-Based Syst* 37:438–450. doi:[10.1016/j.knosys.2012.09.002](https://doi.org/10.1016/j.knosys.2012.09.002)
40. Knees P, Schedl M (2013) A survey of music similarity and recommendation from music context data. *ACM Trans Multimedia Comput Commun Appl* 10(1):2:1–2:21. doi:[10.1145/2542205.2542206](https://doi.org/10.1145/2542205.2542206)
41. Kofler C, Larson M, Hanjalic A (2016) User intent in multimedia search: A survey of the state of the art and future challenges. *ACM Comput Surv* 49(2):36:1–36:37. doi:[10.1145/2954930](https://doi.org/10.1145/2954930)
42. van Leuken RH, Garcia L, Olivares X, van Zwol R (2009) Visual diversification of image search results. In: Proceedings of the 18th international conference on World Wide Web, WWW '09. ACM, New York, pp 341–350
43. Leung CHC, Chan AWS, Milani A, Liu J, Li Y (2012) Intelligent social media indexing and sharing using an adaptive indexing search engine. *ACM Trans Intell Syst Technol* 3(3):47:1–47:27



44. Lew MS, Seve N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: State of the art and challenges. *ACM Comput Surv* 2(1):1–19
45. Li X, Snoek CGM, Worring M, Smeulders AWM (2012) Harvesting social images for bi-concept search. *IEEE Trans Multimedia* 14(4):1091–1104
46. Li X, Uricchio T, Ballan L, Bertini M, Snoek CGM, Bimbo AD (2016) Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Comput Surv* 49(1):14:1–14:39. doi:[10.1145/2906152](https://doi.org/10.1145/2906152)
47. Liu X, Hue B (2013) Heterogeneous features and model selection for event-based media classification. In: *Proceedings of the 3rd ACM conference on international conference on multimedia retrieval, ICMR '13*. ACM, New York, pp 151–158
48. Low Y, Agarwal D, Smola AJ (2011) Multiple domain user personalization. In: *Proceedings of the 17th international conference on knowledge discovery and data mining, KDD '11*. ACM, New York, pp 123–131
49. Mallik A, Ghosh H, Chaudhury S, Harit G (2013) Mowl: An ontology representation language for web-based multimedia applications. *ACM Trans Multimedia Comput Commun Appl* 10(1):8:1–8:21. doi:[10.1145/2542205.2542210](https://doi.org/10.1145/2542205.2542210)
50. Mei T, Rui Y, Li S, Tian Q (2014) Multimedia search reranking: A literature survey. *ACM Comput Surv* 46(3):38:1–38:38. doi:[10.1145/2536798](https://doi.org/10.1145/2536798)
51. Morrison D, Tsirikia T, Hollink V, Vries AP, Bruno É, Marchand-Maillet S (2013) Topic modelling of clickthrough data in image search. *Multimed Tool Appl* 66(3):493–515. doi:[10.1007/s11042-012-1038-8](https://doi.org/10.1007/s11042-012-1038-8)
52. Naaman M (2012) Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. *Multimed Tool Appl* 56(1):9–34. doi:[10.1007/s11042-010-0538-7](https://doi.org/10.1007/s11042-010-0538-7)
53. Nie L, Yan S, Wang M, Hong R, Chua TS (2012) Harvesting visual concepts for image search with complex queries. In: *Proceedings of the 20th ACM international conference on multimedia, MM '12*. ACM, New York, pp 59–68. doi:[10.1145/2393347.2393363](https://doi.org/10.1145/2393347.2393363)
54. Perelman D, Bortnikov E, Lempel R, Sandler R (2012) Lightweight automatic face annotation in media pages. In: *Proceedings of the 21st international conference on World Wide Web, WWW '12*. ACM, New York, pp 939–948
55. Petkos G, Papadopoulos S, Mezaris V, Kompatsiaris Y (2014) Social event detection at mediaeval 2014: challenges, datasets, and evaluation. In: *Mediaeval 2014 workshop*
56. Poblete B, Bustos B, Mendoza M, Barrios JM (2010) Visual-semantic graphs: using queries to reduce the semantic gap in web image retrieval. In: *Proceedings 19th ACM international conference on information and knowledge management (CIKM'10)*. ACM, New York, pp 1553–1556. doi:[10.1145/1871437.1871670](https://doi.org/10.1145/1871437.1871670)
57. Popescu A, Grefenstette G (2011) Social media driven image retrieval. In: *Proceedings of the 1st ACM international conference on multimedia retrieval, ICMR '11*. ACM, New York, pp 33:1–33:8
58. Popescu A, Spyromitros-Xioufis E, Papadopoulos S, Le Borgne H, Kompatsiaris I (2015) Toward an automatic evaluation of retrieval performance with large scale image collections. In: *Proceedings of the 2015 workshop on community-organized multimodal mining: Opportunities for novel solutions, MMCommons '15*. ACM, New York, pp 7–12. doi:[10.1145/2814815.2814819](https://doi.org/10.1145/2814815.2814819)
59. Schedl M, Orio N, Liem CCS, Peeters G (2013) A professionally annotated and enriched multimodal data set on popular music. In: *Proceedings of the 4th multimedia systems conference, MMSys '13*. ACM, New York, pp 78–83. doi:[10.1145/2483977.2483985](https://doi.org/10.1145/2483977.2483985)
60. Schmede S, Xu P, Ferrané I, Eskevich M, Kofler C, Larson MA, Estève Y, Lamel L, Jones GJF, Sikora T (2013) Blip10000: a social video dataset containing spug content for tagging and retrieval. In: *Proceedings of the 4th ACM multimedia systems conference, MMSys '13*. ACM, New York, pp 96–101. doi:[10.1145/2483977.2483988](https://doi.org/10.1145/2483977.2483988)
61. Shen HT, Ooi BC, Tan KL (2000) Giving meanings to WWW images. In: *Proceedings of the 8th international conference on multimedia, MM '00*. ACM, New York, pp 39–47
62. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. *IEEE Trans Pattern Anal Mach Intell* 22(12):1349–1380
63. Song Y, Vallmitjana J, Stent A, Jaimes A (2015) Tvsum: summarizing web videos using titles. In: *IEEE Conference on computer vision and pattern recognition, CVPR '15*. IEEE, pp 5179–5187. doi:[10.1109/CVPR.2015.7299154](https://doi.org/10.1109/CVPR.2015.7299154)
64. Tan HK, Ngo CW (2011) Fusing heterogeneous modalities for video and image re-ranking. In: *Proceedings of the 1st international conference on multimedia retrieval, ICMR '11*. ACM, New York, pp 15:1–15:8

65. Tan S, Ngo CW, Tan HK, Pang L (2011) Cross media hyperlinking for search topic browsing. In: Proceedings of the 19th international conference on multimedia, MM '11. ACM, New York, pp 243–252
66. Tsikrika T, Diou C, de Vries A, Delopoulos A (2011) Reliability and effectiveness of clickthrough data for automatic image annotation. *Multimed Tool Appl* 55(1):27–52. doi:10.1007/s11042-010-0584-1
67. Typke R, Wiering F, Veltkamp RC (2005) A survey of music information retrieval systems. In: Proceedings of the 6th international conference on music information retrieval, ISMIR 2005, pp 153–160
68. Villegas M, Paredes R (2012) Overview of the imageclef 2012 scalable web image annotation task. In: CLEF (Online working notes/labs/workshop)
69. Wang J, Hua XS (2011) Interactive image search by color map. *ACM Trans Intell Syst Technol* 3(1):12:1–12:23
70. Wang XJ, Ma WY, Li X (2004) Data-driven approach for bridging the cognitive gap in image retrieval. In: IEEE International conference on multimedia and expo, ICME '04, vol 3, pp 2231–2234
71. Wang D, Hoi S, Wu P, Zhu J, He Y, Miao C (2013) Learning to name faces: a multimodal learning scheme for search-based face annotation. In: Proceedings of the 36th international conference on research and development in information retrieval, SIGIR '13. ACM, New York, pp 443–452
72. Westerveld T (2000) Image retrieval: Content versus context. In: content-based multimedia information access, RIAO '00, pp 276–284
73. White RW, Roth RA (2009) Exploratory search: beyond the query-response paradigm, vol 1. Morgan & Claypool Publishers, San Rafael
74. Wu L, Hoi S, Yu N (2009) Semantics-preserving bag-of-words models for efficient image annotation. In: Proceedings 1st ACM workshop on large-scale multimedia retrieval and mining, LS-MMRM '09. ACM, New York, pp 19–26
75. Xu S, Jiang H, Lau FCM (2011) Retrieving and ranking unannotated images through collaboratively mining online search results. In: Proceedings of the 20th international conference on information and knowledge management, CIKM '11. ACM, New York, pp 485–494
76. Yang CC, Chan KY (2005) Retrieving multimedia web objects based on pagerank algorithm. In: Special interest tracks and posters of the 14th international conference on World Wide Web, WWW '05. ACM, New York, pp 906–907
77. Yatskar M, Vanderwende L, Zettlemoyer L (2014) See no evil, say no evil: description generation from densely labeled images. *Lexical Comput Semant (\*SEM 2014)*:110
78. Yu J, Tao D, Wang M, Rui Y (2015) Learning to rank using user clicks and visual features for image retrieval. *IEEE Trans Cybern* 4(45):767–779
79. Zhao R, Grosky WI (2002) Narrowing the semantic gap—improved text-based web document retrieval using visual features. *IEEE Trans Multimed* 4(2):189–200



**Teresa Bracamonte** is a Ph.D. candidate at the Department of Computer Science, University of Chile. She is a research assistant of the PRISMA Research Group. Her research interests include Web multimedia retrieval, data mining, and social media analysis.



**Benjamin Bustos** is an Associate Professor at the Department of Computer Science, University of Chile. He is head of the PRISMA Research Group, and he is also a researcher at the Millennium Nucleus Center for Semantic Web Research. He leads research projects in the domain of content-based multimedia information retrieval. His research interests include similarity search, 3D object retrieval, multimedia mining, semantic Web, metric/nonmetric indexing, and pattern recognition. He obtained a doctoral degree in natural sciences from the University of Konstanz, Germany, in 2006.



**Barbara Poblete** is an Assistant Professor at the Computer Science Department [[www.dcc.uchile.cl](http://www.dcc.uchile.cl)] of the University of Chile. She holds a BSc, MEng and MSc from the University of Chile. She received her Ph.D. in October 2009, from the University Pompeu Fabra in Barcelona, Spain, under the supervision of Dr. Ricardo Baeza-Yates and Dr. Myra Spiliopoulou. She worked for 8 years at Yahoo! Labs, first as a long-term Ph.D. Intern in Barcelona and then as a Researcher in the Lab in Santiago. She is currently head of the PRISMA research group and she is also a Young Researcher at the Center for Semantic Web Research [ciws.cl]. Her research interests are in the areas of Web Data Mining, Social Network Analysis, Computational Sociology and Information Retrieval on the Web. She is a PC member of the conferences SIGIR, CIKM, WSDM, ICWSM, ECIR, ICTIR and IAAA and reviewer for the journals TWEB, COMSUR, KAIS, ASOC, among others.



**Tobias Schreck** is a Professor with the Institute for Computer Graphics and Knowledge Visualization at Graz University of Technology, Austria. Between 2011 and 2015, he was an Assistant Professor with the Data Analysis and Visualization Group at University of Konstanz, Germany. Before that, between 2007 and 2011 he was a Postdoc fellow with the Interactive-Graphics Systems Group at TU Darmstadt, Germany. Tobias Schreck obtained a PhD in Computer Science in 2006, and a Master of Science degree in Information Engineering in 2002, both from the University of Konstanz. His research interests include visual-interactive approaches for search and analysis in time-oriented, high-dimensional and 3D object data, with applications in data analysis, digital libraries and cultural heritage.