A survey on Flickr multimedia research challenges

Evaggelos Spyrou
espyrou@iit.demokritos.gr
Computational Intelligence Laboratory (CIL)
Institute of Informatics and Telecommunications
National Center for Scientific Research Demokritos, Athens, Greece

Phivos Mylonas
fmylonas@ionio.gr
Department of Informatics
Ionian University, Corfu, Greece

November 23, 2015

Abstract

Multimedia content sharing within social networks has become one of the most interesting and trending research fields over the last few years. This undoubted emerge of related research works is rather twofold, namely it includes both the analysis and management techniques of the content itself, as well as new ways for its accompanied meaningful interpretation and exploitation. In this paper, we review the recent advances in the above fields in the humanistic framework of the popular Flickr social network. In addition, the major research challenges in the area are demonstrated and discussed, which include current state-of-the-art approaches with respect to interesting humanistic data collection and interpretation research fields, such as multimedia information retrieval, (semi-) automatic tag manipulation, travel applications, semantic knowledge extraction, human activity tracking, as well as related benchmarking efforts. At the end of this survey, we also discuss the main challenges and propose a number of future research directions for interested fellow researchers to continue investigation in the field.

1 Introduction

We all agree that our digital era is characterized and rather dominated by a single, yet very important observation, namely *extremely large amounts of digital multimedia content are being produced everyday and almost instantly shared online by people interacting with others*, within the framework of the social network of their preference. This online social media networking explosion has gone to unprecedented lengths, with people redefining their lives with the aid of social media characteristics. In principle, a social network is considered to be a digital, on-line place, where people create profiles and build a personal network that connects them to other people. During recent years, such networks have emerged into a phenomenon that engages tens of millions of Internet users every day. Moreover, the number of people that use the Internet to share their own generated multimedia content has been continuously increasing. More specifically, 73% of online users having a social networking profile, compared to 8% in 2005, 16% in 2006, and 37% in 2008, respectively [80].

Flickr[1] is an image and video hosting website created by a Vancouver-based company named "Ludicorp" back in 2004 and currently owned by "Yahoo!", an American multinational Internet corporation headquartered in Sunnyvale, California, U.S.A.[2] What makes it special among other multimedia sharing websites is its aspect as an online community, within which users are able to interact by sharing comments about photography and create groups of particular interests. The technology news and media network "The Verge"[3] reported in March 2013 that Flickr had a total of 87 million registered members and 3.5-10 million new photos uploaded daily[164], [65]. Each photo may contain metadata added by its photographer, such as tags that describe either its visual content, its geo-location, or a free text description somehow related to the photo contents (Figure 1). It also contains metadata added by the camera that has been used, such as the actual date

[1] `http://www.flickr.com`

[2] `http://www.yahoo.com`

[3] `http://www.theverge.com`

the photo was taken, specific camera settings, the camera model, etc. Few GPS-enhanced cameras automatically geo-tag the photos they take, but in principal this is still done manually by photographers themselves, after the actual photo has been captured. In principle, textual metadata associated to a photo often serve as a reminder of the context of the image for the photographer and his social circle [112], [147].

Figure 1: Flickr images sample.

The biggest majority of such photos uploaded and shared to the aforementioned popular social network is taken by common users or amateur photographers. The fact that such personal multimedia content has become easy to grasp and capture by rather cheap hardware aided to its mass expansion. In addition, as Van House [147] empirically identified, this content is being used according to four basic humanistic axes, namely: (a) memory, narrative and identity; (b) relationships; (c) self-representation; and (d) self-expression. Eventually, other types of problems arose as well, i.e., it is nowadays acknowledged that user-generated multimedia content is difficult to efficiently get access to, or to be processed and effectively manipulated by humans within a meaningful amount of time and effort spent. As a result tasks have emerged that become day-by-day very difficult and challenging to tackle.

Up to now it is common research knowledge that the overwhelming distribution of such dynamically generated humanistic content over its end users, online communities and consumption devices requires ways for efficient representation and, more importantly, *organization*, in order to be exploited in applications and services. In another empirical study, Zeng and Wei [170] investigated potential relationships between social ties and similarities of the type of digital content that people create and share online. Among their discoveries they found that around the time that a social tie between two individuals is formatted, they begin to create more similar content compared to what they have created before. Interestingly, this similarity tends to evolve in different ways when observing different subgroups of user pairs. In the framework of the latter observation, traditional analysis approaches focusing on analyzing data in terms of objects, and/or concepts and other isolated entities are often quite insufficient, since they do not take into account important properties and relations of online shared multimedia content, the so-called "content metadata". Flickr enables the latter by providing a full set of metadata managing options for its uploaded photos, allowing its users to edit and refine their photo content metadata.

Although in general the application of qualitative and quantitative multimedia content analysis techniques to assess generic metadata records goes back in time and does not advance significantly current research state-of-the-art, it is the nature of today's "instant content capture and sharing" conditions that point out a special interesting role in the process that makes the difference, i.e., the role of a special kind of photo metadata: *geo-tags*. In the dynamic environment of a social network, human behavior and activities are better described and exploited in terms of enriched content metadata. Consequently, geo-tags are considered important for online multimedia analysis and annotation. It is worth noting at this point that the roots of this annotation process lie within the analog photo era, where users wrote some "metadata" information, such as place and date, behind paper photos!

Now, in principle every part of a photo may be tied to a geographic location, but in most typical applications only the position of the photographer is associated with the entire digital photo. As the reader may imagine, this small detail implicates and significantly burdens most humanistic data search and retrieval tasks. In the most typical example, photos of a landmark may have been taken from very different positions apart and in order to identify all photos of this particular landmark within an image database, all photos taken within a reasonable circular distance from it must be considered. Now, when such geo-tagged photos are uploaded to online multimedia content sharing communities like Flickr, which enables the construction of infinite connections among its users [146], a photo can be placed onto a map to view the location the photo was taken. In this way, social network users can browse photos directly from the map, search for photos of a given area, and find related photos of the same place from other users. The aforementioned tasks are considered elementary in order to build additional, ad-hoc value-added digital services on top, like automated route/trip planning or like, to our most recent knowledge, the popular "NOW" app. The latter uses geo-tagged photos to find nearby events happening now[4].

As expected, the act of automatically providing or calculating meaningful photo geo-tags (the so-called "geo-tagging" process) opens a huge research topic for the researchers' community, mainly to the directions of being able to analyze them, identify and determine social patterns amongst them. However, issues of credibility on the volunteered user-generated geo-tagging should become of broader research interest in various areas [38], [135], motivating us to further investigate this topic in the following. At this point, it should be noted that our work clearly differentiates from existing surveys in the field, e.g. the one of Luo et al. [98] and the one of Wu et al. [158]. The former focused on geo-tagged content in general (i.e. originating not only from Flickr but even from other sources) and research within the fields of multimedia and computer vision. Moreover, the latter focused on the social aspect of social media research in general. On the other hand, our survey paper deals with a broader aspect of research using Flickr derived datasets and taking into account the special characteristics of Flickr within the process.

The structure of the rest of our paper is as follows. In the next section 2 we explain in more detail the motivation behind selecting Flickr as the social network under investigation, as well as provide some basic aspects of its functionalities, while in sections 3 - 7 we provide the main details on the research opportunities that exist in the research fields identified. More specifically we start by presenting several approaches in the multimedia content retrieval field (section 3), as well as approaches that focus on the interpretation of multimedia content (section 4), both in the sense of automatic tag/geo-tag generation (subsection 4.1) and of knowledge extraction from Flickr metadata (subsection 4.2). Next section 5 focuses on the humanistic aspect of related applications in a twofold approach, namely by tackling touristic-oriented travel applications (subsection 5.1), and applications dealing with human activity tracking (subsection 5.2). Additional application

[4] `http://techcrunch.com/2013/01/11/now-app/`

domains (section 6) and related benchmarks (section 7) are presented in the following. It is worth noting that these fields form a holistic and complete review approach of the Flickr research community and were imposed de facto by the actual research works that exist in today's literature. Finally, in section 8 we provide a brief overview of remaining challenges and future research directions based on the observations of this survey, whereas in section 9 we briefly summarize our work.

2 Flickr and its Social Aspects

2.1 History and Basic Flickr Functionalities

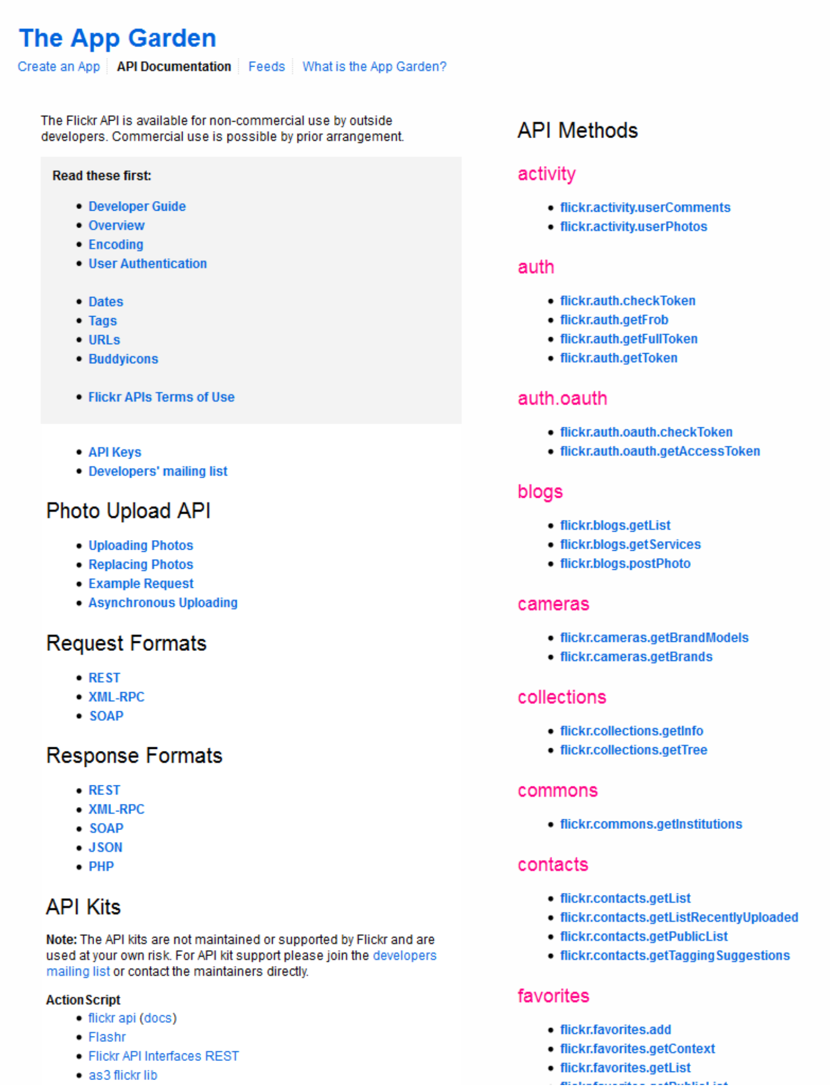
Being the oldest and most popular image and video hosting website for the last decade built around the sharing and organizing of photographs, allowed Flickr to establish a rock-solid position in the social networks era. There exist many photo sharing websites, however, what makes Flickr special, is its interpretation as an online community, within which users are able to interact with others in the following ways: a) they can share comments with other users; b) they can follow other users; and c) they are able to create groups, whose members share the same interests. This intra-user communication has been suggested as the main originality of Flickr [121]. To be more accurate, Flickr focused on the communication among amateur photographers, who although represent the minority in terms of their population within the community, they play the most important role, since they are the main content producers and tend to socialize and encourage new activities. At the basic level, its users may use Flickr as plain online storage for their photos, but at the same time its structure makes it very easy to use tagging as a way of grouping photos by keyword, or into galleries or collections.

Flickr offers also an advanced public Application Programming Interface (API)[5] that enables independent programmers to expand its services and relies on standard HyperText Markup Language (HTML) and Hyper Text Transfer Protocol (HTTP) features, allowing for wide compatibility among platforms and browsers (Figure 2). As a result Flickr has been very popular around the research community during the last few years both for being the largest collection of community collected geo-tagged photos and for offering its public API for accessing these photos along with their textual meta-

[5] https://www.flickr.com/services/api/

data. Thus, the majority of research on community-collected photo metadata and geo-data uses part of its database as a continuously growing test-bench, whose size is larger than the one that the majority of state-of-the-art algorithms is able to handle; more than *14 billion* images have been currently uploaded to Flickr, with most of them containing user generated textual and location metadata.

Figure 2: Flickr API.



The App Garden

Create an App | API Documentation | Feeds | What is the App Garden?

The Flickr API is available for non-commercial use by outside developers. Commercial use is possible by prior arrangement.

Read these first:

- Developer Guide
- Overview
- Encoding
- User Authentication

- Dates
- Tags
- URLs
- Buddyicons

- Flickr APIs Terms of Use

- API Keys
- Developers' mailing list

Photo Upload API

- Uploading Photos
- Replacing Photos
- Example Request
- Asynchronous Uploading

Request Formats

- REST
- XML-RPC
- SOAP

Response Formats

- REST
- XML-RPC
- SOAP
- JSON
- PHP

API Kits

Note: The API kits are not maintained or supported by Flickr and are used at your own risk. For API kit support please join the developers mailing list or contact the maintainers directly.

ActionScript

- flickr api (docs)
- Flashr
- Flickr API Interfaces REST
- as3 flickr lib

API Methods

activity

- flickr.activity.userComments
- flickr.activity.userPhotos

auth

- flickr.auth.checkToken
- flickr.auth.getFrob
- flickr.auth.getFullToken
- flickr.auth.getToken

auth.oauth

- flickr.auth.oauth.checkToken
- flickr.auth.oauth.getAccessToken

blogs

- flickr.blogs.getList
- flickr.blogs.getServices
- flickr.blogs.postPhoto

cameras

- flickr.cameras.getBrandModels
- flickr.cameras.getBrands

collections

- flickr.collections.getInfo
- flickr.collections.getTree

commons

- flickr.commons.getInstitutions

contacts

- flickr.contacts.getList
- flickr.contacts.getListRecentlyUploaded
- flickr.contacts.getPublicList
- flickr.contacts.getTaggingSuggestions

favorites

- flickr.favorites.add
- flickr.favorites.getContext
- flickr.favorites.getList
- flickr.favorites.getPublicList

2.2 Flickr Sharing and Tagging Photos

The act of adding human understandable, descriptive keywords to photos is generally denoted as “tagging”. It is notable that the vast majority of uploaded photos on Flickr has been tagged with a few descriptive keywords, although these may differ significantly among “consumers” and “producers” of digital content [25]. But why do people upload in common view and moreover tag their photos? Angus and Thelwall [6] investigated the main motivations for using Flickr and for tagging photos. Their sample users originated from the United States of America and Denmark and their research indicated that users employ Flickr both as a personal archive and also as a means of sharing photos with friends and family. However, reasons for tagging also lie in the social organization and communication possibilities that Flickr offers. There is much debate concerning the aforementioned reasons for tagging. These reasons have also been categorized by Marlow et al. [101] and Hammond et al. [47] as “organisational/social” and “selfish/altruistic”, respectively. Within the findings of Angus and Thelwall we should also notice that Flickr users do not tend to mix these reasons, but rather adopt one tagging strategy over the other. They conclude that Flickr is treated more as a virtual community and less as a website for commercial gains.

Moreover, Flickr users often interact on purpose, by commenting and asking questions about uploaded photos, as it has been shown by Canningham and Mahui [31]. Additionally, the research of Nov et al. [111] indicated that self and public motivations for tagging are highly correlated with the quantity of tags generated by a user. These motivations include an organizational and a communicational aspect, with the first dealing with the facilitation of photo retrieval and the latter with providing the best detailed description to other users. Finally, Ames and Naaman [4] observed that users may often be encouraged to tag by external factors such as point-of-capture annotation (e.g., on the mobile device) or by tag suggestions. They concluded that it is more likely that users would annotate their photos when they are given certain motivations and affordances.

The procedure of tagging photos within Flickr is towards the aforementioned way. Users are able to tag their photos as part of the uploading process or when they view them. It is also possible to allow their photos to be tagged by other users. An important question herein deals with the accuracy of user-provided textual metadata on Flickr. To this goal, Winget [157] made a study towards the “correct” way of organizing information on the web, with case studies derived from Flickr. Her main conclusions were that users intend on providing accurate textual descriptions, however since these appear rather arbitrary, there is the need of organizing them with structured vocabularies. Moreover, in their survey, Wang et al. [154] concluded that the main open challenges in tagging are the involvement of a large number of humans in the process and the automatic (i.e., computer generated) tagging in large scale collections.

3 Multimedia Content Retrieval Approaches

It should be obvious by now that Flickr is mainly a digital photo sharing website. As a natural result it drew the attention and interest of the image retrieval research community. Challenges derived from Information Retrieval, in general, and multimedia content retrieval, in particular, found prosperous field and datasets to exploit in this new medium. Given the experience and the background of researchers involved, the main approach followed is to use either textual metadata or visual properties of photos and often combine them in an effort to improve the accuracy of their respective algorithms. As depicted in the following, such research efforts vary from textual ones, to ones based on low-level visual, or even hybrid, characteristics.

3.1 Text Retrieval

We ‘ve discussed that the recent growth of social networking services and multimedia content sharing and bookmarking sites boosted the popularity of tagging. Following a formal notation for the latter, one may claim that it allows users to create and manage labels (i.e., “tags”) that categorize associated content using simple keywords; these labels are in principle consisted of non-hierarchical keywords or terms assigned to a piece of information. As a result there exist many approaches, typically earlier, that use a text retrieval-based approach, i.e., the information they use is in the form of text. The fact that images in Flickr are always accompanied by text, in the form of both additional descriptive content and metadata, could be used to accurately identify and match textual queries. In addition, as the amount of user-contributed textual data is growing every day (e.g., by means of comments, reviews, ratings, posts, tagged photos, etc.), and as many of those contributions also include geographical coordinates, there is a vast amount of textual information available for automated mining of geographical and other types of knowledge.

In a first attempt, Ahern et al. [3] analyzed tags associated with georeferenced Flickr images so as to generate knowledge. This knowledge was a set of the most "representative" tags for a specific geographical area, occurring after a TF-IDF approach. They also presented a visualization tool, namely the *World Explorer*, which allowed users explore their results. Among the findings of their qualitative evaluation approach, we should emphasize that users often preferred more level of detail than the one offered by the system, while they generally found data aggregation useful. Their preference was based upon the task they wished to perform. Lerman et al. [82] aimed to personalize text-based search results by adding information about users' relations. They claimed that user preferences may be reflected at their contacts and also that adding such info when searching may filter results in a way that precision is improved. Moreover, they also presented a probabilistic model, which takes advantage of tags in order to mine latent topics in results.

Abbasi et al. [1] aimed to identify landmarks using tags and Flickr groups. They did not exploit geospatial information, as they claimed that the amount of GPS data were not sufficient at that time. They used SVM classifiers, which had been trained on thematical Flickr groups, in order to find relevant landmark-related tags. Their method outperformed state-of-the-art methods that relied on geotagged data and may be also applied in other type of concepts. In a similar context, the work of Serdyukov et al. [128] aimed to place photos on a map, i.e., predict the location they have been taken. To this goal, they rely solely on user collected Flickr metadata and aimed to annotate an image according to these metadata. Their goal was to develop a language model based on the tags users adopt at a specific place, and use this to provide an automatic alternative to manual geo-tagging. Larson et al. [77] detected whether tags correspond to physical objects, and also the scale of these objects, using a natural language approach. They worked on MIRFLICKR[56] data set, which consists of Flickr photos, manually annotated with a crowd-source approach and without the participation of the original photographer.

Following Table 1 provides an overview of the main characteristics of the discussed text-based approaches, focusing on their tasks, features utilized, existence or absence of geo-information, the size of the utilized datasets, as well as their origin. We may observe that the dominating textual feature tackled by all five approaches is "tags", whereas only 40% of them utilize geographical information. The largest dataset is by far the one utilized by [3], whereas a clear diversion is to be observed with respect to the origin of each work's data. Allowing us to go one step further on the analysis of multimedia content retrieval approaches, in the next subsection 3.2 we shall present research works that deal with the visual properties of photos towards efficient content retrieval.

Table 1: Text retrieval

Work	Task(s)	Features		Geo	Size of Data	Origin of Data
		Text	Visual			
[3]	show high-scoring tags on a map (visualization)	location, user, tags		Yes	collected:6M used:4.5M	
[82]	improve search results using user metadata	id, tags, groups		No	13.5K	3 queries, top 100 results
[1]	identify landmark photos rank relevant tags	tags, groups		No	train: 430K test: 232K	50 cities (EU, Asia)
[128]	placing flickr uploaded photos on a map	id, tags		Yes	collected: 400K used: 140K	
[77]	determinate real-world size of tags	tags		No	5K images 500 train 4.5K test	MIRFLICKR[56] dataset

3.2 Visual Retrieval

Traditional visual retrieval and classification problems are all about sufficient visual content representation, as well as the definition of a meaningful metric to measure and quantify similarities or dissimilarities between such content. In this process it is well-acknowledged by researchers that high sensitivity to low-level visual content representing quite different high-level concepts and invariability to visual data that are perceptually alike or belong to the same class lie among the main characteristics of a good representation. Robustness to challenging features, such as geometry or scale invariability are typically tackled, as well the presence of noise in the analysis process. Interesting advances in the field include reducing the set of low-level features considered and learning of the similarity measure directly from a training set. In this category of research, typical visual descriptors include SIFT [94], SURF [13], GIST [114] and Harris [51]. The well-known "Bag-of-Words" (BoW) approach [30] is also typically adopted.

Wang et al.[152] proposed a training algorithm, based on a fast Stochastic Intersection Kernel Machine, which as they claimed, was able to get trained with tens of thousands of examples in a few minutes. They used this algorithm to predict image similarity to a Flickr group. They assumed that two images are considered similar if they belong to the same group. Their results indicated that their approach was able to measure image similarity better than using visual features as a means of similarity/dissimilarity. Yanai et al. ([160], [161]) focused on the analysis of the relationship between words and locations. They used visual features and tried to associate them with certain locations, using an entropy-based approach. The application of their approach was the selection of representative photographs for certain geographical regions, which then helped them to detect cultural differences among different countries and/or locations, e.g., what is considered as a "castle" in different cultures/countries. Avrithis et al. [7] proposed an image clustering scheme that, seen as vector quantization, compressed a large corpus of images by grouping visually consistent ones, while providing a guaranteed distortion bound. This way, they were able to represent the visual content of all thousands of images depicting, e.g., the "Parthenon", in just a few dozens of "scene maps" and were still able to retrieve any single, isolated, non-landmark image, e.g., a graffiti on a wall. They used a geotagged dataset, grouped images geographically and then visually. This way each visual cluster depicted different views of the same scene. All views were then aligned to one reference image and a 2D scene map was constructed by preserving details from all images while discarding repeating visual features. Their indexing, retrieval and spatial matching scheme operated directly on scene maps. Batko et al. [9] presented an experimental CBIR system which used MPEG-7 visual features and developed a set of indexing and searching algorithms search into a set of over 50M photos from Flickr. Their main contribution lied on the technology that underlies the centralized and distributed structures they developed, at an effort to deal with such a large data set. Liu et al. [92] incorporated the social aspect of photos, in order to re-rank search results according to both social and visual relevances, in an effort to produce results that are closer to the users' real intentions. They proposed a novel algorithm, implemented over a hybrid graph, which resulted by the combination of social and visual links.

Joshi and Luo [66] used pre-trained visual detectors of small neighborhoods, incorporated bags-of-geotags within a probabilistic framework and observed the statistical coherence of descriptions. Their goal was to detect certain activities and events in photos. Each category was studied independently to others, since their goal was to separate negative and positive samples. They observed that performance was proportional to the quality of the visual detectors. Luo et al. [97] fused information extracted from both a Flickr data set and a set of satellite images, in order to detect events. The latter were considered as a "third eye" from above. They also made use of both color- and structure-based visual dictionaries and used machine learning approaches to create image classifiers. The aforementioned fusion lead to a significant improvement of results. Philbin and Zisserman [116] focused on the problem of grouping images based on the object they contain. When dealing with very large data sets (>1M photos), this problem appears very challenging, due to significant changes in scale and viewpoint, partial occlusions, the scale of the data and the extreme variation in imaging conditions. They created a matching graph based on visual features, using an approximation of the K-means algorithm. They applied their approach so as to automatically find frequently occurring objects in cities.

Yu and Luo [163] combined visual context with location information in order to detect region-based concepts in photos. The former type of context was based on the analysis of the spatial relationships among the depicted "objects", while the latter on a non precise estimation of the location where pictures had been taken. Both the aforementioned types were then fused using a probabilistic graphical model. Their experimental results indicated that the addition of the location information significantly improved the accuracy of object recognition. Li et al. [88] used SVMs trained on visual features to classify a 30M data set into 500 categories. Their results indicated that such an approach may lead to results comparable to those of humans. They also showed that when a structured SVM is applied in a stream of photos of a single photographer, i.e., when taking into account the temporal contextual information which is provided by Flickr, a dramatic improvement into precision may be achieved. Chatzilari et al. [19] used region level annotations and visual features in an effort to recognize objects with a semi-supervised approach. They started from a set of Flickr photos that contained the same concept, they segmented these photos and from the occurring regions they clustered their feature vectors. They assumed that ideally the most populated cluster should be consisted mainly of regions containing the concept at hand. Their results indicated that the quality of the selected regions was inferior to the optimal, i.e., manual selection, and leads to some cases where the gain in effort compensated for performance loss. Wang et al. [156] proposed the use of "multi-query expansion". More specifically, by extending Latent Dirichlet Allocation, they used the top photos regarding a query of a landmark. Then, they generated a compact pattern set from these photos. They demonstrated an increase in accuracy. Finally, Lu et al. [96] presented a refinement algorithm for social image parsing, i.e. segmenting and identifying each object of a given image. They did not use local (per image segment) annotation, but exploited user-generated tags. They faced the aforementioned problem as a cross-modal data refinement problem and provided promising results.

Table 2 provides an overview of the main characteristics of 13 visual-based approaches, focusing on their core tasks, the nature of features utilized, the existence or absence of geo-information, the size of the utilized datasets, as well as their origin. The reader may observe that 76.92% of them (10/13) utilize both textual and visual information to achieve better results. The most popular set of visual descriptors is SIFT (61.54%), whereas the majority of the research works exploit geo-information in the process (8/13). The largest dataset is the one utilized by [9], followed by the dataset of [88]. The diversity of originating data is again to be noted, making the provision of a large, commonly accepted, comparable dataset an interesting future idea. In the next subsection 3.3 we present the hybrid approach, which incorporates textual metadata in the visual retrieval task.

Table 2: Visual retrieval

Work	**Task(s)**	**Features**		**Geo**	**Size of Data**	**Origin of Data**
		Text	**Visual**			
[152]	learn image similarity from groups	group name	SIFT+BoW, GIST, RGB, gradient Fusion	No	approx. 200K(est.) 38K from Corel (test)	103 Flickr groups
[160], [161]	relationships analysis between real world concepts & geographical locations, mining cultural differences	tags	SIFT+BoW	Yes	230*500	500 noun concepts
[7]	image retrieval using "scene maps"		SURF	Yes	approx. 1M	22 European Cities
[9]	building a CBIR system for web-based data sets	title, description ID, location, tags,comments		Yes	50M	
[92]	image search social visual re-ranking combined social/visual factor	social: group similarity	SIFT+BoW	No	30K	30 queries, e.g., "apple, jaguar, golf, scenery" etc.
[116]	grouping of photos containing the same object, using a matching graph		SIFT, shape, position	No	5K+37K+approx. 1M	Oxford Buildings statue of Liberty Rome
[66]	inferring activities and events	tags	visual detectors	Yes		queries of visual concepts
[163]	concept detection		color, texture	Yes	3K	from bounding boxes "Northeast US", "Florida-Caribbean"
[97]	ground & satellite photos event recognition	tags	color SIFT+BoW	Yes	853+981+720	satellite images 12 queries, Kodak set
[88]	landmark classification	tags	SIFT	Yes	30M	
[19]	extraction of semantically coherent groups of regions, depicting the same objects	groups	SIFT+BoW	No	500*15 Flickr 20K non-Flickr	15 concepts of a publicly available data set
[156]	landmark retrieval	tags	SIFT	Yes		landmarks from 11 cities
[96]	region-level object recognition	tags	color, texture	No	632	VOC2007

3.3 Hybrid Retrieval

In the recent years several research groups attempted to incorporate textual metadata in the visual retrieval task. This approach is based on the fact that an enhanced description of the visual content itself may be provided within its accompanying information. As a result such research efforts try to combine visual content descriptions with textual metadata. Crandall et al [29] used visual, temporal and geospatial information to automatically identify places and/or events at the city and landmark level. They also added temporal metadata information to improve classification performance. They worked with a data set of about 35M photos and extracted spatial relations among photos taken at popular places and concluded that the inclusion of visual and temporal features significantly improves prediction of locations of photos. Gammeter et al. [42] overlaid a geospatial grid over earth and matched pairwise retrieved photos of each tile using visual features. Then they clustered photos into groups of images depicting the same scene. The metadata were used to label these clusters automatically, using a TF-IDF scheme. The proposed method allows for automatic labeling of certain types of objects, e.g., landmark buildings, scenes, pieces of art etc. using holiday snap photos. Its scalability is demonstrated on experiments on millions of images. Moxley et al. [106] classified geo-referenced tags as places, by extending [124]. They also classified landmarks by clustering image datasets considering mutual information and prior knowledge from Wikipedia and visual terms using the mutual information between visual descriptors and tags. They organized their data set using a quadtree structure. Ulges et al. [143] adopted a context-based approach, which assumed that users place semantically similar photos in the same Flickr group. For example a user is likely to create a group and therein upload his/her photos of the same vacation trip. These "batches" were then matched with learned categories and annotated. This way they were able to learn context categories and significantly improved the accuracy when compared to the annotations of individual images. Liu et al. [91] were the first to consider user uploading patterns, geotagging behaviors, and the relationship between the temporal and the spatial gap of two photos from the same user, in order to predict geotags for given photos. They showed that the temporal gaps between the image to be geotagged and historical images are very important for geotagging.

Li et al [84] observed that tagging by amateur photographers is in general uncontrolled, ambiguous, and personalized, thus recognized the problems of unreliable interpretation and linking to visual features of such tags. They also assumed that when different users use the same tags to visually similar images, these tags should semantically reflect the visual content. They proposed a scalable algorithm that learned tag relevance by voting from visually similar neighbors. They did not use geospatial data, nor limited their approach on landmarks/places of interest and aimed to retrieve semantically similar images. Using tag relevance they were able to significantly improve retrieval performance. Barrios et al. [11] presented an image retrieval system that combined textual and visual content. They downloaded and stored locally images from Flickr and used simple color and texture visual descriptors, along with the title, description and tags, for each image. With the same motivation of [29], Quack et al. [123] mined images of landmarks. To this goal, they divided the area of interest into non-overlapping, square tiles, then extracted and used visual, textual and geospatial features. They handled tags by a modified TF-IDF ranking and linked the extracted objects and events to Wikipedia[6]. Their fully unsupervised approach concluded with a verification step, where the Wikipedia article content (both textual and visual) was used for verification. Their approach was applied on urban area photos.

Simon et al. [130] created visual summaries of large image data set based mainly on visual features, but also exploiting tags. They worked on the distribution of images in a collection and aimed to select a set of canonical views to form the scene summary. To this goal they applied clustering techniques on visual features. They showed that although tags are noisy, their role in the construction of summaries is essential. Kennedy and Naaman [74] used visual features and tags, in order to extract the most representative tags and views for landmarks. Their approach was unsupervised and scalable. For its evaluation, they worked on a corpus of 110K Flickr photos from San Fransisco and showed that meaningful and representative tags for locations/landmarks can be mined from frequent tags. Moëllic et al [104] aimed to extract meaningful and representative clusters from large-scale image collections. They proposed a method based on a shared nearest neighbors approach that treats both visual features and tags. They showed that their method was able to provide useful representations of an image set, in terms of representative clusters. Fan et al. [36] proposed a system, namely *JustClick*, which exploited both visual and textual information and after a search and retrieval process, it recommended photos using an interactive interface. To this goal, they ap-

[6] `http://www.wikipedia.org`

plied kernel principal component analysis. Hyperbolic visualization was used to organize and layout the recommendations and users were able to assess the relevance to their initial query intentions. Seah et al. [127] created visual summaries on the results of visual queries on a data set of Flickr images that in contrast to previous works, e.g., the one of [104], attempted to generate concept-preserving summaries. Their method exploited both visual features and tags. The generated summaries aim to maximize the coverage of search results, which were organized into visually and semantically coherent clusters. Qian et al. [122] presented a user-oriented approach for ranking the results of tag-based search, using color and semantic features. They used inter- and intra-user ranking and showed that this way, ranking improved significantly. Finally, Xu et al. [168] presented an approach for discovering latent subspaces shared by multiple features, using a Gaussian process. They integrated the large-margin principle in the Gaussian process and ended up with an effective method for high-dimensional spaces.

Table 3 provides an overview of the main tasks of the aforementioned 15 hybrid retrieval approaches. Their core tasks vary from photo organization/clustering to photos search and retrieval and automatic learning of tag semantics. 11/15 research works value textual "tags" as one of their main sources of information, whereas six of those eleven approaches (i.e., 40.00% of all approaches) combines them with popular SIFT features. Interestingly enough, 10/15 works (66.67%) do not utilize geo-information in the process and base their findings solely on the hybrid analysis. In terms of utilized datasets, the largest dataset is clearly the one discussed in [36], comprising of 1.5 Billion images, followed by a couple works utilizing datasets in the scale of millions, namely: [29], [42], [84], [106], and [122].

Table 3: Hybrid retrieval

Work	Task(s)	Features		Geo	Size of Data	Origin of Data
		Text	Visual			
[29]	organizing a large collection of photos	tags, temporal	SIFT+BoW	Yes	approx. 33M	
[123]	mining images of touristic sites	tags, title, description, ID, time	SURF+BoW	Yes	220K	9 European cities
[42]	automatic tagging of photos referring to landmarks drawing bounding box	(unknown)	SURF	Yes	4M	Eastern USA Europe Japan
[104]	clustering photos	tags	Harris+SIFT+BoW	No	24K+3.6K+8K	Eiffel tower Federer Presidential
[84]	learn tag relevance, without any model training	tags	color correlogram, texture and RGB moments	No	approx. 1M	
[106]	automatic learning of tag semantics	tags	GIST, SIFT+BoW	Yes	approx. 1.7M	[53]
[143]	photo annotation exploiting group structure	tags, groups	SURF+BoW	No	8K+83K+test cases from Corel data set	from a collection of groups

Continued on next page

Table 3 – Continued from previous page

Work	**Task(s)**	**Features**		**Geo**	**Size of Data**	**Origin of Data**
		Text	**Visual**			
[36]	personalized image recommendation via exploratory search		color histogram, Gabor filters, SIFT	No	1.5B	4K image topics
[130]	scene summarization canonical view selection	tags	SIFT	No	500K	Rome
[74]	generate representative tags extract landmark tags	tags, users	color moments, Gabor filters, SIFT	Yes	110K	San Fransisco
[91]	photo search	group similarity	SIFT+BoW	No	30K	30 queries
[11]	photo search and retrieval	title, description, tags	color histogram Gabor filters, edges	No	115K	SAPIR[7]
[127]	generation of summaries of search results			No	270K	30 queries
[122]	ranking of tag-based search results	tags	color moments, wavelet features	No	5.3M	20 tags

Continued on next page

[7] http://sysrun.haifa.il.ibm.com/sapir/index.html

Table 3 – Continued from previous page

Work	Task(s)	Features		Geo	Size of Data	Origin of Data
		Text	Visual			
[168]	discover latent subspaces	tags	SIFT	No	25K images	MIRFLICKR[56] dataset

4 Interpretation of Multimedia Content

Managing multimedia content presents clearly new challenges for the research community that should be addressed in an meaningful way. The main reason for this observation is the fact that humans tend to perceive and characterize multimedia content using solely high-level concepts, such as tags, geo-tags and abstract knowledge-related concepts. The latter are in principle not directly related to the textual or visual attributes or metadata that compose the content itself. Thus, this section investigates the research problem of automatic tag/geo-tag generation, which actually involves a twofold investigation approach: on the one hand there are research works that deal with tag generation, in general, and how to automate the process of producing textual tag recommendations to end-users, in particular, whereas on the other hand there are research efforts that focus only on the so-called content localization by exploiting the prediction of geo-tags. In addition, in an attempt to further improve above described interpretation process, many research efforts take also advantage of automatic knowledge representation and organization techniques, using any available form of intelligence that may be acquired from Flickr textual metadata.

4.1 Automatic Tag/Geo-tag Generation

The emergence of Web 2.0 and the consequent success of social network websites such as Flickr introduced us to new concepts and procedures that may be summoned under the general term “social bookmarking”. The latter can be seen as the action of connecting a relevant user-defined keyword to an image, which aids users to better organize and share their digital content collections. A very interesting research problem emerged, i.e., how to automate the process of making tag recommendations to users when a new resource becomes available. Since the amount of tagged data potentially available is virtually unlimited, interest has emerged in investigating the use of data mining and machine learning methods for automated tag recommendation. In addition, in a slightly different approach, special attention has been given to methods exploiting the prediction of geo-tags, i.e., a special tag category depicting the geographic coordinates where a photo has been taken, a process often called “localization”.

4.1.1 Tag recommendation

Kennedy et al. [73] selected representative tags from urban areas using a multimodal approach. Their method is two-fold: they first collect tags from a geographical region and then extract place and event semantics based on metadata patterns. Their results indicate that the use of visual features can drastically improve precision. Anderson et al. [5] presented a system, namely *TagEz*, which combined both textual and visual features, so as to recommend tags. Their results indicated that the use of textual metadata outperformed both visual and combined feautures. Chen et al. [21] proposed *SheepDog*, a system that automatically recommends tags for photos and also for adding photos into appropriate popular groups. For the latter case, they used SVM predictors in order to identify concepts and used these results to search for groups. Then, they used these groups to harvest more tags and attach them to their photos. Their experiments indicate that the group-level method performed better. Garg and Weber [43], [44] presented a system that dynamically suggested related tags when users tagged their photos. To this goal, it considered similar groups to users' preferences. What differentiates their work upon others is that they did not have available any full text. Instead, every photo was tagged by a single person and suggestions changed dynamically. Moxley et al [105] presented *SpiritTagger* tool in order to recommend tags for Flickr photos of urban regions; the tool is unaware of the user's tags and lies on visual properties and geographic distance, in order to select similar photos. The inclusion of the latter plays a crucial role in the improvement of the results.

Sigurbjörnsson and Van Zwol [129] extracted tag co-occurrence statistics and tag aggregation algorithms, in order to recommend tags by investigating and evaluating four different strategies. Furthermore, they introduced a "promotion" function, whose role was to promote the most descriptive tags. Amongst others, they found that tag frequency distribution follows a power law, while in its mid section contains the best candidates for recommendation. Kleban et al. [76] presented a world scale system for tag recommendation, based on geotags and visual features. This system mines geographically relevant terms and ranks them based on a posterior probability. The authors note that visual features drastically improve performance in densely sampled regions. Popescu and Moëllic [120] presented *Monuanno*, a system that uses visual features to automatically annotate georeferenced landmark images. It first considers neighboring landmarks as potential annotations and after a verification step, it filters outliers. Hsieh and Hsu [55] exploited visual similarity and after a tag expansion process, aim to automatically annotate photos. In order to tackle the problem that tags are in general noisy, they try to fill the gaps by tag expansion based on tagged, visually similar photos and semantic tag consistence. Chaundry et al. [20] presented an approach for tag assignment to geographic areas, using a TF-IDF scheme and logistic regression, for various levels of detail. The performance improvement between the coarsest and the finest level was approx. 25%. In another work of ours [137], we presented an approach whose goal was to extract meaningful content trends and events by exploiting the temporal properties of tags, using a two-level quantization scheme, which was later extended [138] to placing the most important ones on a map, defining a level of importance. With a similar motivation, Zhou et al. [175] focus on identifying travel destinations and distinguishing between places and events. Their approach works on a very large set of Flickr data and uses an RHadoop [8] cloud infrastructure, in order to effectively explore and analyze large data sets. Chen and Shin [24] used both textual and social features of tags and a machine learning approach, in order to extract representative tags that can be related to the users' favorite topics. These social features are introduced in this work and they play a crucial role in extracting tags relevant to the users' interests, while textual features assist finding "correct" tags. Finally, Liu et al. [93] proposed a tag reccomendation system whose goal was to recommend tags for newly uploaded photos, based on the history information in the social communities of users. Their threefold approach comprised of: a) the usage of users' own vocabularies; b) different recommendation methods to different users; and c) the fact that different users are recommended with different number of tags.

Following Table 4 provides an overview of the main tag generation research works gathered in this survey, focusing on the methodology followed, the nature of features and dataset utilized. More specifically, the reader may observe that although the main task of all 14 approaches is similar in nature, they significantly differ in the method each one of them follows to achieve its goals and/or the actual set of features utilized. Most approaches stick to well-known machine learning algorithms and methodologies, like TF-IFD, clustering or Naive Bayes (8/14, 57.14%) and only few of them follow the statistical analysis path. The majority of works uses at least some kind of

[8] https://github.com/RevolutionAnalytics/RHadoop/wiki

textual features (10/14), either combined with visual ones (3/10), or with social ones (3/10). The largest dataset utilized is the one of [175], followed by [129], [43] and [44]. In the following subsection 4.1.2 we shall discuss works focused on content localization by exploiting geo-tagging information.

Table 4: Tag generation

Work	**Task(s)**	**Method**	**Features**	**Dataset**
[73]	representative tags' extraction	clustering	textual, visual	∼110K, San Francisco, USA
[5]	tag recommendation	fusion of learned models	textual, visual	∼1K
[21]	representative tags for users' favorite topics	machine learning	textual, social	∼3.1K
[43], [44]	dynamic tag recommendation	Naive Bayes, TF-IDF, fusion	textual	∼24M
[105]	tag recommendation	photo similarity	visual	∼116K, California, USA
[129]	tag recommendation	statistical	textual	∼52M
[76]	tag recommendation	mining geographical relevant terms	visual, geotags	∼1.2M
[120]	annotation of geotagged photos	clustering	visual, geotags	∼30K
[55]	automatic photo annotation	visual similarity, tag expansion	textual, visual	∼3.3K
[20]	tag assignment to geographic areas	TF-IDF	textual, social	∼135K, Edinburgh, UK
[137], [138]	tag recommendation focusing on trends, events	TF-IDF	tags, temporal	∼18K, Athens, GR
[175]	popular tag extraction focusing on places, events	similarity graph	tags, temporal	∼100M,
[24]	tag recommendation, adding photos to popular groups	Naive Bayes, SVM	textual, social	photos from 177 users
[93]	tag recommendation for newly added photos	nearest neighbors	history information	∼1.9M

4.1.2 Content localization

On the other hand, the task of content localization has gained huge research interest, mainly due to the vast available Flickr database of geo-tagged photos. Hays and Efros [53] presented *IM2GPS*, a system for image localization using visual features. They estimated a probability distribution over the Earth and used it to predict the location of a photo. Although at a first glance their results appear rather poor, we should note that this was the first work to consider such a large data set. The authors also showed that the estimation of geolocation is able to assist other image understanding tasks. Gallagher et al. [41] used a large geotagged corpus from Flickr, extracted several visual features and used the aforementioned location probability maps. They integrated all features and tried to localize photos. They showed that the combination of visual and textual tags may lead to significant improvement on accuracy. Kalogerakis et al. [70] extended the work of [53], by adding temporal information, in an effort to extract information about image sequences and locate images that do not contain any recognizable landmarks. Van Laere et al. [149] divided areas into disjoint regions and then used statistics and a Naive Bayes classifier. Their goal was to predict the area in which previously unseen photos have been taken. They succeeded at the city level and in some cases at the area level. They also found out that given a vocabulary of tags and a photo, the number of tags in the latter that appear in the former, may be considered as a useful indicator for the correctness of the prediction.

In another work of the same authors [148], they trained naive Bayes classifiers at different spatial resolutions. They used only textual features and worked at various spatial resolutions, for a set of 55 European cities. Their technique was able to correctly predict a location within a radius of 1.5km with an accuracy of 80%. They also proposed a similar two-step tag-based approach in [150] that uses language models and similarity search, in order to estimate geo-tags based on a training set. De Rouck et al [32] used language probabilistic models that have been trained on Flickr photos, in order to geo-tag Wikipedia pages. Their approach outperformed Yahoo! Placemaker[9] and their results indicated that the increasing growth of tagged content in Flickr would continuously improve their accuracy. Friedland et al. [39] presented *Video2GPS*. They worked on Flickr videos and used both textual and visual metadata. Their results may seem poor, however they

[9] `http://www.programmableweb.com/api/yahoo-placemaker`

were superior to all other contributions of MediaEval 2010[10]. In another work [40], they combined textual and visual features and worked on the MediaEval 2010 data set. They concluded that solely visual information proved to be inadequate for accurate geo-localization, but when combined with textual, it may assist the improvement of the accuracy. Kalantidis et al. [69] presented *VIRaL*[11], a system aimed to localize photos uploaded by users, based on the visual similarity to already geo-tagged Flickr images. They used a database of more than 2M photos taken from 40 cities. For each query photo and upon a strict visual matching process, a set of geotagged photos was selected. A clustering algorithm predicted then the location of the query photo. They achieved accuracy of about 50m for approx. 88% of the entire data set.

In another effort focused on video sequences, Kelm et al. [72] adopted a hierarchical approach and tried to automatically predict geo-tags for Flickr videos. Their technique lies on nearest neighbor classification, fusing both textual and visual features and also used external resources, such as Geonames[12] and Wikipedia. They correctly localized within a radius of 8km for half of their data set. In [136] we suggested a probabilistic framework which aimed to automatically place Flickr data on a map. The results were manually evaluated by users and indicated that descriptive tags may be produced in the majority of cases. Hauff and Houben [52] added information considering user activities in Twitter[13]. However, even if their results were promising, the median location error was still far from usable. Li et al [86] removed "noisy" photos, i.e., photos that cannot contribute sufficiently to location estimation. They extracted both local and global features and instead of using the whole dataset, they performed clustering and used the resulting centroids, instead. They also proposed an inverted file structure, which improved their results. O' Hare and Murdock [113] presented a statistical language modeling approach, in order to identify locations in arbitrary text. They investigated several ways to estimate models, based on terms and user frequencies. To this goal, they used a set of public, geo-tagged photos in Flickr as ground truth. They were able to predict location within 1 sq. km cell with 17% accuracy, and within 3km radius around such a cell with 40% accuracy. Panteras et al. [115] combine Twitter and Flickr content, in an effort to identify areas that suffered from natural disasters. Twitter content is used so that

[10] `http://www.multimediaeval.org/mediaeval2010/`

[11] `http://viral.image.ntua.gr`

[12] `http://www.geonames.org/`

[13] `http://www.twitter.com`

approximate orientations are derived. Then, using this information, impact area is identified using Flickr content. The advantage of this work is that it leaves out computationally costly visual analysis tasks. The effectiveness of their method was demonstrated using a real-life wildfire incident. Hare et al. [49] estimated a continuous Probability Density Function (PDF) over the Earth and combined textual with a number of weighted visual features. Their approach on tags differed from the others, since they did not filter any of the tags, but they rather used them for evidence, i.e., certain words may be associated with certain countries. Cao et al. [17] worked on the problem of localizing web videos. They followed a near-duplication retrieval strategy between video frames and geo-tagged images. To increase accuracy and remove "noisy" images, they evaluated the consistency of visual features and metadata. However, their approach targeted to identify a landmark, rather than the actual leocation of an image.

Following Table 5 provides an overview of the main content localization techniques, focusing on their main tasks, the kind of features utilized in the process, as well as their achieved accuracy. We may observe that this Table offers the most coherent and focused research task, i.e., that of image localization. In other words most research works in the field have a clear mandate on what to research and their basic difference is how to achieve optimal results. In the process of doing so, half (9/18) of the approaches base their research on visual features, 15/18 on textual, whereas 6/18 utilize both types of features towards a better outcome. One of the most interesting data columns of this Table is, of course, the one depicting the "Accuracy" of each work, since a high value would constitute a good evaluation for a specific work. As a result we observe that the best accuracy is achieved by [149] and [148], utilizing only textual features in the process, followed by the work of [69], which provides slightly worse accuracy rate, when the absolute number is considered (i.e., 98.3% vs. 97%), but with a quite smaller range of 150m. compared to 1.5km, which constitutes it an overall qualitative approach.

Table 5: Content Localization

Work	**Task(s)**	**Features**	**Accuracy**
[53]	image localization	visual	∼50%, 25km
[41]	image localization	visual/textual	∼20%, 50km
[70]	image localization	visual/textual/temporal	∼58%, 400km
[149]	image localization	textual	∼98.3%, 1.5km
[148]	image localization	textual	∼98.3%, 1.5km
[150]	geo-tag prediction	textual	8.82km (median)
[32]	geo-tagging of Wikipedia pages	textual	∼15%, 1.0km
[39]	video localization	visual/textual	∼40%, 1.0km
[40]	video localization	visual/textual	∼45%, 1.0km
[69]	image localization	visual	∼97%, 150m
[72]	image localization	visual/textual, Wikipedia and Geonames	∼35%, 1.0km
[136]	automated tag localization	textual	∼75%, 1.4km
[52]	image localization	textual/Twitter	∼84%, 1.0km
[86]	image localization	visual	
[113]	identification of locations in text	textual	∼48%, 100km
[115]	identification of areas suffering from natural disasters	textual/Twitter	
[49]	image localization	textual/visual user modelling	∼40%, 10km
[17]	image localization	textual/visual	

4.2 Metadata Knowledge Extraction

In an effort to utilize collective intelligence in the process, many research efforts have leaned towards automatic knowledge representation and organization, using intelligence that has been gathered from Flickr textual metadata. As a result, researchers often make use of groups, that may be regarded as a simplistic form of a semantic hierarchy. Schmitz [126] applied a subsumption-based model and used a vocabulary that had been created from Flickr tags and from 9M images, in order to create an ontology. The goal was to provide a faceted ontology as a supplement to a tagging system. However it produced subtrees that reflected distinct facets, but was not able to categorize concepts into facets. Firan et al. [37] used an ontology, Naive Bayes and SVM classifiers, in order to detect events. Using those events, they ended up classifying Flickr photos. Their work indicated that some classes are relatively easy to learn. On the contrary some others may require some kind of special attention or even some level of disambiguation. Lu and Li [95] constructed semantic topic-hierarchies and then mapped Flickr groups onto them, so as to construct group-hierarchies. Their initial experiments indicated that these hierarchies may facilitate the browsing experience of users.

Negoescu et al. [107] grouped Flickr groups using a cluster approach, in order to create “hypergroups”. They showed that homogeneous hypergroups may be created, however, of rather small sizes. Upon manual inspection, they concluded that these hypergroups are indeed meaningful sharing content and/or members. This way, they allowed smaller groups to be easily discovered by potential users. Prangprasopchok and Lerman [117] constructed quite detailed folksonomies by aggregating relations from different users, following a generic approach which can be applied to several other systems. They also proposed a method for the automatic evaluation of such a learned folksonomy, upon comparison to a reference taxonomy. Empirically, they concluded that user-specified relations provide a good source of evidence for the construction of folksonomies. Derrac et al. [33] proposed a method to enrich place types taxonomies with a ternary betweenness relation derived from Flickr. To this goal, they constructed a semantic space of place type and assusmed that natural properties of place types should correspond to convex regions in this space. Their results indicated that their approach outperformed similarity-based methods such as k-NN on place type classiffication tasks and in some cases, humans. Lee et al. [78] first found points of interest and then by using a clustering approach and by applying association rules mining, they tried to detect associative point of interest patterns. They empirically concluded that their approach is able to find interesting patterns. Cai et al. [15] proposed a trajectory pattern mining algorithm and applied it to a Flickr dataset of Australia. Their experiments showed many previously unknown patterns discovered. They also discovered expected landmarks (e.g., cities and attractions) and mined information about sequential movements among them, one of which is that tourists do not necessarily use shortest paths. Finally, Xie et al. [165] proposed an algorithm of an Augmented Folksonomy Graph (AFG), which was used in order to incorporate multi-faceted relations in social media. They also used a novel density-based clustering method so as to discover latent user community from AFG by combining contents and tags of multimedia resources. They showed that their approach outperformed baseline ones with respect to search applications.

Following Table 6 provides a detailed overview of the discussed knowledge extraction research efforts. It categorizes them according to the knowledge type incorporated, illustrates their advantages and disadvantages and reasons on their suitability within the broader research field. More specifically, Table 6 includes the most important types of knowledge representation (ontology, taxonomy, folksonomy, hierarchies and clusters), thus making it clear that it is the research that defines the representation of knowledge inherent within the Flickr data and not vice versa. Still, it is also depicted that many of these representations suffer in terms of objective evaluation methodologies, especially folksonomy and/or travel patterns related. In addition Table 6 provides also a novel last column, where we try to identify the suitability of each research work for future reference.

Table 6: Knowledge extraction

Work	**Knowledge type**	**Pros**	**Cons**	**Suitable for...**
[126]	ontology	provided a faceted ontology	not able to categorize concepts into facets	supplementary to tagging
[37]	ontology	event detection	not all classes were easy to learn	photo classification
[95]	topic hierarchies	constructed Flickr group hierarchies		facilitation of browsing experience
[107]	group clusters (“hypergroups”)	provides meaningful groups	subjective evaluation	easier discovery of small groups
[117]	folksonomy	introduces user-specified relations	subjective evaluation	induction of detailed folksonomies
[33]	taxonomy	works better than similarity-based methods	relies on assumptions	taxonomy augmentation
[78]	travel patterns	rule generation	limited evaluation	understanding user behavior
[15]	travel patterns	discovering unknown patterns	limited evaluation	understanding user behavior
[165]	folksonomy graph	combines tags and visual features	relies on empirical parameters	discovering user communities

5 Humanistic Applications

Current era is with no doubt the one characterized by unsolicited digital multimedia content production, curation and interpretation. To study the effects and consequences of this trend on our human nature and culture, and how we are shaping these technologies according to our cultural needs, digital technology shifted towards the field of humanistic applications. The latter include travel applications that in principle utilize multimedia content to extract preferences of touristic nature with respect to Points of Interest (POIs), landmarks and travelers' routings. In this framework one should also examine the social aspect of Flickr-based research approaches, focusing on works that deal with the relationships among Flickr users, related user activities and their behavior. Both research directions are discussed in the following subsections.

5.1 Travel and Tourism Applications

Undoubtedly a large percentage of multimedia content circulated within social networks is of touristic travel nature, e.g., photos or videos of landmarks, places of interest, events, etc.. As a result sets of geotagged photos on Flickr may explicitly indicate the "trajectories" or so-called "routes" of tourists. The latter can be employed to reveal the tourists' preference on landmarks and routings of tourism. As expected many research efforts focus on the exploitation of such content for applications that target potential travelers and/or tourists. The latter form typically applications you can download on your mobile device to enjoy during your trip. Most such travel apps focus on things like cutting down journey times, pointing you in the direction of authentic local events or simply showing you how to ask after a famous landmark.

In principle we may identify three basic subcategories when dealing with research works examining those applications, namely works that focus on the visual reconstruction of a scene in either two or three dimensions, works that their goal is to recommend meaningful places of interest to users and works that tend to organize the users' schedule by suggesting possible routes or trips to them. An overview of the herein discussed travel applications is provided in the following Table 7, which categorizes them according to their type and illustrates each one's main advantages and drawbacks.

5.1.1 Visual reconstruction

Snavely et al. [132] presented a system based on structure-for-motion (SfM), which aims to allow for interactively browsing and exploring large, unstructured photo collection. Therein, 3D models, photo tours and annotations are created within a system called *Photo Tourism.* The main limitation of this work, as stated by its authors, is that it lied on manually set ground control points. In another work, Snavely et al. [133] extracted viewpoints from photos and created paths that were then used for image-based rendering. Continuously, they automatically created orbits (i.e., lines defined by the set of viewpoints from where photos had been taken), panoramas, canonical views and optimal paths between views. Their application allowed interaction with users for 3D browsing. The most important limitations of this work was the use of a simplistic geometric model for orbits (i.e., a circle). Li et al. [85] presented another approach for modelling landmarks, which was based on iconic scene graphs. They combined 2D appearance with 3D geometric constraints. Then, using iconic scene graphs they created summaries and 3D reconstructions and also recognized landmark in given photos. However, the authors observed that since people tend to take photos of landmarks from a relatively small number of viewpoints, either characteristic or accessible, many areas left practically uncovered. Kaminsky et al. [71] also followed a SfM approach and created visual reconstructions, which with the aid of geotags, were aligned to aerial photos. They observed that significant improvement took place when exploiting geotag and GPS information. However their algorithm quite often failed, in cases where their 2D free-space model was violated.

Agarwal et al. [2] used a large dataset of Flickr photos that had been taken in Rome, Italy and also in other cities and tried to create a 3D reconstructed model in 24 hours using a cloud-based computer architecture of about 500 cores. Their experiment demonstrated that at the time of this work, it was already feasible to reconstruct the core of a city using a dataset of about 150K photos in less than a day. Tuite et al. [142] presented *PhotoCity*, an online game that aimed to train its players so that they become “experts” in taking photos at targeted locations and in great density, with a goal to create 3D building models. They evaluated their approach by reconstructing large portions of two university campuses. The main difference compared to other works was that in this manner they intended to include certain areas that otherwise did not have much photographic coverage on sites such as Flickr. Their experimental results and the comparisons confirmed the aforementioned goal.

5.1.2 Recommendation systems

The second category of tourist applications focuses on the recommendation of places of interest. Typical applications in this area focus on automatically discovering main attractions, letting users decide which to visit. Still they can be further divided into two subgroups, namely the ones that focus on the detection of events or even trends and the ones that try to identify a set of representative images and/or tags and provide interesting search results to users. In this manner, within the first subgroup, Chen and Roy [22] detected events by exploiting tags, date information and geo-tags and used wavelets in order to handle noisy data efficiently. They considered that a typical event consists of tags with similar temporal and location distributions and also visually similar photos. Their approach performed well in periodic events, while being less accurate in non-periodic ones. Van Canneyt et al. [145] proposed a recommendation system for trends in tourist attractions in cities. They dealt with recommendations both as a ranking and as an assignment problem and adopted a probabilistic approach. In the first case they ranked places of interest according to their popularity, by also taking into account temporal information about the user, while in the second, the user selected a few places of interest and timeslots he/she was available and their system proposed the best coverage of these. They concluded that the use of Wikipedia plays a key role in providing the semantics of several places.

Kisilevich et al. [75] used geotagged photos and data clustering techniques in order to determine urban areas of interest, analyzed spatial and temporal distributions so as to identify events and rank places of interest based on their popularity. Their case study took part in five regions of Switzerland. Without any prior knowledge, they were able to detect and annotate several events. Nitta et al. [110] proposed an approach to detect events using tags, geotags and temporal metadata of Flickr photos, by first defining event classes, i.e., semantically related groups of events. They were able to detect events in cases where a small number of photos was available. Jing et al. [64] proposed a photo recommendation scheme, by examining sentiment from the general public towards items. In order to infer photo importance, they use temporal dynamics and local community user ratings. They build on the Probabilistic Matrix Factorization (PMF) framework for photo recommendation and their results indicate significant improvement of the recommendation prediction.

The second subgroup illustrates the research work of Cao et al. [16], where they proposed a tourism recommendation system by using mean shift clustering and by building a set of representative images and tags for each cluster which they then use to match users' input. Users upload photos as queries and the system responds with suggestions. Sang et al. [125] proposed a framework for personalized image search. Therein, results are ranked based on the tags that the user has already used. Their retrieval results indicate significant improvement in search performance. Hollenstein and Purves [54] used a set of 8M Flickr images and a vernacular geography approach, in order to study how accurate user-generated tags are, how may urban areas be described using tags and how may tag images allow the understanding of the location and extent of vernacular regions. Their research indicated that only 0.52% of tags describe indeed city core areas generically. Moreover, the 70% of geotagged images included tags that correspond to places. Similarly, Bartie and Mackaness [12] tried to measure and visualise the visual exposure of city sites, aiming to provide an aid to tasks such as automated way finding, or augmented reality city guides. Their novel algorithm aimed to calculate the visual exposure and the perceived size of buildings, which as they claimed and experimentally verified, was showing the locations from where photos were taken, Finally, Yao et al. [162] presented a graph-based framework which aims in a unified manner to be used for friends reccomendation, image tagging and personalized image search. They used Flickr as a testbed, while claim that their approach may be easily adapted to other social networks.

5.1.3 Trip/Route suggestion systems

Another category of tourist applications extends the former, in the sense that it not only recommends main attractions, but also tries to organize the users' schedule and help them visit as many as they wish in a time efficient way. Popescu and Grefenstette [118] exploited temporal Flickr metadata in an effort to estimate expected visiting times for tourist attractions. This way they tried to deduce what a tourist is able to visit in a city within a day. They evaluate their method using manual estimations for a set of popular attractions for 4 cities. Jain et al. [61] extracted photos around the location of a trip and created a graph. They found tours that start from this location and propose a tour that visits popular places using certain distance constraints.

Their system, *Antourage*, tries to cover all popular landmarks, however it does not consider factors such as time spent at locations and reachability via the existing road network. This work was later extended by Popescu et al. [119] where the authors mined spatial and temporal tourist information from Flickr and tried to discover information about trips in popular cities, e.g., what attractions people visit, how long do they stay at each and what are the best panoramic spots. Their method is also able to "generate" new trips by combining the existent. They found that manual estimations by experts are typically larger than the actual (extracted) times and that visitors often do not enter sites, but are limited to solely sightseeing. Hao et al. [50] presented *Travelscope*, a system that creates virtual tours by mining Flickr data. Its goal was to recommend popular places, given a specific region, annotate some aspects (e.g., as landmarks or activities) and summarize landmarks by providing representative images.

Sun et al. [140] clustered images spatially, identified landmarks within them and ranked them based on their popularity. They aimed to recommend to the user minimum distances with maximum tourism popularity. They empirically evaluated the potential of their approach. Majid et al. [99] presented an approach for personalization and recommendation of tourist locations. They obtained users' preferences from their travel history in a city and used this information to recommend locations in an another, unknown city. They observed that its easier to predict preferences with short and targeted visits, when using methods that are based on popularity. Jiang et al. [63] proposed an Author Topic Collaborative Filtering (ATCF) method to facilitate Points of Interest (POIs) recommendation for social media users. Therein, user preference topics (e.g., cultural, cityscape, or landmark) were extracted based on the textual features. They showed that similar users could still be identified accurately according to the similarity of users topic preferences. Moreover, the corresponding category and user topic preference could be elicited simultaneously.

Lim [89] proposed a system, namely *TourRecInt*, whose goal was to receommend tours based on user interest. This system combined knowledge from Flickr geo-tagged photos and the Wikipedia, and exploited the users' visit history. His approach used constraints for a "distance budget" and a "must-visit" set of places-of-interest (PoIs). At another work of the same author [90], another algorithm, namely "PersTour", who poses the constraints of starting and ending at pre-defined points and having time limitations. One novelty of this work is the "time-based" user interest, i.e., a level of interest based on the time users spent in PoIs. Li et al. [87] proposed three rule-based methods to recommend travel routes for tourists, taking into account their location, while satisfying their personalized demands. They collected metadata from Flickr and events from the chinese application DoubanEvent[14].

Elaborating on Table 7 data, that are grouped according to the nature of the discussed research works, one may distinguish the main advantage of works belonging to the visual reconstruction group as the interaction they offer to the end-user. Recommendation systems provide increased efficiency when dealing with detection of periodic events and may perform very well when exploiting additional textual sources of information like "tags"'. Lastly, route suggestion research is clearly an innovative trending one, whose outcomes could be utilized in everyday life within the framework of current social networks. On the other hand, visual reconstruction approaches clearly suffer from the validity of the utilized models, as well as from high computational costs. In addition, recommender systems rely heavily on the type and nature of utilized multimedia content, as well as its semantics in order to be able to provide meaningful results, whereas route suggestion techniques still have a lot to research on when it comes to proper evaluation methodologies.

[14] http://www.douban.com

Table 7: Travel apps

Work	**App type**	**Pros**	**Cons**
[132]	visual reconstruction	interactive browsing and exploring of large collections	not fully automatic
[133]	visual reconstruction	interaction with users	simplistic model
[85]	visual reconstruction	combined detection and summarization	leaves uncovered areas
[71]	visual reconstruction	alignment of ground and aerial photos	adopted model often fails
[2]	visual reconstruction	full reconstruction of a city center in 24 hours	needs cloud of approx. 500 cores
[142]	visual reconstruction	gamification	limited evaluation
[22]	recommendation systems	good performance in periodic events	medium performance in non-periodic events
[145]	recommendation systems	using Wikipedia	
[75]	recommendation systems	no need for prior knowledge	limited evaluation
[110]	recommendation systems	detection of events given a small set of photos	need of prior knowledge
[64]	recommendation systems	use temporal dynamics, probabilistic	use local user ratings
[16]	recommendation systems	photo queries	only 10 recommendation topics
[125]	recommendation systems	exploiting already used tags	simplistic ranking model
[54]	recommendation systems	determines “good” tags	heavily relies on uploaded content
[12]	recommendation systems	augmented reality city guides	
[162]	recommendation systems	unified approach for friends recommendation tagging and personalized search	
[118]	route suggestion	considering time limitations	manual evaluation
[61]	route suggestion	using distance constraints	not considering reachability
[119]	route suggestion	generation of new “trips” by using existent	
[50]	route suggestion	creation of visual tours	lack of evaluation
[140]	route suggestion	combined distance minimization and popularity	lack of geotagged info
[99]	route suggestion	created recommendations for unknown cities, using previous travel history	user satisfaction not evaluated
[63]	route suggestion	considered user similarity	similarity relied solely on tags
[89]	route suggestion	used distance and must-visit constraints	user satisfaction not evaluated
[90]	route suggestion	used time-based user interest	user satisfaction not evaluated

5.2 Human Activity Tracking

Considering Flickr as a social network, many researchers study the relationships among users, user activities and behavior within it. A summary of these studies is provided in the following Table 8, which classifies them according to the type of user modeling imposed. The Table offers also a brief overview of their pros and cons, as well as positions them within the respective research field. A clear observation from the study of the Table depicts that the main drawback of almost all herein mentioned research works is the lack of objective evaluation methodologies.

To begin with, Marlow et al. [101] presented a taxonomy of tagging systems, with Flickr amongst them. Their goal was to facilite analysis and research of such systems. They outlined possible directions of research in tagging. Negoescu et al. [108] tried to model users and groups with a common tag-based representation, using a probabilistic topic-based analysis. They showed that even if users and groups are conceptually different, a bag-of-tags representation may be used to allow for their representation in a common way. Then, their analysis led to the discovery of similar users and groups. Valafar et al. [144] studied user interactions by analyzing their temporal properties. Their research indicates that a very small fraction of users in the friendship graph is responsible for the vast majority of fan-owner interactions over photos. However, such interactions involve only a small fraction of photos in Flickr. They also showed that most of the photos gain the majority of their fan base within the first week upon publication. Gelli et al. [45] make use of sentiment and context features, to the goal of predicting the popularity of photos. They propose three novel context features and demonstrate a novel analysis of the correlation of sentiments to popularity. Stvilia and Jorgensen [139] investigated Flickr member activities and how these may assist on the automatic metadata creation. They worked on a set of historical photographs and their respective comments and discussions. They used two knowledge organization systems, namely the Thesaurus for Graphic Materials (TGM) and the Library of Congress Subject Headings (LCSH). Among their results we should note that more than 50% tags were not found neither in the TGM nor in the LCSH. They concluded that the extension of the aforementioned systems could allow them to be more accessible to different user communities.

Lerman and Jones [81] investigated "social browsing", i.e., the strong correlations users have with their contacts in Flickr. They showed that the primary way that users follow, when they aim to discover photos, is not by tag-based search or subscription to specialized groups, but rather by browsing through photo streams of their contacts. Cha et al. [18] collected and analyzed large-scale traces of information dissemination in Flickr. They empirically showed that photos spread due to social links and this propagation process is limited within close connections of users and also slower than their initial expectations. They also concluded that the popularity of content is generally localized in the network and also the popularity of photos continuously increases. Similarly, Jung [68] studied information propagation within Flick. More specifically, under the assumption that given a certain tag, a social "pulse" may be established by counting: a) the number of users, and (b) the number of resources over time, he showed that information propagation takes place by inducibility from other tags by comparing social pulses. To this goal he also implemented a searching system, namely *Tagoole*, which operates on tags. McParlane et al. [102] aimed to predict the popularity of photos in a "cold start" scenario (i.e. cases where no or limited user interaction exists). They considered the image context, the visual features and the user context and tried to predict the number of views and/or comments a given photo has. They showed that it is feasible to overcome the problems occuring in cold start scenarios.

Mislove et al. [103] used empirical data and investigated the link formation process. They showed that links tend to be created by users that have many links, and also users tend to link to those users that are close to them in the social network structure. Finally, Cox et al. [27] interviewed Flickr users so as to understand the use of the website in conjunction with the users' practices in photography, i.e., they considered Flickr from the scope of hobbyist photographers. They concluded that Flickr should be analyzed as a social network of amateur photographers, rather than a simple photo storage facility. At this point the interested reader may have identified the fact that there are indeed a lot of Flickr-related research works out there that do not fall under any of the previously mentioned groups/categorization. As a result, in the next section 6 we shall briefly attempt to summarize interesting research efforts from so far not discussed application domains.

Table 8: Humanistic approach

Work	**User modeling**	**Pros**	**Cons**	**Suitable for...**
[101]	taxonomy	outlined possible directions of research		analysis and research of tagging systems
[108]	user-groups	unified representation of users and groups	empirical evaluation	search
[144]	fans-owner	indirect measurement of popularity	relies on heuristics	social network dynamics
[45]	user-views	context and sentiment features		predicting photo popularity
[139]	user activities	exploitation of prior knowledge outside Flickr	limited evaluation	indexing and retrieval
[81]	user-contacts	learning users' habits	empirical evaluation	search
[18]	social network graph	understanding information dissemination	empirical evaluation	information propagation analysis
[68]	folksonomies	definition of "social pulses"	limited evaluation	information propagation analysis
[102]	comments	works on "cold start"		popularity prediction
[103]	user-contacts	understanding how users make new contacts	empirical evaluation	social network dynamics
[27]	photographers	understanding how photographers use Flickr	empirical evaluation	learning amateur photographers' habits

6 Other Application Domains

The variety of Flickr content allows also for several other research efforts on application domains that do not fall under previously discussed categories and are summarized in the following Table 9, according to their application domain, pros and cons, and respective problem solving. Apart from the apparent organization of related works, the main novelty of Table 9 is its last column, where the interested reader could easily track and identify the application domain of each work. The latter, together with the brief inline description of each methodology that follows, would act as a single point of reference for future or fellow researchers in the field.

For instance, Jin et al. [62] applied regression- and diffusion-based prediction models on certain Flickr textual and visual features and used them for social studies such as politics, economics and marketing. They experimented on the prediction of product sales and on the American presidential election of 2008. In the first case, they showed that Flickr may monitor the worldwide adoption of products, while for the latter it provided hints that may assist for the prediction of the election results. Similarly, Singh et al. [131] combined Tweets and Flickr posts in order to study spatio-temporal events. This way, they were able to extract semantic information, e.g., about political events and seasonal characteristics. Zhang et al. [172] aimed to discover and visualize tag relationships from spatial and temporal similarities. They showed that photo tags may be clustered based on their temporal and geodata distributions and provided appropriate visualizations. Their approach was empirically evaluated and concluded that the aforementioned visualizations of tag semantics may help users intuitively capture subtle geo-temporal relationships among tags.

Taking related research efforts a step further, Lei et al [79] adopted a multimodal methodology, based on both acoustic and textual features and aimed to identify cities using machine learning approaches. They observed that in some cases acoustic features are enough for correct classification. Clements et al. [26] used only geo-tags, defined a similarity scheme among their distributions and proposed a weighting scheme, in order to identify similar places at world and city level. They found that a user's favorite landmarks in a city she/he has not visited in the past may be predicted by re-ranking the most popular ones among other users that share similar preferences. However, they concluded that this task is very difficult to achieve, except from cases where users have "clear" travel preferences. Leung and Newsam [83] used datasets from Flickr and Geograph[15], in an effort to derive maps of "what-is-where" on the surface of the earth. They observed that maps generated using Geograph data were more accurate than those generated using Flickr data. They explained that this happened due to the fact that photographers in Geograph intent to annotate, while in Flickr do not have such intentions. Zerr et al. [171] presented a system, namely *NicePic!*, able to classify images in a photo stream into two categories: a) most and b) least attractive. They used several visual and textual features to achieve this goal.

Wu et al. [159] proposed a novel distance scheme, namely *Flickr distance.* They used it as a means of measuring relations between semantic concepts, constructed a concept network and applied it to concept clustering and image annotation. They experimentally showed that Flickr distance appears more coherent to human perception than previous approaches. Similarly, Cox et al. [28] proposed a set of metrics in order to characterize and compare Flickr groups. Among their findings we should mention that additionally to very large groups (in terms of both members and photos), there exist many small groups with low activity. Also most large groups are not dominated by a few individuals. Xu et al. [167] proposed a set of metrics for measuring the semantic relatedness between tags of images. In order to remove noise and redundant tags, they used a bipartite graph. They also considered higher order tags as most important. They evaluated their approaches on clustering and searching tasks.

Jaffe et al. [60] generated summaries by selecting the most representative photos, using geo-tags and a clustering approach. These summaries could be biased by the specific user, the query content and context. Their approach was empirically evaluated. Wang et al. [153] proposed a generative probabilistic model and used it for group recommendation. It works on both users and groups, it jointly discovers the latent interests and simultaneously learns a recommendation function. They showed that both the prediction function and the latent topic learning may benefit from each other. Group recommendation was also the focus of Wang et al. [155], who assumed that different types of relationships in heterogeneous information network may be used to improve the recommendation results. They applied a non-negative matrix factorization (NMF) method regularized with user-user similarity via heterogeneous information networks. Xie et al. [166] proposed a method for mining user interests based on personal photos. They combined a hierarchi-

[15] `http://www.geograph.com`

cal structure for image representation, along with a user image latent model. This way they were able to identify some user interests (e.g. some concepts such as car, plane, tree) and distinguish them from “background”-interests (e.g. street scenes). Bojic et al. [14] investigated several methods for home location and applied them to Flickr datasets, comparing to bank transaction records. They concluded that the choice of the appropriate home definition method should take into account the unique characteristics of the dataset it is applied on. At another work of the same group, i.e. the one of Sobolevsky et al. [134], the authors aim to subjectively measure the attractiveness of cities in Spain, based on bank card transactions, geotagged photographs and tweets. Moreover, their analysis showed that visitors tend to spread upon different smaller destinations over summer time. On the contrary they are more concentrated at major destinations during the rest of the year.

The aforementioned problem of home location is also tackled by Zheng et al. [174]. In this work, the authors aim to predict the locations of home and vacations of users, using visual and spatiotemporal features. Chen et al. [23],aimed to create visual summaries to be used as tourist maps, by capturing the most important points of interest. They also selected a representative photo for each landmark. They observed that in many cases this representative photo was not the one “expected” due to the nature of the data set. Hao et al. [48] presented a methodology for the automatic creation of travelogues, i.e., a kind of a travel-related experience logging. They retrieved geo-tagged photos from Flickr and embedded them in these travelogues, whose goal was to facilitate other tourists trip planning. This way and despite the noise and the lack of structure within typical travelogues, they were able to effectively generate recommendations, create summarizations of places and enrich travelogues with images. Baber et. al. [8] asked tourists to capture photos of a monument, with the goal to be able to support subsequent question-asking. The results of their study indicate that “much tourist photography represents a special form of image capture” in which tourists tend to gravitate towards the best vantage points to take their own versions of photos seen in brochures. You et al. [169] exploited the learning capabilities of deep neural networks to perform sentiment analysis on weakly labeled large scale photo collections. They proposed a novel architecture, able to be applied to other domains. Finally, Donaire et al. [34] performed a case study using uploaded photos of tourists. They were able to identify different groups using cluster analysis. Their results indicated that groups differentiate in the selection of potential sights, however they share a particular way of looking at sights.

Table 9: Other Domains

Work	**Domain**	**Pros**	**Cons**	**Suitable for...**
[62]	predictions in politics/marketing	extended models for predictions	limited evaluation	social studies
[131]	studying spatio-temporal events	combination with Twitter	empirical evaluation	semantic information extraction
[172]	tag relationships	clustering and visualizations of tags	empirical evaluation	geo-temporal relationships' extraction among tags
[79]	classification using textual/audio features	audio features' exploitation		video localization
[26]	geotag distributions	travel recommendations for unvisited cities	limited evaluation	travel recommendations
[83]	content localization	combination with Geograph	solely Flickr data are not sufficient	land cover classification
[171]	photo attractiveness classiffication		heuristic approach	recommendations
[159]	semantic similarity measure	coherent to human perception		search
[28]	semantic similarity measure	understanding of the role of smaller groups	empirical evaluation	Flickr groups similarity
[167]	semantic similarity measure	noise removal using bipartite graph	limited evaluation	clustering/search
[60]	landmark summarization		empirical evaluation	travel recommendations

Continued on next page

Table 9 – Continued from previous page

Work	Domain	Pros	Cons	Suitable for...
[153]	studying social network graphs	latent user interests' discovery		group recommendation
[155]	learning user-group relationships	latent representations of users and groups		group recommendation
[166]	mining user interests	image content representation model	limited set of interests	recommendations
[14]	home location definition	comparison of many methods	difficult to perform well on Flickr data	understanding human activity
[134]	city attractiveness measure	big data	based on some assumptions	understanding human activity
[174]	home vs. vacations	good performance		understanding human activity
[23]	landmark summarization	extraction of popular representations of landmarks	often selecting indoor representations instead of outdoor	tourist maps
[48]	information extraction	enriching of travelogues with photos	empirical evaluation	automatic creation of travelogues
[8]	sense-making	real life application	limited evaluation	context-based image capturing
[169]	sentiment analysis	works on weakly labeled data		big data analytics
[34]	tourist group analysis	interesting approach		understanding human activity

While most of the challenges in dealing with Flickr data may be classified into the aforementioned considerations, still there are also additional areas that need to be addressed. At this point it is worth noticing that one identifiable characteristic of all works presented in this section, but also in other sections of this manuscript as well, is the diversity of the Flickr datasets utilized for research purposes. Consequently, in the following section 7 of this survey we ought to present the flipside of the coin, i.e., a short, yet indicative list of research works dealing with the construction, utilization and/or exploitation of Flickr datasets for benchmarking purposes towards the provision of objective evaluation tools.

7 Benchmarks

As it has been seen in previous sections, many heterogeneous datasets originating from Flickr have been used for research. In most cases, these datasets are not shared by their authors, or in some cases have been tailored to facilitate the proposed method/algorithm/use case. However, it would be a serious omission to leave out of this survey a collection of works in the benchmarking field with respect to Flickr data.

Among the ones worth mentioning herein is the MIRFLICKR Retrieval Evaluation [56], which begun with a collection of 25K photos collected from Flickr. Its goal was to provide a benchmarking test set for the image retrieval community targeting to fulfill what the authors consider to be the four main requirements for such a data set, namely: a) to be representative of an area; b) to provide accurate ground truth; c) to be freely redistributable; and d) to be accompanied with standardized tests for evaluation. The authors proposed a number of standardized challenges, namely: a) visual concept/topic recognition; b) tag propagation; and c) tag suggestion. The set has been later on extended to 1M photos and a a number of content-based visual descriptors has been supplied for the entire collection [57]. In 2006, CLEF[16] introduced a new task aiming towards interactive multilingual search of images in Flickr, as part of the ImageCLEF benchmark [46]. In the aforementioned task, participants had to build a front-end to Flickr, a) by using its search API and b) able to support multilingual search. The goal was to study user behaviour, given a set of searching tasks, emphasizing on the study of the process, rather than the evaluation of its outcome. Also in 2006, the PASCAL Visual Object

[16] `http://www.clef-initiative.eu/`

Classes Challenge (VOC2006) [35] focused on object detection and recognition on an annotated data set of photos, where a subset was collected by Flickr.

An interesting task has been initiated in 2012 by MediaEval[17], namely the "Placing Task". Therein participants are asked a) to explore several aspects of multimedia documents (e.g., textual, social, visual, etc.), in order to estimate the location where a given Flickr media item has been captured and b) to optionally decide whether a media item may be objectively placeable. Also in 2012, the well-known TRECVID benchmark[18] evaluated automatic and interactive retrieval systems at a task named "Instance Search", using among others a set of Flickr videos. The goal of the task was given a visual example, to find more video segments of a person/object/place, depicted within it. In 2014 Yahoo! released the "Flickr Creative Commons" dataset [141], aimed for research use. This dataset consists of approx. 100M photos and 700K videos and is one of the largest public multimedia datasets. They have also computed a few open audiovisual features, using a supercomputer. They claim that this dataset may host a variety of research studies and challenges and they plan to create such challenges or expand current ones. A first work, (under progress at the time of the submission of this paper) which makes use of this dataset in order to learn semantic concepts is the one of Ni et al. [109]. Mao [100] also worked using this data set and described the information contained, while trying to estimate the most visited US place by Americans. Similarly, Izadinia et al. [59] investigated direct learning of image classification from tags "in the wild" (i.e., not filtered). Finally, in 2015, Ionescu et al. [58] introduced a dataset and its evaluation tools. This dataset, namely *Div150Cred*, consists of 300 landmark locations represented by approx. 45K Flickr photos. It also contains 16M photo links for around 3K users, metadata, Wikipedia pages and content descriptors for text and visual modalities.

8 Challenges and Future Research Directions

It should have been evident by now that based on the above discussion, there has been a dramatic advance in the research and development of Flickr-related multimedia systems. In particular, it seems that given its technical

[17] `http://www.multimediaeval.org`

[18] `http://trec.nist.gov/`

facilities to developers (i.e., an open, enriched API) and its increasing popularity among online users worldwide, Flickr tends to be the next big thing with respect to multimedia analysis researchers. As expected the process still faces several main challenges and in this section, we attempt to list these challenges and point out the corresponding future research directions.

In our opinion the future lies among others in the introduction and exploitation of supervised and unsupervised models for characterizing the various facets of Flickr images combined with their metadata. Quite often there is a lot of structured and unstructured data available together with the uploaded images that can be potentially exploited further through joint modeling, clustering, and classification techniques that will - to an extend - bridge the so-called semantic gap and benefit future multimedia system implementations in numerous ways.

In the field of text-based query processing the main challenge is the presence/absence of reliable textual metadata with images. Although significant efforts have been introduced in the past for large-scale collection of high-level manual annotations, like the quite popular among researchers ESP game [151], we still believe there is room for improvement with respect to efficient collection of manual tags for images. The latter presents the twofold advantage of facilitating text-based queries, as well as building reliable training datasets for content-based multimedia analysis and automatic annotation algorithms. So, we expect to see in the future a bridging of keyword- and content-based search through a unified framework and Flickr data form the ideal candidate towards this goal.

It is also reasonable to hope that in the near future, the technology will utilize the vast amounts of Flickr datasets in order to diversify to other challenging domains. It is evident at this point that the future of real-world image retrieval within the social networks framework lies in exploiting both tag- and visual content-based search. Given the rather easy to implement tools that Flickr provides to interested researchers, combined with the research-friendly Creative Commons[19] licensing (CC) scheme that most Flickr photos follow by default, we believe that the potential from combining the two worlds is great and that this endeavor will hopefully be actualized in the years to come.

Tempted to take a risk and further specify more concrete research directions, we would like to see a paradigm shift in the near future, with research focus being shifted more on application-oriented, domain-specific works that

[19] `http://creativecommons.org/`

will have considerable impact in everyday life. A long-term goal of research could therefore also include the ability to predict peoples movements by analyzing the embedded timestamp and geographic information within photos [10]. In other words researchers will be able to accurately predict where a person is most likely currently located and where she/he may be headed in the near future. Another interesting research field with great future potential forms revealing of spatial and temporal patterns from Flickr photos, either with or without the use of additional sources of information, like POI databases. This would easily lead to touristic applications that would enable people's smartphones with the ability to instantly download and suggest the best options for their touristic route when arriving in an unknown city, based on previous recommendations or experiences of their social networks' friends.

As expected, the vast amount of Flickr data collected so far, as well as its impressive increasing rate, attracted the interest of the deep learning community. Being a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers with complex structures or otherwise, composed of multiple non-linear transformations, deep learning methodologies and corresponding research approaches are ideal for Flickr data manipulation and extraction of interesting trending classification and/or recommendation results in the future. As a result we are currently observing a shift of research interest towards this direction (e.g., works similar to the one of Zhao et al. [173]) - and we expect to see a lot more in the near future. Last but not least, content localization will definitely be a trend of the future, both in the sense of developed services and applications, as well as its combination with traditional software engineering efforts. In principle localizing multimedia content involves several modalities, but we expect to see a major advance in research efforts dealing with digital images when combined with geo-tagging metadata information, like for instance the 2014 MediaEval Placing Task: Multimodal Location Estimation [20], thus the role and availability of Flickr image repository could be crucial.

Apart from the future directions there are also clear rather practical challenges to be tackled by researchers. The diversity of utilized Flickr datasets is considered to be an "evaluation plague" and only recently there are some efforts to deal with it. More specifically the MediaEval Challenge emerged whose core task is to retrieve diverse social images [21] given a predefined

[20] http://www.multimediaeval.org/mediaeval2014/placing2014/

[21] http://www.multimediaeval.org/mediaeval2015/

dataset typically derived from Flickr social network. Online sharing of large Flickr datasets among researchers could also be a collective solution to the problem, although the complexity and the decentralized nature of research groups worldwide significantly hinder this task. In conjunction to this statement, the fact that most online multimedia content shared within Flickr and other social networks suffers from at least indistinctive or obscured licensing schemes, constitutes its utilization for research purposes another challenging task. In this direction Flickr is considered to be a pioneer by providing huge amounts of people's photos under the default CC license, still other social networks should follow and simplify access to information.

Clearly all the above discussion justifies our intention to identify the trends in the surveyed areas and organize them in a practical way, so as for fellow researchers to be able to identify an efficient point of future reference. In doing so we went down both the formal, in the sense of introducing specific Tables within each group, and the novel way, in the sense of introducing a "suitable for" table column for many rather hard to grasp or quantify groups, such as the knowledge representation group of works. Last but not least, among our personal future tasks remains the close follow-up and update of this survey according to the future research directions to come, combined possibly together with its extension to other popular social networks, like Twitter, Facebook, Instagram and Youtube, provided their publicly available API will allow us to do so. Finally, an interesting and challenging task would be to extend this review to other multimedia content types, such as text snippets and/or video sequences.

All in all, herein we have presented a comprehensive survey highlighting current progress, emerging directions, some innovative ideas, and methods for efficient evaluation relevant to the framework of the popular Flickr social network. From the outcomes of our review it seems that although social networks form a "young" notion, the benefits of social networking are largely associated with the participatory humanistic nature of our "digital world" and have significantly influenced in a positive manner the corresponding research efforts over the recent years. We consider that this review captures only a snapshot of the exciting research field of Flickr data analysis and retrieval and cast it as an overview of its early years; we expect it to be at least revised in the near future to include specific future directions alongside. Meanwhile, we do hope that the quest for efficient multimedia content management methodologies will continue and that similar to Flickr paradigms will emerge to the benefit of the research community.

9 Conclusions

Due to the popularity of social networks such as Flickr and the increasing large amount of digital multimedia data uploaded and shared among its users, enabling a thorough, yet compact, record of related multimedia research efforts is a rather challenging task. In this paper, we reviewed the latest advances in research efforts that use Flickr social network as the source of both their data and analysis. These works focus mainly on the broader areas of humanistic data collection and interpretation, as well as the semantic and social, user-generated, multimedia content adaptation. More specifically, we emphasized on algorithms, procedures and methodologies dealing with three identified aspects of multimedia content retrieval, namely text, visual and hybrid retrieval, issues of automatic tag/geo-tag generation, reconstruction approaches, recommendation systems and travel/route suggestion systems, as well as on techniques that aim to extract knowledge from metadata and on approaches that track human activities and related benchmarks. Based on the challenge discussions and interpretation of each group, future research directions may be identified.

References

[1] R. Abbasi, S. Chernov, W. Nejdl, R. Paiu and S. Staab, *Exploiting Flickr Tags and Groups for Finding Landmark Photos.* Advances in Information Retrieval, Springer, 2009.

[2] S. Agarwal, Y. Furukawa, N. Snavely,I. Simon, B. Curless, S.M. Seitz and R. Szeliski, *Building rome in a day.* Communications of the ACM, vol.54, no.10, pp. 105–112, 2011.

[3] S. Ahern, M. Naaman, R. Nair and J.H.-I. Yang, *World explorer: visualizing aggregate data from unstructured text in geo-referenced collections.* Proc. of the ACM/IEEE-CS JCDL, 2007.

[4] M. Ames and M. Naaman, *Why we tag: motivations for annotation in mobile and online media.* In Proc. of the ACM CHI, 2007.

[5] A. Anderson, K. Ranghunathan and A. Vogel, *Tagez: Flickr tag recommendation.* Association for the Advancement of Artificial Intelligence, 2008.

[6] E. Angus and M. Thelwall, *Motivations for image publishing and tagging on flickr*. In Proc. of 14th Elpub, 2010.

[7] Y. Avrithis, Y. Kalantidis, G. Tolias and E. Spyrou, *Retrieving landmark and non-landmark images from community photo collections*. In Proc. of ACM MM, 2010.

[8] C. Baber, J. Cross, T. Khaleel and R. Beale, *Location-based photography as sense-making*. In Proc. of BCS HCI, 2008.

[9] M. Batko, F. Falchi, C. Lucchese, D. Novak, R. Perego, F. Rabitti, J. Sedmidubsky and P. Zezula, *Building a web-scale image similarity search system*. Multimedia Tools and Applications, vol.47, no.3, pp.599–629, Springer, 2010.

[10] D. Barchiesi, T. Preis, S. Bishop and H. S. Moat, *Modelling human mobility patterns using photographic data shared online*. Royal Society Open Science 2.8: 150046, 2015.

[11] J.M. Barrios, D. Dıaz-Espinoza and B. Bustos, *Text-Based and Content-Based Image Retrieval on Flickr: DEMO*. In Proc. of IEEE SISAP, 2009.

[12] P. Bartie and W. Mackaness *Mapping the visual magnitude of popular tourist sites in Edinburgh city*. Journal of Maps, Taylor & Francis, 2015.

[13] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, *Speeded-up robust features (SURF)*. In Computer Vision and Image Understanding, vol.110, no.3, pp.346–359, Elsevier, 2008.

[14] I. Bojic, E. Massaro, A. Belyi, S. Sobolevsky and C. Ratti, *Choosing the right home location definition method for the given dataset*. arXiv:1510.03715 [cs.SI], 2015.

[15] G. Cai, C. Hio, L. Bermingham, L. Kyungmi and I. Lee, *Mining Frequent Trajectory Patterns and Regions-of-Interest from Flickr Photos*. In Proc. of HICSS, 2014

[16] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, and T.S. Huang, *Aworld-wide tourism recommendation system based on geotaggedweb photos*. In Proc. of IEEE ICASSP, 2010.

[17] L. Cao, X.M. Liu, W. Liu, R. Ji and T. Huan, *Localizing web videos using social images.* Information Sciences 302, pp. 122–131, Elsevier, 2015

[18] M. Cha, A. Mislove and K.P. Gummadi, *A measurement-driven analysis of information propagation in the flickr social network.* In Proc. of ACM WWW, 2009.

[19] E. Chatzilari, S. Nikolopoulos, S. Papadopoulos, C. Zigkolis and Y. Kompatsiaris, *Semi-supervised object recognition using flickr images.* In Proc. of IEEE CBMI, 2011.

[20] O. Chaudhry and W. Mackaness, *Automated extraction and geographical structuring of Flickr tags.* Proc. of GEOProcessing, 2012.

[21] H.M. Chen, M.H. Chang, P.C. Chang, M.C. Tien, W.H. Hsu and J.L. Wu, Ja-Ling *SheepDog: group and tag recommendation for flickr photos by automatic search-based learning.* In Proc. of ACM MM, 2008.

[22] L. Chen and A. Roy, *Event detection from flickr data through wavelet-based spatial analysis.* In Proc. of ACM CIKM, 2009.

[23] W.-C. Chen, A. Battestini, N. Gelfand and V. Setlur, *Visual summaries of popular landmarks from community photo collections.* In Proc. of IEEE ASILOMAR, 2009.

[24] X. Chen and H. Shin, *Tag recommendation by machine learning with textual and social features.* Journal of Intelligent Information Systems, vol.40, no.2, pp.261–282, Springer, 2013.

[25] J. Cheng and D. Cosley, *How annotation styles influence content and preferences.* In Proc. of ACM HT, 2013.

[26] M. Clements, P. Serdyukov, A.P. de Vries and M.J.T. Reinders, *Using flickr geotags to predict user travel behaviour.* In Proc. of ACM SIGIR, 2010.

[27] A.M. Cox, P.D. Clough and J. Marlow, *Flickr: a first look at user behaviour in the context of photography as serious leisure.* Information Research, vol.13, no.1, 2008.

[28] A. Cox, P. Clough and S. Siersdorfer, *Developing metrics to characterize Flickr groups.* Journal of the American Society for Information Science and Technology, vol.62, no.3, pp.493–506, Wiley Online Library, 2011.

[29] D.J. Crandall, L. Backstrom, D. Huttenlocher and J. Kleinberg, *Mapping the world's photos.* In Proc. of of ACM WWW, 2009.

[30] G. Csurka, C. Dance, L.X. Fan, J. Willamowski and C. Bray, *Visual categorization with bags of keypoints.* In Proc. of ECCV, 2004.

[31] S. J. Cunningham and M. Mahoui, *Interacting with and through a digital library collection: commenting behavior in flickrs the commons.* In Proc. of the ACM/IEEE-CS JCDL, 2013.

[32] C. De Rouck, O. Van Laere, S. Schockaert and B. Dhoedt, *Georeferencing Wikipedia pages using language models from Flickr.* In Proc. of Terra Cognita 2011, in conjuction with ISWC, 2011.

[33] J. Derrac and S. Schockaert, *Enriching Taxonomies of Place Types Using Flickr.* Lecture Notes in Computer Science, Vol.8367, pp.174–192, Springer, 2014.

[34] J.A. Donaire, R. Camprubi1 and N. Gali, *Tourist clusters from Flickr travel photography* Tourism Management Perspectives, Vol.11, pp 26–33, 2014.

[35] M. Everingham, A. Zisserman, C. Williams and L. Van Gool. *The pascal visual object classes challenge 2006 (voc 2006) results.*

[36] J. Fan, D.A. Keim, Y. Gao, H. Luo and Z. Li, *JustClick: Personalized image recommendation via exploratory search from large-scale Flickr images.* IEEE Trans. on Circuits and Systems for Video Technology, vol.19, no.2, pp.273–288, 2009.

[37] C.S. Firan, M. Georgescu, W. Nejdl and R. Paiu, *Bringing order to your photos: event-driven classification of flickr images based on social knowledge.* In Proc. of ACM CIKM, 2010.

[38] A.J. Flanagin and M.J. Metzger, *The credibility of volunteered geographic information.* GeoJournal, vol.72, no.3–4, pp.137–148, Springer, 2008.

[39] G. Friedland, J. Choi and A. Janin, *Video2GPS: a demo of multimodal location estimation on flickr videos*. In Proc. of ACM MM, 2011.

[40] G. Friedland, J. Choi, H. Lei, Howard and A. Janin, *Multimodal location estimation on Flickr videos*. In Proc. of ACM WSM, 2011.

[41] A. Gallagher, D. Joshi, J. Yu and J. Luo, *Geo-location inference from image content and user tags*. In Proc. of IEEE CVPR, 2009.

[42] S. Gammeter, L. Bossard, T. Quack and L. Van Gool, *I know what you did last summer: object-level auto-annotation of holiday snaps*. In Proc. of IEEE ICCV, 2009.

[43] N. Garg and I. Weber, *Personalized tag suggestion for flickr*. In Proc. of ACM WWW, 2008.

[44] N. Garg and I. Weber, *Personalized, interactive tag recommendation for flickr*. In Proc. of ACM RecSys, 2008.

[45] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo and S.F. Chang, Image Popularity Prediction in Social Media Using Sentiment and Context Features. In Proc. of ACM MM, 2015.

[46] J. Gonzalo, J. Karlgren and P. Clough, *iCLEF 2006 Overview: Searching the Flickr WWW photo-sharing repository*. Evaluation of Multilingual and Multi-modal Information Retrieval, Springer,2007.

[47] T. Hammond, T. Hannay, B. Lund, and J. Scott, *Social bookmarking tools (i): A general review*. D-lib Magazine, 2005.

[48] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang and L. Zhang, *Equip tourists with knowledge mined from travelogues*. In Proc. of ACM WWW, 2010.

[49] J. Hare, J. Davies, S. Samangooei and P.H. Lewis, *Placing Photos with a Multimodal Probability Density Function*. In Proc. of ACM ICMR, 2014.

[50] Q. Hao, R. Cai, J.-M. Yang, R. Xiao, L. Liu, S. Wang and L. Zhang, *Travelscope: standing on the shoulders of dedicated travelers*. In Proc. of ACM MM, 2009.

[51] C. Harris and M. Stephens, *A combined corner and edge detector*. In Proc. of Alvey vision conference, 1988.

[52] C. Hauff and G.-E. Houben, *Geo-Location estimation of flickr images: social web based enrichment*. Adv. in Information Retrieval, LNCS vol. 7224, pp. 85–96, Springer, 2012.

[53] J. Hays and A.A. Efros, *IM2GPS: estimating geographic information from a single image*. In Proc. of IEEE CVPR, 2008.

[54] L. Hollenstein and R. Purves *Exploring place through user-generated content: Using Flickr tags to describe city cores*. Journal of Spatial Information Science, no.1, pp.21–48, 2013.

[55] L.C. Hsieh and W.H. Hsu, *Search-Based Automatic Image Annotation via Flickr Photos Using Tag Expansion.*. In Proc. of IEEE ICASSP, 2010.

[56] M.J. Huiskes and M.S. Lew, *The MIR Flickr Retrieval Evaluation*. In Proc. of ACM MIR, 2008.

[57] M.J. Huiskes, B. Thomee and M.S. Lew, *New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative*. In Proc. of ACM MIR, 2010.

[58] B. Ionescu, A. Popescu, M. Lupu, A.L. Ginsca, B. Boteanu and H. Muller, *Div150Cred: A social image retrieval result diversification with user tagging credibility dataset*. In Proc ACM MMSys, 2015.

[59] H. Izadinia, B.C. Russell, A. Farhadi, M.D. Hoffman and A. Hertzmann, *Deep Classifiers from Image Tags in the Wild*. In Proc. of ACM MM-Commons, 2015.

[60] A. Jaffe, M. Naaman, T. Tassa and M. Davis, *Generating summaries for large collections of geo-referenced photographs*. In Proc. of ACM WWW, 2006.

[61] S. Jain, S. Seufert and S. Bedathur, *Antourage: Mining Distance-Constrained Trips from Flickr*. In Proc. of ACM WWW, 2010.

[62] X. Jin, A. Gallagher, L. Cao, J. Luo and J. Han, *The wisdom of social multimedia: using flickr for prediction and forecast*. In Proc. of ACM MM, 2010.

[63] S. Jiang, X. Qian, J. Shen and T. Mei *Travel Recommendation via Author Topic Model Based Collaborative Filtering*. Lecture Notes in Computer Science, Vol. 8936, pp.392–402, Springer, 2015.

[64] Y. Jing, X. Zhang, L. Wu, J. Wang, Z. Feng and D. Wang *Recommendation on Flickr by combining community user ratings and item importance*. In Proc. of IEEE ICME, 2014

[65] J. Jordan, *What's Flickr all about?*, `http://www.digicamhelp.com/processing-photos/photo-hosting/all-about-flickr/`, retrieved 2015-04-29.

[66] D. Joshi, and J. Luo, *Inferring generic activities and events from image content and bags of geo-tags*. In Proc. of ACM CBIVR, 2008.

[67] D. Joshi, A. Gallagher, J. Yu and J. Luo, *Exploring user image tags for geo-location inference*. In Proc. of IEEE ICASSP, 2010.

[68] J.J. Jung, *Understanding information propagation on online social tagging systems: a case study on Flickr*. Quality & Quantity, vol.48, no.2, pp. 745–754, Springer, 2014.

[69] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, Ph. Mylonas and S. Kollias, *VIRAL: Visual image retrieval and localization*. Multimedia Tools and Applications, vol.51, no.2, pp. 555–592, Springer, 2011.

[70] E. Kalogerakis, O. Vesselova, J. Hays, A.A. Efros and A. Hertzmann, *Image sequence geolocation with human travel priors*. In Proc. of IEEE ICCV, 2009.

[71] R.S. Kaminsky, N. Snavely, S.M. Seitz and R. Szeliski, *Alignment of 3D point clouds to overhead images*. In Proc. of IEEE CVPR, 2009.

[72] P. Kelm, S. Schmiedeke and T. Sikora, *A hierarchical, multi-modal approach for placing videos on the map using millions of flickr photographs*. In Proc. of ACM SBNMA, 2011.

[73] L. Kennedy, M. Naaman, S. Ahern, R. Nair and T. Rattenbury *How flickr helps us make sense of the world: context and content in community-contributed media collections*. In Proc. of ACM MM, 2007.

[74] L.S. Kennedy and M. Naaman, *Generating diverse and representative image search results for landmarks.* In Proc. of ACM WWW, 2008.

[75] S. Kisilevich, D. Keim, N. Andrienko and G. Andrienko, *Towards Acquisition of Semantics of Places and Events by Multi-perspective Analysis of Geotagged Photo Collections.* Geospatial Visualisation, pp. 211–233, Springer, 2013.

[76] J. Kleban, E. Moxley, J. Xu and B.S. Manjunath, *Global annotation on georeferenced photographs.* In Proc. of ACM CIVR, 2009.

[77] M. Larson, C. Kofler and A. Hanjalic, *Reading between the tags to predict real-world size-class for visually depicted objects in images.* In Proc. of ACM MM, 2011.

[78] I. Lee, G. Cai and K. Lee, *Exploration of geo-tagged photos through data mining approaches.* Expert Systems with Applications, vol.41, no.2, pp.397–405, Elsevier, 2014.

[79] H. Lei, J. Choi and G. Friedland, *Multimodal city-verification on flickr videos using acoustic and textual features.* In Proc. of IEEE ICASSP, 2012.

[80] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr, *Social media and mobile Internet use among teens and young adults*, Retrieved from `http://www.pewinternet.org/Reports/2010/Social-Media-and-Young-Adults.aspx`, 2010.

[81] K. Lerman and L. Jones, *Social browsing on flickr.* arXiv preprint cs/0612047, 2006.

[82] K. Lerman, A. Plangprasopchok and C. Wong, *Personalizing Image Search Results on Flickr.* CoRR, vol. abs/0704.1676, 2007.

[83] D. Leung and S. Newsam, *Proximate sensing: Inferring what-is-where from georeferenced photo collections.* In Proc. of IEEE CVPR, 2010.

[84] X. Li, C.G.M. Snoek and M. Worring, *Learning tag relevance by neighbor voting for social image retrieval.* In Proc of ACM MIR, 2008.

[85] X. Li, C. Wu, C. Zach, S. Lazebnik and J.-M. Frahm, *Modeling and recognition of landmark image collections using iconic scene graphs*. In Proc. of ECCV, 2008.

[86] J. Li, X. Qian, Y.Y. Tang, L. Yang and T. Mei, *GPS Estimation for Places of Interest From Social Users' Uploaded Photos*. Transactions on Multimedia, vol.15, no.8, pp. 2058–2071, IEEE, 2013.

[87] J. Li, Y. Yang and W. Liu, *Exploring Personalized Travel Route Using POIs*. Int. Journal of Computer Theory and Engineering, vol. 7, no. 2, pp.126–131, 2015.

[88] Y. Li, D.J. Crandall, and D.P. Huttenlocher, *Landmark classification in large-scale image collections*. In Proc. of IEEE ICCV, 2009.

[89] K.H. Lim, *Recommending Tours and Places-of-Interest based on User Interests from Geo-tagged Photos*. In Proc. of SIGMOD PhD Symposium, 2015.

[90] K.H. Lim, J. Chan, C. Leckie and S. Karunasekera *Personalized Tour Recommendation based on User Interests and Points of Interest Visit Durations*. In Proc. of IJCAI, 2015.

[91] B. Liu, Q. Yuan, G. Cong and D. Xu, *Where your photo is taken: Geolocation prediction for social images*. Journal of the Association for Information Science and Technology, Wiley Online Library, 2014.

[92] S. Liu, P. Cui, H. Luan, W. Zhu, S. Yang and Q. Tian, *Social-oriented visual image search*. Computer Vision and Image Understanding, vol. 118, pp. 30–39, Elsevier, 2014.

[93] X. Liu, X. Qian, D. Lu, X. Hou and L. Wang, *Personalized tag recommendation for Flickr users*. In Proc. of IEEE ICME, 2014.

[94] D.G. Lowe, *Object recognition from local scale-invariant features*. In Proc. of IEEE ICCV, 1999.

[95] D. Lu and Q. Li, *Exploiting semantic hierarchies for flickr group*. Active Media Technology, pp.74–85, Springer, 2010.

[96] Z. Lu, X. Gao, S. Huang, L. Wang and J.R. Wen, *Social Image Parsing by Cross-Modal Data Refinement* Proc. of IJCAI, 2015.

[97] J. Luo, J. Yu, D. Joshi, Dhiraj and W. Hao, *Event recognition: viewing the world with a third eye*. In Proc. of ACM MM,2008.

[98] J. Luo, D. Joshi, J. Yu, and A. Gallagher, *Geotagging in multimedia and computer visiona survey*. Multimedia Tools and Applications, vol.51, no.1, pp.187–211, Springer, 2011.

[99] A. Majid, L. Chen, Ling, G. Chen, H.T. Mirza, I. Hussain and J. Woodward, *A context-aware personalized travel recommendation system based on geotagged social media data mining*. Int. Journal of Geographical Information Science, vol.27, no.4, pp. 662–684, Taylor & Francis, 2013.

[100] T. Mao, *Mining One Hundred Million Creative Commons Flickr Images Dataset to Flickr Tourist Index* Int. Journal of Future Computer and Communication, vol.4, no.2, 2015.

[101] C. Marlow, M. Naaman, D. Boyd, and M. Davis, *Ht06, tagging paper, taxonomy, flickr, academic article, to read*. In Proc. of ACM HT, 2006.

[102] P.J. McParlane, Y. Moshfeghi and J.M. Jose, *"Nobody comes here anymore, it's too crowded"; Predicting Image Popularity on Flickr*. Proc. of ACM ICMR, 2014.

[103] A. Mislove, H.S. Koppula, K.P. Gummadi, P. Druschel and B. Bhattacharjee, *Growth of the flickr social network*. In Proc. of ACM WOSM, 2008.

[104] P.-A. Moëllic, J.-E. Haugeard and G. Pitel, *Image clustering based on a shared nearest neighbors approach for tagged collections*. In Proc. of ACM CIVR, 2008.

[105] E. Moxley, J. Kleban and B.S. Manjunath, *Spirittagger: a geo-aware tag suggestion tool mined from flickr*. In Proc. of ACM MIR, 2008.

[106] E. Moxley, J. Kleban, J. Xu and B.S. Manjunath, *Not all tags are created equal: Learning Flickr tag semantics for global annotation*. In Proc. of IEEE ICME, 2009.

[107] R.-A. Negoescu, B. Adams, D. Phung, S. Venkatesh and D. Gatica-Perez, *Flickr Hypergroups*. in Proc. of ACM MM, 2009.

[108] R.-A. Negoescu and D. Gatica-Perez. *Modeling flickr communities through probabilistic topic-based analysis*, IEEE Trans. on Multimedia, vol.12, no.5, pp.399–416, 2010.

[109] K. Ni, R. Pearce, K. Boakye, B. Van Essen, D. Borth, B. Chen and E. Wang, *Large-Scale Deep Learning on the YFCC100M Dataset.* arXiv preprint arXiv:1502.03409, 2015.

[110] N. Nitta, Y. Kumihashi, T. Kato and N. Babaguchi, *Real-World Event Detection Using Flickr Images.* MultiMedia Modeling, pp. 307–314, Springer, 2014.

[111] O. Nov, M. Naaman, and C. Ye, *What drives content tagging: the case of photos on flickr.* In Proc. of ACM CHI, 2008.

[112] O. Nov, M. Naaman, Mor and C. Ye, *Analysis of participation in an online photo-sharing community: A multidimensional perspective.* Journal of the American Society for Information Science and Technology, vol.61, no.3, pp.555–566, Wiley Online Library, 2010.

[113] N. OHare and V. Murdock, *Modeling locations with social media.* Information retrieval, vol.16, no.1, pp. 30–62, Springer, 2013.

[114] A. Oliva and A. Torralba, *Modeling the shape of the scene: a holistic representation of the spatial envelope.* International Journal of Computer Vision, vol.42, no.3, pp.145–175, Springer, 2001.

[115] G. Panteras, S. Wise, X. Lu, A. Croitoru, A. Crooks and A. Stefanidis, *Triangulating Social Multimedia Content for Event Localization using Flickr and Twitter.* Transactions in GIS, Wiley, 2014.

[116] J. Philbin and A. Zisserman *Object mining using a matching graph on very large image collections.* In Proc. of IEEE ICVGIP, 2008.

[117] A. Plangprasopchok and K. Lerman, *Constructing Folksonomies from User-Specified Relations on Flickr.* In Proc. of ACM WWW, 2009.

[118] A. Popescu and G. Grefenstette, *Deducing trip related information from flickr.* In Proc. of ACM WWW, 2009.

[119] A. Popescu, G. Grefenstette and P.-A. Moëllic, *Mining tourist information from user-supplied collections.* In Proc. of ACM CIKM, 2009.

[120] A. Popescu and P.-A. Moëllic, *MonuAnno: automatic annotation of georeferenced landmarks images*. In Proc. of ACM CIVR, 2009.

[121] C. Prieur, D. Cardon, J. S. Beuscart, N. Pissard, and P. Pons, *The strength of weak cooperation: A case study on flickr*. arXiv preprint arXiv:0802.2317, 2008.

[122] X. Qian, D. Lu and X. Liu, *Image Retrieval by User-oriented Ranking*. In Proc. of ACM ICMR, 2015.

[123] T. Quack, B. Leibe and L. Van Gool, *World-scale mining of objects and events from community photo collections*. In Proc. of ACM CIVR, 2008.

[124] T. Rattenbury, N. Good and M. Naaman, *Towards automatic extraction of event and place semantics from flickr tags*. In Proc. of ACM SIGIR, 2007.

[125] J. Sang, C. Xu and D. Lu, *Learn to personalized image search from the photo sharing websites*. IEEE Trans. on Multimedia, vol.14, no.4, pp. 963–974, 2012.

[126] P. Schmitz, *Inducing ontology from flickr tags*. In Proc. of ACM WWW, 2006.

[127] B.-S. Seah, S.S. Bhowmick and A. Sun, *Summarizing social image search results*. In Proc. of ACM WWW Companion, 2014.

[128] P. Serdyukov, V. Murdock and R. Van Zwol, *Placing flickr photos on a map*. Proc. of ACM SIGIR, 2009.

[129] B. Sigurbjörnsson and R. Van Zwol, *Flickr tag recommendation based on collective knowledge*. In Proc. of ACM WWW, 2008.

[130] I. Simon, Ian and Snavely, Noah and Seitz, Steven M *Scene Summarization for Online Image Collections.*. In Proc. of ICCV, 2007.

[131] V.K. Singh, M. Gao and R. Jain, *Social pixels: genesis and evaluation*. In Proc. of ACM MM, 2010.

[132] N. Snavely, S.M. Seitz, and R. Szeliski, *Photo tourism: exploring photo collections in 3D*. ACM Transactions on Graphics (TOG), vol.25, no.3, pp.835–846, 2006.

[133] N. Snavely, R. Garg, S.M. Seitz and R. Szeliski, *Finding paths through the world's photos*. ACM Transactions on Graphics (TOG), vol.27, no.3, pp.11-21, 2008.

[134] S. Sobolevsky, I. Bojic, A. Belyi, I. Sitko, B. Hawelka, J.M. Arias and C. Ratti, *Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity*. arXiv:1504.06003 [cs.SI], 2015.

[135] S.E. Spielman, *Spatial collective intelligence? Credibility, accuracy, and volunteered geographic information*. Cartography and Geographic Information Science, vol.41, no.2, pp.115-124, Taylor & Francis, 2014.

[136] E. Spyrou and Ph. Mylonas, *Placing User-Generated Photo Metadata on a Map*. In Proc. of IEEE SMAP, 2011.

[137] E. Spyrou and Ph. Mylonas, *From Tags to Trends: A First Glance at Social Media Content Dynamics*. In Proc. of MHDW, 2012.

[138] E. Spyrou and Ph. Mylonas, *Analyzing Flickr metadata to extract location-based information and semantically organize its photo content*. Neurocomputing, Vol. 172, pp. 114–133, 2016.

[139] B. Stvilia and C. Jörgensen, *Member Activities and Quality of Tags in a Collection of Historical Photographs in Flickr*. Journal of the American Society for Information Science and Technology, vol.61, no.12, pp.2477–2489, 2010.

[140] Y. Sun, H. Fan, M. Bakillah and A. Zipf, *Road-based travel recommendation using geo-tagged images*. Computers, Environment and Urban Systems, Elsevier, 2013.

[141] B. Thomee, D.A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth and L.-J. Li, *The New Data and New Challenges in Multimedia Research*. arXiv preprint arXiv:1503.01817, 2015.

[142] K. Tuite, N. Snavely, D.-Y. Hsiao, N. Tabing and Z. Popovic, *PhotoCity: training experts at large-scale image acquisition through a competitive game*. In Proc. of ACM SIGCHI Conf. on Human Factors in Computing Systems, 2011.

[143] A. Ulges, M. Worring and T. Marcel, *Learning visual contexts for image annotation from flickr groups*. IEEE Trans. on Multimedia, vol.13, no.2, pp.330–341, IEEE, 2011.

[144] M. Valafar, R. Rejaie and W. Willinger, *Beyond friendship graphs: a study of user interactions in Flickr*. In Proc. of ACM WOSN, 2009.

[145] S. Van Canneyt, S. Schockaert, O. Van Laere and B. Dhoedt, Bart, *Time-dependent recommendation of tourist attractions using Flickr*. In Proc. of BNAIC, 2011.

[146] J. Van Dijck, *Flickr and the culture of connectivity: Sharing views, experiences, memories*. Memory Studies, vol.4, no.4, pp.401–415, SAGE Publications, 2011.

[147] N.A. Van House, *Flickr and public image-sharing: distant closeness and photo exhibition*. In Proc. of ACM CHI, 2007.

[148] O. Van Laere, S. Schockaert and B. Dhoedt, *Combining multi-resolution evidence for georeferencing Flickr images*. Scalable Uncertainty Management, pp. 347–360, Springer, 2010.

[149] O. Van Laere, S. Schockaert and B. Dhoedt, *Towards automated georeferencing of flickr photos*. In Proc. of ACM GIR, 2010.

[150] O. Van Laere, S. Schockaert and B. Dhoedt, *Finding locations of flickr resources using language models and similarity search*. In Proc. of ACM ICMR, 2011.

[151] L. Von Ahn and L. Dabbish, *Labeling images with a computer game*. In Proc. of SIGCHI conf on Human Factors in Computing Systems, ACM, 2004.

[152] G. Wang, D. Hoiem and D. Forsyth, *Learning image similarity from flickr groups using stochastic intersection kernel machines*. Proc. of IEEE ICCV, 2009.

[153] J. Wang, Z. Zhao, J. Zhou, H. Wang, B. Cui and G. Qi, *Recommending Flickr groups with social topic model.* Information retrieval, vol.15, no.3–4, pp.278–295, Springer, 2012.

[154] M. Wang, B. Ni, X. S. Hua, and T. S. Chua, *Assistive tagging: A survey of multimedia tagging with human-computer joint exploration.* ACM Computing Surveys (CSUR), 44(4):25, 2012.

[155] Y. Wang, Y. Xia, S. Tang, F. Wu and Y. Zhuang, *Flickr Group Recommendation via Heterogeneous Information Networks* In Proc. of ACM ICIMCS, 2015.

[156] Y. Wang, X. Lin, L. Wu and W. Zhang, *Effective Multi-Query Expansions: Robust Landmark Retrieval.* In Proc. of ACM MM, 2015.

[157] M. Winget, *User-defined classification on the online photo sharing site flickr or, how i learned to stop worrying and love the million typing monkeys.* Advances in Classification Research Online, 17(1):116, 2006.

[158] J. Wu, H. Sun and Y. Tan, *Social media research: A review.* Journal of Systems Science and Systems Engineering, vol.22, no.3, pp.257–282, Springer, 2013.

[159] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma and S. Li, *Flickr distance.* In Proc. of ACM MM, 2008.

[160] K. Yanai, H. Kawakubo and B. Qiu, *A visual analysis of the relationship between word concepts and geographical locations.* In Proc. of ACM ICIVP, 2009.

[161] K. Yanai, K. Yaegashi and B. Qiu, *Detecting cultural differences using consumer-generated geotagged photos.* In Proc. of ACM LOCWEB, 2009.

[162] T. Yao, Y. Liu, C.-W. Ngo and T. Mei, *Unified entity search in social media community.* In Proc. of ACM WWW, 2013.

[163] J. Yu and J. Luo, *Leveraging probabilistic season and location context models for scene understanding.* In Proc. of ACM CBIVR, 2008.

[164] *The man behind Flickr on making the service awesome again*, The Verge, 20.03.2013, retrieved 2015-04-17.

[165] H. Xie, Q. Li, X. Mao, X. Li, Y. Cai and Q. Zheng, *Mining Latent User Community for Tag-Based and Content-Based Search in Social Media.* The Computer Journal, vol.57, no.9, pp.1415–1430, British Computer Society, 2014.

[166] P. Xie, Y. Pei, Y. Xie and E. Xing *Mining User Interests from Personal Photos.* In Proc. of AAAI Conference on Artificial Intelligence, 2015.

[167] Z. Xu, X. Luo, Y. Liu, L. Mei and C. Hu *Measuring Semantic Relatedness between Flickr Images: From a Social Tag Based View.* The Scientific World Journal, vol. 2014, Hindawi, 2014.

[168] C. Xu, D. Tao, Y. Li and C. Xu *Large-margin multi-view Gaussian process.* Multimedia Systems vol.21, pp.147–157, Springer, 2015.

[169] Q. You, J. Luo, H. Jin and J. Yang, *Robust image sentiment analysis using progressively trained and domain transferred deep networks.* In Proc. of AAAI Conference on Artificial Intelligence, 2015.

[170] X. Zeng and L. Wei, *Social ties and user content generation: Evidence from Flickr.* Information Systems Research, vol.24, no.1, pp.71–87, 2013.

[171] S. Zerr, S. Siersdorfer, J. San Pedro and J. Hare, *NicePic! A system for extracting attractive photos from Flickr streams.* In Proc. of ACM SIGIR, 2014.

[172] H. Zhang, M. Korayem, E. You and D.J. Crandall, *Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities.* In Proc. of ACM WSDM, 2012.

[173] F. Zhao, Y. Huang, L. Wang, T. Tan, *Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval.* arXiv:1501.06272 [cs.CV], 2015.

[174] D. Zheng, T. Hu, Q. You, H. Kautz and J. Luo, *Towards Lifestyle Understanding: Predicting Home and Vacation Locations from Users Online Photo Collections* In Proc. of AAAI Conference on Web and Social Media, 2015.

[175] X. Zhou, C. Xu and B. Kimmons, Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform Computers, Environment and Urban Systems, vol.54, pp.144–153, Elsevier, 2015