Invited Paper

A Review of Web Image Mining

Keiji Yanai †

Abstract In this paper, we review works related to big visual data on the Web in the literature of computer vision and multimedia research regarding the following points: (1) Web image acquisition for construction of visual concept database for image/video recognition, (2) Web image application for visual concept analysis and data-driven computer graphics, and (3) real-world sensing through Web images to detect location-dependent and event-related visual information.

Key words: Web image mining, visual concepts, social media mining, visual event detection

1. Introduction

Nowadays there exist an enormous number of photos and videos on the Web, which are sometimes called “big visual data”, and their amount is increasing every second. Because of wide spread of social media sites where people can share photos and videos such as Facebook, Twitter, YouTube and Flickr as well as of camera devices connected to the Internet such as smartphones and tablet devices, people can easily post the photos and the video taken by themselves anytime from anywhere. In most of the cases, they add textual descriptions or keywords when posting images or videos to the Web to make them searchable. Thus, most of the photos and the videos on the web accompany textual information related to their contents such as textural tags or descriptions in the social media sites and surrounding texts in the Web pages. Therefore, they can be regarded as “weakly labeled data”. Such big visual data with weakly label information is really helpful for computer vision and multimedia research. The most representative usage of it is training data for object recognition. Recent progress of object recognition using deep learning heavily relies on big visual data on the Web. As other major applications of weakly labeled images and videos, they are used for analysis of the relation between words and images/videos, and for event photo detection to get to know what happens over the world intuitively and visually.

In this paper, we overview works on big visual data on the Web in the literature of computer vision and multimedia research regarding the following points: (1) Web image acquisition for construction of visual concept database for image/video recognition, (2) Web image application for visual concept analysis and data-driven computer graphics, and (3) real-world sensing through Web images to detect location-dependent and event-related visual information. Table 1 shows the detail of the structure of this survey paper.

2. Web Mining for Visual Concepts

For these ten years, object recognition technology has been progressed greatly. Local features and bag-of-features representation pioneered the possibility of generic object recognition, and the recent advent of deep convolutional neural network (DCNN) pushed up its limit greatly. Especially, the state-of-the-art DCNN-based methods enable computer vision on large-scale object recognition to outperform human object recognition. In addition to the advent of the new method, in fact, great recent advance on object recognition heavily relies on big photo/video data on the Web, most of which are uploaded by a great many number of common people. The biggest organized image database, ImageNet, was created by collecting Web images and selecting relevant images to each of in-
In this section, we overview the works on collecting visual data from the Web from the early Web image search works to the recent crowd-sourced image collection and large-scale visual concept gathering.

2.1 Progress of Web Image Search Engines

Commercial Web image search engines such as AltaVista and Google Image Search launched around 2000. Before then, only Web text search engines existed. Because the early Web image search engines relied only on surrounding texts of image-embedding tags (<IMG>) in Web pages, the preciseness of their search results was not so good. On the other hand, current commercial Web image search engines such as Google Image Search and Bing Image Search have been improved greatly by integrating users’ click-through statistics and visual feature analysis with the conventional surrounding text analysis and link analysis. Figure 1(a) and 1(b) shows the search results of Google Image Search for the query, “lion animal”, in 2003 and 2015, respectively. The difference in the quality is quite obvious.

Note that click-through log data was difficult to obtain except for the companies running commercial image search engines such as Google and Microsoft, although it was proved to be helpful to improve re-ranking accuracy. Hua et al.\(^5\) claimed that a large-scale click log of commercial Web image search engine could bridge both the semantic and intent gap. By taking advantage of click logs, recent Web image search engines always return images corresponding to the most prevailing meaning of the given query. For example, not any person photos but only ramen photos are always returned for the query “Jiro (丼)”, which is both a very common person name in Japan and the name of a popular ramen noodle chain. Recently, Microsoft Research Asia has released the click-through logs of Bing Image Search as “Clickture” dataset for research purpose. As an early work introducing click-through data into Web image search, Cheng et al.\(^2\) proposed to use click-through log as relevance feedback instead of user’s actual feedback.

2.2 Re-ranking of Web Image Search Results

Before commercial Web image search launched, some experimental Web image search engines employing both textual-based search and content-based image retrieval were proposed for research purpose. WebSeer\(^4\), Web-SEEK\(^1\), Image Rover\(^3\), and AMORE\(^6\) are represented works. These systems search for images based on the query keywords, and then a user selects query images from search results. After image selection by the user, the systems search for images similar to the query images based on image features. These systems carry out their search in an interactive manner.

Because interactive search needs human intervention, it is not appropriate for gathering a large number of images associated with many kinds of keywords. Then, Yanai proposed an automatic image gathering system, the Image Collector\(^1\), which enabled users to collect images associated to the given keywords without any manual labor. It adopted pseudo relevance feedback by HTML analysis, although it also had the image selection step based on image features. In addition, it did not need to prepare large-scale Web index in advance, because it intended to “re-rank” the outputs of commercial Web text search engines such as Google Web Search regarding the images embedded to the returned Web pages. Therefore, their system was able to collect real Web images on a large scale. In fact, they collected Web images associated with 150 kinds of adjectives for visual concept analysis\(^1\).

Yanai\(^1\) also proposed to use the images gathered from the Web as training samples for image classification, and proved that Web could be used as visual knowledge source for image classification by the experiments. However, because their image classification method relied on content-based image retrieval employing low-level visual features, the classification accuracy was not good. After that, they integrated their work with the image annotation method proposed by

Barnard et al.\textsuperscript{5,31} and improved the precision of the images gathered automatically from the Web\textsuperscript{143}. Figure 2 shows “waterfall” images. The label “waterfall” in each image indicates the regions with high probability of “waterfall”. Because it learns a region annotation model from the gathered images, the probability of “water” of each region in a given image can be estimated.

As the first work which introduced a state-of-the-art object recognition method into a Web image collecting task, Fergus et al.\textsuperscript{38} applied the constellation model\textsuperscript{37} to re-ranking of the results of Google Image Search. Their method was able to train object categories from imperfect training data containing outliers by employing RANSAC\textsuperscript{39}. They\textsuperscript{36} also proposed an extended model of pLSA (Probabilistic Latent Semantic Analysis)\textsuperscript{49} to train object categories from Google Image Search’s results.

Feng et al.\textsuperscript{34} introduced a co-training method\textsuperscript{11} into Web image re-ranking which employs a textual feature based classifier and a visual feature based classifier alternately. They used TFIDF-based word vectors as textual features, wavelet-based texture features and color histograms as visual features, and SVM as classifiers for both kinds of the features. Their method, however, needs small amount of annotation for bootstrapping.

Schroff et al.\textsuperscript{104,105} examined the effectiveness of combination of bag-of-features\textsuperscript{45} and SVM, which is one of the standard methods for object categorization. Their work proved that SVM with soft margin worked well even with noisy training data.

Li et al.\textsuperscript{75} proposed a Web image collecting system, OPTIMOL, employing Hierarchical Dirichlet Processes\textsuperscript{120}. Vijayanarasimhan et al.\textsuperscript{127} proposed to employ Multiple Instance Learning\textsuperscript{45} to handle noisy Web data for Web image re-ranking.

Jing et al.\textsuperscript{56} proposed a modified method of PageRank\textsuperscript{13} to re-rank images based on visual similarity between images. Although it was simple but effective, it had the drawback that it did not take into account diversity of the results. RankCompete\textsuperscript{14} improved VisualRank by simultaneous ranking and clustering for diverse results.

The above-mentioned work were oriented for generic purpose. On the other hand, some works on Web image collection for the specific kinds of targets were proposed at the same time. Song et al.\textsuperscript{146} proposed to collect facial photos of celebrities from the Web with Multiple Instance Learning. Berg et al.\textsuperscript{8} proposed to collect facial photos of celebrities from the Web news photos with clustering-based estimation of associations between names and faces. The same authors\textsuperscript{10} also proposed to gather animal photos from the Web. The advantage to restrict their target images to the specific kinds is that the specialized recognition methods for the targets such as face and animal are available.

2.3 Recognition with Web-scale Image Data

Although image re-ranking aimed to filtering out irrelevant images, Torralba et al.\textsuperscript{121} showed that very huge-scale visual data consisting of weakly-labeled 80 million images collected from the Web, which are called as “80 Million Tiny Images”, could be used as visual knowledge for nearest neighbor search-based image recognition, although the data includes noise images. They proved that Web-scale visual data could solve category-level image recognition by instance-level image search.

Similarly, by using two billion image dataset, which is a part of the image database of Bing Image Search, and a near duplicate image search method, Wang et al.\textsuperscript{130,131} confirmed that the data-driven approach was valid.

In the previous two works, raw Web images gathered without re-ranking are used as visual knowledge. On the other hand, Microsoft researchers and Google researchers independently tried to organize billion-scale raw Web image data employing image re-ranking methods with their huge parallel computing resources. Both the works can be regarded as trials to build an organized large-scale image database such as ImageNet\textsuperscript{26} automatically. Although the objectives of both were the same, the approaches were slightly different. The Microsoft group employed instance-level image recognition, while the Google group employed category-level image recognition.

Wang et al.\textsuperscript{129}, researchers in Microsoft Research, made much larger-scale experiment to build an image
knowledge (ImageKB) database consisting of 235.3 million images associated with half million concepts which are selected from two billion Bing images used in their previous works\textsuperscript{117,118}. They employed image annotation based on near duplicate detection, text-feature-based image filtering with one-vs-rest SVM, and re-ranking by inferring the authorities of images from their nearest neighbors. They indexed all the images based on a hierarchical word ontology which was automatically created from Web data\textsuperscript{109}.

Tsai et al.\textsuperscript{124}, researchers in Google, created a large-scale Web image database by re-ranking 200 million images associated with 300 thousand kinds of keywords which include compound words and proper nouns such as product names and place names. Regarding each keyword, they collected one thousand images at most from Google Image Search, then partitioned them into Visual Synsets in which the member images are visually similar and semantically similar to each other by using affinity propagation\textsuperscript{41}, and re-ranked images within each Visual Synsets employing linear SVM with image features. Since most of the images are shared with several Visual Synsets, each image has been annotated with several words as a result. In the experiments on image annotation, they used two million linear SVMs trained with Visual Synsets using a cluster of 2000 nodes. The obtained results outperformed simple instance-level search with raw Web images.

2.4 Visual Concept Acquisition with Crowd-sourcing

Although the works on Web image collection and re-ranking intended to collect as many relevant images to given keywords as possible without human labor, creating perfect database which includes no noise images was impossible. Even Visual Synsets\textsuperscript{124} and ImageKB\textsuperscript{129} contains some extent of noise images, which are expected to be much less than 80 Million Tiny Images\textsuperscript{211}. Since perfect image database is usually needed for training and evaluation of object recognition models, some datasets such as Caltech-101/256\textsuperscript{129} were created by researchers’ hands (mostly by PhD students for their PhD degrees). However, creating the larger datasets than Caltech-256 by the limited number of hands is too time-consuming and almost impossible.

To solve this problem, crowd-sourcing was introduced into manual image annotation tasks. The LabelMe project\textsuperscript{102} was the early form of the crowd-sourced image annotation, which asked people (mostly researchers, students and their families) to draw object regions and give labels on the displayed images through the Web interface as volunteers. Since the voluntary labor of image annotation is boring and less motivated in general, no one (except for the mother of Torralba, the project leader of LabelMe) was not willing to do so. It took one year to annotate 20,000 regions, which was very slow pace for data collection.

The paid crowd-sourcing service, Amazon Mechanical Turk (AMT)*, has turned this situation completely. AMT mediates tasks which can be completed via the Web interface between task requesters and workers over the world. Sorokin et al.\textsuperscript{117} proposed to use AMT for image annotation and showed that AMT enabled a large-scale and very rapid annotation at low cost. Fei-Fei et al.\textsuperscript{26}, inspired by Sorokin’s work, started building a very large-scale hand-labeled image database, ImageNet**, for computer vision research. Currently it contains 14,197,122 images associated with 21,841 nouns. To build ImageNet, they obtained image URLs by sending noun words in several languages as queries to Web Image Search engines and Flickr via their officially-provided Web APIs, and crawled all the images of the obtained URLs. Then they sent images to AMT to ask anonymous crowd workers to select only relevant images to the given noun words.

The ImageNet database contributed the research community of computer vision very greatly. In fact, 1.2 million images of 1000 categories which are a subset of ImageNet were released as the dataset for Large Scale Visual Recognition Challenge\textsuperscript{***}. This is used as a standard large-scale object recognition dataset to train Deep Convolutional Neural Network\textsuperscript{\textsuperscript{70,107}}.

By using crowd-sourcing and Web image gathering via the Web APIs of commercial Web image search engines and photo sharing social media, various kinds of image datasets have released such as 200 kinds of “birds”\textsuperscript{132}, 500 kinds of “birds”\textsuperscript{7}, 102 kinds of “aircrafts”\textsuperscript{84}, 120 kinds of “dogs”\textsuperscript{64} and 256 kinds of “foods”\textsuperscript{125}. They were intended to be built for fine-grained visual categorization research.

AMT is also used for other tasks than building fine-grained datasets. For example, Patterson et al.\textsuperscript{95} added attributes to a large-scale SUN scene image database\textsuperscript{35} using AMT, and Bell et al.\textsuperscript{6} created the OpenSurface database consisting of thousands of examples of surfaces

\textsuperscript{*} https://www.mturk.com/
\textsuperscript{**} http://image-net.org/
\textsuperscript{***} http://www.image-net.org/challenges/LSVRC/
segmented from Flickr photos of interiors, and annotated with material parameters (reflectance, material names), texture information (surface normals, rectified textures), and contextual information (scene category, and object names) with AMT.

In some works, AMT was incorporated into object recognition procedures, which was called “humans in the loop”. Vijayanarasimhan et al.\textsuperscript{128} proposed to combine active learning of object detectors and AMT crowd-sourcing tasks to draw bounding boxes as a loop procedure to raise accuracy of object detection gradually. On the other hand, Branson et al.\textsuperscript{12} proposed complementary use of AMT with object classifiers by giving AMT workers simple easy questions to tackle difficult fine-grained object classification. Kawano et al.\textsuperscript{61} proposed to use AMT after applying automatic image re-ranking based on domain-adaptation which utilizes the knowledge of existing image categories for automatic extension of food image database.

Note that not all the AMT workers are trusty, and some of them try to cheat requesters to save their time. Figure 3 shows the cheated results (red boxes) by an irresponsible AMT worker for the task to annotate bounding boxes on food regions. To prevent requesters from being cheated, tasks should be asked to multiple works and integrate multiple results to get final results. More importantly, requesters need to reject inappropriate results and cancel to pay rewards for untrustworthy workers properly.

2.5 Visual Concept Learning from the Web

As mentioned in Section 2.1, the precision of the search results of commercial Web image search has been improved greatly. However, researches on Web image re-ranking are still actively being explored. Differently from the previous old works, the recent works focus more on the aspect on visual concept learning, and employ the recent new methods such as mid-level discriminative patches\textsuperscript{110} and Deep Convolutional Neural Network (DCNN) activation features\textsuperscript{107}. Therefore, evaluation measures on the collected images are sometimes not only their precision but also classification accuracy by visual concept classifiers trained with gathered images on common dataset such as PASCAL VOC. Their objectives also have changed from collecting training data associated with general noun concepts such as “a dog” and “a car” to gathering images associated with fine-grained or compound words such as “a Saint Bernard dog” and “a classic car”. Moreover, some of them contain very large-scale experiments such as collecting 33,240 visual concepts\textsuperscript{97} and continuously collecting 1700 relations between visual concepts for 2.5 month with 200 core clusters\textsuperscript{20}.

Cheng et al.\textsuperscript{203} proposed NEIL (Never Ending Image Learner) which mined not only images but also relations between visual concepts at the same time from the Web. As relations, NEIL can learn object-object relations including part-of and instance-of relations, and object-attribute, scene-object, and scene-attribute relations. The same authors\textsuperscript{21} proposed to integrate object discovery and segmentation with NEIL.

Divvala et al.\textsuperscript{27} proposed LEVAN (Learning EVerything about ANything) which automatically collect compound visual concepts with bounding boxes. First they expanded a query word into the given word with various modifiers using Google Books Ngram. For example, “horse” was expanded into “eating horse”, “horse head” and “racing horse” after image-based pruning of non-visual compound words using corresponding images gathered from the Web. Then, they also merged compound concepts which are visually similar. Finally concept models are trained with modified version of a weakly supervised method\textsuperscript{93}, for training deformable part-based model (DPM)\textsuperscript{33}.

Yang et al.\textsuperscript{146} proposed supervised re-ranking using pre-built tagged image database and adaptive ranking-SVM. Tang et al.\textsuperscript{119} proposed IntentSearch which re-ranks a Web image search engine result by estimating a topic intended by a user from clicking of only one image. Yang et al.\textsuperscript{145} introduced bag-of-objects model into Web image re-ranking employing salient object detection\textsuperscript{39} and link analysis\textsuperscript{65} for object detection.

Xia et al.\textsuperscript{134} devised a method to select seeds used as training data of the classifier for image re-ranking. For seed selection, they employed rank-order distance\textsuperscript{154}.
which takes into account density in the visual feature space, and used state-of-the-art DCNN features\(^{(107)}\) for image representation.

Qiu et al.\(^{(97)}\) made a large-scale visual semantic complex network consisting of ten million Web images and 33,240 visual concept nodes based on textual and visual features. The objective is the same as\(^{(124)}\)\(^{(129)}\), although this work more focused on building a concept network representing relations between visual concepts.

In general, visual concept learning from the Web needs to collect relevant Web images with high accuracy. However, in case of using visual concepts learned from the Web as mid-level representation, high accuracy is not always needed. Torresani et al. proposed “Classomnes”\(^{(122)}\) which is a new visual feature to represent an image as an aggregation of outputs of a large number of weakly trained visual concept classifiers. They collected images associated with 2659 categories from the Web, and built 2659 classifiers without removing noise images. By the experiments, they showed the effectiveness of the “classome” representation. This is the same idea as Object Bank\(^{(76)}\) which used noise-free training data. Li et al.\(^{(77)}\) improved it by adding salient object detection method\(^{(35)}\) and mi-SVM\(^{(2)}\) to select representative parts.

As a slightly different work from the above-mentioned ones, Kim et al.\(^{(66)}\) proposed a method to model and analyze temporal dynamics on Web images employing the sequential Monte Carlo network, which had never been explored before.

2.6 Web Image and Video Mining for Video Recognition

The works on Web image mining we had explained so far mainly aimed at collecting images associated with noun words or compound words of nouns and adjectives. In this subsection, we overview works on Web image and video mining for video recognition which aims at gathering images and videos associated with verbs, although such works are not so common among Web multimedia mining works.

As an earlier work, Niebles et al.\(^{(90)}\) proposed a method to extract human action sequences from unconstrained Web videos by employing a human body detector. Cinbis et al.\(^{(51)}\)\(^{(52)}\) proposed a method to learn action models automatically from Web images gathered via Web image search engines, and recognize actions for the same video dataset as 90). In these two works, they used a people detector based on HOG (Histogram of Oriented Gradient)\(^{(25)}\) to extract human body regions, which restricted recognizable human actions to body actions such as “running”, “walking” and “dancing”.

Ballan et al.\(^{(4)}\) proposed a method to add tags to Web video shots\(^{*}\) by using Web images as training samples. They assumed several word tags are attached to a whole given video, and gathered images by using combination of the tags as query words for Web image search engines such as Google, Bing and Flickr.

ShotTagger\(^{(74)}\) also aims at shot-level tag annotation to Web videos. They employed Multiple Instance Learning\(^{(85)}\) to detect the shots associated to the given tags.

Do et al.\(^{(28)}\)\(^{(29)}\) proposed a method for automatic video shot collection from Web videos. They adopted two-step video shot selection consisting of tag-analysis-based video ranking using Web 2.0 Dictionary\(^{(147)}\) and visual-analysis-based video shot ranking based on VisualRank\(^{(56)}\). Differently from the previous two works which used not motion visual features but only static visual features, this work employed the state-of-the-art spatio-temporal features\(^{(91)}\) and focused on mining video shots corresponding to verbs. Figure 4 shows some collected video shots of “surf wave”, the precision of which within the top 100 video shots was 75%.

Ulges et al.\(^{(126)}\) proposed TubeTagger to train video concept detectors from YouTube videos for TRECVID semantic indexing task\(^{(112)}\), the detection target of which includes various kinds of visual concepts such as objects, scenes, events and actions. Kordumoba et al.\(^{(68)}\) also used both videos and images on social media to learn video concept detectors for TRECVID, and reached the

\* “Video shots” are short temporal segments of a video. In general, they are generated by dividing a video on each of scene change points.
conclusion that tagged Web images are a better source for learning video concepts than tagged videos.

3. Web Image Application

In this section, we overview works related to visual concept analysis and data-driven computer graphics. Both were impossible without big visual data on the Web.

3.1 Visual Concept Analysis

Web multimedia data are helpful not only as training data of object recognition but also for analysis on visual concepts, since they usually accompany textual tags or messages. By collecting photos or videos with texts on a large scale, we can analyze the relation between visual concepts and words.

Yanai et al.\cite{yanai2009} proposed image region entropy to measure “visualness” of concepts, that is, to what extent concepts have visual characteristics using a large number of Web images. To know which concept has visually discriminative power is important for deciding visual concepts to be mined from the Web. Visual concepts with low entropy are more appropriate for image recognition. This initial work on entropy analysis for visual concepts in terms of distribution of visual features has been followed by several others such as Koskela et al.\cite{koskela2011}, which used entropy to analyze the large-scale multimedia ontology, LSCOM\cite{lu2009}. For the same objective, Lu et al.\cite{lu2009} proposed a framework to develop a lexicon of high-level concepts with small semantic gaps (LCSS) from a large-scale web image dataset. Concepts with small semantic gaps are equivalent to concepts with high visualness. Berg et al.\cite{berg2010} used visualness measure to discover visual attributes from Web image data.

Sun et al.\cite{sun2011} proposed to quantify tag representativeness of visual content of social images, which is equivalent to visualness. Jeong et al.\cite{jeong2011} proposed a method to compute visualness based on clustering, which took account of inter- and intra-cluster purity as well as the entropy. Xu et al.\cite{xu2012} clustered Web images into small visually coherent clusters, “visualsets” which are similar to Visual Synsets\cite{yanai2009}, and then evaluated visualness of each visualset.

On the other hand, several methods to measure distance (dis-similarity) between visual concepts have been proposed. Wu et al.\cite{wu2011} proposed the Flickr distance, which is measured by the square root of Jensen-Shannon (JS) divergence between the corresponding visual language models of two visual concepts. The latent topic based visual language models are learned with the images gathered from Flickr. Katsurai et al.\cite{katsurai2011} proposed a cross-modal approach based on probabilistic CCA\cite{wan2011} to measure distance between concepts using a large number of tagged photos.

3.2 Large-scale Visual Data for Computer Graphics and 3D Reconstruction

Large-scale visual data on the Web enabled data-driven approach not only in computer vision and multimedia research but also in computer graphics. In general, there are two kinds of data-driven computer graphics studies: 2D graphics and 3D graphics.

Regarding data-driven 2D graphics, million-scale Web images and large-scale image search technique are used instead of conventional computer graphics modeling technique. The following works are representative ones: scene photo completion\cite{xu2011}, sketch-based photo composition, Sketch2Photo\cite{shen2011}, real-time guide for free-hand drawing, ShadowDraw\cite{zhang2011}, generating fluid animation from an image using video database\cite{shen2011}, and interactive exploration of photo collections, AverageExplorer\cite{tan2011}.

In data-driven 3D graphics, 3D structure of scenes are reconstructed from Web photos based on structure-from-motion technique. Some representative works are as follows: 3D photo browsing of user-generated photos, Photo Tourism\cite{loizou2011}, 3D reconstruction from the photos taken by many different people\cite{sun2011}, large-scale 3D reconstruction and modeling of whole cities\cite{xu2011}, and 3D reconstruction of interiors of famous architectures\cite{yang2011}. Although these works assume all the photos taken within the target places, Exploring Virtual Space\cite{tan2011} seamlessly stitches similar images which are taken at completely different places, and generates realistic but impossible virtual space tour.

4. Web Mining for Real-World Sensing

In this section, as works on mining real-world images to know the current status of the world, we overview geotagged photo mining and event photo mining from social media.

4.1 Geotagged Images

Until ten years ago, researches on geotagged images focused on only location-based photo browsing for a personal geotagged photo collection\cite{wang2007,zhu2009}, since it is almost impossible to obtain a large number of geotagged images. However, the situation has been changed after photo sharing social media, Flickr, started to handle geotagged photos in 2006. Flickr has become the
most common geotagged photo database for the multimedia studies on geotagged photos, because the Flickr database is open to everyone via FlickrAPI which allows users’ program to search the whole Flickr photo database for geotagged images without any restriction. Then, many works on geotagged image recognition with huge Flickr geotagged image database have been proposed after that.

Cristani et al.\textsuperscript{23} and Cao et al.\textsuperscript{15} proposed methods on event recognition of geotagged images by integrating visual features and geographical information. In general, a geotag represents a pair of values on latitude and longitude. It is a just 2-dimensional vector. To convert a 2-d vector into richer representation, Luo et al.\textsuperscript{83} and Yaegashi et al.\textsuperscript{137} converted geotags into visual information from the sky using aerial images, and Joshi et al.\textsuperscript{57} transformed geotags to words using reverse geo-coding technique. On the other hand, Yuan et al.\textsuperscript{150} used GPS trace data which is a series of geotags instead of using just a pinpoint geotag in order to classify images into several pre-defined events. Yu et al.\textsuperscript{149} used time and seasons for geotagged image recognition in addition to visual information and geo-location data.

While event or scene recognition on geotagged images had been studied actively, no place recognition of an image was studied before Hayes et al.\textsuperscript{46} proposed “IM2GPS”, a data-driven approach, to estimate places where non-geotagged images are taken by nearest-neighbor search for six million geotagged images gathered from Flickr. MediaEval Placing task\textsuperscript{101} has taken over and extend data-driven photo location estimation employing accompanying textual information and other metadata such as user IDs in addition to visual features.

As extension of location-based photo browsing, several works had considered the problem of selecting representative or canonical photographs for online image collections. Jaffe et al.\textsuperscript{54} selected a summary set of photos from a large collection of geotagged photographs based on only tags and geotags. By analyzing the correlations between tags and geotags, a map-based visualization “Tag Map” was developed to help indicate the most important regions and the concepts represented in those regions.

Simon et al.\textsuperscript{109} proposed a method to select canonical views for the landmarks by clustering images based on the visual similarity between two views. Similarly to this work, Kennedy et al.\textsuperscript{63} attempted to generate representative views for the world’s landmarks based on the clustering and on the generated link structure. Li et al.\textsuperscript{78} proposed a method to collect landmark photos using landmark classification. Zheng et al.\textsuperscript{152} also collected the photos of landmarks over the world from the Web using local feature matching, and built a worldwide landmark search engine.

Cristani et al.\textsuperscript{23} proposed Location Dependent pLSA (LD-pLSA) which is a probabilistic topic model for geo-tagged image analysis using latent representations. LD-pLSA is able to represent location-dependent image topics. Yanai et al.\textsuperscript{144} proposed to detect cultural differences among different locations using a large number of geotagged photos. Figure 5 shows that wedding cakes in US and Europe have different characteristics.

As a similar work, Doersch et al.\textsuperscript{30} proposed to detect geo-informative patches which exhibit location-dependent characteristics from geotagged photos employing mid-level discriminative patches proposed by Singh et al.\textsuperscript{110}. In the paper, they showed large-scale experimental results on Paris, London, Milan, Barcelona and Prague. Based on the similar idea, Lee et al.\textsuperscript{71} proposed to detect patches representing style differences on objects such as car, house and tree which depend on time and location.

Leung et al.\textsuperscript{73} proposed “proximate sensing” to create land condition maps on the surfaces of the earth by using Flickr geotagged photos, which is the opposite meaning of “remote sensing” that captures land photos from the satellite. They classified land condition into being developed or undeveloped.

There are so many works on other works on geotagged photos than above-mentioned ones. Luo et al.\textsuperscript{82} and Zheng et al.\textsuperscript{151} have published well-organized survey papers on geotagged photo processing.

\textsuperscript{*} http://multimediaeval.org/
4.2 Event Detection from Social Media

Thanks to recent popularization of social media and smartphones, it is very common for people to take photos and post them to social media, especially when they experienced “events” which are slightly different from scenes in the usual everyday life. For examples, many people post photos when visiting festivals, sport games and exhibitions, or experiencing some seasonal events and large-scale natural phenomena. Thus, the social media users can be regarded as “socially distributed cameras” which report various kinds of events visually. Regarding the events many people experienced at the same time, many related photos are posted to social media. By mining such event-related photos, we are able to get to know local events which are not reported by the mass media or the state of damage by natural phenomena more quickly than from the mass media news.

To this end, recently, “visual event detection” from social media such as the photo sharing site, Flickr, and the microblog service, Twitter, has been actively studied. In this subsection, we overview social media visual mining for events in the real world.

In the multimedia community, an “event” is used in various contexts. Some work defined it as an activity in which people participate and take pictures such as hiking, playing sports at park and wedding party\(^\text{17}\), while in the TRECVID Multimedia Event Detection task it was defined as an abstract concept of “action” or complex actions, and includes more personal activities such as making a sandwich, repairing an appliance and marriage proposal\(^\text{122}\). As another work on activity events, anomaly detection from video/image streams has been studied before\(^\text{153}\). The objective is to detect anomaly events such as invasions and accidents from fixed camera video streams. These works focus on “event classification/recognition/detection” which is a kind of image/video recognition.

On the other hand, in case of “event detection from social media stream”, an “event” tends to become more public and to gather many people, since a certain number of photos or tweets related to a certain event are needed to detect the corresponding event. The MediaEval Social Event Detection (SED) Task\(^\text{101}\) defined “events” strictly as public events the schedules of which were announced on the Web event database, last.fm, such as music events and sport events, while in some event detection works the definition of “events” was broader and they allowed more personal events such as wedding to be regarded as “events”\(^\text{17,99}\). In this review, in addition to scheduled social events such as sport games and festival, we regard natural phenomena happening in the limited areas and the limited periods as “events” such as typhoon, heavy rain/snow, and beautiful sunset which exhibit uncommon scenes and draw attention of many people.

Many works on event detection have been proposed in the multimedia community so far. Most of the works including the MediaEval SED task\(^\text{96,100,101}\) used Flickr photos and tags as a target data from which events were detected, while the number of the works on Twitter photo data is limited. We describe some works on Flickr event detection first, and then we explain works on Twitter event detection.

(1) Event Detection from Flickr

Rattenbury et al.\(^\text{99}\) proposed one of the pioneer works on event detection from Flickr tags. They proposed Scale-structure Identification which is a burst detection method with multiple time scales. They used only event frequency along the temporal direction to detect event tags, and used no geotags and no visual features. Chen et al.\(^\text{17}\) extended it by taking into account the spatial direction in addition to the temporal direction.

On the other hand, Quack et al.\(^\text{98}\) proposed an object and event photo mining method which relies on visual features, and spatial information. Since their main objective was landmark detection, temporal information was not used. The detected event photos are very similar to each other in the same event cluster, because they used the number of matched SURF keypoints\(^\text{48}\) as visual similarity. Therefore, the detected events are more personal than scheduled social events.

Papadopoulos et al.\(^\text{94}\) also proposed a method on landmark and event photo mining which employs graph-based clustering with hybrid similarity of both visual similarity and tag similarity. Since they employed hybrid similarity and bag-of-features\(^\text{24}\), this method outperformed the method proposed by Quack et al.\(^\text{98}\) as event photo detection.

The approach by Liu et al.\(^\text{79}\) was different from the other works. They selected the venues where the scheduled events were regularly held in advance, and monitored the statistics of the number of photos shared to detect events. Although the method was simple, the result was promising. This indicates that event photo detection do not always need sophisticated methods, and a simple method is enough when the sufficient number of related photos are available.
The MediaEval Social Event Detection Task\cite{101} is a representative benchmark task on social event detection. Because the training data is available in the task, supervised methods are commonly used by the participants. Reuter et al.\cite{100} employed SVM with temporal, geographical and textual features, while Petkos et al.\cite{96} proposed multi-modal clustering with supervisory signal employing visual features as well.

(2) Event Detection from Microblog

Because microblogs such as Twitter and Weibo have unique characteristics which are different from other social media in terms of timeliness and on-the-spot-ness, they include much information on various events in the real world. Especially, the photos posted to microblogs are useful to understand what happens in the world visually and intuitively.

Although there exist many works related to Twitter mining using only text analysis such as the work by Sakaki et al.\cite{103}, only a limited number of works exist on Twitter mining using image analysis currently. However, this research area is expected to become more active, because it has many unexplored topics and it is a promising photo data source due to its growing speed and availability of photos, especially for geotagged photos.

As the early works on microblog photos, Yanai have proposed “World Seer”\cite{141} which can visualize geotagged photo tweets on the online map in real-time by monitoring the Twitter stream. This system can store geo-photo tweets to a database as well. They have been gathering geo-photo tweets from the Twitter stream since January 2011 with this system. On the average, they gathered about half million geo-photo tweets a day, about one third of which are hosted at Instagram. Because the number of uploaded photos to Flickr a day in 2014 was officially announced as 1.5 million and only 10 to 20 percent of them are estimated to have geotags, Twitter can be regarded as a more promising data source of geotagged photos.

To utilize their Twitter image database, Nakaji et al.\cite{88} proposed a system to mine representative photos related to the given keyword or term from a large number of geo-tweet photos. They extracted representative photos related to events such as “typhoon” and “New Year’s Day”, and successfully compared them in terms of the difference on places and time. However, their system needs to be given event keywords or event term by hand. Kaneko et al.\cite{58,59} extended it by adding event keyword detection to the visual Tweet mining system. As results, they detected many photos related to seasonal events such as festivals and Christmas as well as natural phenomena such as snow and typhoon including extraordinary beautiful sunset photos taken at the specific area as shown in Figure 6. All of these works focused on only geotagged tweet photos.

Chen et al.\cite{19} treated with photo tweets regardless of geo-information. They analyzed relation between tweet images and messages, and defined the photo tweet which has strong relation between its text message and its photo content as a “visual” tweet. In the paper, they proposed the method which is based on the LDA topic model to classify “visual” and “non-visual” tweets. However, because their method was generic and assumed no specific targets, the classification rate was only 70.5% in spite of two-class classification.

Recently, Yanai et al. proposed Twitter Food Photo Mining\cite{145} which takes advantage of the characteristics of Twitter that many meal photos are uploaded in the time of meals everyday. They used a realtime food recognition engine of the mobile food photo recognition application, FoodCam\cite{62}, to detect one hundred kinds of foods from the Twitter stream. They claimed they had already collected more than half million ramen noodle photos, which would be helpful for researches on large-scale fine-grained food image classification. Some
sites and social curation service such as Pinterest\textsuperscript{47}, textual Web data resources and open data by governments and companies will be one of the promising directions as visual analysis and event detection on social media.

**Acknowledgment**

We thank the helpful comments from Dr. Do Hang Nga and anonymous reviewers.

**References**


19. T. Chen, D. Lu, M.-Y. Kan, and P. Cui. Understanding and

---

Gao et al.\textsuperscript{43} proposed a method to mine brand product photos from Weibo which employs supervised image recognition in the same ways as 145. They integrated and used visual features and social factors (users, relations, and locations) as well as textual features. The same authors proposed to use hypergraph construction and segmentation for event detection\textsuperscript{44}.

5. **Conclusions**

In this survey, we reviewed various kinds of works related to Web image mining including Web image re-ranking, visual concept acquisition, visual concept analysis, image recognition by Web-based visual data, Web-image-driven computer graphics and real-world visual event detection from Flickr and Twitter.

The amount of image/video data on the Web is increasing day by day. Accordingly, the possibility of Web image/video mining is increasing. As mentioned in this survey, there are two promising directions of Web image/video mining. One is visual concept mining, and the other is social media mining.

Regarding visual concept mining, much unexplored topics are still left such as mining more complex compound concepts, non-object concepts including verbs and adjectives, and relation between images and sentences for automatic image description generation. Combining visual concept mining and the state-of-the-art deep convolutional neural network (DCNN) is also one the promising directions\textsuperscript{45}.

Regarding social media mining, currently, works on microblogs and works on Flickr are being studied independently. Integrating multiple data sources including various social media, microblogs, photo/video sharing


67) A. Kimura, K. Ishiguro, M. Yamada, A. Marcos Alvarez,


Keiji Yanai received B.Eng., M.Eng. and D.Eng degrees from the University of Tokyo in 1995, 1997 and 2003, respectively. From 1997 to 2006, he was a research associate at Department of Computer Science, the University of Electro-Communications, Tokyo, and he was an associate professor until 2015. Currently, he is a professor at Department of Informatics, the University of Electro-Communications, Tokyo. His recent research interests include object recognition and Web multimedia processing. He is a member of IEEE Computer Society, ACM, IPSJ, IEICE and JSIA.