Noname manuscript No.
(will be inserted by the editor)

Automatic Image Semantic Interpretation using Social Action and Tagging Data

NEELA SAWANT
JIA LI · JAMES Z. WANG

Received: date / Accepted: date

Abstract The plethora of social actions and annotations (tags, comments, ratings) from online media sharing Websites and collaborative games have induced a paradigm shift in the research on image semantic interpretation. Social inputs with their added context represent a strong substitute for expert annotations. Novel algorithms have been designed to fuse visual features with noisy social labels and behavioral signals. In this survey, we review nearly 200 representative papers to identify the current trends, challenges as well as opportunities presented by social inputs for research on image semantics. Our study builds on an interdisciplinary confluence of insights from image processing, data mining, human computer interaction, and sociology to describe the folksonomic features of users, annotations and images. Applications are categorized into four types: *concept semantics, person identification, location semantics* and *event semantics*. The survey concludes with a summary of principle research directions for the present and the future.

Keywords Web 2.0 · social media · collaborative annotation · image semantics · folksonomic features · survey

The material was based upon work supported in part by the National Science Foundation under Grant No. IIS-0949891 and IIS-0347148, and by The Pennsylvania State University.

N. Sawant (corresponding author)
College of Information Sciences & Technology,
The Pennsylvania State University, University Park, PA, USA
E-mail: nks125@psu.edu

J. Li
Statistics Department,
The Pennsylvania State University, University Park, PA, USA
E-mail: jiali@psu.edu

J. Z. Wang
College of Information Sciences & Technology,
The Pennsylvania State University, University Park, PA, USA
E-mail: jwang@psu.edu

1 Introduction

Modern scientific progress is driven largely by our ability to make sense of enormous data collections, and harness the findings in a continued sense-making loop. Aptly, humans are termed *informavores*: species that consume information to accelerate their technical evolution [127]. The level of information consumption is limited to the extent permitted by the organization of underlying data, and for this reason, it is important to devise systematic and meaningful methods for data storage and retrieval. Till date, significant progress has been made in the general area of information retrieval, with numerous models, algorithms and systems governing large text collections. Multimedia, on the other hand, still remains obscure due to ill-understood theories of perception and cognition. In this work, we focus on images - a prevalent form of multimedia, and study the progress in semantic understanding of large image collections. Semantics, with respect to images, represents the association between low-level visual features and high-level concepts that can be described in words. Such knowledge possibly arises from the awareness of the context in which photographs are shot. Thus, the quest for image understanding encompasses traditional research on object detection and scene interpretation as well as capturing abstract notions of events, locations, and personalized references that situate images beyond the realm of visual features.

The task of translating raw pixels into abstract semantics is not a well-formed process. Even simple scenes contain a complex arrangement of objects that can be described using a variety of color, texture, shape and position features. Ambiguities result from *intra-class variability* and *inter-class similarity* that objects exhibit under different photographic conditions. The disparity between low-level visual features and high-level concept semantics, known as the *semantic gap* [169], severely limits the ability of automated systems to discern the nuances of visual semantics. The mainstream research on visual semantics primarily resorts to machine learning techniques, that given a concept and a large corpus of manually annotated exemplars, build concept models useful for future annotation of unlabeled images. A practical constraint occurs from difficulties in the acquisition of manually annotated high-quality training corpora. Therefore, demonstrable applications of most learning based techniques are still restricted to relatively small-scale datasets [15, 59, 74, 110, 201]. Comprehensive surveys in the area of content based organization and retrieval of images are published early on by Smeulders et al. [169] and more recently by Datta et al. [44].

Our survey analyzes the paradigm shift in the semantic understanding of images, brought by recent Web 2.0 phenomenon of social networking and collaborative media annotation. Newer methods that partially automate training data selection by harnessing Web images [37, 50] or using crowd-sourcing options [154, 174] are on the rise. Also, concept modeling techniques using fewer exemplars [58, 109] have become desirable. We discuss how the voluntary actions and annotations by millions of Web users that contribute to astronomical image collections also pose new research questions and afford new opportunities of semantics extraction. Two major sources of large labeled image collections are considered: *collaborative image labeling games* and *tagging in media sharing social networks*. We analyze the characteristic settings in which labels are contributed with attention to user idiosyncrasies, image and tag properties. The approaches to social image semantic extraction marry traditional image processing with techniques and models in the purview of social information retrieval. Fig. 1 highlights important aspects of this review, where applications of semantics and knowledge extraction are divided into four categories:

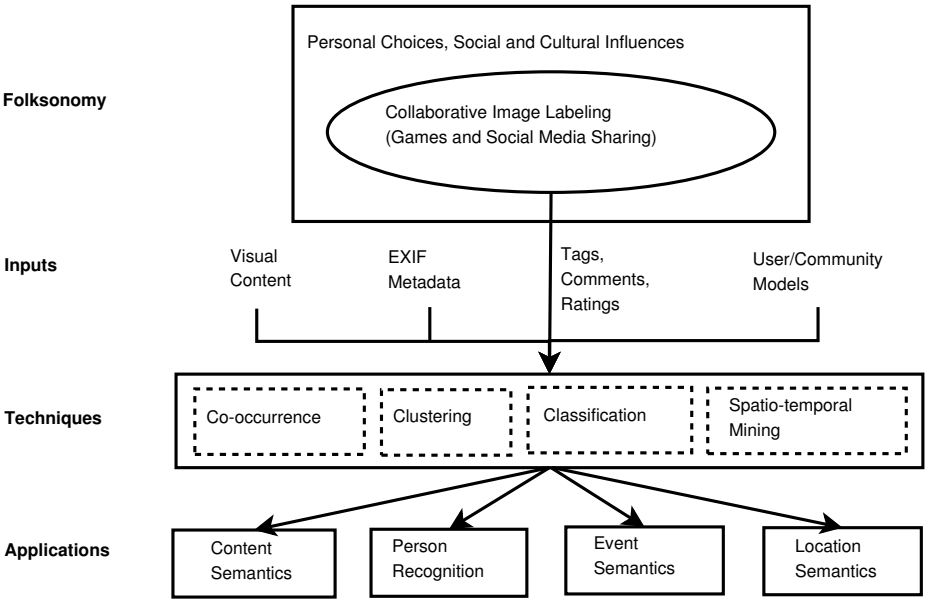

Fig. 1 Focal points of the survey.

1. What does the picture portray? or **Content semantics**: Content semantics is the goal of mainstream research in visual semantic interpretation. We discuss how social data has molded research directions in tag relevance estimation, concept/aesthetics modeling and image annotation.
2. Who is in the picture? or **Person recognition**: Whereas content semantics may establish presence of people in images, person recognition gets to the specifics of labeling each person with an identifier. Such identification helps in organization of personal image collections, for celebrity picture search on the Web and for social network discovery.
3. When is the picture taken? or **Event semantics**: Pictures are often a snapshot of an event or an occasion. Event semantics involves identification of person-specific, community-specific or global events associated with the visual content.
4. Where is the picture taken? or **Location semantics**: Location semantics correspond to geographically-grounded places (such as *Paris, Greece*) or non-grounded entities (such as *museum, library*). Inferring location semantics is useful for discovering potential landmarks and tourism related information.

2 Social Sources of Image Labels

Recent systems like LabelMe [104, 154] and Amazon mechanical turk [126, 174] distribute image annotation and evaluation tasks to Internet users. The volume of annotations generated from such crowd-sourcing techniques helps reduce the burden on experts without significantly sacrificing the quality of annotations. The annotators are provided with detailed instructions on how to best select labels that can be directly used for concept modeling. This ensures that relatively good quality annotations are

generated for object detection and relevance estimation tasks. It is shown that crowdsourcing is a reasonable substitute for repetitive expert annotations, when there is high agreement among annotators [82,140]. Other sources of image annotations are collaborative games and social media sharing which undoubtedly represent the fastest growing labeled image collections in the world. In this section, we analyze the characteristic settings in which collaborative games and social media help generate image labels and other metadata.

2.1 Collaborative Games

Collaborative games are a channel for human computing through which hundreds of thousands of players contribute perceptual and cognitive information about multimedia objects[1]. Recent advances in gaming interfaces utilize simple gestures, making games accessible even in transit [73]. As the games utilize recreational activities of users to generate inputs for artificial intelligence research, they are termed as *Games with a Purpose*, or GWAP [6,75]. Three popular games, namely, Google ImageLabeler, Peekaboom and Phetch are described below.

– *Google ImageLabeler*: Formerly known as the ESP game [5,150], Google ImageLabeler is a game where randomly paired players attempt to guess labels given by each other to a common input image (snapshot in Fig. 2). Players score points when a label is agreed upon. When a word has been used frequently (decided by a threshold), it is marked off-limit to encourage new labels. As the player engagement increases, so does the volume of labeled images.

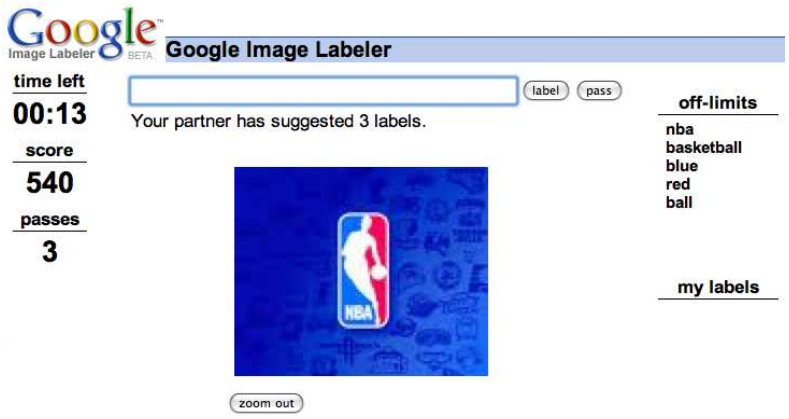

Fig. 2 Snapshot of Google ImageLabeler interface.

– *Peekaboom*: Peekaboom is played between randomly paired players, where one plays the part of a *'describer'* and the other is a *'guesser'*. The describer is given an image

[1] More than 200,000 players of the ESP game (later renamed Google ImageLabeler) contributed over 50 million image labels as a number of players spent more than 40 hours a week playing the game. Peekaboom recorded more than 500,000 human-hours of play [6].

completely masked from the guesser. Given a word, the describer progressively reveals parts of the input image, where the object corresponding to the word is present. Points are scored when the other player correctly guesses that word. Thus, a successful game of Peekaboom yields rough object localizations. Fig. 3 shows a snapshot of the game interface.

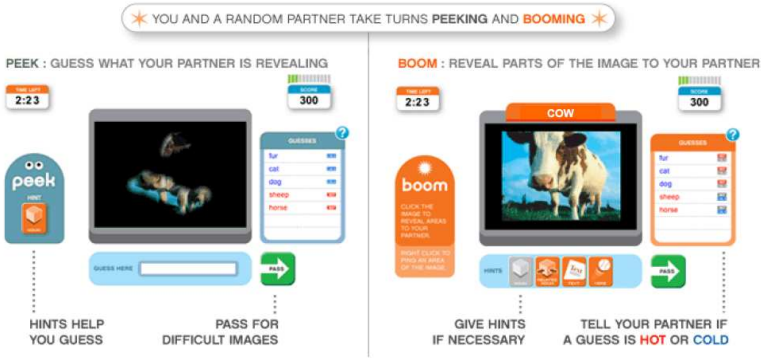

Fig. 3 Snapshot of Peekaboom interface [7].

– *Phetch*: Similar to Peekaboom, Phetch pairs random players in *'describer-guesser'* roles. The role of the describer is to furnish a multi-word description of an input image, that the guesser uses to locate the original image from a reference image collection. Points are scored when the input image is located by the guesser. With each successful round, Phetch gains a validated input image description.

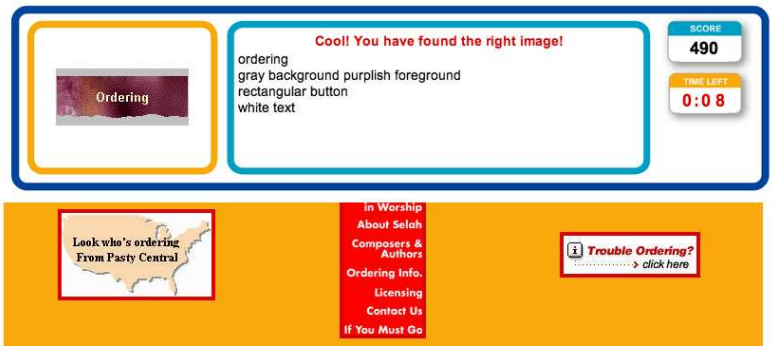

Fig. 4 Snapshot of Phetch interface.

Games with a purpose share a number of features designed to appeal to human psychology and to encourage annotation through incentives of fun and competition [6]. The motivation to score increasingly more points within a time limit, improving game skill level and competing with other players increases user engagement in annotation tasks. Pairing random players increases the likelihood that players do not cheat and

annotations are indeed relevant to the visual content. von Ahn et al. showed that, within this general framework, computational problems can be converted to collaborative games of one of three categories: *output agreement, input agreement* and *inversion problem* [6]. The output agreement games (such as Google ImageLabeler) present a common input (such as an image) to user pairs who score points when they are able to generate the same output (such as descriptive words). In problem inversion games (such as Peekaboom and Phetch), only one player is shown the input and is required to generate a description. The other player has to infer the original input from the description. Finally, the input agreement games require players to make a decision about the commonality of the input presented to them, after viewing the descriptions that they independently generate. An example is a music related game Tag-a-tune, where pairs of users determine if they are listening to the same piece of music by inspecting the description independently created by each other.

2.2 Social Media Sharing

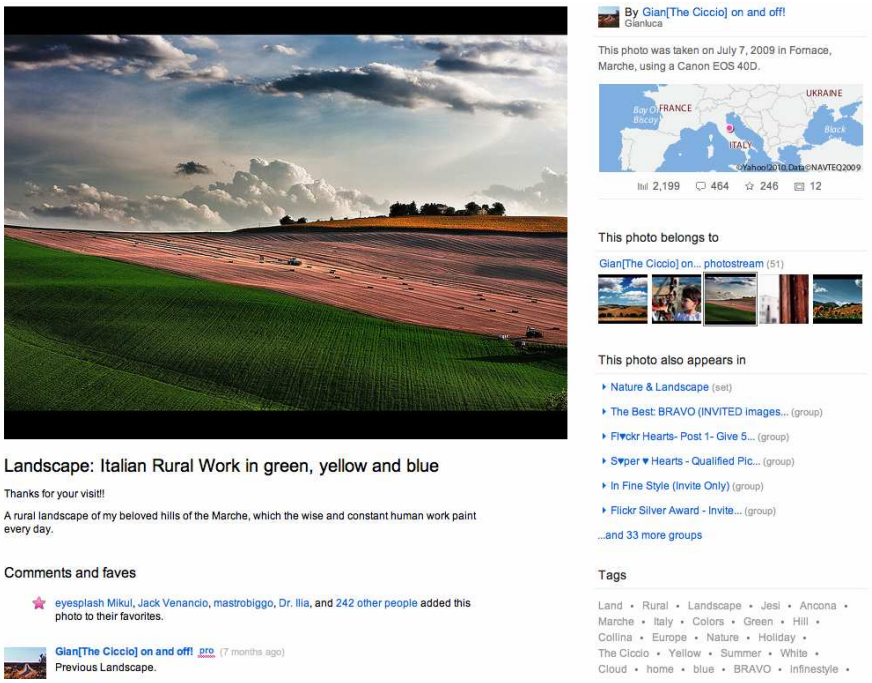

Fig. 5 Snapshot of social media Website Flickr. A typical image is associated with comments, tags and ratings (mark as a favorite).

Basic social networking mechanisms equip users with the ability to form online contacts, join special interest groups, upload and share documents (textual and multimedia). Users can also contribute annotations (tags), comments and ratings as depicted in Fig. 5. These actions and annotations have become a fast-growing source of image

descriptions[2]. Tags are essentially personal keywords which impose a soft organization on data. As opposed to taxonomies that are restricted by rigid definitions and relationships, tags are continuously influenced by popular trends and colloquial vocabulary. Therefore, the organization imposed by tags is popularly referred to as **folksonomy** (folk + taxonomy).

The idea of using words from personal context for filing purposes has been in practice for decades. In the *memory extender* system of W. P. Jones [91], users associated a set of personalized keywords to their file collections. The benefits were two-fold: a) *recallability*, where specification of some of the associated keywords helped retrieve a particular file from a large personal file collection, and b) *recognition*, where given a file's keywords, the user could be easily reminded of the file creation context. As the keywords originated from personal context, they guided a user's *mental model* during exploratory search, thus controlling search efficiency [63].

With the introduction of social media sharing, files are tagged for the benefit of self as well as others. Moreover, within the limit of user-specified permissions, other users may tag one's personal resources. Accordingly, the motivations behind tagging evolve beyond personal benefits to accommodate for social influences. Ames et al. proposed a taxonomy to describe motivations along two dimensions: a) *function* and b) *sociality* (target audience) [10]. Functions are categorized into 'organization' and 'communication', whereas sociality is categorized into 'self' and 'social'. The organization function corresponds to the use of tags for future retrieval and for contribution to public resources. The communication function refers to the provision of contextual information to self and to others. Also included is *social signaling*, whereby a user's interest and credibility is communicated through substantial tagging contribution. Marlow et al. described a more detailed set of motivations comprising of: future retrieval, contribution and sharing, attention seeking, play and competition, opinion expression, and self presentation [124]. Expansion along the sociality dimension has also resulted in finer categories. For example, Ames et al. divided the social category into two subcategories: 'friends & family' and 'public'. Kustanowitz and Shneiderman used three sub-categories: 'family & friends', 'colleagues & neighbors', and 'citizens & markets' [103]. It has been shown that function and sociality strongly affect the level and usefulness of tagging [139].

2.3 Folksonomic Challenges

Social annotations differ from expert annotations as they are contributed from personal, often unknown motivations and not directed towards specific computational tasks. Further, personal tendencies and community influences affect the quality of tags [161]. In this section, we summarize main folksonomic challenges that need to be addressed before social annotations can be suitably utilized.

- **Motivations**: Motivations have a direct influence on the usability of tags for scientific purposes [139]. Tags that arise from the need of future retrieval and contribution, in particular for the benefit of external audience, are likely to be visually more relevant compared to tags used for personal references. Images within special interest groups are very likely to be specifically annotated and heavily monitored

[2] Flickr [196], acquired over four billion user photos within six years after its launch [61]. Tagged Flickr images are widely being used to drive research in visual concept detection.

by the group members, as they are motivated to contribute and connect with other users having similar interests.

- **Cultural influences**: Cultural differences guide perception and cognition differently [136]. For example, an analysis of image tags created by European American and Chinese participants concluded that whereas Westerners focus more on foreground objects, the Easterners have a more holistic way of viewing images early on [51]. This was discovered through the analysis of tag assignment order. For Easterners, the specificity of tags increased from holistic scene description to individual objects. On the other hand, the tags given by the Westerners focused on individual objects first and then on overall scene content.
- **Vocabulary problem**: The spontaneous choice of words to describe the same content varies among different people, and the probability of two users using the same term is very little. Known as the *vocabulary problem* [64], this issue is often cited as a common characteristic of folksonomic annotations. The different word choices introduce problems of *polysemy* (one word with multiple meanings), *synonymy* (different words with similar meanings) and *basic level variation* (use of general versus specialized terms to refer to the same concept) [72].
- **Specialized knowledge**: Certain user tags containing special characters, numbers and personal references can be considered as specialized knowledge if they are not meaningful to the general audience. For example, tags such as 'me' or 'd20' may not have any significance for global audience. Such tags can be filtered out as stop-words with the help of usage statistics.
- **Semantic loss**: An annotator in folksonomies is not obliged to associate all relevant tags with an image, leading to semantic loss in the textual descriptions [193]. The batch-tag option provided by most photo sharing sites adds to this problem by allowing users to annotate an entire collection of photos with a set of common tags. Even if such tags are potentially useful to provide a broad personal context, they cannot be used to identify image-level differences, thus leading to semantic loss. One consequence of this fact is that the absence of a tag from an image description cannot be used to confirm the absence of the concept in that image. Therefore, such images cannot be directly used as negative examples for training.

Owing to the above challenges, the labeled data from Web 2.0 does not readily substitute for expert annotations in the identification of visual semantics. Large scale studies show that nearly half of tag applications on the Web or social media collections are irrelevant for general audience [96, 175]. Such tags need to be pruned to be able to harness tagged images effectively.

2.4 Statistical Semantics

In stable social tagging systems, tag vocabulary reaches statistical regularity and forms patterns [72]. Usage statistics can be used to segregate patterns from noise and to detect emergent semantics beneficial to understand what people mean, at least to a level sufficient for information access [157]. Such study of statistical patterns of human word usage for semantic interpretation is referred to as *statistical semantics* [65].

Usage statistics computed over multiple independent users play a pivotal role in a number of information retrieval applications. Using the *relevance feedback* mechanism [153], user clicks on Web search results are used to tune future result ranking [3, 194].

Aggregation of clicks by user sessions has been utilized to cluster results with similar semantics, particularly for disambiguation of polysemous queries [181]. In collaborative games, aggregation takes the form of a requirement that a large number of user pairs must agree on image descriptions before they can be deemed relevant to visual content. Overall, the phenomena of social actions and annotations share many similarities with the proposition of *wisdom of crowds* [178], that the aggregated verdict of a group of independent people is closer to the truth than that of any individual in the group. The origin of this theory goes back several decades. At a country fair in 1906, Sir Francis Galton observed that when hundreds of people were asked to guess the weight of an Ox, none of the individuals - even the cattle experts, could correctly guess the weight. On the other hand, the average of all estimates was closer to the real weight of the Ox. This occurrence has since then become a famous anecdote for wisdom of crowds. In case of social media annotations, similar analogy exists. When a tag is applied by a large number of users to similar visual content, such relationship is significant from the point of view of tag visual semantics. Drawing a parallel with wisdom of crowds, four main characteristics must be discussed [178]:

- **Diversity of opinion** - Every person is entitled to a personal opinion. In case of social image tagging, each person is entitled to their own subjective interpretation of image content and corresponding use of annotations.
- **Independence** - A person's opinion is not influenced by that of the others. In case of social image tagging, each person can independently provide zero or more tags to zero of more images belonging to self and others. However, complete independence cannot be guaranteed when social influences are strong.
- **Decentralization** - Each person can operate in a local setting and have a different view of the system. In case of social image tagging, decentralization is ensured as users have control of their own tagging activity, without being exposed to tags given by other users to content-similar images.
- **Aggregation** - A mechanism to convert the opinions into an aggregated verdict must exist. As the population size increases, the confidence in the verdict increases as well. Consider for example, the task to compute similarity between two tags. One simple mechanism is to count the number of images tagged with both tags. Other mechanisms can be devised by considering complex relationships of tags with other tags and users in folksonomy.

The assumption of statistical semantics is that a typical user makes rational choices. In such case, the actions and annotations of a few idiosyncratic users are reduced to noise when a large number of users are considered.

3 Representation of Folksonomy

The popular view of folksonomy is as a ternary relationship between users, resources and tags [76, 86, 105, 160]. We resort to a definition that describes folksonomy as *'a tuple* $F := (U, T, I, A)$ *where* U*,* T*, and* I *are finite sets representing users, tags and images (documents in general) respectively, and* A *is a ternary relation between them, i.e.* $A \subseteq U \times T \times I$*, whose elements are called tag assignments'* (adapted from [86]). Further, users may be connected to other users through social relationships.

Folksonomy is generally represented as a tripartite hypergraph and its elements are represented using the *vector space model* [156]. Let a folksonomy have m users

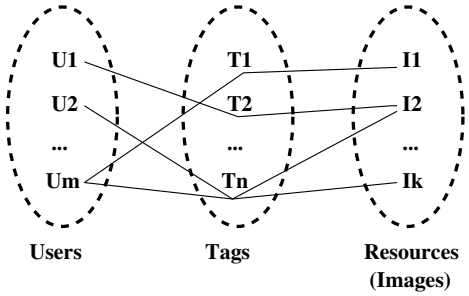

Fig. 6 Tripartite hypergraph representing folksonomy (adapted from [76]).

denoted as U, n tags denoted as T, and k images denoted as I (Fig. 6). Then one type of vector space model for a tag t is a m-dimensional frequency vector, where the value at position p equals to the number of times the $p^{th}(p \in (1, \ldots, m))$ user U_p has used the tag. Another representation is a k-dimensional frequency vector, where the value at position q equals to the number of times the tag has been assigned to the $q^{th}(q \in (1, \ldots, k))$ image I_q. These two vector models capture the distribution of each tag over all users and all images respectively. For socially shared images (as in Flickr), a tag can be applied to an image at most once. Therefore, the $k - dimensional$ vector space model for tags computed over images is binary. Collaborative image labeling games, on the other hand, permit a frequency vector representation for tags, since multiple players are allowed to apply the same tag to a common image.

Stacking vector representations creates an association matrix. For example, the $k - dimensional$ tag vectors over images can be stacked into a tag-image matrix TI similar to the term-document matrix representation in the field of text information retrieval. The tripartite folksonomy graph can be modeled as a three-dimensional matrix UTI, by stacking together the tag-image matrices for all users (dotted box at the left of Fig. 7). Returning to the two-mode (bipartite) representations is easier, and requires aggregation over one of the folksonomic dimensions of the matrix. For example, the tag-image matrix TI is obtained by collapsing the 3-D matrix UTI along the user dimension. User vocabulary or the tag-user matrix TU is derived by collapsing the 3-D matrix UTI along the image dimension. These bipartite representations can be further used to extract distributional similarity between different elements such as user-user, image-image, and tag-tag. Fig. 7 shows that the folding of a two-mode matrix results in a one-mode matrix where the element at a position (i, j) measures the distributional similarity between element i and element j of the outer dimension of the input two-mode matrix. For example, one folding of TI computed as $(TI)(TI)'$ results in a one-mode matrix where element at position (i, j) is the distributional similarity of tags T_i and T_j. This similarity is useful for modeling and comparing tags. Folksonomic matrices are often rather sparse and elements may be correlated. To address these issues, efficient low-rank approximations can be computed using techniques such as latent semantic analysis (LSA) [106] and probabilistic latent semantic analysis [79].

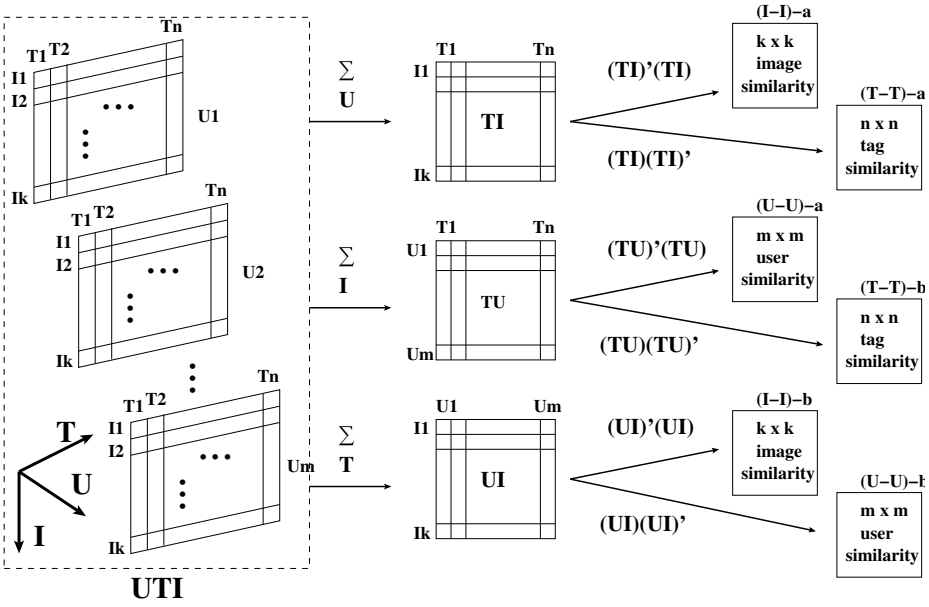

Fig. 7 The dotted box at the left represents the tripartite nature of folksonomy. The tripartite matrix can be collapsed along each of the three dimensions, resulting in a two-mode association matrix. Folding each two-mode matrix generates distributional similarity matrices.

3.1 Features

The quality of tag assignments depends on the features of images, users, tags and their inter-relationships. In this section, we summarize prominent features that are useful for social image analysis. Image features include visual descriptors and EXIF metadata. Abstract user qualities are captured into features such as reliability and expertise. Tags and keywords extracted from photo comments are represented using features such as frequency and entropy. The apparent heterogeneity in features underscores the need for effective multi-modal data and decision fusion techniques.

3.1.1 Image Features

Two types of features can be used to describe the content and context of social images:

- **Visual features**: Over the years, a large number of visual features are proposed to describe color, texture and shape in images. Descriptors can be computed over local image regions (interest points, corners, blobs) or entire images (histograms, moments). Basic features can also be described as per their composition in the images, using features such as layout and correlogram. Detailed discussion about various local and global image features are available in [120, 169, 182].
- **EXIF metadata**: EXIF data provides information about camera parameters such as aperture setting, exposure time and focal distance [57]. Additionally, time and geo-location information of when and where a photo was taken, is also available.

3.1.2 User Features

The variability in tag assignment, and consequently the quality of assigned tags is affected by a user's personal choices and the context of the user's social network. For example, the type of tagging motivation is correlated with the number and types of tags by the user. Also, the number of tags is proportional to the size of the user's network and the number of social groups to which he belongs [138, 139]. An estimation of the idiosyncrasies helps assess the quality level of a user's annotations. In this section, we summarize a number of descriptive features such as expertise, reputation and reliability.

- **Expertise**: An expert is a provider of high-quality annotated resources. Topic experts can be identified by a substantive contribution of relevantly tagged resources [137] or by a membership to special interest groups related to that topic. Noll et al. defined experts using the *Hyperlink-Induced Topic Search (HITS)* algorithm [97] and distinguished tag spammers from experts [137]. Members of special interest groups are expected to possess specialized knowledge as compared to non-members. It is possible to identify the topic of expertise and vocabulary by jointly analyzing visual content and tagging behavior of group members using techniques like probabilistic latent semantic analysis [134].
- **Reputation**: Expertise is a topic-specific feature. Reputation, on the other hand, is a more general property that assimilates overall activities of networked users into a *social order* [170]. The degree to which a member's work is recognized in the network and a user's social influence can be used as an indicator of reputation [179]. For example, in the computation of *Flickr Interestingness* [62], a user's tagging and social activity plays a major role, such that professional and active members are qualitatively ranked higher. Tags, comments and views by high ranked users are considered more useful and can be employed in determining image interestingness. The prestige of special interest groups in which the photo appears is also a contributing factor [49].
- **Reliability**: A tag assignment is considered reliable if similar associations are consistently observed over a large user collection. Unreliable tag assignments should be treated carefully in relevance ranking applications [9, 17, 18, 100, 101]. The reliability of a specific user's annotations can be modeled using game-theoretic techniques as well [163].
- **Other network measures**: A user and his social network can be represented using traditional network measures such as characteristic path length, clustering coefficients, cliquishness and connectivity [28].

3.1.3 Tag and Metadata Features

Quality assessment of textual annotations is an important research question in the general area of information retrieval. In this section, we discuss features for tags (and in general textual metadata) such as frequency, entropy, clarity, and linguistic properties.

- **Frequency** - The distribution of tag frequency for stable media sharing sites can be described with a power law distribution [76]. The power law nature indicates that a few tags are chosen by a large number of users, and the long tail of distribution corresponds to tags that are rarely used. A tag that is applied to a very large proportion of items is too general to be useful, while another tag that is applied very few times is useless due to its obscurity [162, 175]. A frequency threshold can

be applied to remove potentially low quality tags [123]. Variants such as *Term Frequency - Inverse Document Frequency* (TF-IDF) can also be used to obtain unbiased estimation of tag representativeness [155]. The search frequency of tags can also be used as a quality indicator [160].

- **Entropy** - Typically, the total number of tags in a collective vocabulary is much less than the total number of objects being tagged [35]. Also, the frequency of different tags is different. Therefore, the capacity of each tag in isolating a single document or identifying certain documents as more important than others, varies. This behavior can be captured using measures of information entropy. Chi et al. proposed an information theoretic approach to tag quality assessment, defined as the reduction in entropy or uncertainty in retrieving a particular document using that tag [35]. Zhang et al. compared tag frequency with entropy and concluded that frequency is a better predictor of tag quality, which can be refined using the measure of information entropy [205]. Low entropy tags are likely to have a skewed distribution in the sense that they appear more frequently in images belonging to only specific categories. Such tags can be treated as more content-descriptive and used to design visual category classifiers [125].
- **Clarity** - Image tag clarity is a measure of the descriptive power of a tag. It is computed as the KL divergence between language models computed from tag-specific image collection and the entire image collection [177]. As such, tags that are too general, end up having low clarity values.
- **Linguistic properties** - The number of characters, and especially those of special characters can be used as simple linguistic measures to filter low quality tags [160].

Similarity Computation for Tags: Tags can be compared by subjecting their vector space models to the measures of cosine similarity, Euclidean distance or a combination of statistical and structural information from lexical taxonomies [88]. We present three other prominent techniques that resort to large knowledge and document corpora.

- **WordNet distance**: WordNet is a manually constructed lexical database of English that contains over 155,000 words arranged in hierarchical groups of related words called synsets [60, 191]. A number of measures are proposed to compute the semantic similarity and relatedness between words using the information contents of synsets and the shortest path distance between synsets [143]. Budanitsky and Hirst presented an evaluation of five WordNet based measures indicating the superiority of information content based measures [22].
- **Google distance**: Google distance is an automated word-similarity measure computed using the frequency of occurrence and co-occurrence of words [38]. Normalized Google distance between image tags x and y can be computed as:

$$\frac{\max\{\log f(x), \log(f(y))\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \tag{1}$$

  where $f(x)$ denotes the number of images tagged using x and $f(x,y)$ denotes the number of images tagged using both x and y. N is a normalization factor computed through summation of co-occurrence counts over all tag pairs (x,y).
- **Flickr distance**: WordNet similarity and Google distance are both purely text based measures. Flickr distance is an automated measure that computes word association through the similarity of visual models associated with words [192]. For each tag, the collection of Flickr images tagged with that word is used to build a

a latent topic-based visual language model. The Flickr distance between two tags is measured as the square root of Jensen-Shannon divergence between the corresponding visual language models.

4 Techniques

The diverse visual, textual and EXIF features associated with social media can be processed using a variety of techniques. In this section we present four prominent techniques: co-occurrence statistic, clustering, classification, and spatio-temporal mining.

4.1 Co-occurrence Statistic

Two attributes co-occur if both are associated with a common instance. Let $P(v_i)$ denote the probability of an attribute v_i, computed as the ratio n_i/n where n is the total number of instances and n_i is the number of instances associated with v_i. Let $P(v_i, v_j)$ denote the co-occurrence probability of attributes i and j, computed as $n_{i,j}/n$, where $n_{i,j}$ is the frequency of co-occurrence of v_i and v_j. Then, one measure of co-occurrence is as the count $n_{i,j}$, possibly normalized using independent attribute frequencies. Another measure called the point-wise mutual information is computed using log of probability ratio as $\log \frac{P(v_i,v_j)}{P(v_i)P(v_j)}$.

Co-occurrence statistic helps measure association between different types of social media attributes. This is particularly useful for word sense disambiguation [133]. For example, the *distributional hypothesis* of statistical semantics states that, words that occur in the same contexts tend to have similar meanings. Therefore, co-occurrence of words that appear in the vicinity of a word can be used to disambiguate the meaning of the target word [27,76,107]. Furthermore, subsumption relationships (hierarchy of general to specific tags) can be captured using co-occurrence properties. Such knowledge is important for automatic creation of dynamic dictionaries and ontologies [198, 38]. Yang et al. proposed games to sustain and validate such dictionaries [199].

4.2 Clustering

Clustering is a statistical data analysis technique that groups observations in an unsupervised manner such that the intra-cluster element similarity is higher than the inter-cluster element similarity. Clustering is beneficial for a number of applications: as a pre-processing step for classification, for visualization and to facilitate browsing in large data collections. Clustering of multi-modal information such as visual features, tags and hyperlinks can be used to identify structural relationships in content [24, 39,188]. The top-down scheme generated using hierarchical clustering helps navigate from general to more specific concepts and explore large databases, as exemplified by systems like scatter-gather [41]. Clustering is also well suited for spatio-temporal data that inherently yield to meaningful groupings corresponding to geographical regions and time durations [135]. In this section, we restrict our discussion to prominent clustering techniques for social media analysis. For detailed technical discussion of clustering techniques, please refer to [55].

- **Agglomerative hierarchical clustering**: Agglomerative clustering is a type of hierarchical clustering where small clusters are iteratively merged into larger clusters to create a tree-like organization called *dendogram*. The leaves of a complete dendogram correspond to individual elements, whereas the root node consists of an aggregation of all elements. During each iteration of clustering, two candidate clusters with the smallest inter-cluster distance are merged to generate a higher-level cluster. The inter-cluster distance is typically measured using the minimum (*single-linkage*), maximum *(complete-linkage)* or average *(average-linkage)* pairwise distance between two clusters. The usage of pairwise distances makes hierarchical clustering suitable for categorical social data (such as tags, photos, comments) where pair-wise similarity is easy to compute using distributional measures.
  Hierarchical structure is useful for multi-scale visualization as well as for browsing large databases that involve coarse-to-fine exploratory navigation of the underlying data. The hierarchy is especially meaningful for geo-located photos, as clusters often correspond to actual geographic regions. Identifying representative tags over hierarchical clusters of geo-located data have helped discover location tags and their subsumption relationships (such as *'Golden gate bridge'* as a part of *'San Francisco'*). Hierarchies can also be used to effectively capture event semantics from temporal document collections [200].
- **Partitional clustering**: In partitional clustering, all cluster centroids are initialized in the beginning itself. k-means is a popular partitional clustering technique for feature approximation and image segmentation. In k-means, k cluster centroids are randomly initialized and iteratively refined to optimize the overall coherence of cluster assignments. Partitional clustering is heavily used for visualization of search results, where different senses associated with a multi-sense query are segregated into groups that enhance search experience. For example, clustering image search results of a query *'jaguar'* may highlight two unrelated groups - one corresponding to the animal and the other, a car. An extension called *search result clustering* that generates clusters with highly readable names [203] has become popular in recent years. The IGroups application based on this technique, computes clusters of image search results and represents each cluster with representative thumbnails [188]. Spectral clustering is another clustering technique for grouping folksonomic information represented in a graphical format [122]. A graph is represented as an association matrix, where the value at position (i, j) denotes the connectivity or the pair-wise similarity between elements i and j. Clusters are computed based on the Eigen decomposition of the (possibly normalized) similarity matrix. Leskovec et al. used the spectral clustering approach to partition interaction graph of users into meaningful communities [108].

4.3 Classification

Classification is a supervised learning technique that trains on labeled (classified) data to predict labels for unseen data. In this section, we discuss three supervised classification techniques for social images: nearest neighbor approach, Bayesian classification and support vector machines. For detailed technical discussions, please refer to [55].

- **Nearest neighbor approaches**: Nearest neighbor classification is an instance based supervised learning approach where the label for an instance is determined

from the labels of similar instances - termed as 'neighbors'. Also known as *lazy learning*, nearest-neighbor approaches defer classification until the prediction time. A popular approach is the k-nearest neighbor or k-NN technique where the candidate annotation is determined by a majority vote over k number of nearest instances. Aggregation by a regression analysis can also be used to compute a weighted mean of labels. When instances are associated with multiple labels (bag of tags representation), multi-label ranking approaches can be used to aggregate labels [21]. The nearest neighbor approach is very useful for applications such as social tag recommendation where similarity between a labeled and unlabeled document is used to propagate tags from the labeled document to the unlabeled document. This idea is similar in principle to item based collaborative filtering strategies where content similarity is utilized to mine annotations from already labeled data [158]. The search for visual neighbors can be conducted within personal or community data collections to attain personalization.

- **Bayesian classifiers**: Bayesian classifiers are a popular choice for generative concept modeling [111,159]. The probability of a concept C (hypothesis) given a set of visual features F (observations), i.e. $P(C|F)$ is computed using Bayesian formulation $\frac{P(F|C)P(C)}{P(F)}$. In this formula, $P(C)$ represents the concept prior, $P(F)$ represents the marginal and $P(F|C)$ is the probability with which features F are observed in the training data of the concept. The issue of data sparsity is often addressed using a Naïve Bayes formulation, where individual features $f_i, i \in (1, \ldots, |F|)$, are considered independent such that $P(F|C) = \prod_{i=1}^{|F|} P(f_i|C)$.
- **Support vector machines**: Support vector machine (SVM) is a supervised classification technique that uses discriminative modeling. The SVM for a binary classification problem is a hyperplane in the associated feature space that optimally separates positive training instances of a class from negative instances. Multi-class classification problems are decomposed into multiple binary classification problems using a one-versus-all or one-versus-one (pairwise) classification. SVMs have been used for computational aesthetic modeling from user-generated photos [42] and domain-specific classification tasks [31].

4.4 Spatio-Temporal Mining

Identification of events and locations provide a meaningful browsing modality by mining spatio-temporal information such as geo-coordinates, photo capture time, system upload time and times of user interactions (comments, tags, ratings) associated with folksonomic data. As such spatio-temporal mining is a vast subject and cannot be discussed here in sufficient details. Additional references can be found in [77,151] for temporal analysis and in [56,99,128] for spatial analysis. In this section we briefly touch upon a subset that is most relevant to social media.

Spatio-temporal models can be constructed by applying clustering or classification techniques within geographic or temporal constraints. For geographical data, structural relationships coherent with locations are important, and distance information corresponds to the geographic distance between location coordinates. In case of temporal analysis, the inherent periodicity (such as recurrence of hours, weeks, and months) is important and three phenomena are of interest - *trend* (increasing or decreasing growth

pattern), *seasonality* (periodicity of activity) and *anomalies* (deviation from predicted activity). Detecting data patterns helps anticipate future activity, such as search volume for a query that can be easily optimized with help of early predictions. Trend and seasonality can be deduced using time domain techniques (such as auto-correlation and cross-correlation) or frequency-domain techniques (such as Fourier transform and wavelet analysis). Detecting anomalies such as sudden bursts in activity may help uncover important events[3]. Burst detection was first addressed by Kleinberg where he used an infinite automaton to model topic bursts in document streams [98]. Vlachos et al. applied discrete Fourier transform and power spectral density measurement to identify periodicity and bursts in data [184]. Ihler et al. used a time-varying Poisson model modulated by a hidden Markov process to capture bursty events [83].

5 Applications

Image semantics has a variety of facets as outlined in Section 1. In this section, we discuss representative work in content semantics, person semantics, event semantics and location semantics. The discussion of event and location semantics is combined as the co-existence of location and temporal information is typically analyzed using joint spatio-temporal mining techniques.

5.1 Visual Semantics

Traditionally, the interpretation of visual semantics is conducted using supervised concept learning from labeled data. With the advent of Web 2.0, image annotation or tag recommendation techniques can be be classified into two types: a) model-based and b) model-free techniques. Model-based techniques are conceptually similar to the traditional concept learning techniques, except that the input training data is mined from labels in social media and collaborative games. The modeling error when using noisy folksonomic annotations is typically higher than the error obtained while using expert labels. Nonetheless, the scalability achieved using folksonomic data is attractive and efforts to auto-select high-quality training data are on the rise. Also, runtime annotation is usually automatic and fast. The second type of annotation systems are model-free and often semi-automatic in the sense that they require one or more tags or camera metadata to bootstrap the annotation process. Such systems are loosely referred to as tag completion strategies. Without the restriction of statistical modeling, model-free systems potentially scale to arbitrarily large vocabularies, albeit at an increased runtime complexity. This section deals with three aspects of visual semantics: tag relevance assessment, concept modeling, and image annotation. Also, studies on computational aesthetic modeling are presented.

5.1.1 Tag Relevance Estimation

The primary use of tags is for content retrieval, where the retrieved content may be used as training data for concept modeling. However, tags associated with an image

[3] This problem has gained focus since late 1990s, with the introduction of *topic detection and tracking* challenge for event-based information organization [8, 200].

may or may not be relevant to the visual content of the image. In a study of Flickr tags, Liu et al. showed that the tag input sequence does not have a strong correlation with the relevance of tags. The evidence of a trend that top tags are more relevant is weak and the most relevant tag is ranked at the top in less than 10% images [117]. Hence, it becomes necessary to determine and prune noisy tags. Noise mitigation has been extensively studied for text classification [52,149,147] and for outlier detection tasks [29,78]. In case of social tags, similar approaches are applicable.

Co-occurrence of tags with visual features can be used to prune irrelevant keywords. Jin et al. used a fusion of multiple WordNet similarity measures and visual features to remove tags that were weakly correlated with visual features [90]. Li et al. employed a K-nearest neighbor voting method to determine the relevance between an image and an associated tag [114]. Common tags of visually similar images were ranked according to votes collected from visual neighbors. In the TagProp algorithm, a weighted nearest neighbor model was used to predict relevance from annotations of labeled visual neighbors [183]. Kennedy et al. demonstrated that original photographers can be treated as reliable sources of high quality annotations due to their first-hand knowledge of the content-matter [95]. They utilized agreement between photographers on labels of visually similar images to select reliable tag-feature association.

The tripartite nature of folksonomies has inspired ranking strategies to assess and retrieve quality resources. The folkrank algorithm, provided a ranking such that frequently appearing users, tags and resources appeared early on in results [80,81]. Jin et al. computed the tripartite relationships between groups, tags and images in social networks to modify relevance ranking to group-based social image search [89]. Tang et al. presented a sparse graph-based semi-supervised learning approach for removing weak pairing between image features and tags [180]. To refine the relevance ranking, Liu et al. used a probabilistic approach that first initialized relevance scores for tags and computed a refinement using a random walk procedure over a tag similarity graph [119]. They further proposed a unified framework using visual and semantic similarity of images to measure the compatibility of tags before and after the approach. Wang et al. proposed a random walk with restarts method leveraging co-occurrence-based similarity as well as the ranking and confidence information of original annotations [185]. In another extension, feature noise was reduced using a canonical correlation analysis that created a compact representation of features [13].

5.1.2 Concept Modeling

A number of statistical techniques are developed to model the association between words and visual features, given a large labeled training set of concept exemplars [14, 87,110,111]. In Web 2.0, it is possible to obtain labels from collaborative games such as Peekaboom and ESP game [6] or from socially tagged images. Learned models are useful for high-speed automatic annotation of unlabeled images.

Wang et al. developed Gaussian mixture model representations for concepts by clustering keypoint code vectors in tagged images [187]. Chatzilari et al. clustered region-level MPEG-7 descriptors from social images for object modeling [31]. The descriptors for an object were extracted from semantically coherent image groups using the SEMSOC framework [69] that combines knowledge from WordNet, tag co-occurrence and visual features information. One limitation on these approaches, as with traditional concept learning, is that the prediction can be made only for concepts specified in training. Datta et al. described a PLMFIT model where a limited set of trained concepts

can be scaled to arbitrarily large vocabularies. Sawant et al. extended this idea using content-based annotations as meta-features and translating them to the vocabulary of a user's local interaction network [159]. This approach is useful especially when the user's network indulges in thematic image sharing (resembling special interest groups), so that with the knowledge of the network's interests, appropriate inference can be made about a user's own tag preferences. The Sheepdog image annotation system also presented a translation method by detecting pre-determined concepts in images and identifying relevant communities for the concepts [32]. Using tag models built over the community vocabularies, annotations were recommended.

5.1.3 Image Annotation

Model-free semi-automatic annotation systems require users to supply an initial set of tags for the images to be annotated. Additional tags are recommended using techniques such as co-occurrence statistic, content based retrieval, nearest neighbor matching, and clustering [1, 112].

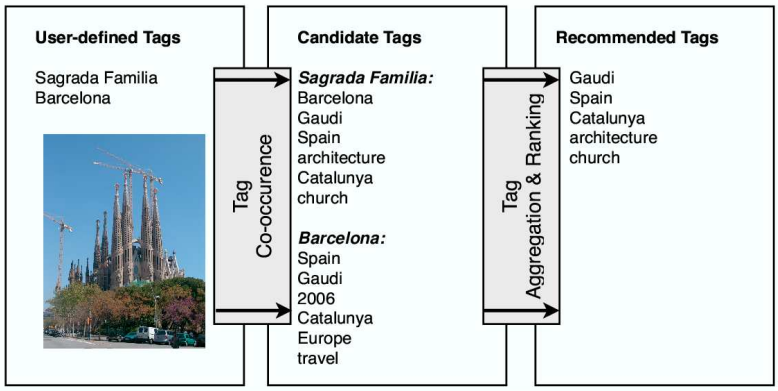

Fig. 8 Using co-occurrence and aggregation for tag recommendation [166].

Co-occurrence of two tags over a large number of photos indicates that a significant relationship exists between the tags. In practice, co-occurrence value is normalized by the individual tag occurrence frequencies to account for bias from popular tags. Tags having a high co-occurrence with existing tags are selected as natural candidates for additional annotations. Various aggregation strategies are applied to rank the candidate annotations. One example is shown in Fig. 8, where candidate tags are weighted using a voting mechanism [166]. Garg et al. selected tags using global and personal level tag co-occurrence and aggregated them using a weighted combination [67]. Weinberger et al. used a probabilistic framework to model tag co-occurrence and identified tags that disambiguate the different temporal, location or semantic contexts in which a polysemous tag appears [190]. If an image associated with a tag set T contained an ambiguous tag t, then additional tags t_i and t_j were selected so as to increase the divergence $p(t|T \cup t_i)$ and $p(t|T \cup t_j)$. Wang et al. selected candidate annotations using item-based co-occurrences with existing annotations [186]. Candidate annotations were ranked using image visual similarity in a Markov model formulation. Wu et al. used

tag visual correlation, image conditioned tag correlation and tag co-occurrence to create separate rankings of candidate tags [193]. The rankings were combined through a Rankboost algorithm.

Nearest-neighbor search methods are useful to identify visual neighbors from which annotations can be mined. Search can be optimized using high dimensional indexing techniques [113]. The Tag Suggestr [102] system based on a combination of visual and text features, required a user to annotate a photo with an initial set of tags. The initial tags were used to retrieve additional tags from related photos that had some of the initial tags in their tag lists. Ivanov et al. proposed a system, where tags are propagated from images containing duplicate objects as in the query image [84]. Objects were represented using visual word vectors learned from hierarchical k-means clustering over sparse local region features. A classification step was applied to categorize images before the nearest neighbor search [115]. Using location based nearest neighbors, labels from labeled photos were propagated to unlabeled images [131].

Clustering is an effective technique to identify related tags. When one tag in a cluster is relevant to a photo, other tags from that cluster can be treated as potentially relevant as well. In Annosearch tool [189] annotations associated with content-similar images were clustered using search result clustering. Shepitsen et al. developed a personalized recommendation system that incorporated user profiles and previous tag clusters to re-rank the tags suggested by a non-personalized recommendation algorithm [165]. Clustering can also be used to power smart batch-tagging techniques where images are clustered and only the cluster representatives are tagged. Image-level tags are propagated using measures of visual similarity and temporal consistency [118].

A number of research methods also harness the rich metadata associated with photos. The camera metadata associated with images can be used for image classification and annotation [20,167] as well as for creating new browsing experiences. Cao et al. suggested annotations from a set of event/activity and scene/location tags, using the similarity of GPS and time information among visually similar images [26]. Time, location and visual similarity between photos was exploited to propagate high-confidence labels of photos to similar photos [25]. The SpiritTagger tool utilized GPS coordinates and visual content to annotate photos with other geographically relevant tags [129]. The tool ZoneTag also used GPS locations to identify related tags from a user's social network [130]. Lindstaedt et al. presented a *Tagr* system based on a mash up of different image and user features [116].

5.1.4 Computational Aesthetics Modeling

Aesthetics is *the branch of philosophy which deals with questions of beauty and artistic taste* [2]. Image aesthetics is an abstract notion of quality influenced by features like hue, saturation, sharpness, contrast, colorfulness and edge distribution. The subjectivity of aesthetics interpretation makes computational modeling a challenge. However, community contributed photo ratings and comments, have become partially successful in meeting the challenge. Ratings from peer-rated photo sharing communities such as Photo.net [144] and DPChallenge.com [53] are used as reasonable ground truth data for computational aesthetic modeling [42,93,141]. A threshold on average ratings is used to classify photos into high quality and low quality categories. Datta et al. studied the correlation of over fifty visual features and developed aesthetic classifiers using support vector machines [42]. They also demonstrated an application of SVM regressor to predict a continuous aesthetic quality score. Ke et al. used Bayesian classification

to predict the perceptual high or low quality of images using features such as edge and color distribution, blur, and hue. Pere et al. computed an overall aesthetic measure using features such as color contrast, sharpness, and noise [141].

One other approach to aesthetics and emotion modeling is *opinion mining and sentiment analysis* [142] of tags and comments associated with images. For example, words like 'bravo', 'beautiful', 'bestshot' are indicators that people appreciate visual content of photos. Additionally, an analysis of user actions such as ratings and page views may help determine positive and negative sentiments. Solli et al. have presented a web image dataset associated with emotion-related labels [172]. Such datasets can be crawled from community contributed collections or specifically obtained by designing aesthetic rating interfaces. Fig. 9 shows a snapshot of a system that obtains ratings from users and trains aesthetic inference models [45].

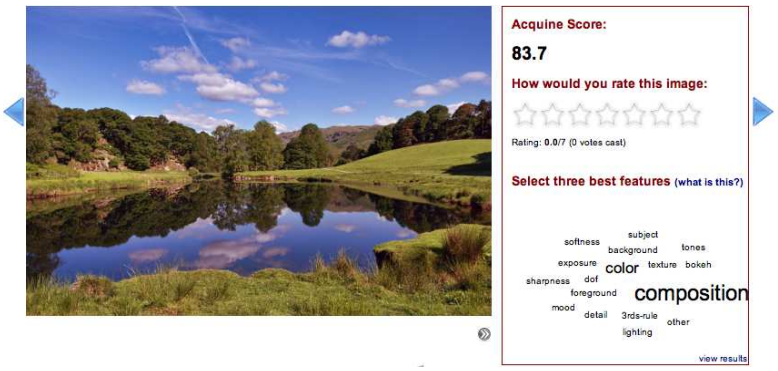

Fig. 9 Snapshot from *ACQUINE* image aesthetic rating Website.

5.2 Person Recognition

Identification of individuals in personal photographs is helpful for management of personal photo collections. Detected co-appearances of people can be used to build event chapters [30] and to infer links in social networks [71]. Current research on person identification draws heavily on face recognition technology [70, 207] as well as clustering and classification of supplemental information such as clothes, body, and hair appearance [12, 36, 168, 173, 176, 195, 206]. The repeated temporal and spatial occurrences as well as frequency of individual appearances, can be used to build belief models that help prediction in case of unlabeled photographs [132]. GPS information can be used for person recognition and propagation of such annotations in personal albums [46, 204].

5.3 Event and Location Semantics

The discovery of spatio-temporal patterns and representative tags has attracted a great interest in the research community. In addition to event and location semantics, bursts in spatio-temporal tags often coincide with names of personalities and communities.

Patterns in tagging behavior provide additional dimensions for semantic descriptions and can be discovered using mining as well as interactive visualization techniques [11, 54]. Multi-scale clustering is a popular technique for spatio-temporal mining. Rattenbury et al. applied multi-scale clustering to extract event and location semantics from geotagged images [148]. Chen et al. applied wavelet analysis to a joint representation of time and location metadata of Flickr photos to detect periodic events [33]. Zunjarwad et al. applied the hyperlink-induced topic search algorithm [97] to visual and social network information to characterize events (and its facets such as who, where, when, and what) [208]. Joshi et al. used supervised learning techniques to model the association between event categories (such as hiking and skiing) and bags of geo-tags computed from the spatial neighborhoods of Flickr photographs [92]. An alternate approach was proposed by Yuan et al. by utilizing a fusion of GPS trace features with compositional visual features [202]. The information about addresses and points of interest was mined using a large online location database called GeoNames [68]. Such location databases are also useful for geo-recognition or identification of location semantics when no prior information is available. Cristani et al. proposed an approach for geo-category prediction by clustering geographically proximal images [40]. Serdyukov et al. used Flickr tags for geographical location prediction of photos lacking geo-references [164].

5.3.1 Landmark Recognition

People, in particular tourists, are often interested in viewing photos of landmarks across world-wide locations. To automatically identify landmarks from existing photo collections, researchers have resorted to methods based on only metadata, only content or a fusion of both. Methods such as [47, 171] used supervised learning to select distinguishing features to identify landmarks in a pre-determined list of locations. Berg et al. employed computer vision techniques to select *iconic* images for selected locations [16]. This system is limited to processing photos from a single viewpoint and the relationships among different views of a landmark cannot be identified.

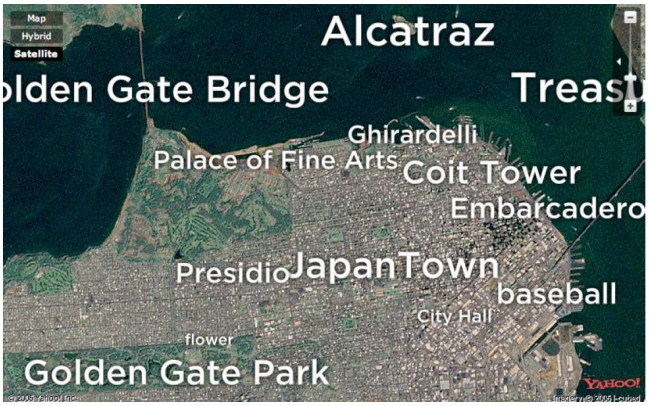

Fig. 10 Visualization of points of interest in *San Francisco* using Flickr image tags [94]

.

Social tags with associated spatio-temporal information can be readily plugged into general techniques for landmark recognition. Jaffe et al. used clustering techniques to

generate summaries for large collections of geo-referenced photographs [85]. Kennedy et al. used a multi-modal fusion of location information, tags and computer vision inputs [94] to predict landmarks as depicted in Fig. 10. Tags occurring over a concentrated geographical area were considered more specific to location-semantics than tags diffused over large regions. TF-IDF based frequency measures were used to assess representativeness. Predictions were coupled with visually representative images computed using clustering of salient visual features. Clustering helps discover multiple views of the location, each of which can be described with representative tags. A tool named World Explorer built around this idea, helps generate visualization for arbitrary areas in the world [4]. It can also be used to construct tourist maps from geo-tagged Flickr images [34].

6 Epilogue

Semantic interpretation from social images is a young field that is growing at a tremendous pace. With millions of users, billions of images and associated metadata, the raw inputs available for the discovery of semantic and structural knowledge are immense. From the literature review in this paper, we observed that current research resorts to applications of mature algorithms in the area of data mining and information retrieval. However, uncertainties associated with human choices of metadata and subjective interpretation of visual content pose new challenges that cannot be fathomed using these areas alone. To be able to handle the transition of multimedia research from small expert labeled datasets to astronomical collections with noisy labels requires support from other research fields that have not been part of mainstream multimedia research so far. Studies of human perception, cognition, linguistics, psychology, and social sciences represent important aspects addressing the *human element* in social multimedia. Secondly, heterogeneous information cues from visual, textual and behavioral data need to be effectively combined using techniques of multi-modal data and decision fusion. Finally, a strong boost is needed in technology development that can efficiently handle the increased burden on time and space requirements. Thus we envision people, information and technology to be the three fundamental and equally important components of future research in semantic interpretation of social media. These components need to be independently or jointly analyzed, wherever applicable, to advance the state-of-the-art of semantic understanding from images. In this section, we briefly touch on some aspects of interest to current and future research directions.

– **Accessible information**: Current understanding of folksonomic features is skewed. User modeling, in particular, is an enigma and requires effective analysis of a user's actions and annotations, possibly within the context of his social network. The information is available from a variety of sources such as visual features, tags, comments, ratings, page views, and social networks. The representation and fusion of such heterogeneous data is central to effective ingestion of information. Representation should be considered as an art of making data accessible such that the semantic and structural properties of elements become apparent. Statistical modeling and visualization techniques need to complement each other to *detect needles in haystacks*. We envision multi-lingual image folksonomies as a bridge that will join knowledge and ontologies from different languages. By matching visual features, multi-lingual information associated with images can be connected and easily ac-

cessed. Development of semantic web is a possible extension to make information accessible not only to humans but also to automated agents.

- **Interface design**: Manual tagging consumes time and effort. The information foraging theory predicts that individual tag production rates can be increased by lowering the effort of tag production [145, 146]. Alleviating the tagging burden can push more individuals to be participative, thus increasing the net production of quality tags. Current interfaces partially alleviate the effort of tagging by providing automated recommendations, easy tag selection [23], and faster manual annotation methods [197]. However, external tag recommendation may influence user decisions and force folksonomies to artificially converge. The implications and usability of such convergence has not been studied yet. Support for manual annotations can also be provided by making annotations fun. In this respect, interface design and usability analysis of social media and collaborative games is an important research direction. A good design should be able to prompt users to provide more quality metadata as well as validate it. Personalization can also be incorporated in the design to tailor the experience to user interests and expertise.
- **Adaptable technology**: Adaptability is an important aspect of folksonomies as they continuously evolve with the underlying user-base and vocabulary. In this regard, the paradigm of change mining is important for analysis of concept drifts [19]. Another aspect of adaptation arises from the variable performance of social inputs in concept modeling. The potential of folksonomic annotations for visual concept modeling and image annotation is evident. However, the high error rates experienced by current systems are too large for the systems to be of wide-spread practical use. The success of automation may be greater for certain concepts, such as those with low semantic gap [121]. Kennedy et al. have conducted initial experiments to predict which visual concepts might benefit from human annotations versus obtaining images from Web collections [96]. This presents an important suggestion that concept modeling approaches need to adapt to different concept types - as opposed to a '*one size fits all*' approach. Effective personalization techniques need to be developed to further adapt techniques to individual users and groups.
- ***Scalable technology*** - The rapid increase in social collections presents scalability challenges for existing data mining techniques. It is necessary to develop fast, lightweight and effective techniques that scale to the rapid influx of data. Considering the spatio-temporal nature, social data may be subjected to models under the larger umbrella of *data stream mining* techniques [66]. Processing of raw data can be replaced by mining higher order patterns as in *higher order mining* [152, 201]. Fast computing platforms such as the Map-Reduce framework [48] need to be heavily adopted in information extraction tasks.

In conclusion, we find that the research on image semantics is undergoing a paradigm shift with the introduction of social media sharing and collaborative games. The added context has broadened the scope of semantic interpretation beyond visual features. To harness the plethora of information requires interdisciplinary research and a confluence of insights from areas like image processing, data mining, human computer interaction and sociology. In this paper, we have reviewed nearly 200 representative papers and summarized current efforts according to four application categories: concept semantics, person identification, location semantics and event semantics. We have also pointed out challenges and future research directions. We hope that the readers find this survey as a concise summary of the rapidly evolving field of social image semantics.

References

1. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6), 734–749
2. Aesthetics, definition from oxford dictionaries. `http://www.oxforddictionaries.com/definition/aesthetics`
3. Agichtein E, Brill E, Dumais S (2006) Improving web search ranking by incorporating user behavior information. Proc Res Dev Inf Ret, 19–26
4. Ahern S, Naaman M, Nair R, Yang J (2007) World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. Proc Jt Conf Digit Libr, 1–10
5. von Ahn L, Dabbish L (2004) Labeling images with a computer game. Proc Hum Factors Comput Syst, 319–326
6. von Ahn L, Dabbish L (2008) Designing games with a purpose. Commun. ACM 51(8), 58–67
7. von Ahn L, Liu R, Blum M (2006) Peekaboom: A game for locating objects in images. Proc Hum Factors Comput Syst, 55–64
8. Allan J, (ed) (2002) Topic detection and tracking: event-based information organization. Kluwer Academic Publishers
9. Allan M, Verbeek J (2009) Ranking user-annotated images for multiple query terms. Proc Br Mach Vis Conf
10. Ames M, Naaman M (2007) Why we tag: Motivations for annotation in mobile and online media. Proc Hum Factors Comput Syst, 971–980
11. Andrienko N (2003) Exploratory spatio-temporal visualization: an analytical review. J Vis Lang & Comput 14(6), 503–541
12. Anguelov D, Lee K, Gokturk S, Sumengen B (2007) Contextual identity recognition in personal photo albums. Proc Comput Vis Pattern Recognit, 1–7
13. Bailloeul T, Zhu C, Xu Y (2008) Automatic image tagging as a random walk with priors on the canonical correlation subspace. Proc ACM Multimed Inf Ret, 75–82
14. Barnard K, Duygulu P, Forsyth D, Freitas ND, Blei DKJ, Hofmann T, Poggio T, Shawe-taylor J (2003) Matching words and pictures. J Mach Learn Res 3, 1107–1135
15. Barnard K, Fan Q, Swaminathan R, Hoogs A, Collins R, Rondot P, Kaufhold J (2008) Evaluation of localized semantics: Data, methodology, and experiments. Int J Comput Vis 77(1-3), 199–217
16. Berg T, Forsyth D (2007) Automatic ranking in iconic images. Tech rep, University of California, Berkeley
17. Bian J, Liu Y, Agichtein E, Zha H (2008) A few bad votes too many?: towards robust ranking in social media. Proc Workshop on Advers Inf Ret on the Web, 53–60
18. Bian J, Liu Y, Zhou D, Agichtein E, Zha H (2009) Learn to recognize reliable users and content in social media with coupled mutual reinforcement. Proc World Wide Web, 51–60
19. Böttcher M, Höppner F, Spiliopoulou M (2008) On exploiting the power of time in data mining. SIGKDD Explor Newsl 10(2), 3–11
20. Boutell M, Luo J (2004) Photo classification by integrating image content and camera metadata. Proc Pattern Recognit, 901–904
21. Brinker K, Hüllermeier E (2007) Case-based multilabel ranking. Proc Int Jt Conf Artifical Intell, 702–707
22. Budanitsky A, Hirst G (2006) Evaluating wordnet-based measures of lexical semantic relatedness. Proc Res Comput Linguist 32(1), 13–47
23. Budiu R, Pirolli P, Hong L (2009) Remembrance of things tagged: How tagging effort affects tag production and human memory. Proc Hum Factors Comput Syst, 615–624
24. Cai D, He X, Li Z, Ma WY, Wen JR (2004) Hierarchical clustering of www image search results using visual, textual and link information. Proc ACM Multimed, 952–959
25. Cao L, Luo J, Huang TS (2008) Annotating photo collections by label propagation according to multiple similarity cues. Proc ACM Multimed, 121–130
26. Cao L, Luo J, Kautz H, Huang T (2008) Annotating collections of photos using hierarchical event and scene models. Proc Comput Vis Pattern Recognit
27. Cattuto C, Benz D, Hotho A, Stumme G (2008) Semantic analysis of tag similarity measures in collaborative tagging systems. Proc Workshop Ontol Learn & Popul, 39–43
28. Cattuto C, Schmitz C, Baldassarri A, Servedio V, Loreto V, Hotho A, Grahl M, Stumme G (2007) Network properties of folksonomies. AI Commun 20(4), 245–262
29. Chandola V, Banerjee A, Kumar V (2009) Outlier detection: A survey. ACM Surv

30. Chandramouli K, Izquierdo E (2010) Semant structuring and retrieval of event chapters in social photo collections. Proc ACM Multimed Inf Ret, 507–516
31. Chatzilari E, Nikolopoulos S, Kompatsiaris I, Giannakidou E, Vakali A (2009) Leveraging social media for training object detectors. Proc Digit Signal Proc, 1–8
32. Chen HM, Chang MH, Chang PC, Tien MC, Hsu WH, Wu JL (2008) Sheepdog: group and tag recommendation for Flickr photos by automatic search-based learning. Proc ACM Multimed, 737–740
33. Chen L, Roy A (2009) Event detection from Flickr data through wavelet-based spatial analysis. Proc ACM Inf Knowl Manag, 523–532
34. Chen WC, Battestini A, Gelfand N, Setlur V (2009) Visual summaries of popular landmarks from community photo collections. Proc ACM Multimed, 789–792
35. Chi E, Mytkowicz T (2008) Understanding the efficiency of social tagging systems using information theory. Proc ACM Hypertext and Hypermedia, 81–88
36. Choi JY, Yang S, Ro YM, Plataniotis K (2008) Face annotation for personal photos using context-assisted face recognition. Proc ACM Multimed Inf Ret, 44–51
37. Chua TS, Tang J, Hong R, Li H, Luo Z, Yan-Tao Z (2009) Nus-wide: A real-world web image database from national university of singapore. Proc ACM Image and Video Ret, 1–9
38. Cilibrasi R, Vitanyi P (2007) The Google similarity distance. IEEE Trans Knowl Data Eng 19(3), 370–383
39. Cristani M, Perina A, Castellani U, Murino V (2008) Content visualization and management of geo-located image databases. Ext Abstr On Hum Factors Comput Syst, 2823–2828
40. Cristani M, Perina A, Castellani U, Murino V (2008) Geo-located image analysis using latent representations. Proc Comput Vis Pattern Recognit
41. Cutting DR, Karger DR, Pedersen JO, Tukey JW (1992) Scatter/gather: A cluster-based approach to browsing large document collections. Proc Res and Dev Inf Ret, 318–329
42. Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. Lect. Notes in Comput Sci: Proc Eur Conf Comput Vis 3953(3), 288–301
43. Datta R, Joshi D, Li J, Wang JZ (2007) Tagging over time: Real-world image annotation by lightweight meta-learning. Proc ACM Multimed, 393–402
44. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: Ideas, influences, and trends of the new age. ACM Comput Surv 40(2), 1–60
45. Datta R, Wang JZ (2010) Acquine: Aesthetic quality inference engine - real-time automatic rating of photo aesthetics. Proc ACM Multimed Inf Ret, Demo, 421–424
46. Davis M, Smith M, Canny J, Good N, King S, Janakiraman R (2005) Towards context-aware face recognition. Proc ACM Multimed, 483–486
47. Davis M, Smith M, Stentiford F, Bamidele A, Canny J, Good N, King S, Janakiraman R (2006) Using context and similarity for face and location identification. Proc Symp on Electron Imaging Sci and Tech
48. Dean J, Ghemawat S (2004) Mapreduce: simplified data processing on large clusters. Proc Symp on Opear. Syst Design & Implement., 10–10
49. Deconstructing Flickr interestingness. `http://wes2.wordpress.com/2006/05/12/deconstructing-flickrs-interestingness`
50. Deng J, Li K, Do M, Su H, Fei-Fei L (2009) Construction and analysis of a large scale image ontology. Vis Sci Soc
51. Dong W, Fu WT (2010) Cultural difference in image tagging. Proc Hum Factors Comput Syst, 981–984
52. Donmez P, Carbonell J, Schneider J (2009) Efficiently learning the accuracy of labeling sources for selective sampling. Proc ACM Knowl Discov Data Mining, 259–268
53. DPChallenge - a digital photography contest. `http://www.dpchallenge.com`
54. Dubinko M, Kumar R, Magnani J, Novak J, Raghavan P, Tomkins A (2006) Visizing tags over time. Proc World Wide Web, 193–202
55. Duda R, Hart P, Stork D (2000) Pattern classification, 2nd edn. Wiley-Interscience
56. Ester M, Kriegel H, Sander J (1999) Knowledge discovery in spatial databases. Proc Ger Conf Artif Intell, 61–74
57. Exif and related resources. `http://www.exif.org`
58. Feifei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE Trans Pat Anal Mach Intell 28(4), 594–611

59. Feifei L, Fergus R, Perona P (2007) Learn generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput Vis Image Underst 106(1), 59–70
60. Fellbaum C, (ed) (1998) WordNet: An electronic lexical database (language, speech, and communication). The MIT Press
61. 4,000,000,000. http://blog.flickr.net/en/2009/10/12/4000000000/
62. Flickr interestingness. http://www.flickr.com/explore/interesting
63. Fu WT (2008) The microstructures of social tagging: A rational model. Proc ACM Comput Support Co-op Work, 229–238
64. Furnas G, Landauer T, Gomez L, Dumais S (1987) The vocabulary problem in human-system communication. Commun ACM 30(11), 964–971
65. Furnas G, Landauer T, Gomez L, Dumais S (1984) Statistical semantics: analysis of the potential performance of keyword information systems. Proc SIGCHI Conf Hum Factors Comput Syst, 187–242
66. Gaber M, Zaslavsky A, Krishnaswamy S (2005) Mining data streams: a review. ACM SIGMOD Rec 34(2), 18–26
67. Garg N, Weber I (2008) Personalized, interactive tag recommendation for Flickr. Proc ACM Recomm Sys, 67–74
68. Geonames. http://www.geonames.org
69. Giannakidou E, Kompatsiaris I, Vakali A (2008) Semsoc: Semantic social and content-based clustering in multimedia collaborative tagging systems. Proc IEEE Int Conf Semant Comput, 128–135
70. Girgensohn A, Adcock J, Wilcox L (2004) Leveraging face recognition technology to find and organize photos. Proc ACM SIGMM Int workshop Multimed Inf Ret, 99–106
71. Golder S (2008) Measuring social networks with digital photograph collections. Proc ACM Hypertext and Hypermedia, 43–48
72. Golder S, Huberman B (2006) Usage patterns of collaborative tagging systems. J Inf Sci 32(2), 198–208
73. Gonçalves D, Jesus R, Correia N (2008) A gesture based game for image tagging. Ext abstr on Hum Factors Comput Syst, 2685–2690
74. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset. Tech Rep 7694, California Institute of Technology
75. Games with a purpose. http://www.gwap.com/gwap/about
76. Halpin H, Robu V, Shepherd H (2007) The complex dynamics of collaborative tagging. Proc World Wide Web, 211–220
77. Hamilton JD (1994) Time series analysis, 1st edn. Princeton University Press
78. Hodge V, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22(2), 85–126
79. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42(1-2), 177–196
80. Hotho A, Jäschke R, Schmitz C, Stumme G (2006) Inf retrieval in folksonomies: Search and ranking. Proc Eur Semant Web Conf, vol 4011, 411–426
81. Hotho A, Jäschke R, Schmitz C, Stumme G (2006) Folkrank: A ranking algorithm for folksonomies. Proc Conf Fachgruppe Inf Ret, 111–114
82. Huiskes M, Lew M (2008) The mir Flickr retrieval evaluation. Proc ACM Multimed Inf Ret, 39–43
83. Ihler A, Hutchins J, Smyth P (2007) Learn to detect events with markov-modulated poisson processes. ACM Trans Knowl Discov Data 1(3), 13
84. Ivanov I, Vajda P, Goldmann L, Lee J, Ebrahimi T (2010) Object-based tag propagation for semi-automatic annotation of images. Proc ACM Multimed Inf Ret, 497–506
85. Jaffe A, Naaman M, Tassa T, Davis M (2006) Generating summaries and visualization for large collections of geo-referenced photographs. Proc ACM Workshop on Multimed Inf Ret, 89–98
86. Jäschke R, Marinho L, Hotho A, Schmidt-Thieme L, Stumme G (2007) Tag recommendations in folksonomies. Proc Eur Conf Princ. and Pract of Knowl Discov Databases, 506–514
87. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. Proc Res Dev Inf Ret, 119–126
88. Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. Proc Res Comput Linguist, 19–33

89. Jin X, Luo J, Yu J, Wang G, Joshi D, Han J (2010) iRIN: image retrieval in image-rich information networks. Proc World wide web, 1261–1264
90. Jin Y, Khan L, Wang L, Awad M (2005) Image annotations by combining multiple evidence & wordnet. Proc ACM Multimed, 706–715
91. Jones W (1986) The memory extender personal filing system. Proc Hum Factors Comput Syst, 298–305
92. Joshi D, Luo J (2008) Inferring generic activities and events from image content and bags of geo-tags. Proc Content-based Image and Video Ret, 37–46
93. Ke Y, Tang X, Jing F (2006) The design of high-level features for photo quality assessment. Proc Comput Vis Pattern Recognit, 419–426
94. Kennedy L, Naaman M, Ahern S, Nair R, Rattenbury T (2007) How Flickr helps us make sense of the world: Context and content in community-contributed media collections. Proc ACM Multimed, 631–640
95. Kennedy L, Slaney M, Weinberger K (2009) Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases. Proc Workshop Web-scale Multimed Corpus, 17–24
96. Kennedy L, Chang S, Kozintsev I (2006) To search or to label?: predicting the performance of search-based automatic image classifiers. Proc ACM Workshop on Multimed Inf Ret, 249–258
97. Kleinberg J (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5), 604–632
98. Kleinberg J (2003) Bursty and hierarchical structure in streams. Proc Knowl Discov and Data Mining 7(4), 373–397
99. Koperski K, Adhikary J, Han J (1996) Spatial data mining: Progress and challenges - survey paper. Proc Workshop Res Issues Data Mining Knowl Discov, 1–10
100. Koutrika G, Effendi F, Gyöngyi Z, Heymann P, Garcia-Molina H (2007) Combating spam in tagging systems. Proc Workshop Advers Inf Ret Web, 57–64
101. Krestel R, Chen L (2008) Using co-occurence of tags and resources to identify spammers. Proc Conf Pract Knowl Discov Databases
102. Kucuktunc O, Sevil S, Tosun A, Zitouni H, Duygulu P, Can F (2008) Tag suggestr: Automatic photo tag expansion using visual information for photo sharing websites. Proc Semant Digit Media Tecnol: Semant Multimed
103. Kustanowitz J, Shneiderman B (2005) Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives. Tech rep, University of Maryland
104. LabelMe: The open annotation tool. `http://labelme.csail.mit.edu`
105. Lambiotte R, Ausloos M (2006) Collaborative tagging as a tripartite network. Proc Comput Sci, 1114–1117
106. Landauer T, Foltz P, Laham D (1998) An introduction to latent semantic analysis. Discourse Proc. 25, 259–284
107. Lee L (1999) Measures of distributional similarity. Proc Comput Linguist, 25–32
108. Leskovec J, Lang K, Dasgupta A, Mahoney M (2008) Statistical properties of community structure in large social and information networks. Proc World Wide Web, 695–704
109. Levi K, Weiss Y (2004) Learn object detection from a small number of examples: The importance of good features. Proc Comput Vis Pattern Recognit 2, 53–60
110. Li J, Wang JZ (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans Pat Anal Mach Intell 25(9), 1075–1088
111. Li J, Wang JZ (2008) Real-time computerized annotation of pictures. IEEE Trans Pat Anal Mach Intell 30(6), 985–1002
112. Li Q, Lu S (2008) Collaborative tagging applications and approaches. IEEE Multimed 15(3), 14–21
113. Li X, Chen L, Zhang L, Lin F, Ma WY (2006) Image annotation by large-scale content-based image retrieval. Proc ACM Multimed, 607–610
114. Li X, Snoek CG, Worring M (2008) Learn tag relevance by neighbor voting for social image retrieval. Proc ACM Multimed Inf Ret, 180–187
115. Lindstaedt S, Mörzinger R, Sorschag R, Pammer V, Thallinger G (2009) Automatic image annotation using visual content and folksonomies. Multimed Tools App 42(1), 97–113
116. Lindstaedt S, Pammer V, Mörzinger R, Kern R, Mülner H, Wagner C (2008) Recommending tags for pictures based on text, visual content and user context. Proc Internet & Web Appl & Serv, 506–511
117. Liu D, Hua XS, Yang L, Wang M, Zhang HJ (2009) Tag ranking. Proc World Wide Web, 351–360

118. Liu D, Wang M, Hua XS, Zhang HJ (2009) Smart batch tagging of photo albums. Proc ACM Multimed, 809–812
119. Liu D, Wang M, Yang L, Hua XS, Zhang H (2009) Tag quality improvement for social images. Proc IEEE Multimed and Expo, 350–353
120. Liu Y, Zhang D, Lu G, Ma WY (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recognit 40(1), 262–282
121. Lu Y, Zhang L, Tian Q, Ma WY (2008) What are the high-level concepts with small semantic gaps?. Proc Comput Vis Pattern Recognit
122. Luxburg U (2007) A tutorial on spectral clustering. Stat and Comput 17(4), 395–416
123. Manning C, Raghavan P, Schutze H (2008) Index compression. In: Introduction to information retrieval, Cambridge university press, 85–108
124. Marlow C, Naaman M, Boyd D, Davis M (2006) Ht06, tagging paper, taxonomy, Flickr, academic article, to read. Proc Conf Hypertext Hypermedia, 31–40
125. Marques O, Lux M (2008) An exploratory study on joint analysis of visual classification in narrow domains and the discriminative power of tags. Proc ACM Workshop Multimed Semant, 40–47
126. Amazon mechanical turk. `http://www.mturk.com/mturk`
127. Miller G (1983) Informavores. In: The study of information: Interdiscipiinary messages, Wiley-Interscience, 111–113
128. Miller H, Han J (2001) Geographic data mining and knowledge discovery, Taylor & Francis
129. Moxley E, Kleban J, Manjunath BS (2008) Spirittagger: A geo-aware tag suggestion tool mined from Flickr. Proc ACM Multimed Inf Ret, 24–30
130. Naaman M, Nair R (2008) Zonetag's collaborative tag suggestions: What is this person doing in my phone?. IEEE Multimed 15(3), 34–40
131. Naaman M, Paepcke A, Garcia-Molina H (2003) From where to what: Metadata sharing for digital photographs with geographic coordinates. On The Move to Meaningful Internet Syst: CoopIS, DOA, and ODBASE, 196–217
132. Naaman M, Yeh R, Garcia-Molina H, Paepcke A (2005) Leveraging context to resolve identity in photo albums. Proc Jt Conf Digit Libr, 178–187
133. Navigli R (2009) Word sense disambiguation: A survey. ACM Comput Surv 41(2), 1–69
134. Negoescu RA, Gatica-Perez D (2008) Analyzing Flickr groups. Proc Content-based Image & Video Ret, 417–426
135. Ng R, Han J (1994) Efficient and effective clustering methods for spatial data mining. Proc Very Large Databases, 144–155
136. Nisbett RE, Peng K, Choi I, Norenzayan A (2001) Culture and systems of thought: Holistic versus analytic cognition. Psychol Rev 108(2), 291–310
137. Noll M, Au Yeung C, Gibbins N, Meinel C, Shadbolt N (2009) Telling experts from spammers: expertise ranking in folksonomies. Proc Res Dev Inf Ret, 612–619
138. Nov O, Naaman M, Ye C (2008) What drives content tagging: the case of photos on Flickr. Proc Hum Factors Comput Syst, 1097–1100
139. Nov O, Ye C (2010) Why do people tag? motivations for photo tagging. Commun. ACM 53(7), 128–131
140. Nowak S, Rüger S (2010) How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. Proc ACM Multimed Inf Ret, 557–566
141. Obrador P, Anguera X, deOliveira R, Oliver N (2009) The role of tags and image aesthetics in social image search. Proc SIGMM Workshop Social Media, 65–72
142. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Ret 2(1-2), 1–135
143. Pedersen T, Patwardhan S, Michelizzi J (2004) Wordnet::Similarity - Measuring the Relatedness of Concepts. Demo Papers North Am Chap Assoc for Comput Linguist - Hum Lang Tecnol, 38–41
144. Photography community, including forums, reviews, and galleries from photonet. `http://photo.net`
145. Pirolli P (2007) Information foraging theory: Adaptive interaction with information. Oxford University Press
146. Pirolli P (2009) An elementary social information foraging model. Proc Hum Factors Comput Syst, 605–614
147. Ramakrishnan G, Chitrapura KP, Krishnapuram R, Bhattacharyya P (2005) A model for handling approximate, noisy or incomplete labeling in text classification. Proc Mach Learn, 681–688

148. Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from Flickr tags. Proc Res Dev Inf Ret, 103–110
149. Rebbapragada U, Brodley C (2007) Class noise mitigation through instance weighting. Proc Eur Conf Mach Learn, 708–715
150. Robertson S, Vojnovic M, Weber I (2009) Rethinking the ESP game. Proc Ext Abstr Hum Factors Comput Syst, 3937–3942
151. Roddick J, Spiliopoulou M (2002) A survey of temporal knowledge discovery paradigms and methods. IEEE Trans Knowl and Data Eng 14(4), 750–767
152. Roddick J, Spiliopoulou M, Lister D, Ceglar A (2008) Higher order mining. SIGKDD Explor Newsl 10(1), 5–17
153. Rui Y, Huang TS, Ortega M, Mehrotra S (1998) Relevance feedback: a power tool for interactive content-based image retrieval. Proc Circuits Syst Video Tech 8(5), 644–655
154. Russell B, Torralba A, Murphy K, Freeman W (2008) LabelMe: A database and web-based tool for image annotation. Int J Comput Vis 77(1), 157–173
155. Salton G, Buckley C (1987) Term weighting approaches in automatic text retrieval. Tech rep, Cornell University
156. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun. ACM 18(11), 613–620
157. Santini S, Gupta A, Jain R (2001) Emergent semantics through interaction in image databases. IEEE Trans Knowl Data Eng 13(3), 337–351
158. Sarwar B, Karypis G, Konstan J, Reidl J (2001) Item-based collaborative filtering recommendation algorithms. Proc World Wide Web, 285–295
159. Sawant N, Datta R, Li J, Wang JZ: Quest for relevant tags using local interaction networks and visual content. Proc ACM Multimed Inf Ret, 231–240
160. Sen S, Harper F,M, LaPitz A, Riedl J (2007) The quest for quality tags. Proc Int ACM Support Group Work, 361–370
161. Sen S, Lam S, Rashid A, Cosley D, Frankowski D, Osterhouse J, Harper F, Riedl J (2006) Tagging, communities, vocabulary, evolution. Proc Comput Support Co-op Work, 181–190
162. Sen S, Vig J, Riedl J (2009) Learn to recognize valuable tags. Proc Intell User Interfaces, 87–96
163. Seneviratne L, Izquierdo E (2010) An interactive framework for image annotation through gaming. Proc ACM Multimed Inf Ret, 517–526
164. Serdyukov P, Murdock V, van Zwol R (2009) Placing Flickr photos on a map. Proc Res Dev Inf Ret, 484–491
165. Shepitsen A, Gemmell J, Mobasher B, Burke R (2008) Personalized recommendation in social tagging systems using hierarchical clustering. Proc ACM Recomm Sys, 259–266
166. Sigurbjörnsson B, van Zwol R (2008) Flickr tag recommendation based on collective knowledge. Proc World Wide Web, 327–336
167. Sinha P, Jain R (2008) Classification and annotation of digital photos using optical context data. Proc Content-based Image & Video Ret, 309–318
168. Sivic J, Zitnick C, Szeliski R (2006) Finding people in repeated shots of the same scene. Proc Br Mach Vis Conf, 909–918
169. Smeulders A, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pat Anal Mach Intell 22(12), 1349–1380
170. Smith M, Kollock P (1999) In: Communities in cyberspace, chap 5–6, Routledge 107–165
171. Snavely N, Seitz S, Szeliski R (2006) Photo tourism: Exploring photo collections in 3d. ACM Trans Graph
172. Solli M, Lenz R (2010) Emotion related structures in large image databases. Proc ACM Image and Video Ret, 398–405
173. Song Y, Leung T (2006) Context-aided human recognition: Clustering. Proc Eur Conf Comput Vis, 382–395
174. Sorokin A, Forsyth D (2008) Utility data annotation with amazon mechanical turk. Comput Vis Pattern Recognit Workshops
175. Suchanek F, Vojnovic M, Gunawardena D (2008) Social tags: Meaning and suggestions. Proc ACM Inf Knowl Manag, 223–232
176. Suh B, Bederson B (2007) Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition. Interact Comput 19(4), 524–544
177. Sun A, Bhowmick S (2009) Image tag clarity: in search of visual-representative tags for social images. Proc SIGMM Workshop Social Media, 19–26

178. Surowiecki J (2004) The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, Economies, Societies and Nations. Doubleday
179. Sylvan E (2010) Predicting influence in an online community of creators. Proc Hum Factors Comput Syst, 1913–1916
180. Tang J, Yan S, Hong R, Qi GJ, Chua TS (2009) Inferring semantic concepts from community-contributed images and noisy tags. Proc ACM Multimed, 223–232
181. Truran M, Goulding J, Ashman H (2005) Co-active intelligence for image retrieval. Proc ACM Multimed, 547–550
182. Tuytelaars T, Mikolajczyk K (2008) Local invariant feature detectors: a survey. Found Trends Comput Graph Vis 3(3), 177–280
183. Verbeek J, Guillaumin M, Mensink T, Schmid C (2010) Image annotation with tagprop on the MIRFlickr set. Proc ACM Multimed Inf Ret, 537–546
184. Vlachos M, Meek C, Vagena Z, Gunopulos D (2004) Identifying similarities, periodicities and bursts for online search queries. Proc ACM SIGMOD Int Conf Manag Data, 131–142
185. Wang C, Jing F, Zhang L, Zhang HJ (2006) Image annotation refinement using random walk with restarts. Proc ACM Multimed, 647–650
186. Wang C, Jing F, Zhang L, Zhang HJ (2007) Content-based image annotation refinement. Proc Comput Vis Pattern Recognit, 1–8
187. Wang M, Yang K, Hua XS, Zhang HJ (2009) Vis tag dictionary: interpreting tags with visual words. Proc Workshop Web-scale Multimed Corpus, 1–8
188. Wang S, Jing F, He J, Du Q, Zhang L (2007) Igroup: Presenting web image search results in semantic clusters. Proc Hum Factors Comput Syst, 587–596
189. Wang X, Zhang L, Jing F, Ma WY (2006) Annosearch: Image auto-annotation by search. Proc Comput Vis Pattern Recognit, 1483–1490
190. Weinberger K, Slaney M, Van Zwol R (2008) Resolving tag ambiguity. Proc ACM Multimed, 111–120
191. Wordnet. `http://wordnet.princeton.edu`
192. Wu L, Hua XS, Yu N, Ma WY, Li S (2008) Flickr distance. Proc ACM Multimed, 31–40
193. Wu L, Yang L, Yu N, Hua XS (2009) Learn to tag. Proc World Wide Web, 361–370
194. Xue GR, Zeng HJ, Chen Z, Yu Y, Ma WY, Xi W, Fan W (2004) Optimizing web search using web click-through data. Proc ACM Inf Knowl Manag, 118–126
195. Yacoob Y, Davis L (2006) Detection and analysis of hair. IEEE Trans Pat Anal Mach Intell 28(7), 1164–1169
196. Yahoo! Flickr. `http://www.flickr.com`
197. Yan R, Natsev A, Campbell M (2009) Hybrid tagging and browsing approaches for efficient manual image annotation. IEEE Multimed 16(2), 26–41
198. Yang Q, Chen X, Wang G (2008) Web 2.0 dictionary. Proc Content-based Image and Video Ret, 591–600
199. Yang Q, Jian B, Chen X (2010) Tag dictionary and its applications. Proc ACM Multimed Inf Ret, 397–400
200. Yang Y, Pierce T, Carbonell J (1998) A study of retrospective and on-line event detection. Proc Int ACM SIGIR Conf Res and Dev in Inf Ret, 28–36
201. Yao B, Yang X, Zhu SC (2007) Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks. Proc Energy Minimization Methods Comput Vis Pattern Recognit, 169–183
202. Yuan J, Luo J, Kautz H, Wu Y (2008) Mining gps traces and visual words for event classification. Proc ACM Multimed Inf Ret, 2–9
203. Zeng HJ, He QC, Chen Z, Ma WY, Ma J (2004) Learn to cluster web search results. Proc Res Dev Inf Ret, 210–217
204. Zhang L, Hu Y, Li M, Ma W, Zhang H (2004) Efficient propagation for face annotation in family albums. Proc ACM Multimed, 716–723
205. Zhang S, Farooq U, Carroll JM (2009) Enhancing information scent: Identifying and recommending quality tags. Proc ACM Support Group Work, 1–10
206. Zhao M, Liu S (2006) Automatic person annotation of family photo album. Proc Image & Video Ret, 163–172
207. Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: A literature survey. ACM Comput Surv 35(4), 399–458
208. Zunjarwad A, Sundaram H, Xie L (2007) Contextual wisdom: social relations and correlations for multimedia event annotation. Proc ACM Multimed, 615–624