Social Media Driven Concept Detection

Tim Althoff

TU Kaiserslautern
timalthoff@web.de

Abstract. Concept detection aims at linking video scenes with semantic concepts (like beach, soccer, or airplanes-flying) appearing in them for the purpose of future retrieval. However, current systems are strongly limited as they require annotated training examples for each concept to be learned. Social inputs are abundantly available and may represent, with their added context, a substitute for expert annotations. This paper summarizes current trends in designing algorithms to fuse visual features with noisy social labels and behavioral signals as well as challenges and opportunities in this area. A particular focus lies on a recent approach to use comment-based local interaction networks for concept detection in images as presented in [5].

Keywords: social media, web 2.0, concept detection, tag recommendation, image annotation, collaborative annotation

1 Introduction

Over the last few years the amounts of digital data has been growing exponentially [2]. The main reason is the vast increase of sharing image and video data over the Internet. YouTube[1] as the mostpopular online video platform where users can upload, share, and watch video clips exceeds more than 3 billion views a day. This is more than double the prime-time audience of all three major U.S. broadcast networks combined. Additionally, 48 hours of video are uploaded to YouTube every minute. In other words, more video is uploaded to YouTube in 30 days than all three major US networks created in 60 years [9]. Furthermore, the combination of better search and discovery (in addition to more content) has driven the daily time each user spends on YouTube on average up 55% in the year 2009 [11]. One of the most popular image hosting websites is Flickr[2], which hosts more than 4 billion images. The social networking website Facebook[3] even experiences around 2.5 billion image uploads to the site each month [1].

Summarizing, this large amount of data offered by online portals like YouTube, Flickr, or Facebook is mainly generated by users sharing their photos and videos. It has become easy to acquire huge volumes of digital data because digital acquisition devices (still image and video cameras) are already wide-spread. Additionally, more visual content is produced due to commercial efforts of companies

[1] http://www.youtube.com
[2] http://www.flickr.com
[3] http://www.facebook.com

like Google indexing the world's documents, maps, or cityscapes, and large-scale digitization efforts of broadcasting networks [10].

Multimedia content becoming a source of information and entertainment to millions of users, and video databases growing at enormous rates, create the need for effective and efficient multimedia retrieval facilities. Especially, users with specific retrieval needs targeting at retrieval of specific video scenes (e.g. a lawyer evaluating copyright infringement) are hardly served by present-day video retrieval applications [8].

1.1 Labeling Multimedia Content

As humans perceive video as a complex interplay of cognitive concepts, the common goal in video retrieval is to provide access at the semantic level. In principle, this can be achieved by labeling all combinations of people, objects, settings, and events appearing in the audiovisual content [8]. Labeling video content is regarded as a challenge of our time as humans use approximately half of their cognitive capacity to achieve such tasks [4]. There are two types of semantic labeling solutions: Labels are either manually assigned by humans after audiovisual inspection, or assigned automatically by machines.

Manual labeling of (broadcast) video has traditionally been the realm of professionals [8]. In cultural heritage institutions, for example, library experts label archival videos for future disclosure using controlled vocabularies. Because expert labeling is tedious and costly it typically results in a brief description of a complete video only. In contrast to expert labor, social tagging refers to a recent trend to let amateur consumers label (mostly personal) visual content on web sites like YouTube, Flickr, and Facebook. These labels are not meant to meet professional standards and are known to be ambiguous, overly personalized, and limited [8]. In this paper, Snoek and Worring [8] come to the conclusion that manual labeling, whether by experts or amateurs, is always geared toward one specific type of use and, therefore, inadequate to cater for alternative video retrieval needs (especially when scene-based).

Machine-driven labeling aims to derive meaningful descriptors from video data. Often, the labels are based on the filename, surrounding text, social tags, closed captions, or a speech transcript. While text-based video search has proven itself effective for scene-retrieval from (English) broadcast news and interviews, it results in disappointing retrieval performance when the audiovisual content is neither mentioned, nor properly reflected in the associated text. Therefore, the available information of the visual content should be included in the labeling process.

A common denominator of content-based multimedia retrieval systems is their dependence on low-level visual features such as color, texture or shape. However, users often do not understand similarity expressed in low-level visual features as they expect semantic similarity [8]. The source of this problem lies in the *semantic gap*: "The lack of coincidence between the information that one can extract from the data and the interpretation of the same data for a user in a given situation [7]." This gap exists for various reasons. One reason

is that different users interpret the same video differently, including subjective interpretations related to feelings or emotions. However, also for objective interpretations, developing automatic methods is still difficult. Mainly, this is due to the large variations of appearance of visual data corresponding to one semantic concept (e.g. different models, shapes, and colors of a windmill). As these causes are inherent to the problem, it must be the aim of multimedia retrieval to bridge the semantic gap [8].

1.2 Concept Detection

Several search strategies for content-based multimedia retrieval systems have been proposed to support the user. One example is *query-by-image*, where the user queries an archive based on an example picture to retrieve visually similar content. Another is *query-by-text*, where the user enters a few keywords and retrieves content that is linked with these terms. The latter approach is considered standard practice but requires an indexing that links the images and videos in a database with descriptive keywords (or tags).

In this context, *concept detection* aims at a machine indexing by automatically linking video scenes with semantic concepts appearing in them [10]. Here, a semantic concept is defined as "an objective linguistic description of an observable entity" [8]. This task is also referred to as automatic tagging, image/video annotation, or high-level feature extraction in the literature [10].

1.3 Limitations

Concept detection systems are usually based on supervised machine learning techniques, which require training examples for any concept to be learned. Since target concepts can be visually complex, each one might require hundreds of sample views. So far, this problem has been overcome to some extent by acquiring ground truth labels in joint efforts of the research community [10]. However, this time-consuming manual labeling restricts concept detection in a number of ways:

1. It limits the number of concepts that can be learned. Hauptmann et al. give an outlook on what numbers of concepts might ultimately be required for practical high-quality video search. It lies in the range of 3.000 - 5.000 concepts (and has been restricted to the domain of news video) [3]. Current sizes of current detector vocabularies are a magnitude below these quantities [10]. Also, content providers like YouTube count 48 hours of uploaded video content per minute [9]. Obviously, for these amounts of data manual annotation is infeasible.
2. It has been pointed out that detectors overfit to small manually acquired training sets and generalize poorly [10].
3. Keeping track of dynamic changes of users' information needs is infeasible as new concepts of interest emerge (such as "Barack Obama" or "E. coli") [10].

1.4 User-annotated Content as Training Data

To reduce the manual annotation effort associated with concept learning, web video has been proposed as an alternative source of training data [10]. From web video portals like YouTube, large quantities of video content can be obtained automatically together with descriptive user-generated tags. By employing these tags as class labels, web data can complement manually annotated training sets or even substitute them completely, such that a concept learning free of manual supervision is performed. This offers vital advantages in terms of scalability (more concepts can be learned) and flexibility (adaptation to changing and newly emerging concepts can take place) [10]. Furthermore, it has been recognized that valuable information which could be used to improve concept detection is abundantly available in the social web [6]. For example, a user's tagging preferences can be approximated by the collective tagging behavior of the user's social network [5].

However, such data is considered *weakly labeled* which can only substitute expert annotation when certain challenges are addressed (see Section 2.3).

1.5 Outline

The rest of this paper explores in which ways social media could be utilized for concept detection. The second section describes what can be learned from social data, which sources there are, and which challenges are imposed by using social data. In the third section, a recent approach is presented that uses the structure of social interaction and the collective tagging behavior to extend content-based annotations to arbitrary folksonomic settings [5]. The paper is concluded in Section 4 with a summary and discussion of the results.

2 Use of Social Media for Concept Detection

The recent phenomenon of social networking and collaborative media annotation in the social web has induced a paradigm shift in the semantic understanding of images [6]. Social actions and annotations (tags, comments, ratings) from online media sharing websites represent a strong substitute for expert annotations as they may bring additional context. In this section, novel approaches are summarized that fuse visual features with noisy social labels and behavioral signals based on a survey by Sawant et al. [6].

2.1 What Can We Learn from Social Data?

Image semantics correspond to 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. Therefore, image understanding does not only relate to research on object detection and scene interpretation but also to capturing abstract notions of events,

locations, and personalized references that situate images beyond the realm of visual features [6]. Recent approaches in concept detection utilize large image collections on the web or crowd-sourcing options to semi-automate training data acquisition. Sawant et al. divide the applications of semantics extraction based on voluntary actions and annotations by millions of web users into four categories [6]:

1. What does the picture portray? or **Content semantics**: Content semantics is the goal of mainstream research in visual concept detection. Applications include tag relevance estimation, concept modeling, image annotation, and computational aesthetics modeling.
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.

2.2 Social Sources of Image Labels

Tasks like image annotation and evaluation can be distributed to Internet users using crowd-sourcing techniques like LabelMe[4] and Amazon mechanical turk[5]. This helps to reduce the burden on experts without significantly sacrificing the quality of annotations. Another source is collaborative games, often referred to as *Games with a Purpose* [6], a channel for human computing through which players contribute perceptual and cognitive information about multimedia objects. This paper does not focus on the latter but instead analyzes the characteristic setting in which *social media sharing* helps generate image labels and other metadata.

Basic social networking mechanisms allow users to form online contacts, join special interest groups, upload and share documents (textual and multimedia). Furthermore, users can also contribute annotations (tags), comments and ratings. These social inputs have become a rapidly-growing source of image descriptions. 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. For this reason, the organization imposed by tags is popularly referred to as *folksonomy* (folk + taxonomy) [6].

[4] LabelMe: the open annotation tool. http://labelme.csail.mit.edu
[5] http://www.mturk.com/mturk

The benefits from using words from personal context for filing purposes are two-fold: first, they help to retrieve particular files from a large personal file collection (recallability), and second, given a file's keywords, the user could be easily reminded of the file creation context (recognition). With the advent of social media sharing, files are tagged for the benefit of self as well as others. Additionally, other users may tag one's personal resources (within the limit of user-specified permissions). Accordingly, the motivations behind tagging evolve beyond personal benefits to accommodate for social influences: future retrieval, contribution and sharing, attention seeking, play and competition, opinion expression, and self presentation [6].

2.3 Folksonomic Challenges

Social annotations are different from expert annotations in the sense that they are contributed from personal, often unknown motivations without a specific computational task in mind. Furthermore, the quality of tags is affected by personal tendencies and community influences. Before social annotations can be suitably utilized, some folksonomic challenges need to be addressed as summarized by Sawant et al. [6]:

- **Motivations**: Motivations directly influence the suitability of tags for scientific purposes. Tags that are contributed for the purpose of future retrieval and contribution, especially for external audiences, are likely to be visually more relevant compared to tags used for personal references.
- **Cultural influences**: Perception and cognition is guided by cultural differences, e.g. whereas Westerners focus more on foreground objects, Easterners have a more holistic view of viewing images early on.
- **Vocabulary problem**: Different people choose varying words when spontaneously describing the same content resulting in a low probability of two users using the same term. This is known as the *vocabulary problem*, 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).
- **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 (e.g. "me" or "d20").
- **Semantic loss**: Annotators in folksonomies are not required to provide all relevant tags with an image, leading to semantic loss in the textual description. The batch-tag option provided by most photo sharing sites allowing users to annotate an entire collection of photos with a set of common tags adds to this problem. Such tags are potentially useful as they provide a broad personal context, but they cannot be used to identify image-level differences, thus leading to semantic loss. Consequently, the absence of a tag from an image description cannot be used to confirm the absence of the concept in that image.

Due to these challenges, collaboratively labeled data does not directly substitute for expert annotations in the identification of visual semantics. In fact, studies have shown that nearly half of the used tags on the web are irrelevant for general audience and have to be pruned to effectively harness images [6].

2.4 Representation of Folksonomy

Folksonomy is usually viewed as a ternary relationship between users, tags, and resources. Sawant et al. describe 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" [6]. Additionally, users may be connected to other users through social relationships. Folksonomy is generally represented as a tripartite hypergraph (see Figure 1) and its elements are represented using the vector space model. This tripartite folksonomy graph can also be modeled as a three-dimensional association matrix. Then, bipartite representations (e.g. a tag-image matrix) can be obtained by aggregation over one of the folksonomic dimensions of the matrix. These bipartite representations can be further used to extract distributional similarity between different elements such as user-user, image-image, and tag-tag (see [6] for more detail).

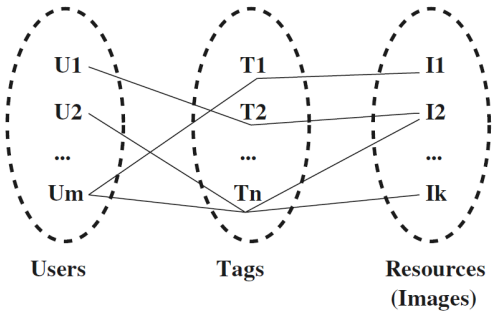

Fig. 1. Tripartite hypergraph representing folksonomy as a ternary relationship between users, tags, and resources (images). In addition, users may be connected to other users through social relationships. Figure from [6].

2.5 User Features

Apart from visual image features and textual features known from information retrieval, user features play a major role in the context of social images. An estimation of the idiosyncrasies helps assess the quality level of a user's annotations. Descriptive user features include expertise, reputation and reliability [6]:

- **Expertise**: An expert is provider of high-quality annotated resources. Topic experts can be identified by a substantial contribution of relevantly tagged resources or by a membership to special interest groups related to that topic.
- **Reputation**: While expertise is a topic-specific feature, reputation is a more general property that assimilates overall activities of networked users into a social order. 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.

Another contributing factor is the prestige of special interest groups in which the photo appears. Furthermore, tags, comments, and views by high-ranked users are considered more useful and can be employed in determining image interestingness.
- **Reliability**: A tag assignment is considered reliable if similar associations are consistently observed over a large user collection.
- **Other network measures**: A user and his social network can also be represented using traditional network measures such as characteristic path length, clustering coefficients, cliquishness, and connectivity [6].

3 Utilizing Photo Sharing Platforms for Concept Detection

In this section, an approach by Sawant et al. to use the social network and photo sharing platform Flickr for concept detection in images is summarized [5].

The success of text-based retrieval on platforms like Flickr depends on the individual choices to annotate images, and therefore, is affected by the personal tendencies, community influences, and generally folksonomic issues as presented in Section 2.3. To deal with these challenges, multimedia researchers have been concentrating on automatically annotating images using techniques such as visual content analysis, user personalization modeling, collaborative filtering, or a combination thereof [5].

Pure visual content based methods involve learning the association between visual features and words to automatically generate textual annotations for images. These methods are limited to a relatively small number of visual concepts and do not scale to the millions of folksonomic tags limited by high training efforts to manually create high-quality training sets. Alternatively, collaborative filtering-based methods can handle very large folksonomic tag set, but typically require at least one initial tag per image. Obviously, for completely untagged images, this strategy cannot produce annotations. In addition, when computing related tags using a global co-occurrence, interesting and locally contextual tags may be suppressed. Lastly, personalization-based methods that model the user's personal tagging preferences by observing the tagging history, can produce locally relevant tags. However, for users without sufficient tagging history or having many idiosyncrasies, personalization may not be very helpful [5].

In contrast to these limited approaches, Sawant et al. propose an image tag recommendation strategy that is fully automatic, incorporates visual content and yet scales to large folksonomic vocabularies. They address the problem of insufficient knowledge about the user's personal tagging preferences by approximating it by the tagging behavior of other users with whom the user has interacted. This set of other users is referred to as the *local interaction network* (LIN). They believe that the users belonging to a social network may share similar interests, vocabulary and behavior. Thus, it should be possible to, at least partially, characterize a user by analyzing the collective photo tagging behavior of her social network. It is argued that this collective vocabulary is much less susceptible to

personal idiosyncrasies as noisy tag applications can be suppressed by trends observed over multiple users. Furthermore, while the context in which tags are applied and local semantics may get obscured in a global analysis, local networks may still be able to capture these [5].

3.1 Local Interaction Networks

In general, a social network consists of a large number of user pair connections (u_i, u_j), whereby a user u_i and a user u_j are connected at some level. More explicitly, this connection can be a pro-actively specified "friendship" between u_i and u_j, or representing the co-membership of both users in some special interest groups. Other modes of interaction include photo or profile views, action of liking/favoriting a photo, or tagging collaboration. However, the authors opt for a local comment-based interaction network of users, where a connection between users u_i and u_j is created when either u_i comments on the photos of u_j or vice versa. Connections formed through comments are more easily traceable (as opposed to photo or profile views) and quite abundantly available (as opposed to favoriting or tagging collaboration) [5].

For later analysis and evaluation, Sawant et al. created a dataset using Flickr with 890 local interaction networks with 27,708 total users comprising ∼5.5M images with ∼49.4M tag-to-image assignments (∼1.25M unique tags). The user whose photos are to be tagged is also denoted as the seed user. In an exploratory analysis, they show that a user's local interaction network has the capability to emulate the seed user's preferences at least partially. The mean vocabulary overlap between a seed's vocabulary and that of its network is 51%, indicating that the local interaction networks are significantly related to the seed user [5].

Furthermore, four different strategies of selecting a practical set of tags from the network vocabulary were explored using a signal-to-noise-analogy: ALL (complete network vocabulary), POP (popular network tags), ATLST2 (tags used by at least 2 users), and COMPOSITE (modified ATLST2+POP). Experimental results show that when ATLST2 has less than 3000 tags it outperforms the rest, and COMPOSITE being the best strategy otherwise because it truncates after the first 3000 tags. This indicates that one does not have to consider arbitrarily large networks and that up to 3000 tags are sufficient to get a useful characterization [5].

3.2 Recommendation Framework

Now, the idea is to combine the network vocabulary with the predictions of a visual content-based annotation system. These are produced by ALIPR, a fully automatic and real-time annotation system trained to detect 332 concepts such as man-made, art, sky, water, modern and flower. Then, the notion of inductive transfer or transfer learning is used to extend these limited annotations to the network vocabulary. The goal is to infer the user tags with the help of ALIPR generated tags (using a Naïve Bayes formulation). For example, if ALIPR is able to recognize the concept "dog" based on visual features, but the

photos of the network associate a related concept "puppy" to that function, then the frequent co-occurrence of "dog" in the ALIPR tag set and "puppy" in the network vocabulary indicates that an ALIPR prediction of a tag "dog" should be translated to the tag "puppy" for a better estimate of the actual user tags (for mathematical details see [5]).

3.3 Results and Discussion

The proposed tag recommendation framework, denoted as C+LIN was compared to alternative approaches both quantitatively and qualitatively. For quantitative analysis standard precision and recall measures were used with the tags given by the seed user to his or her images being the ground-truth. The alternative approaches were: the top K predictions of the content (ALIPR) annotation baseline (C), the top-ranked K tags of the local interaction network (LIN), the seed's K most popular tags (S), and the seed's vocabulary and content (C+S) which works like C+LIN where the network vocabulary is substituted by the seed's own vocabulary. The quantitative and qualitative results are depicted in Figure 2 and 3, respectively.

	% Precision		% Recall	
	@5	@10	@5	@10
C	2.59	2.14	2.05	3.19
LIN	10.03	7.13	8.12	11.46
C+LIN	**12.02**	**9.57**	**10.85**	**17.40**
S	32.42	23.59	26.25	36.07
C+S	5.33	4.56	5.43	8.66

Fig. 2. Quantitative Results Summary. The proposed combination C+LIN outperforms the baselines of pure content based annotation (C) and the local network (LIN). Figure from [5].

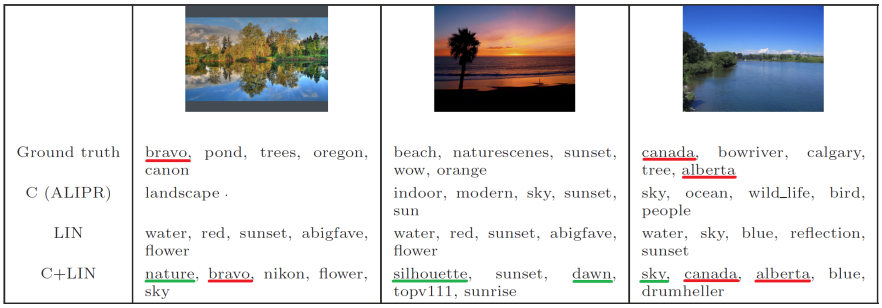

Fig. 3. Qualitative Results. These results show that the C+LIN combination can produce interesting and relevant annotations (green), beyond the tags contained in the ground truth (red). Figure from [5].

The ground truth tags given by the seed user can suffer from folksonomic idiosyncrasies. It may be incomplete (user may choose to not include a relevant tag that would otherwise be found in other similar images) or visually irrelevant (e.g., uncle, zzzzzzz1232zz). Therefore, even some meaningful predictions get penalized. Hence, the results need to be assessed on a qualitative basis, too [5].

Based on quantitative results, the strategy S is the best performing method. This points to the idiosyncratic habit of the users to tag their images similarly, habits which are encouraged by the convenient batch tagging facilities, where users can annotate entire image collections with the same set of tags. Therefore, often in folksonomic settings, a user's previously used tags can be the best determinants of her future uploads. The strategy C+S performs poorly, because either the tagging batch process yields annotations with much diverse (and possibly irrelevant) visual features, which hampers the model learning, or the user's own tag history may simply be insufficient to train a useful personalization model. However, the combination C+LIN outperforms the baselines of pure content based annotation (C) and the local network (LIN) as well as the model trained using the combination of content and the user's own tags (C+S). This demonstrates that when the seed users personal tagging preferences are not known (or when insufficient data is available), a suitable approximation can be made using the tagging behavior of the user's local interaction network. Furthermore, the qualitative results show that the C+LIN combination can produce interesting and relevant annotations, beyond the tags contained in the ground truth [5].

4 Conclusion

In this paper, it was shown that there is a need for video and image retrieval on a semantic level. Concept detection aims at linking video scenes with semantic concepts appearing in them. However, current systems are strongly limited due to the need for large collections of annotated training examples for each concept to be learned. To reduce the burden of manual annotation, web multimedia content has been proposed as an alternative source for training data offering vital advantages in terms of scalability and flexibility. Furthermore, additional context from social actions and annotations (tags, but also comments, and ratings) can represent a strong substitute for expert annotations. However, before they can be suitably utilized, several folksonomic challenges need to be addressed. In the third section, a recent approach was summarized that utilizes the social network Flickr for concept detection on images. The main idea was to approximate a user's personal tagging preferences by the tagging behavior of other users with whom the user has interacted. This so-called local interaction network was built based on comments on Flickr. Their approach is particularly interesting as they extent limited annotations from a traditional concept detection system to large folksonomic vocabularies using the collective tagging behavior of the user's social network. A key result was that when the seed users personal tagging preferences are not known (or when insufficient data is available), a suitable approximation can be made using the tagging behavior of the user's local interaction network. Since the modes of interaction on YouTube are very similar to the ones on Flickr (in particular, users are allowed to comment on other videos) the presented results should easily generalize to the video domain in concept detection.

Using social images for concept detection is a rapidly-growing field. With millions of users, billions of images and associated metadata, the raw inputs

available for the discovery of semantic and structural knowledge are immense [6]. The literature review by Sawant et al. concludes that current research resorts to applications of mature algorithms in the area of data mining and information retrieval. Nevertheless, the new challenges posed by human choices of metadata and subjective interpretation of visual content cannot be met using these areas alone. Studies of human perception, cognition, linguistics, psychology, and social sciences represent the human element in social multimedia and need to be addressed to drive multimedia research from small expert labeled datasets to very large collections of noisy labels. Additionally, techniques of multi-modal data and decision fusion are required to effectively combine heterogeneous information cues from visual, textual, and behavioral data. Lastly, Sawant et al. point at the demand for scalable technology that can efficiently handle increased performance requirements.

In conclusion, the paradigm shift with the introduction of social media sharing has broadened the scope of concept detection beyond visual features. Utilizing the plethora of available information requires interdisciplinary research of image processing, data mining, human computer interaction, and sociology [6].

References

1. EConsultancy: 20+ mind-blowing social media statistics revisited. http://econsultancy.com/blog/5324-20+-mind-blowing-social-media-statistics-revisited (January 2010), retrieved June 16, 2010.
2. Gantz, J., Reinsel, D.: Extracting value from chaos. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (June 2011), retrieved June 30, 2011.
3. Hauptmann, A., Yan, R., Lin, W.: How many high-level concepts will fill the semantic gap in news video retrieval? In: Proceedings of the 6th ACM international conference on Image and video retrieval. pp. 627–634 (2007)
4. Palmer, S.: Vision science: Photons to phenomenology, vol. 1. MIT press Cambridge, MA. (1999)
5. Sawant, N., Datta, R., Li, J., Wang, J.: Quest for relevant tags using local interaction networks and visual content. In: Proceedings of the international conference on Multimedia information retrieval. pp. 231–240 (2010)
6. Sawant, N., Li, J., Wang, J.: Automatic image semantic interpretation using social action and tagging data. Multimedia Tools and Applications pp. 1–34 (2011)
7. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1349–1380 (2000)
8. Snoek, C., Worring, M.: Concept-based video retrieval. Foundations and Trends in Information Retrieval 2(4), 215–322 (2008)
9. The Official Google Blog: Thanks, YouTube community, for two BIG gifts on our sixth birthday! (May 2011), available from http://googleblog.blogspot.com/2011/05/thanks-youtube-community-for-two-big.html, retrieved June 16, 2011
10. Ulges, A.: Visual Concept Learning from User-tagged Web Video. Ph.D. thesis, University of Kaiserslautern, Kaiserslautern, Germany (2009)
11. Website Monitoring: Youtube facts & figures (history & statistics). http://www.website-monitoring.com/blog/2010/05/17/youtube-facts-and-figures-history-statistics/ (May 2010), retrieved June 16, 2010.