Video Annotation Through Search and Graph Reinforcement Mining

Emily Moxley, Tao Mei, *Member, IEEE*, and B. S. Manjunath, *Fellow, IEEE*

***Abstract*—Unlimited vocabulary annotation of multimedia documents remains elusive despite progress solving the problem in the case of a small, fixed lexicon. Taking advantage of the repetitive nature of modern information and online media databases with independent annotation instances, we present an approach to automatically annotate multimedia documents that uses mining techniques to discover new annotations from similar documents and to filter existing incorrect annotations. The annotation set is not limited to words that have training data or for which models have been created. It is limited only by the words in the collective annotation vocabulary of all the database documents. A graph reinforcement method driven by a particular modality (e.g., visual) is used to determine the contribution of a similar document to the annotation target. The graph supplies possible annotations of a different modality (e.g., text) that can be mined for annotations of the target. Experiments are performed using videos crawled from YouTube. A customized precision-recall metric shows that the annotations obtained using the proposed method are superior to those originally existing for the document. These extended, filtered tags are also superior to a state-of-the-art semi-supervised technique for graph reinforcement learning on the initial user-supplied annotations.**

***Index Terms*—Data mining, graph theory, video annotation, video content analysis.**

I. Introduction

THE current multimedia boom demands effective, yet quickly adaptive, organization for efficient user retrieval and browsing. Tagging, or annotation, enables text-based querying which is the most common way to search for multimedia documents using current technology. Tagging drives the search process in all online media repositories like Flickr [5] and YouTube [33]. Annotation also summarizes content and enables surfing. However, the automatic detection of events or objects using computer vision techniques and low-level features remains an open problem. The tagging therefore is done on sites like YouTube and Flickr by users who supply keyword annotations for their files. This results in annotations dependent on user interpretation as well as the current vocabulary for and understanding of a subject. This vocabulary may change over time, as, for instance, in the case where "swk" became a nickname for the "star wars kid" whose video propelled him to Internet stardom. Therefore, web video repositories such as YouTube have user-generated annotations that have not been quality-controlled. These annotations are typically incomplete and noisy since they result from one-time annotations from single users. The annotation set typically contains incorrect keywords and is missing quite a few relevant ones. An automated method that provides both high recall and precision of tags is needed before these large databases can be effectively searched and viewed.

Annotation has largely been done in the research community by building a model for each keyword. Each model is then applied to every document to test for each annotation. However, this process largely ignores the work that has already been done by humans for comparable documents. The manual annotations are particularly useful when a similar or identical document has already been annotated. Since online media networks are driven by large communities with similar interests, the content tends to grow virally and a video can be easily tagged by collecting and filtering the tags of similar videos. Motivation for an approach that leverages this information is given by the YouTube example shown in Fig. 1 which demonstrates the redundant quality of many repositories. In sites such as YouTube, each video is tagged independently, but using the overlap we can learn better annotations that improve retrieval and browsing of the site. This paper motivates a new form of video annotation afforded by an online media community where similar variations of the same video frequently exist and user annotations are noisy and incomplete.

Using knowledge that modern information and documents have a repetitive nature as indicated both by the literature [2], [30] as well as the empirical example in Fig. 1, and inspired by the graph fusion technique used for annotation in [28], we propose to leverage graph theory techniques on information from individual community users for *annotation with an unlimited vocabulary*. This paper tackles the annotation problem from a novel perspective, i.e., a *data mining* approach that focuses on mining frequent terms out of documents relevant to a particular target. By using the annotations provided from independent tagging instances, it furthermore addresses an opportunity uniquely afforded by online communities and one that has not been explicitly tested. It is a new or complementary direction for computer vision-based automatic annotation, because it leverages the large-scale data on the Internet to mine repetitive patterns and find new concepts which cannot be discovered using existing approaches. This technique amounts to mining correct annotations out of a collection of possible annotations found for similar documents.

A diagram of the proposed annotation system aggregating user information is shown in Fig. 2. First, a query is issued to

Manuscript received June 01, 2008; revised March 13, 2009. First published January 26, 2010; current version published March 17, 2010. This work was performed when the first author visited Microsoft Research Asia and was supported in part by NSF IGERT Grant #DGE-0221713. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lexing Xie.

E. Moxley and B. S. Manjunath are with the University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: emoxley@ece.ucsb.edu; manj@ece.ucsb.edu).

T. Mei is with Microsoft Research Asia, Beijing 100190, China (e-mail: tmei@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2041101

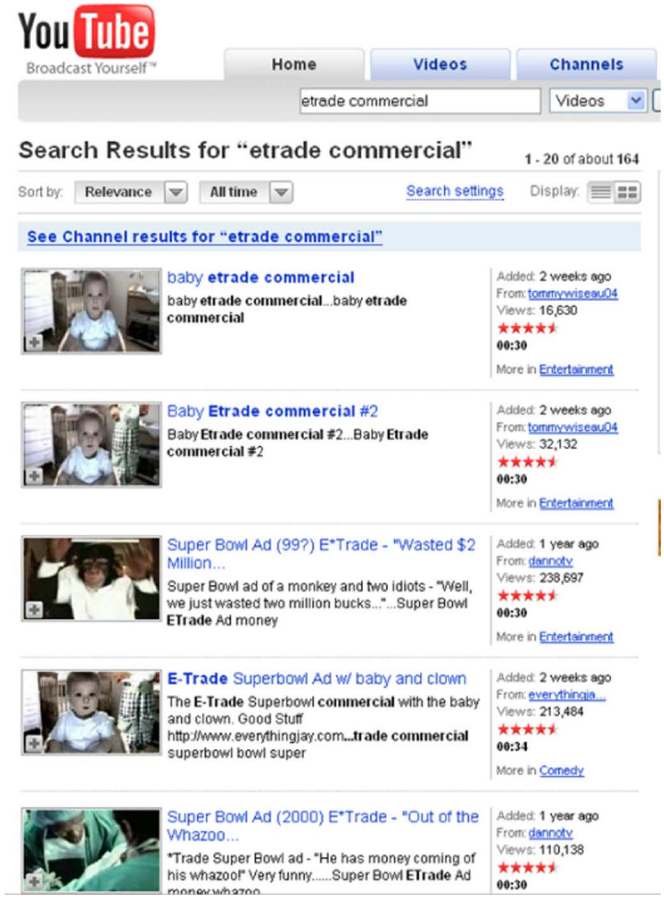


Fig. 1. Top results for "etrade commercial" on YouTube. Duplicates and similar videos can be found in top results, with repetition of tags and annotations. In this example, the "superbowl" tag appears for the third and fourth return, a term often used in describing the video motivated by the debut forum of the clip. Extending this tag to the first two returns may be helpful in retrieval or description of the video.

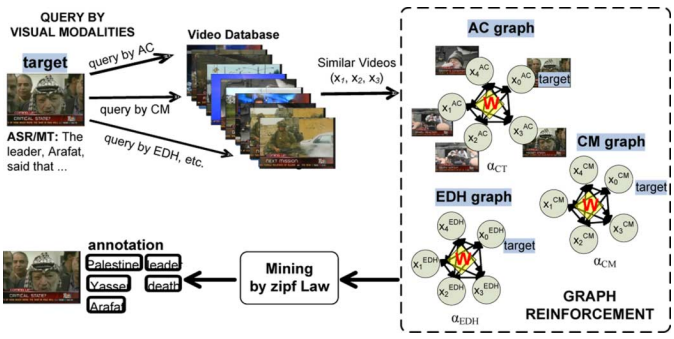


Fig. 2. System diagram. The system first collects similar videos and generates small graphs for different visual features. For example the visual features may be AutoCorrelogram (AC), Color Moment (CM), Edge Distribution Histogram (EDH), and so on. A stable graph is found by iterative diffusion using similarity matrix $\mathbf{W}$ over the connected nodes and then annotations are mined from the weighted $[\alpha_{\mathrm{AC}}, \alpha_{\mathrm{CM}}, \alpha_{\mathrm{EDH}}]$ annotations of each stable graph. A zipf-based cutoff is used to determine the relevant annotations for a particular video.

the database crawled from YouTube to search with a modality independent of the annotation type. For instance, if text annotations are sought, then the search is done using a non-textual feature such as a visual feature, e.g., Color Moment (CM), Edge Distribution Histogram (EDH), or AutoCorrelogram (AC). Then a stable graph is found that emphasizes frequent labels in the result set of similar videos, $[x_1, \ldots, x_N]$. The stability induces a smoothness on the document annotations so that similar documents have similar annotations; that is, annotations roughly correlate with the distance between documents represented by the similarity matrix $\mathbf{W}$. The stable graph can be directly analyzed for keywords or combined using weights $[\alpha_{\mathrm{CM}}, \ldots, \alpha_{\mathrm{EDH}}]$ with other graphs to annotate the target. This approach addresses the shortcoming of limited vocabulary size in previous methods since the process is not restricted to machine learning of certain annotations or categories. Instead, a general unsupervised process is used where visual features are used to find similar videos, and then analysis of their user tags determines the annotations for the target. The dictionary size is limited only to the union of all annotations used anywhere in the system, rather than by those with sufficient training data to learn a model based on feature primitives. The mining of similar videos improves the completeness and accuracy of the annotation process.

In the remainder of this paper, Section II presents previous work in video and image annotation as well as research on online media sites. Section III explains the proposed stable graph creation using user annotations. Section IV details the subsequent frequent term mining for annotation. Section V gives experimental results, and Section VI concludes the paper.

II. Related Work

This paper addresses the problem of multimedia annotation of images and videos. Simultaneously, its unique focus on online community-based data uses principles from research on collaborative tagging and social media sites, such as YouTube, Flickr, and other multimedia repositories. The research in these areas is explained below in Sections II-A and II-B.

A. Multimedia Annotation

The annotation problem has received significant attention in image and video realms, since annotation helps bridge the semantic gap that results from querying using one mode (e.g., text) for returns of another mode (e.g., images). Multimedia annotation algorithms can be said to be *supervised* or *unsupervised* based on whether it uses known training data. An annotation method can also be described as a *computer vision* approach if it builds word-specific models from low-level visual features, or a *data mining* approach if it mines correlations among annotations or propagates existing information.

Typical supervised methods for image and video annotation developed through the TRECVID collaboration [24] began using supervised learners, specifically support vector machines (SVM), to learn a pre-defined, small set of concepts from labeled training images. Extending from this single SVM for annotation, work performed by Yan *et al.* trains two SVMs using different features, adding the most confident predictions to the labeled training set and then training and classifying using the other SVM iteratively until all targets have been annotated [22].

More recent work extends single annotation SVM models by mining for correlations in the SVM-produced annotation predictions. For example, Natsev *et al.* employ correlation mining

using model vectors created from a low-level visual method [18]. Qi *et al.* also mine annotation correlations, such as "mountain" and "outdoor," as well as cross-modal correlations between visual and textual features [19]. Lavrenko *et al.* employ both vision and mining to construct a joint probability of visual region-based words with text annotations, incorporating co-occurrent visual features and co-occurrent annotations [12], while Yan *et al.* use decision trees to mine correlations in annotating around 30–40 concepts [32]. Tseng *et al.* present a fused model that combines rule-mining in temporal, speech, and visual models with the model in [12] to improve results using on the TRECVID annotation set [25]. Data mining is used in these supervised methods exclusively for correlation discovery within a small annotation set. The largest vocabulary in these works has 120 words.

Recently, techniques combining computer vision and data mining derived from graph theory have attracted extensive attention by mining co-occurrence data revealed by the structure of vision-based graphs. The graph techniques used for annotation are typically considered to be *semi-supervised* learning, a special case of supervised learning where the distribution of unlabeled points is used in the learning process. Chapelle [3] and Zhu [36] provide a survey of semi-supervised learning and associated graph theory. Jing shows how graph learning can be used to rerank image search results [10]. Zhou introduces the smooth manifold ranking theorem on a single graph as a solution to the semi-supervised labeling problem [34]. Tong *et al.* provide a solution for combining two graphs for label propagation [23], and Wang extends this method to an arbitrary number of graphs [28]. These studies use data mining techniques to examine neighborhood properties and guide propagation. As supervised techniques, they are limited to a fixed lexicon; the largest lexicon is limited to 39 words since the initial labeling is performed using a computer vision model for each label.

Attempts using a larger vocabulary in an unsupervised framework have been made. In image annotation, Rui *et al.* present a bipartite graph theory technique that has an unlimited vocabulary [20]. They identify candidate annotations in the text around an image, and then extend these annotations using surrounding text from images simultaneously close in visual and semantic space. Velivelli *et al.* use automatic speech recognition (ASR) results to mine a corpus for annotations [26]. However, in this approach data mining creates a vocabulary for the entire database rather than finds specific annotations for a video. Xing *et al.* focus on topic discovery in a video using multi-wing harmoniums, where each harmonium is derived from a different mode [31]. Xing's work does not explicitly address the accuracy of annotations associated with the general topics they discover, despite an indication that certain topic words can be extracted. The specificity of these topic words is limited, however, in that only general topics with a large number of positive examples are extracted.

It is observed that existing annotation methods usually suffer from two problems: 1) *supervised* approaches do not address tag discovery, since previously unseen annotations lack training data and as a result are limited to a pre-defined concept set; and 2) *unsupervised* approaches such as [26] fail to use mining when annotating a specific target, limiting the mining step to initial dictionary creation before a target is chosen or general topic annotation. This paper proposes an unsupervised video annotation algorithm based on data mining annotations from a large vocabulary.

B. Social Media and Collaborative Tagging

Simultaneously, research into online media communities has motivated an exploration of the effectiveness of user annotation and the learning possible from independent annotations freely offered by community users. Shirky has indicated the importance of the annotation process that can occur in online media websites as it allows for a dynamic, evolving understanding of the world [21]. Kennedy *et al.* provide an example of effective learning from an online media site [11], using community image annotations to create geographic tags (e.g., landmarks and neighborhoods) and temporal tags (e.g., event tags). A moderate body of literature exists identifying statistics, trends, and taxonomies for online media databases. Golder and Huberman provide an overview of collaborative tagging, that is, the user-generated tagging fostered in community media sites [8].

The Flickr case study performed by Golder and Huberman analyzes tag statistics of the photo-sharing site to motivate problems such as combating spam [8], an issue that Cho *et al.* also cite [4]. The Google Image Labeler represents an application that addresses this issue by using repeated tagging instances [9]. It turns robust image annotation into a game in which each party independently labels an image and receives points only when an annotation overlaps with that of another user. Tag collaboration can protect a site against exploitative attacks that occur when a user uploads advertising media and tags the disguised ad with popular search terms to attract hits. The repeated tagging instances afforded by aliased data in online media sites [2] enable tag *filtering* in social media computing to prevent annotation spam.

Simultaneously, Furnas *et al.* emphasize tag *extension* for such sites by relating that a large number of keywords can be generated only by a large group of users [6]. The work in this paper harnesses the large amount of freely contributed overlapping annotation [2] to address the issues of consistency and spam [4], [7], [14] in an automated learning process. This question of effective learning using user-supplied information has not been explicitly tested.

The contributions of this work are two-fold. In comparison to [28] which focuses on multi-graph semi-supervised learning, inspiring the proposed graph learning, this work does not rely on a training set that limits the annotation set to those labels for which there is training data. Rather, it discovers the appropriate annotations in tagging instances of similar documents using a graph technique without annotation modeling from feature primitives. Secondly, the annotation scheme presented in this paper addresses a timely problem put forth by the research community [4], [6], [7], gleaning trustworthy keywords from the tagging synergy of multiple online community users, that has not been previously attempted.

III. Graph Reinforcement on Similar Documents

The graph reinforcement technique represents an inductive learning process that uses the weak predictions afforded by each similar video to create a stronger prediction of appropriate annotations for the set of videos. This graph reinforcement formulation can also be seen as combining a collection of weak clas-

TABLE I
ANALOGY BETWEEN GRAPH THEORY TERMINOLOGY AND COLLABORATIVE TAGGING ALGORITHM, WITH NOTATION

Physical Analogy	YouTube Analogy	Notation
vertex or node	video	x_i
weighted edges	similarity between videos i and j	W_{ij}
initial attribute w at node i	user-supplied annotation w for video	Y_i^w
subsequent attribute w at node i	annotation w updated by neighboring video annotations	f_i^w
attribute w relevance over nodes	annotation w relevance for all videos in graph	$f^w = f^w = \left[f_1^w f_2^w ...\right]^{\mathsf{T}}$
stable attribute over nodes	annotation relevance for stable graph	f^{w*}

sifiers to produce a stronger one. Each video in the graph formulation is an error-prone annotator or a weak classifier. These weak classifiers from the individual documents are then combined using a graph stabilization technique that produces a reliable annotation.

The collaborative tagging algorithm formulated as a graph theory problem creates a graph from the near neighbors of the target. This process is done in our case by querying the YouTube database using a particular visual feature and then forming a graph using the closest videos for the reinforcement step. By allowing correlations in near neighbors to reinforce each other, we are able to extract better annotations for the documents. A term that does not appear in annotations for a document, but appears frequently in a collection of similar documents, may be extended as a "new" annotation for the original document. Additionally, annotation "spam" will be filtered in the instances where an annotation does not also occur in the set of similar documents.

We begin by showing stable graph reinforcement of a single graph. The well-established, state-of-the-art semi-supervised case of graph reinforcement annotation is presented in Sections III-A1 and III-B1, and then our extension to the unsupervised case in Sections III-A2 and III-B2. The flexibility of using more than one graph allows the system to use multiple modalities and distance metrics that can be combined using an adaptive weighting.

A. Single Graph Reinforcement

First let us consider the technique using a single graph. The problem is formulated by a set of vertices (or nodes) and edge weights. Vertices, in our case, are YouTube videos, and the edges represent the similarity between node videos, as shown in Fig. 2. Similar YouTube videos, as found in repeated versions of the original or remixed/edited copies of the original, will have heavier edges. This similarity can be measured using any singular feature distance; it may be affinity in text, visual, audio, or concepts/semantic space, for instance. The graph analogy is summarized in Table I.

Once the vertices have been extracted and edge weights calculated, the problem becomes finding the most stable graph structure. In finding a "stable" structure, the idea is to smooth the attributes (annotations) of each node over the space: the attributes of one node should be similar to those of a node nearby. Before stabilization, this may not be the case as the attributes are the result of a noisy initial state formed by the uploader's annotations. Each node is to influence its neighbors such that the node attributes vary proportionally with the distance in space. Finding the stable graph structure designates stable feature attributes, f_i^{w*}, at each of the vertices, x_i, that vary proportionally with edge weights.

The similarity matrix of the graph, $\mathbf{W}$, has elements W_{ij} that are the edge weights of the fully-connected graph between vertices i and j. Typically this distance is normalized, or shaped, using a radius parameter σ. The elements of $\mathbf{W}$ are given by the standard similarity metric

$$W_{ij} = \begin{cases} \exp\left(-\frac{d(x_i,x_j)}{\sigma}\right), & \text{if } i \neq j \\ 0, & \text{else} \end{cases} \tag{1}$$

where $d(x_i, x_j)$ is some distance function between x_i and x_j. The shaping parameter σ was set to be the standard deviation of a particular feature distance. This affinity metric, proposed in [34], is strictly positive and has the quality that close features receive a high weight.

The problem of defining the most stable graph diverges in the case where the attributes of some nodes are *known*, in which case supervised or semi-supervised learning is preferable, versus the case in nearly all practical annotation instances, where the vocabulary is so extensive that groundtruth for all annotations does not exist. Rather, a few videos have a limited number of annotation predictions that positively identify only a few feature attributes but do not negatively identify any annotations. We will next describe graph reinforcement via semi-supervised and unsupervised learning.

1) Semi-Supervised Graph Reinforcement: Semi-supervised learning typically outperforms the classification of supervised learning methods when only a handful of annotations exist in a large collection of data [3], [35]. In semi-supervised learning, assumptions made on the large set of unlabeled data are built into models to improve performance.

The authors of [34] phrase the annotation problem in terms of a regularization framework. Finding the most stable graph amounts to solving

$$f^{w*} = \arg\min_{f^w} \left\{ \begin{array}{c} \sum_{i,j} W_{ij} \left| \frac{f_i^w}{\sqrt{D_{ii}}} - \frac{f_j^w}{\sqrt{D_{jj}}} \right|^2 + \\ \mu \sum_i \left| f_i^w - Y_i^w \right|^2 \end{array} \right\} \tag{2}$$

where Y_i^w is the initial labeling of node i for annotation w: $+1$ for positive, -1 for negative, and 0 for unlabeled regarding annotation w; $\mathbf{D}$ is the diagonal normalizing matrix given by $D_{ii} = \sum_j W_{ij}$; and f_i^w is the label value, or confidence that a particular label is applicable, at node x_i. The first term in (2) can be considered a "smoothness constraint" implying a cost for labels that change too quickly over the space. The second term is a "sticking constraint" that implies a cost for changing the initial labeling. The closed-form solution to (2) is found to be

$$f^{w*} = \left(\mathbf{I} + \frac{1}{\mu}\mathbf{L}\right)^{-1} \mathbf{Y}^w \tag{3}$$

where $\mathbf{L}$ is the normalized graph Laplacian given by the equation $\mathbf{L} = \mathbf{D}^{-(1/2)}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-(1/2)}$ and $\mathbf{Y}^w = [Y_1^w \quad Y_2^w \quad \ldots]^{\mathrm{T}}$ is the initial confidence or predictions made by a user or a computer vision model of the label w on each node. Alternatively, we can solve the problem sequentially. Applying the update

$$f^w_{\tau=t+1} = \frac{1}{1+\mu}(\mathbf{I} - \mathbf{L})f^w_{\tau=t} + \frac{\mu}{1+\mu}\mathbf{Y}^w \tag{4}$$

iteratively results in convergence at f^{w*} [28]. The reinforcement of the initial labeling $\mathbf{Y}$ is apparent in this equation, and may not be ideal in the case where initial labels are unknown. In the next section, we explore an unsupervised approach that extends from this case.

2) Unsupervised Graph Reinforcement: In the case of unsupervised learning where a training set has not been given, the iterative process described in (4) process reduces to a trivial case. Without initial labeling, $Y_i^w = 0\ \forall i, w$, and (2) reduces to

$$f^{w*} = \arg\min_{f^w}\left\{\sum_{i,j} W_{ij}\left|\frac{f_i^w}{\sqrt{D_{ii}}} - \frac{f_j^w}{\sqrt{D_{jj}}}\right|^2 + \mu\sum_i |f_i^w|^2\right\} \tag{5}$$

The second summation, the "sticking constraint," becomes a constant positive value, and minimizing the expression over f^w results in a trivial solution where $f_i^w = f_j^w$. Furthermore, "supposing" or "imposing" an initial labeling $\mathbf{Y}^w$ results in a repeated reinforcement of noisy, often incorrect, labels which is apparent when considering (4). "Supposing" an initial labeling using the owner annotations results in incorrect reinforcement since the lack of an initial label does not necessarily imply that label is inappropriate, as it may have only been overlooked. Experiments in this study show that supposing either a negative annotation, $Y_i^w = -1$, or an unlabeled value, $Y_i^w = 0$, for missing annotations underperforms a more explicitly unsupervised graph propagation technique.

Instead, a different tack must be taken to find a stable graph over the feature space. The weights in $\mathbf{W}$ are used to update the nodes according to the properties of its neighbors. Consider $\mathbf{W}'$ the row-normalized $\mathbf{W}$, that is $W'_{ij} = (W_{ij})/(\sum_j W_{ij})$. $\mathbf{W}'$ is a non-generative diffusion kernel on a fully-connected graph. By "non-generative," we simply mean that the final weighted number of annotations of the graph at convergence is the same as the initial number of annotations, but has been redistributed. Effectively, the potential annotations of each node diffuse to the other nodes, weighted by affinity between the nodes.

We consider the keyword f_i^w of a particular term w for the video x_i, and update it according to f_j^w, the annotation confidence in video j. Additionally, a term is included such that each step does not completely transfer its current labels to neighbors, but retains them weighted by some factor μ. For each possible term w for the video, the distribution of the term over the graph nodes after one iteration is described by the term frequency vector f^w

$$f^w_{\tau=t+1} = [(1-\mu)\mathbf{W}' + \mu\mathbf{I}]^{\mathrm{T}} f^w_{\tau=t} \tag{6}$$

where $\mathbf{I}$ is the identity matrix and $f^w = [\, f_1^w \quad f_2^w \quad \ldots \quad f_N^w\,]^{\mathrm{T}}$ is the feature vector indicating classification annotation w for each of the N documents.

As the update matrix, $[(1-\mu)\mathbf{W}' + \mu\mathbf{I}]$, is row-normalized, the total number of annotations does not grow but rather redistributes. The ergodic theorem indicates that the process will converge [13]. Intuitively, if a particular keyword is only found for the target, it will be distributed to the nodes and its importance minimized (tag *filtering*). On the other hand, if the other similar documents have an annotation that is not contained in the target, that keyword will be added to the attributes of the target node (tag *extension*). Additionally, the graph will reinforce a particular term in the target that is also present in other documents.

The formulation thus far has been limited to discussion of reinforcement using a single graph. Extension to the multi-graph case, which allows for use of multiple features, modalities, and distance metrics, is discussed in the next section.

B. Multiple Graph Reinforcement

Multi-graph learning has been used for learning situations in which more than one feature, mode, or distance measure can typify similarity. For instance, perhaps the system has a metric for visual similarity and audio similarity, but it is not intuitively obvious how to effectively combine these two disparate metrics. Combination of multiple metrics or features can be done through a weighted combination of the stable graph generated by different cases. In some situations, this amounts to a weighted bagging technique, where each graph represents a classifier and they are combined to create a stronger prediction.

The problem in multi-graph learning becomes solving for α_g, the weighting terms used when combining the individual graphs. In a totally naive case, $[\,\alpha_1, \quad \ldots, \quad \alpha_G\,]$ can be set to equal values, or in a case of perfect knowledge set to 1 for the best graph and 0 for the others. In all other cases, a smoothness measurement can be calculated on each of the graphs to provide an estimate of the quality of the graph. Intuitively, smoother graphs, which show a great degree of intra-similarity in term space for a great degree of visual feature similarity, are better models of the particular node x_i or the particular term w.

1) Semi-Supervised Learning With Multiple Graphs: In terms of Zhou's regularization framework [34], the semi-supervised multi-graph problem is taken to be solving

$$f^{w*} = \arg\min_{f^w}\sum_{g=1}^{G}\sum_{i,j}\left\{\begin{array}{c}\alpha_g W_{g,ij}\left|\frac{f_i^w}{\sqrt{D_{g,ii}}} - \frac{f_j^w}{\sqrt{D_{g,jj}}}\right|^2 \\ +\alpha_g\mu\sum_i |f_i^w - Y_i^w|^2\end{array}\right\}. \tag{7}$$

Wang [28] shows that solving (7) by optimizing over both f and α results in a trivial solution where $\alpha_{g_{\mathrm{best}}} = 1$ for the smoothest graph ($g_{\mathrm{best}} = \arg\min_g\{f^{w\mathrm{T}}\mathbf{L}_g f^w\}$) and $\alpha_g = 0$ otherwise. This weighting amounts to doing single graph reinforcement on only the smoothest graph. Instead, Wang suggests that by relaxing α_g to α_g^r, we can solve for α_g, in the case of fixed f^w, with

$$\alpha_g = \frac{\left(\frac{1}{f^{w\mathrm{T}}\mathbf{L}_g f^w + \mu|f^w - Y|^2}\right)^{\frac{1}{r-1}}}{\sum_{g=1}^{G}\left(\frac{1}{f^{w\mathrm{T}}\mathbf{L}_g f^w + \mu|f^w - Y|^2}\right)^{\frac{1}{r-1}}} \tag{8}$$

and solve for f^w, in the case of fixed α, with

$$f^w = \left(\mathbf{I} + \frac{1}{\mu}\frac{\sum_{g=1}^{G}\alpha_g^r\mathbf{L}_g}{\sum_{g=1}^{G}\alpha_g^r}\right)^{-1}. \tag{9}$$

An iterative EM algorithm that alternates updating α according to (8) and then f^w according to (9) converges to a unique solution [28].

2) Unsupervised Learning for Multiple Graphs: The aforementioned weighting formulation is limited when negative samples do not exist as explained in Section III-A2. Equation (8) requires that close nodes should have similar annotations with term ${f^w}^{\mathrm{T}}\mathbf{L}_g f^w$. However, this relies on a normalization of feature attributes $f_i^w \in [-1, 1]$ so that diverging decisions provide a negative contribution to ${f^w}^{\mathrm{T}}\mathbf{L}_g f^w$. Yet the nature of user annotation is such that we cannot assume an annotation is irrelevant to a video simply because it is omitted from the user-supplied labels. Building on the idea of smoothness formulated in (8), we derive a smoothness function for the unsupervised case when negative examples do not exist.

Using the same principle as in the semi-supervised learning scenario, we stress that close nodes in the modality used during graph creation should also be close in annotations. Therefore, a greater "cost" should be assigned when annotations differ for close neighbors. We assign a "cost" function per annotation

$$C_g(f^w, \mathbf{W}_g) = \sum_{i,j} W_{g,ij} \left| f_i^w - f_j^w \right|. \tag{10}$$

For each of the G graphs, it makes the most sense to average the graph's smoothness over the possible annotations. The graph smoothness for a particular graph sums the cost function from each word found in (10) but we also include a relaxation parameter r similar to that in (8). By summing the cost over all annotations and then normalizing, we find weights that sum to 1

$$\alpha_g = \frac{\left[\frac{1}{\sum_w C(f^w, \mathbf{W}_g)}\right]^{\frac{1}{r-1}}}{\sum_g \left[\frac{1}{\sum_w C(f^w, \mathbf{W}_g)}\right]^{\frac{1}{r-1}}}. \tag{11}$$

This smoothness measurement can be thought of as the "goodness" of the particular graph, and can be used to deemphasize poor ("unsmooth") graphs. Intuitively, as $r \to \infty$, $\alpha_g \to (1/G)$, equally weighting all graphs. As $r \to 1$, only the smoothest graph with small $\sum_w C(f^w, \mathbf{W}_g)$ is kept.

IV. Annotation Identification

After the stable graphs have been formed and the appropriate weights calculated, the target node's annotations have been updated to include a weighted proportion of the annotations for similar videos. Now annotations must be found in the stable graph. The *relevant* words describing the target must be gleaned from the set of labels associated with each video in the G graphs. We identify the words most confidently tied to the target video and supply them as annotations of the video. Using a zipf-based cutoff for keyword extraction has been used elsewhere for annotation [17], [27], and we describe and motivate this process for our annotation algorithm.

The *zipf curve* has been cited as an approximate model of word distributions in natural language [1]. The zipf curve is an instance of the power law family. The curve is defined using a shape parameter s which uniquely defines the distribution, defined for the kth most frequent word in the case of a dictionary of size N as

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}. \tag{12}$$

Using a threshold based on a zipf curve allows the algorithm to keep a variable number of keywords, rather than, say, a fixed number $\bar{K}$ most important keywords. It is not dependent on a pre-specified threshold or a pre-specified number of annotations, which is important because some videos require more annotations than others. For each stable graph, a best-fit zipf curve is found by solving for the shape parameter s for a vector of the target node's attributes or annotations, f_i^*. Then, keywords appearing more often than the theoretical K'th-ranked word, $f(K'; s, N)$, are flagged as annotations. A range of K' provides instances that can create a precision/recall-like curve allows us to see the tradeoffs as we include fewer (greater precision) or more (greater recall) keywords.

V. Experiments

Experiments were conducted to evaluate the effectiveness of the proposed unsupervised approach for automatic collaborative tagging. Specifically, a comparison is made between the unsupervised approach, the semi-supervised approach in [28] and summarized in Sections III-A1 and III-B1, and the original tags provided for a YouTube video by an individual user. An analysis of annotation extension and filtering of the proposed technique follows in Section V-D and the performance without aid of initial tags in Section V-E. We conclude with a study of system parameters graph size, N, and weighting parameter, r, in Sections V-F and V-G.

A. Data

For our experiments, YouTube videos were used as the annotation targets since such data has the necessary qualities of 1) repetitiveness, and 2) independent tagging instances. The repository was crawled to extract 728 videos. It was ensured that some of the videos had overlapping and similar content in the total set. For instance, several e-trade commercials featuring a talking baby were extracted, along with the clips YouTube has identified as "duplicates." These duplicates and similar videos were not explicitly marked in the database. The local database was grown to 728 videos such that performance gain was not random, resulting from inclusion of the most common annotations. The inclusion of similar videos is reasonable since it is believed that 85% of YouTube videos have such overlap in the online database [2]. Complicating instances were explicitly included, such as commercials with similar themes or revisions/edits of an original video, along with other randomly crawled YouTube videos. These instances were expected to make the task more difficult, but representative of YouTube videos which often undergo iterations or edits.

A keyframe was extracted on average once every 10 s. They were not regularly spaced in time but were the frames closest to the centroids of clustered CLD features. Using the CLD centroid frames allows the keyframes to capture different scenes or views from the video. The similarity between two videos was

considered to be the maximum of any pairwise keyframe similarity between the two videos. Certainly, this is a limitation of the system and a more sophisticated method for similarity estimation can be adopted. The annotation dictionary was limited only by the total set of user-supplied labels. The total lexicon dictated by the 728 videos consisted of 3326 words.

B. Performance Metric: Relevance-Coverage

A standard precision-recall metric does not accurately reflect the performance of this algorithm since annotations do not fit neatly into a binary true/false categorization. Rather, they fit into a range between "relevant" to "irrelevant," and can also be "incorrect." Our scoring provides a tag w with a score, $c_w = +1$ for a relevant tag, 0 for an irrelevant tag, and -1 for an incorrect tag. A similar three-class scoring method has been adopted in our previous work on video annotation from transcripts [17] as well as studies on image annotation [16], [29]. The "relevance" and "irrelevance" were judged from belief that a typical user may use that word in a query seeking the video, and thus, whether applying such a keyword would help in querying for that video. Relevance was gauged by a group of judges that had viewed the videos and were given these instructions. A modified precision metric, called *relevance*, is defined as the average score of the K extracted tags that occur more frequently than the K'th most frequent word in the best-fit zipf curve, $P = (1/K)\sum_{i=1}^{K} c_i$.

Furthermore, the set of relevant tags for a particular video is not strictly limited, and therefore a standard recall metric cannot be used. Instead, a running list was kept of all "relevant" annotations for a video encountered using any set of parameters found in this paper or previous work by the authors. Then, we adopted a recall-like metric, called *coverage*, that indicates the percentage of all seen positive annotations $\mathcal{A}$ for a certain video covered by the method: $R = (|\mathcal{S} \cap \mathcal{A}|)/(|\mathcal{A}|)$, where $\mathcal{S}$ is the set of tags extracted using the particular method. $\mathcal{A}$ is the union of all positive annotations found through any method during the experiments. The best metric has the greatest area under the relevance/coverage curve created by varying K', exhibiting high relevance without expending coverage.

C. Unsupervised Learning versus Semi-Supervised Learning

An experiment was done to compare the proposed approach to the semi-supervised approach (OMG-SSL) from Wang [28]. For OMG-SSL, we use the owner's initial labels for the videos as training data for the semi-supervised case; that is, $Y_i^w = 1$ if the keyword w appeared in the annotations in the online database. If a label w is not present, we give it an "unlabeled" value of $Y_i^w = 0$. An experiment not shown was done giving it an "incorrect" $Y_i^w = -1$, initializing a missing annotation as not relevant, which showed worse performance.

For these studies, only visual modalities were used, and only an L_2 norm is used for distance measurement between features. However, it is notable that these specifications are adaptable. The following visual modalities were used which have proved effective for video annotation [15]:

- **Edge Distribution Histogram** (EDH) 75-dimensional;
- **5 × 5 Color Moment** (CM5 × 5) 225-dimensional, based on 5 × 5 block division of images in Lab color space;
- **3 × 3 Color Moment** (CM3 × 3) 81-dimensional, based on 3 × 3 block division;

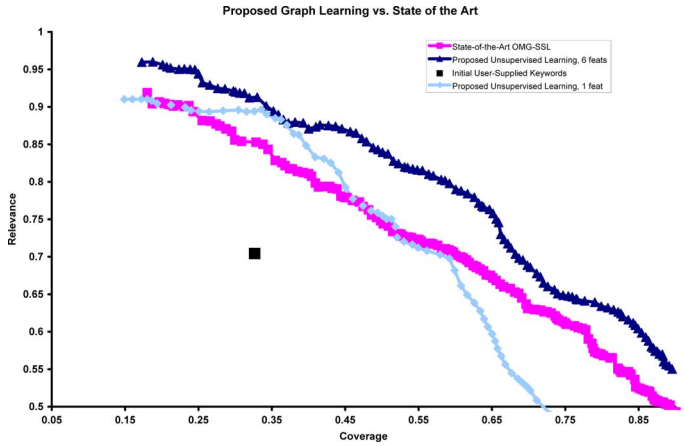


Fig. 3. Graph showing proposed unsupervised approach versus semi-supervised approach and the relevance/coverage point of the owner tags. Unsupervised multi-graph reinforcement performs the best. Unsupervised singular graph reinforcement offers higher relevance but lower coverage than semi-supervised learning since it considers fewer videos.

- **Wavelet Texture** (WT) 128-dimensional;
- **Color AutoCorrelogram** (AC) 144-dimensional, based on 36-bin color histogram and four distances;
- **HSV Color Histogram** (HSV) 64-dimensional;
- **Co-occurrence Texture** (CT) 75-dimensional.

The single graph case creates a graph with nodes x_i formed by videos, and edge weights, W_{ij}, the linear combination of the normalized distance between two nodes in each modality. For the single-graph learning case, the typical similarity metric

$$W_{ij} = \begin{cases} \sum_{\text{feat}} \exp\left(-\frac{d_{\text{feat}}(x_i,x_j)}{\sigma_{\text{feat}}}\right), & \text{if } i \neq j \\ 0, & \text{else} \end{cases} \tag{13}$$

is used for feat $\in \{\text{EDH}, \text{CM5} \times 5, \text{CM3} \times 3, \text{WT}, \text{AC}, \text{HSV}, \text{CT}\}$. Shaping parameter σ_{feat} was set to the standard deviation of d_{feat}. This choice of σ ensures that the variance in any singular feature space does not dominate the distance between two videos. The multi-graph learning case takes each of these features and creates a separate graph, and combines them using the weighting scheme found in Sections III-B1 and III-B2. The similarity between two videos is taken as the *maximum* pairwise similarity between two keyframes. Regrettably, this reduces the power of the algorithm since a clip of a video is taken as identical to the full video. A feature based on the video as a whole rather than a singular keyframe is desirable and can be used in future work.

Fig. 3 highlights some results of this study. The proposed approach performs better than the state-of-the-art OMG-SSL. OMG-SSL assumes prior labels from the incomplete initial tags and repeatedly emphasizes them, a factor identified as problematic in Section III-A2. The unsupervised approach significantly outperforms the original annotations, even at the same recall level. Additionally, the graph shows that singular graph reinforcement has higher initial relevance, but lower coverage at low relevance, than the semi-supervised approach. This result arises from singular graph reinforcement's consideration of only a few very similar videos. Fig. 4 shows an example set of initial tags, as well as tags from both semi-supervised and unsupervised graph reinforcement and mining.

ID: Title	BJDdwZjB4z8: "star wars kid"	35p4PIXeMMU: "Women know your limits"	6q3HP4vFd7g: "E-Trade Superbowl XLII (42) 2008 'Baby' Commercial !!!"
Frame from Video			
Initial Tags	star, wars, kid	comedy, dinner, fun, limits, women	42, ad, advert, advertisement, baby, clown, commercial, e, england, etrade, giants, new, patriots, superbowl, trade, xlii, york
OMG-SSL	star, wars, kid, starwars	women, know, your, limits, fun, dinner, comedy	trade, baby, commercial, e, superbowl, clown, etrade, xlii, ad, 42, advertisement, england, giants, york, patriots, advert, new
Proposed Unsupervised Learning	star, wars, kid, lightsaber, starwars, fat, jedi, funny	women, know, your, limits, funny, woman, comedy, dinner, fun	e, trade, baby, superbowl, commercial, super, bowl, clown

Fig. 4. Example test videos and annotations, with red bold text for correct, plain black text for irrelevant, and italics for incorrect. The left two examples exhibit the tag *extension* afforded by unsupervised and semi-supervised learning. The right example shows the tag *filtering* possible using unsupervised learning, that is not done properly by the semi-supervised algorithm. Collaborative tagging done using graph mining techniques generally improves the tags, providing new tags and filtering irrelevant ones.

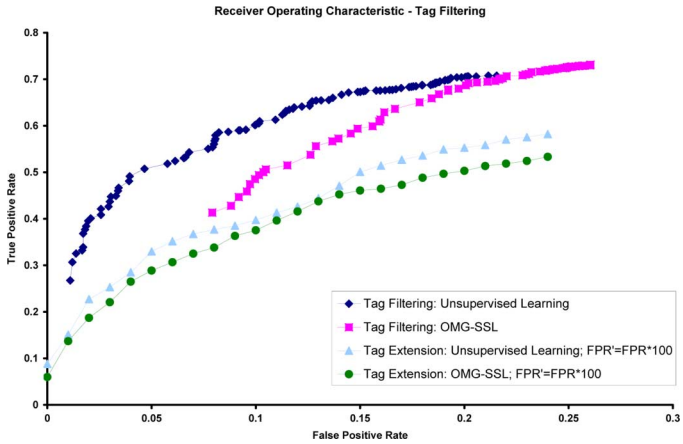


Fig. 5. ROC curve for each algorithm's treatment of tag *extension* and tag *filtering*. **Tag extension**: TP for annotation extended by algorithm to video and judged correct, FP for extended but judged incorrect, TN for not extended and judged incorrect, and FN for not extended but judged correct. Unsupervised annotation has better tag extension qualities because of elimination of "stickiness" on initial labels. **Tag filtering**: TP for annotation judged correct and kept by algorithm, FP for judged incorrect but kept, TN for judged incorrect and discarded, and FN for judged correct but discarded. Unsupervised learning has more area under ROC and therefore better treatment of owner annotations.

D. Collaborative Annotation Extension and Filtering

In Section II, the possibility of tag extension and tag filtering afforded by online communities was emphasized. Here, we explicitly test OMG-SSL's and our algorithm's performance in each of these areas separately. One of the biggest contributions of this annotation process that leverages similar multimedia documents is in supplying *new* annotations. Considering all relevant annotations found in this study, on average 69.6% were discovered only after mining neighbors; that is, on average, the initial labeling had 30.4% of the relevant annotations that can be discovered by collecting the annotations of similar videos. The ability to correctly find new annotations from weakly labeled training data without building a distinct model for a particular annotation is a unique feature of the system.

Fig. 5 plots the receiver operating characteristic for tag extension. A false alarm is a tag that is extended that is incorrect, and a miss is a correct tag that is not extended. These extended annotations represent terms not found in the original document and therefore impossible to discover without mining similar documents. Both the unsupervised and semi-supervised technique are successful at annotation discovery, though the unsupervised technique shows a slight advantage arising from the "sticking" constraint of the semi-supervised technique.

In the effort to combat spam, owner annotations can be filtered by the algorithm. Fig. 5 shows the receiver operating characteristic for tag filtering. A false alarm is an incorrect initial tag that is kept, and a miss is a correct initial tag that is discarded by the algorithm. The unsupervised technique has a better ROC since it does not repeatedly reinforce noisy, incorrect initial labels as explained in Section III-A2 and as is clear from (4). The unsupervised learning algorithm proposed here is better suited for gleaning signal from noise in cases of with large amounts of noisy initial data.

E. Annotating Without Initial Keywords

A special study was done to examine the performance of the algorithm when annotating a new video that had not already been given initial keywords by the owner. This scenario corresponds to annotations at the time of upload that could be suggested to the user. As shown in the relevance/coverage graph in Fig. 6, both algorithms perform quite well annotating totally unlabeled documents, a consequence of the presence of aliased or similar data. When annotating without initial keywords, the "stickiness" constraint explained in Section III-A1 has been removed. Thus, the proposed algorithm only offers a slight advantage specifically when the video has not already been annotated but performs on par with OMG-SSL.

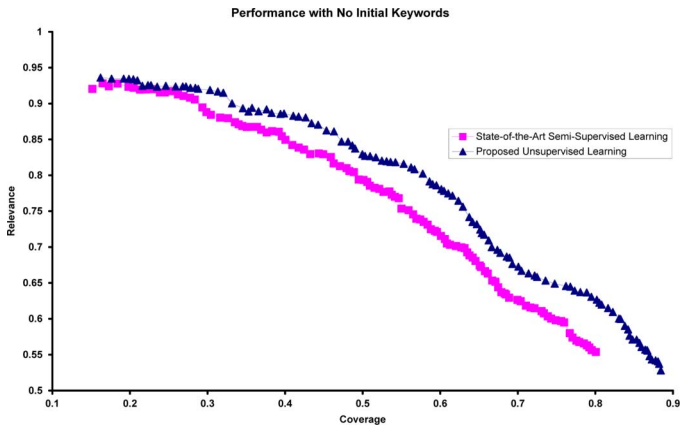


Fig. 6. Performance of proposed algorithm and state-of-the-art when presented with an unlabeled video. Great performance results from redundancy in social media sites. OMG-SSL performs on par with the proposed method since the sticking constraint of the initial labeling is not a hindrance in this scenario.

F. Evaluation of Graph Size, N

An experiment was conducted to find the effect of graph size on annotations. "Graph size" refers to the number of near neighbors that construct each of the graphs formed from different modalities, denoted as N in Section III-B2. Intuitively, the multimedia annotation technique should be robust to graph size, as

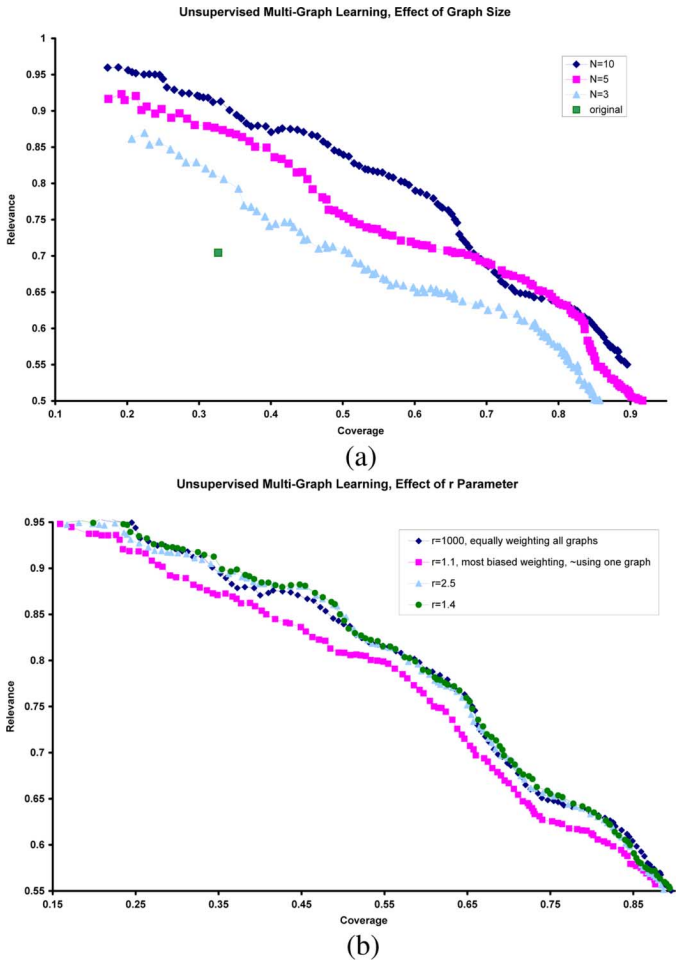


Fig. 7. (a) Effect of graph size on performance. Performance tends to improve with larger graphs because more nodes result in positive reinforcement of commonalities without reinforcing irrelevant annotations from the additional nodes. (b) Effect of r parameter. A biased weighting ($r = 1.4$) that significantly weights the smoothest graphs rather than equally weighting all graphs ($r = 1000$) has the most area under relevance/coverage graph. Smoothing cost function derived from graphs effectively combines the graphs for multi-graph reinforcement. (a) By graph size, N. (b) By weighting parameter, r.

the graph diffusion step incorporates document similarity into the annotation decision process.

The results using graph sizes 3, 5, and 10 for the unsupervised multi-graph case are shown in Fig. 7(a). Clearly, larger N values improve performance. Intuitively, as the graph becomes larger we are using more dissimilar videos. However, the system is able to robustly handle them by reinforcing the similar elements without contributing new incorrect ones. Only the commonalities are positively reinforced.

G. Evaluation of Weighting Parameter, r

The value of parameter r can be used to vary the gradation between smoothness differences on the multiple graphs, as described in Section III-B2. As $r \rightarrow 1$, the α_g values render an effect where only the smoothest graph is considered, $\alpha_{g_{\text{best}}} = 1$ and otherwise $\alpha_g = 0$. Increasing $r \rightarrow \infty$ results in equal weighting, $\alpha_g \rightarrow (1/G)\ \forall g$. The higher the visual correlation between videos, the more their tags should correlate. As the tags and visual feature similarity are from independent sources but both provide information about the video, tag correlation serves as validation that the choice of visual feature is appropriate. Results from an experiment which evaluated choice of parameter r can be seen in Fig. 7(b).

Effectively using only one graph with $\mathrm{r} = 1.1$, as expected, results in the worst performance. We found that using a small value of r, 1.4, which incorporated all graphs at varying weights performed best. However it was only slightly better than using all graphs at equal weights because all of the visual features chosen are known to be relevant in video search. The performance would be more profound if using features that were only relevant for some videos; for instance, audio features are very relevant in annotating music videos but less relevant perhaps in commercials. Still, the slight improvement offered by using weights rather than equal weighting validates the notion of using graph smoothness to derive weights.

VI. Discussions and Conclusions

This paper presents the gains achievable through collaborative tagging in community media sites. Sites such as YouTube are populated by repeated, duplicate, and related documents because of the viral nature of Internet media. We have presented a robust method for automatically annotating documents in such an environment. The algorithm creates stable graphs that are found in one medium (visual) that supplement the annotations in another relevant media form (text). This method is robust and trainable to particular qualities of the target data as well as annotation goals, with strong performance from a range of graph sizes and weighting parameters. It is most powerful when performing annotation using a collection of individual tagging instances on identical documents (e.g., del.icio.us) or similar documents (e.g., YouTube).

Experiments were conducted only on annotating video with text labels, but this system is fully capable of finding *visual* annotations or annotations in other modes, and also could be used on webpages, songs with lyrics, or images with description. It is fully combinable with an annotation method based on computer vision modeling, and can be extended using lexical relations. The social network links that underlie online media websites can be formulated into a graph and used in the multi-graph annotation learning. Future research should be performed to compare this method with alternate forms of collaborative tagging, such as simple annotation frequency in some immediate neighborhood, as well as its performance on alternate sites. Besides just simple relevance/coverage analysis as performed here, usefulness may also be gleaned from user response when it is used as a tag suggestion agent.

Acknowledgment

The authors would like to thank Dr. M. Wang and Dr. X.-S. Hua for their insightful discussions.

References

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval*. Reading, MA: ACM Press/Addison-Wesley, 1999.

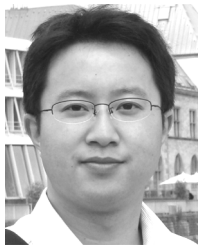
[2] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM Conf. Internet Measurement*, New York, 2007, pp. 1–14.

[3] , O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[4] J. Cho and A. Tomkins, "Guest editors' introduction: Social media and search," *IEEE Internet Comput.*, vol. 11, no. 6, pp. 13–15, Nov.–Dec. 2007.
[5] *Flickr*. [Online]. Available: http://www.flickr.com/.
[6] G. W. Furnas, C. Fake, L. von Ahn, J. Schachter, S. Golder, K. Fox, M. Davis, C. Marlow, and M. Naaman, "Why do tagging systems work?," in *Proc. Conf. Human Factors in Computing*, 2006, pp. 36–39.
[7] G. Geisler and S. A. Burns, "Tagging video: Conventions and strategies of the YouTube community," in *Proc. Int. Conf. Digital Libraries*, 2007, p. 480.
[8] S. Golder and B. A. Huberman, The Structure of Collaborative Tagging Systems, Information Dynamics Lab, HP Labs, Tech. Rep., Aug. 2005.
[9] *Google Image Labeler*. [Online]. Available: http://images.google.com/imagelabeler/.
[10] Y. Jing and S. Baluja, "Pagerank for product image search," in *Proc. Int. World Wide Web Conf. (WWW 2008)*.
[11] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr helps us make sense of the world: Context and content in community-contributed media collections," in *Proc. ACM Multimedia*, New York, 2007, pp. 631–640, ACM.
[12] V. Lavrenko, S. L. Feng, and R. Manmatha, "Statistical models for automatic video annotation and retrieval," in *Proc. ICASSP*, 2004.
[13] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[14] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, Flickr, academic article, to read," in *Proc. Hypertext and Hypermedia*. New York: ACM Press, 2006, pp. 31–40.
[15] T. Mei *et al.*, "MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search," in *Proc. TRECVID 2007 Workshop*, Nov. 2007.
[16] E. Moxley, J. Kleban, and B. Manjunath, "SpiritTagger: A geo-aware tag suggestion tool mined from Flickr," in *Proc. ACM Multimedia Information Retrieval*, Oct. 2008.
[17] E. Moxley, T. Mei, X.-S. Hua, W.-Y. Ma, and B. Manjunath, "Automatic video annotation through search and mining," in *Proc. ICME*, Jun. 2008.
[18] A. P. Natsev, M. R. Naphade, and J. R. Smith, "Semantic representation: Search and mining of multimedia content," in *Proc. SIGKDD 2004*, ACM, 2004, pp. 641–646.
[19] G.-J. Qi, X.-S. Hua, Y. Rui, T. Mei, J. Tang, and H.-J. Zhang, "Concurrent multiple instance learning for image categorization," in *Proc. CVPR*, 2007.
[20] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu, "Bipartite graph reinforcement model for web image annotation," in *Proc. ACM Multimedia*, 2007.
[21] C. Shirky, Ontology is Overrated—Categories, Links, and Tags. [Online]. Available: http://shirky.com/writings/ontology overrated.html.
[22] Y. Song, X.-S. Hua, L.-R. Dai, M. Wang, and R.-H. Wang, "An automatic video semantic annotation scheme based on combination of complementary predictors," in *Proc. ICASSP*, 2006.
[23] H. Tong, J. He, M. Li, C. Zhang, and W.-Y. Ma, "Graph based multimodality learning," in *Proc. ACM Multimedia*, 2005.
[24] *TRECVID Retrieval Evaluation*. [Online]. Available: http://trecvid.nist.gov/.
[25] V. S. Tseng, J.-H. Su, J.-H. Huang, and C.-J. Chen, "Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 260–267, Feb. 2008.
[26] A. Velivelli and T. S. Huang, "Automatic video annotation by mining speech transcripts," in *Proc. CVPR Workshop*, 2006.
[27] A. Villacorta, "Spheres of influence," in *Proc. ACM SIGGRAPH 2007*, New York, 2007, p. 254.
[28] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proc. ACM Multimedia*, 2007, pp. 862–871.
[29] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "Annosearch: Image auto-annotation by search," in *Proc. CVPR*, 2006, pp. 1483–1490.
[30] X. Wu, A. Hauptmann, and C.-W. Ngo, "Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts," in *Proc. ACM Multimedia*, 2007, pp. 168–177.
[31] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images with dual-wing harmoniums," in *Proc. UAI*, 2005, pp. 633–641.
[32] R. Yan, J. Tesic, and J. R. Smith, "Model-shared subspace boosting for multi-label classification," in *Proc. SIGKDD 2007*, ACM, 2007.
[33] *YouTube*. [Online]. Available: http://www.youtube.com/.
[34] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press, 2004.
[35] X. Zhu, "Semi-supervised learning with graphs," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 2005.
[36] X. Zhu, Semi-Supervised Learning Literature Survey, Univ. Wisconsin, Madison, Tech. Rep., 2005.

Emily Moxley received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, in 2005 and the M.S.E. and Ph.D. degrees in electrical and computer engineering from University of California, Santa Barbara, in 2007 and 2009, respectively.

Her research interests include image annotation, video annotation, multimedia analysis, and social media analysis.



Tao Mei (M'07) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He joined Microsoft Research Asia, Beijing, China, as an Associate Researcher in 2006. His current research interests include multimedia content analysis, computer vision, pattern recognition, and online multimedia applications such as multimedia search, advertising, recommendation, presentation, and social media. He has authored five book chapters and over 70 journal and conference papers in these areas, and holds more than 20 filed international and U.S. patents or pending applications.

Dr. Mei serves as an Editorial Board Member of the *Journal of Multimedia* and *Advances in Multimedia*, a Guest Editor of *ACM Multimedia Systems Journal* for the Special Issue on Multimedia Intelligent Services and Technologies, a Technical Program Committee Member for a dozen international conferences, and a Reviewer for over ten prestigious international journals. He received the Best Paper and Best Demonstration Awards in the ACM International Conference on Multimedia 2007, and the Best Poster Award in the IEEE International Workshop on Multimedia Signal Processing 2008. He is a member of the Association for Computing Machinery (ACM).

B. S. Manjunath (F'05) received the B.E. degree (with distinction) in electronics from Bangalore University, Bangalore, India, in 1985, the M.E. degree (with distinction) in systems science and automation from the Indian Institute of Science, Bangalore, in 1987, and the Ph.D. degree in electrical engineering from University of Southern California, Los Angeles, in 1991.

He is now a Professor of electrical and computer engineering and Director of the Center for Bio-Image Informatics at the University of California, Santa Barbara. His current research interests include image processing, data hiding, multimedia databases, and bio-image informatics. He is a co-editor of *Introduction to MPEG-7* (Wiley, 2002).

Dr. Manjunath was a recipient of the National Merit Scholarship (1978–1985) and was awarded the University gold medal for best graduating student in Electronics Engineering in 1985 at Bangalore University. He was an associate editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and is currently an associate editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON MULTIMEDIA.