SEMANTIC ANNOTATION OF PERSONAL VIDEO CONTENT USING AN IMAGE FOLKSONOMY

Hyun-seok Min[1], JaeYoung Choi[1], Wesley De Neve[1], Yong Man Ro[1], and Konstantinos N. Plataniotis[2]

[1]Image and Video Systems Lab, Korea Advanced Institute of Science and Technology (KAIST), Yuseong-gu, Daejeon, 305-732, Korea
[2]Multimedia Lab, The Edward S. Rogers Sr. Dept. of Electrical and Computer Engineering (ECE), University of Toronto, Toronto, Ontario M5S 3GA, Canada

ABSTRACT

The increasing popularity of user-generated content (UGC) requires effective annotation techniques in order to facilitate precise content search and retrieval. In this paper, we propose a new approach for the semantic annotation of personal video content, taking advantage of user-contributed tags available in an image folksonomy. Video shots and folksonomy images are first represented by a semantic vector. Next, the semantic vectors are used to measure the semantic similarity between each video shot and the folksonomy images. Tags assigned to semantically similar folksonomy images are then used to annotate the video shots. To verify the effectiveness of the proposed annotation method, experiments were performed with video sequences retrieved from YouTube and images downloaded from Flickr. Our experimental results demonstrate that the proposed method is able to successfully annotate personal video content with user-contributed tags retrieved from an image folksonomy. In addition, the size of our tag vocabulary is significantly higher than the size of the tag vocabulary used by conventional annotation methods.

Index Terms—Folksonomy, semantic annotation, user-generated content, video indexing

1. INTRODUCTION

Nowadays, users can be considered both consumers and producers of multimedia content. This has resulted in a significant increase of the amount of user-generated content (UGC) in the past few years [1]. Social media applications such as Flickr and YouTube currently provide tools that allow users to manually tag image and video content using their own vocabulary. The result of personal free tagging of multimedia content for the purpose of retrieval and organization is called a folksonomy [2].

It is well-known that manual tagging is a time-consuming and cumbersome task. As a result, users either do not annotate their multimedia content or they only make use of a limited number of tags [3]. This hampers search performance. An alternative for manual tagging consists of automatically annotating UGC with semantic tags that are generated by content analysis techniques. That way, more semantic metadata can be made available for the purpose of keyword matching, thus improving search performance.

Semantic annotation of multimedia content has been a long-time issue in the field of multimedia search and retrieval [4][5]. Previous research efforts have mainly focused on model-based annotation using a finite number of classifiers or concept detectors. In model-based annotation, concepts not learned during training cannot be detected during testing due an absence of knowledge about these concepts. The lack of 'generalization' power demonstrated by conventional annotation techniques is a critical problem when annotating personal video content. Indeed, personal video content typically contains a wide range of semantic concepts.

The NIST-sponsored TRECVID (TREC Video Retrieval Evaluation) evaluation meetings are an on-going series of workshops focusing on several challenges in the field of content-based retrieval of video [6]. In particular, the TRECVID high-level feature extraction task aims at finding semantic concepts in video content. Most of the TRECVID participants make use of classifiers such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs). However, these binary classifiers only work well with a limited set of predefined concepts, thus only partly bridging the semantic gap [7]. Specifically, the TRECVID efforts have been mainly focused on detecting a lightweight set of 39 concepts defined in LSCOM-Lite [8]. The use of 39 concepts is clearly not sufficient to cover the wide range of semantic concepts present in personal video content.

In this paper, we propose a folksonomy-based method for the semantic annotation of personal video content. In particular, the proposed annotation method makes use of tags assigned to folksonomy images. These folksonomy images are similar to the video shots that need to be annotated. The more folksonomy images that are similar to a query video shot, the better the quality of the tags used to annotate the video shot. The proposed method differs from conventional model-based annotation approaches. Specifically, our method takes advantage of the wide variety of user-supplied tags in an image folksonomy for the purpose of annotating video content, compared to conventional annotation approaches that rely on a limited number of trained concept models.

The remainder of the paper is organized as follows. Section 2 presents an overview of the proposed method for annotating personal video content. This section also explains the creation of semantic vectors in order to represent video shots and folksonomy images. In addition, this section discusses how to compute the similarity between a video shot and folksonomy images by making use of semantic vectors. Experimental results are provided in Section 3. Finally, conclusions are drawn in Section 4.

2. SEMANTIC VIDEO ANNOTATION

2.1. Trained concepts versus folksonomy concepts

Thus far, semantic video tagging using trained concepts has mainly focused on the modeling and classification of these concepts [9]. A subset of the visual concepts that appear within a video shot is typically detected using a frame-to-concept matching process, between low-level features of key frames on the one hand and a finite set of concept classifiers on the other hand. The concepts that are most likely to occur in the key frames are selected and then used for the purpose of semantic tagging. However, it should be clear that the number of generated tags is limited due to the fact that the number of concept detectors is limited.

Personal video content may contain a wide variety of semantics. It is necessary to differentiate between the semantic concepts that can be reliably trained using concept classifiers and the semantic concepts that are represented by the user-generated tags in an image folksonomy. A major difference lies in the fact that the set of user-generated tags is unlimited, while the set of trained concepts is restricted to a limited number of concepts. Moreover, the set of trained concepts typically includes common visual semantic concepts that can be reliably trained by machines, while the set of user-generated tags allows more freedom in the representation of semantics, compared to the set of trained concepts. The main contribution of this paper consists of a method that is able to exploit the set of user-generated tags in an image folksonomy for the purpose of annotating personal video content.

2.2. General overview

Fig. 1 shows the proposed approach for folksonomy-based annotation of personal video content. Our approach essentially consists of two major steps: a video analysis step and a semantic annotation step.

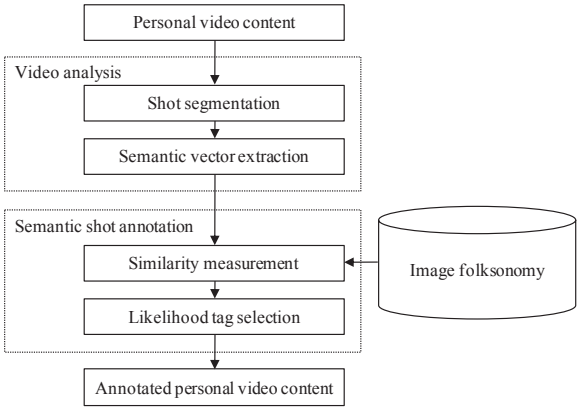

Fig. 1. Overall approach for folksonomy-based annotation of personal video content

In the video analysis step, the input video sequence is first segmented into shots. Next, representative key frames are extracted for each shot. Low-level visual features are subsequently extracted from the selected key frames. For each shot, a semantic vector is generated by classifying the visual features of the key frames into one of the predefined concept classes.

In the semantic annotation step, the similarity between each video shot and the folksonomy images is measured. By ranking the folksonomy images according to their similarity with the query video shot, a set of popular user-generated tags can be created. These tags can then be used for annotating a video shot.

2.3. Video shot representation

In our research, annotation is done at the level of video shots. Each shot is represented by a number of key frames [10][11]. Each key frame is in its turn represented by two vectors: a low-level feature vector $\mathbf{x}$ of dimension N and a semantic vector $\mathbf{v}$ of dimension M. We assume that the concept detectors used to create the semantic vector $\mathbf{v}$ have been defined and trained in advance.

A semantic vector $\mathbf{v}$ is generated by a semantic mapping α from low-level feature space $\mathfrak{X}$ to semantic vector space $\mathfrak{V}$ as follows:

$$\alpha : v_m = P(w_m \mid \mathbf{x}), \tag{1}$$

where the posterior probability $P(w_m|\mathbf{x})$ denotes the semantic relevance of the concept w_m for a particular key frame, given the low-level feature vector $\mathbf{x}$ of the key frame in question. Using the Bayesian theorem, the posterior probability can be computed based on the concept conditional probability $P(\mathbf{x}|w_m)$ and the prior probability $P(w_m)$ [12]. In this paper, concept conditional probabilities are derived using SVM models [13].

To capture the visual semantics that appear in local spatial regions, we employ the block segmentation strategy explained in [13]. As such, each key frame I is segmented into H sub-regions $\{I_1, \ldots, I_H\}$, where I_h represents the h^{th} sub-region of key frame I. The concept conditional probability of $\mathbf{x}_h$ to w_m is then denoted as $v_{h,m} = P(\mathbf{x}_h|w_m)$, where $\mathbf{x}_h$ represents the low-level feature vector of the h^{th} sub-region of key frame I and where $v_{h,m}$ is used to construct $\mathbf{v}_h$, the semantic vector of the h^{th} sub-region of I.

We employ K key frames to represent each shot. The concept conditional probabilities computed for each key frame are used to generate a summarizing semantic vector for each sub-region in the corresponding video shot. The summarizing concept conditional probabilities are computed as follows:

$$v_{h,m}^q = median(v_{h,m}^k \mid k = 1,2,..., K), \tag{2}$$

where $v_{h,m}^q$ is the concept conditional probability for the h^{th} sub-region of video shot q and *median()* is the median value operator. The concept conditional probabilities $v_{h,m}^q$ are then used to construct the semantic vector $\mathbf{v}_h^q$ for the h^{th} sub-region of shot q.

2.4. Folksonomy image representation

Let $\mathbf{F} = \{I_1, \ldots, I_F\}$ be a set of F images in the folksonomy $\mathbf{F}$, where I_f denotes the f^{th} image in $\mathbf{F}$. Similar to the key frames of a video shot, each folksonomy image is divided into a set of H sub-regions, and each sub-region is represented by a low-level feature vector and a semantic vector. The semantic vector for the h^{th} sub-region of the f^{th} folksonomy image is denoted as $\mathbf{v}_h^f$.

2.5. Similarity measurement

The proposed method makes use of the user-provided tags in an image folksonomy for annotating personal video content. This requires finding folksonomy images similar to the key frames of a video shot. The semantic similarity d_k between a video shot q and the f^{th} image in the folksonomy $\mathbf{F}$ can be measured as follows:

$$d_k = 1 - \frac{1}{H \cdot M} \cdot \sum_{h=1}^{H} \sum_{m=1}^{M} \sqrt{\left(v_{h,m}^q - v_{h,m}^f\right)^2}. \tag{3}$$

2.6. Selection of likelihood tags

The tags generated by the proposed annotation method frequently occur in the top-ranked l images in the folksonomy, i.e. the top l images that are semantically similar to the video shot that needs to be annotated. Indeed, the number of times a tag was assigned to the top-ranked l images can be used to determine how likely a tag is related to the video shot that needs to be annotated. The selection of likelihood tags for a particular video shot is explained in more details below.

First, we construct a set that unites all tags associated with the top-ranked l images. This tag set, which consists of T tags, can be defined as $\mathbf{T}_u = \{t_1, \ldots, t_T\}$. Note that t_i is unique in the tag set $\mathbf{T}_u$. Next, we measure the frequency of all tags in $\mathbf{T}_u$. This results in a set of frequencies that is denoted as $\mathbf{C}_u = \{c_1, \ldots, c_T\}$, where c_i stands for the frequency of tag t_i in the top-ranked l images.

To compute the likelihood of tag t_i, we average the semantic similarity values of the selected images tagged by t_o. This is done as follows:

$$d_i^{avg} = \frac{1}{c_i} \cdot \sum_{k=1}^{l} \left[d_k \cdot B(\mathbf{T}_k) \right], \tag{4}$$

where d_i^{avg} represents the average semantic similarity for tag t_i, and $\mathbf{T}_k$ is the tag set of the k^{th} image in the list of top-ranked l images. $B(\mathbf{T}_k)$ is a Boolean function that outputs True (1) when t_i belongs to $\mathbf{T}_k$ and False (0) otherwise.

The likelihood of tag t_i is dependent on both the average semantic similarity between a video shot and the folksonomy images tagged with t_i, and the frequency c_i of tag t_i in the selected l images. As such, the tag likelihood can be calculated by combining the tag frequency c_i and the corresponding average semantic similarity d_i^{avg} as follows:

$$z_i = \omega_c \cdot \left(\frac{c_i}{l} \right) + \omega_d \cdot d_i^{avg}, \tag{5}$$

where z_i is the likelihood of tag t_i and c_i/l is the normalized tag frequency of t_i over l. Further, ω_c and ω_d (= 1 - ω_c) are weighting parameters for the tag frequency and the semantic similarity, respectively. For each tag in $\mathbf{T}_u$, we can then generate a set of tag likelihood values denoted as $\mathbf{Z} = \{z_1, \ldots, z_T\}$. Finally, a set of likelihood tags $\mathbf{T}_q$ is selected from $\mathbf{T}_u$ by selecting tags that have a high likelihood value.

$$\mathbf{T}_q = \{ t_i \,|\, i = 1, \cdots, N_{lik}^q \} \text{ s.t. } (z_1 > z_2 > \ldots > z_{th}) \tag{6}$$

where N_{lik}^q is the number of tags with a likelihood value higher than a threshold value z_{th}.

3. EXPERIMENTAL RESULTS

To verify the effectiveness of the proposed method for semantic video tagging, a number of experiments were performed. In order to construct an image folksonomy, images were retrieved from Flickr using the publicly available Flickr API. This resulted in a dataset consisting of a total of 1,500 images, annotated with tags provided by anonymous users. In our folksonomy, the number of tags for each image was ranging from 1 to 15. In addition, each image was annotated with four different tags on average.

Further, 70 different personal video sequences were retrieved from 'YouTube'. This resulted in a set of 1,015 video shots. In order to be able to evaluate the performance of the proposed annotation technique, a ground truth was created for all video shots by three participants, using the tag vocabulary associated with the 1,500 images downloaded from Flickr. A tag was added to the ground truth when it was agreed upon by all participants. As such, 6,426 ground truth tags were generated for the 1,015 shots. Among the 6,426 ground truth tags, 200 different tags could be identified.

To compute the similarity between a shot and the folksonomy images, we used semantic vectors of dimension 34. This is, the semantic vectors represented 34 different semantic concepts: 'gravel', 'park', 'pavement', 'road', 'rock', 'sand', 'sidewalk', 'face', 'people', 'field', 'peak', 'wood', 'flowers', 'leaves', 'trees', 'indoor', 'indoor-light', 'night', 'street-light', 'high-wave', 'low-wave', 'still water', 'mirrored water', 'ice (snow)', 'cloudy', 'sunny', 'sunset', 'sunset-on-mountain', 'beach', 'buildings', 'windows' 'brick', 'arch', and 'wall'. Two criteria were used for the selection of the 34 semantic concepts: their frequency in the folksonomy images and their ease of machine detection.

To train the concept classifiers for the 34 semantic concepts, a total of 1,597 home photo images from the MPEG-7 VCE2 dataset were used. These concept classifiers are needed to construct the semantic vectors for the video shots and the folksonomy images. MPEG-7 color and texture descriptors were used to represent the low-level visual features during the training process [13].

The performance of our annotation method was measured using the average true positive (TP) and false positive (FP) rate:

$$TP = \frac{1}{Q} \sum_{q=1}^{Q} \frac{N_{TP}^q}{N_{GTS}^q} \text{ and } FP = \frac{1}{Q} \sum_{q=1}^{Q} \frac{N_{FP}^q}{N_{lik}^q}, \tag{7}$$

where Q denotes the total number of video shots, N_{GTS}^q is the number of ground truth tags for the q^{th} video shot, N_{TP}^q is the number of correctly tagged video shots, and N_{FP}^q is the number of incorrectly tagged video shots. In our experiments, five key frames were extracted for each shot. Further, the parameters in Eq. (5) and Eq. (6) were initialized as follows: w_c=0.5, w_d=0.5, and z_{th}=0.5.

For the 1,015 video shots to be annotated, we obtained likelihood tags using the proposed annotation method. Next, we measured the tagging performance. Fig. 2 shows the average TP and FP rates, obtained by varying the number of selected folksonomy images (i.e., by varying l).

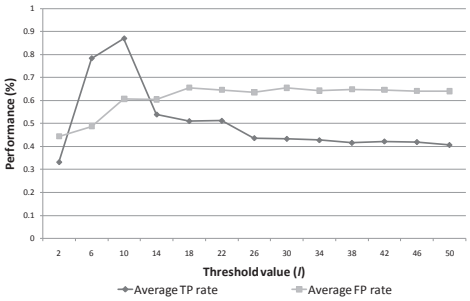

Fig. 2. Average TP and FP rate

As shown in Fig. 2, 10 folksonomy images are sufficient to achieve a high tagging accuracy rate. In particular, when 10 images are selected from the folksonomy, we can see that the ground truth tags are classified with a TP rate of 87% and a FP rate of 61%. The high TP rate can be explained by the observation that the number

of likelihood tags is higher than the number of ground truth tags. In particular, for the 1,015 video shots, the number of likelihood or candidate tags was 14,212, while the number of ground truth tags was 6,426. The high FP rate can be attributed to folksonomy images that are annotated with visually incorrect tags. Further, 8 likelihood tags were predicted on average for each video shot.

Among the 14,212 likelihood tags predicted by the proposed method, 5,590 tags are considered to be correct (according to the ground truth). In the set of 14,212 likelihood tags, 253 different tags could be identified. Out of the 253 different likelihood tags, 128 likelihood tags could also be found in the 200 ground truth tags (the ground truth tags were all visually related to the images). The remaining 125 tags mainly consisted of specific nouns (e.g., 'Netherlands') and subjective keywords (e.g., 'old'). As such, the experimental results show that the proposed annotation scheme could predict 128 different ground truth concepts for the personal video sequences, despite the fact that semantic vectors with a dimension of 34 concepts were used to compute the semantic similarity between video shots and folksonomy images. Compared to a semantic annotation approach based on trained concepts (e.g., the LSCOM-Lite vocabulary consisting of 39 concepts), the proposed method can predict a significantly higher number of tags for the purpose of annotating personal video content.

Key frame		
Ground truth tags	architecture, street, tree, landscape, building, staircase, city	waterside, river, bridge, tree, landscape, terrain, lake, forest
Likelihood tags	architecture, old, street, tree, landscape, building, house, staircase, art, ancient, rome, window	waterside, river, green, bridge, tree, landscape, sunset, clouds, sky, terrain, travel, lake
Key frame		
Ground truth tags	snow, landscape, snowboard, sky, mountain, outdoor	sunset, waterside, sun, beach, clouds, sea, ocean
Likelihood tags	snow, scene, landscape, waterside, ski, snowboard, sky, light, sun	sunset, terrain, lake, river, waterside, sun, beach, netherlands

Fig. 3. Example key frames with ground tags and likelihood tags

Fig. 3 contains a number of annotated key frames that illustrate the subjective performance of the proposed video annotation algorithm. The underlined tags are considered to be correct. Although visually unrelated tags such as 'ancient', 'art', 'travel', and 'old' are not considered to be correct, a number of these tags can still be used to provide additional context. This observation illustrates another important aspect of the proposed tagging scheme, namely that user-supplied tags can also be used to annotate video content with subjective and abstract tags, next to tags related to the visual semantics present in the personal video content.

4. CONCLUSIONS AND FUTURE WORK

This paper discussed a new technique for semantic tagging of personal video content, making use of the wide variety of user-generated tags available in an image folksonomy. Our approach is significantly different from traditional annotation techniques using a limited number of trained concepts. The experimental results, although preliminary, demonstrate that the proposed technique is able to successfully annotate video shots with user-supplied tags retrieved from an image folksonomy. Specifically, for the datasets used, we are able to achieve a TP rate of 87%, coming with a FP rate of 61%. In addition, the size of our tag vocabulary (128 correctly predicted ground truth tags) is significantly higher than the size of the tag vocabulary used by conventional annotation methods (e.g., the 39 LSCOM-Lite concepts). The latter observation makes our approach particularly suited for the annotation of personal video content.

Future research will aim at improving the annotation accuracy by investigating techniques for reducing the number of incorrect tag assignments in an image folksonomy.

5. REFERENCES

[1] OECD Study on the Participative Web: User Generated Content, 3 October 2007.

[2] R. Ramakrishnan and A. Tomkins, "Toward a People Web," IEEE Computer, Vol. 40, No. 8, pp. 63-72, Aug. 2007.

[3] M. Ames and M. Naaman, "Why We Tag: Motivations for Annotation in Mobile and Online Media," ACM CHI 2007, pp. 971-980, 2007.

[4] M. Tkalcic and J. Tkalcic, "Convergence of Web 2.0 and Semantic Web: A Semantic Tagging and Searching System for Creating and Searching Blogs," IEEE Int'l Conf. on Semantic Computing, pp. 201-208, Sep. 2007.

[5] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," IEEE Trans. on Pattern Anal. and Mach. Intelligence, Vol. 29, No. 3, pp. 394-410, Mar. 2007.

[6] W. Kraaij and G. Awad, "TRECVID 2008 High-Level Feature Task: Overview," In Proceedings of TRECVID 2008, 2008.

[7] A.G. Hauptmann, M.G. Christel, and Rong Yan, "Video Retrieval Based on Semantic Concepts," Proceedings of the IEEE, vol. 96, no. 4, pp. 602-622, April, 2008.

[8] M. Naphade, J.R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-Scale Concept Ontology for Multimedia," IEEE Multimedia, vol. 13, no. 3, pp. 86-91, 2006.

[9] M. Boutell, C. Brown, and J. Luo, "Survey on the State of the Art in Semantic Scene Classification," Technical Report 799, Univ. of Rochester, Rochester, NY, Dec. 2002.

[10] Z. Li, G.M. Schuster, and A. Katsaggelos, "MINMAX Optimal Video Summarization," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 10, pp. 1245–1256, Oct. 2005.

[11] C. Panagiotakis, A. Doulamis, and G. Tziritas, "Equivalent Key Frames Selection Based on Iso-Content Principles," IEEE Trans. on Circuits and Systems for Video Technology, Vol. 19, No. 3, pp. 447 - 451, 2009.

[12] C. Yang, M. Dong, and F. Fotouhi, "Image Content Annotation Using Bayesian Framework and Complement Components Analysis," Proc. of ICIP, pp. 1193-1196, Oct. 2005.

[13] S. Yang, S.K. Kim, and Y.M. Ro, "Semantic Home Photo Categorization," IEEE Trans. on Circuits and Systems for Video Technology, Vol. 17, No. 3, pp. 324-335, Mar. 2007.