AUTOMATIC SEMANTIC ANNOTATION FOR VIDEO BLOGS

Xiaoyu Zhang[†], *Changsheng Xu*[‡], *Jian Cheng*[†], *Hanqing Lu*[†], *Songde Ma*[†]

[†] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
{xyzhang, jcheng, luhq, masd}@nlpr.ia.ac.cn

[‡] Institute for Infocomm Research, 119613, Singapore
xucs@i2r.a-star.edu.sg

ABSTRACT

In recent years, weblogs (or blogs) have received great popularity worldwide, among which video blogs (or vlogs) are playing an increasingly important role. As vlogs gain in population, how to make them more easily accessible has become a hot research topic. In this paper, we propose a novel automatic annotation model for vlogs. We extract informative keywords from both the target vlog itself and external resources which are semantically and visually relevant to it. We also present a new evaluation criterion, which assigns a score to an annotation according to its accuracy and completeness in representing the vlog's semantics. Experimental results demonstrate the effectiveness of both the annotation model and evaluation criterion.

Index Terms — Video blog, vlog, multi-label annotation, annotation expansion, evaluation criterion

1. INTRODUCTION

With the rapid development of the World Wide Web, *weblogs* (or *blogs* for short) have emerged as a brand-new communication and publication medium over the past few years. Using blogs, people can describe events, provide opinions and promote conversation with ease. Traditionally, blogging is a purely textual activity because of the predominant use of text. In recent years, with the explosion in the amount of digital multimedia information on the web, new genres of blogs have arisen, among which one of the most popular is *video blog*. According to the online encyclopedia Wikipedia[1], a video blog (shortened to *vlog*) is a blog which uses video as the primary content, often accompanied by supporting text, image, and additional metadata to provide context. Because of the richness of expression of videos, vlogs are much more powerful and compelling than text-only blogs, and thus gain much attention nowadays.

With the number of vlogs growing exponentially, how to effectively index the vlogs and make them more easily searchable has become a challenging problem. In the field of multimedia management, semantic annotation is a promising approach for effective video search; and because manual annotation is labor-intensive and time-consuming, automatic annotation has become the major direction of research efforts. As a result, an effective method for automatic vlog annotation is obviously the key to solving the problem mentioned above. However, to the best of our knowledge, no published research work has so far been directed specifically to vlog annotation.

From the annotation point of view, vlog has the advantage over general video in that vlog is video-centered, which means almost all the textual content in a vlog is used to describe the video. As a result, if we use the text in vlog to annotate the video, the annotation will be more pertinent and less redundant than that of a general video. However, vlog annotation also has its tough side. Because vlogs are created by vloggers from all over the world, it is inevitable that the words used in vlog texts are arbitrary and nonstandard. Therefore, the annotation extracted directly from the vlog text is, in most cases, of low quality, which will consequently jeopardize the performance of vlog search.

Vlog annotation is essentially a *multi-labeling* process [1], as a vlog can usually be annotated with multiple words. There exist many effective approaches for multi-label image/video annotation, and it has become a trend that the annotation should be extracted not only from the target image/video itself, but also from other images/videos which are relevant to it. For example, in [2][3][4], annotation expansion methods are proposed, which perform both text-based and content-based search within a labeled database to find semantically and visually similar images and acquire extra annotations from them.

In this paper, we propose an effective way for automatic vlog annotation. In order to acquire high-quality annotation for a vlog, we first extract intrinsic annotation from the original text of a vlog; then, using external resources, we improve the intrinsic annotation by context-based annotation expansion. Encouraging performance of our solution is achieved from the experiments on our vlog database.

The main contributions of this paper are as follows:

- We utilize the rich web resources and convenient web searchers to facilitate the annotation expansion.
- We propose a novel *context histogram* to describe the semantics of a word in a specific context.
- We define a new score-based evaluation criterion for multi-label annotation problems, which is not merely confined to vlog annotation.

The rest of this paper is organized as follows. In Section 2, we introduce our automatic vlog annotation model in detail. In Section 3, we describe our score-based evaluation criterion for multi-label annotation. Experimental results are reported in Section 4. Finally, we conclude the paper with future work in Section 5.

[1] http://www.wikipedia.org

2. AUTOMATIC VLOG ANNOTATION

In our vlog annotation model, the annotation of a vlog consists of two parts: the intrinsic annotation extracted from the text of the target vlog and the expanded annotation from relevant external resources.

2.1. Intrinsic Annotation Extraction

Since a vlog often has supporting text in itself, we can extract informative keywords as its intrinsic annotation. The textual content in a vlog mainly comprises the title, description, and comments, among which the title and description are closely related to the semantics of the vlog video, while the comments are often filled with irrelevant words and thus too noisy to be used. As a result, only the title and description are kept for annotation extraction.

As the title indicates the main topic of the whole vlog, it is of the greatest importance for understanding the semantics of the vlog. Therefore we first extract annotation words from the title. After stop word removal, important words are reserved in the word set W_{title}.

For the textual description, we also remove the stop words beforehand. Then, using the standard text processing technique such as tf-idf, we can acquire the important words, and create another word set, $W_{description}$. Since in the description not all the words are relevant to the semantics of the central video, $W_{description}$ can be rather noisy. Considering the fact that in an article, keywords are usually used to reveal the main subject, or the title, we assume that if an annotation word is a good one, it should be highly correlated with at least one word in W_{title}. Therefore, we delete from $W_{description}$ the words which have low correlation with all the words in W_{title}.

Existing approaches to measuring word correlation mainly fall into two categories: the lexical and statistical approaches. The lexical approaches are typically based on lexicons such as WordNet, and prove effective for some words. However, as pointed out in [5], lexical approaches suffer from the problem of word ambiguity which can make the word correlation unreliable. Besides, as the number of word correlations is innumerable and ever-growing, there will always be word correlations that are not included in a lexicon. The statistical approaches are data-driven and attempt to discover word correlation based on term cooccurrence, which are more general and flexible than the fixed lexicons. For example, in [6], a Normalized Google Distance (NGD) is proposed to measure the relevance of two words, which leverages the Google [2] search engine to get the words' cooccurrence in the web database. In this paper, we adopt the hypothesis of NGD that the relative frequency of two words appearing in the same documents on the web is a good reflection of their semantic correlation, and define the correlation (or similarity) of two words w_1 and w_2 as follows:

$$\mathrm{Sim}_{word}(w_1, w_2) = \log \frac{\mathrm{n}(w_1, w_2)}{\mathrm{n}(w_1)\,\mathrm{n}(w_2)}, \tag{1}$$

where $\mathrm{n}(w_1)$ is the number of pages a Google search for w_1 returns, while $\mathrm{n}(w_1, w_2)$ is the number of pages returned with both w_1 and w_2 submitted as a query. The larger the value of Sim_{word}, the more relevant the two words are on semantics. Moreover, in order to make the similarity measurement more suitable for vlog words, we use Google Video [3] searcher instead to acquire the returned page numbers.

After annotation extraction and irrelevant word removal, we merge word sets W_{title} and $W_{description}$, and obtain the intrinsic annotation $W_{intrinsic}$:

$$W_{intrinsic} = W_{title} \bigcup W_{description}\,. \tag{2}$$

2.2. Context-Based Annotation Expansion

As mentioned above, the intrinsic annotation is far from enough to represent the vlog's semantics sufficiently. In order to get a high-quality annotation, we perform annotation expansion based on the specific context of the vlog.

2.2.1. External annotation candidate extraction

Inspired by the search-based annotation methods in [2][3][4], we conduct annotation expansion for the target vlog through a search-based mode, where a labeled database is indispensable. As we know, YouTube [4] is one of the most popular video sharing websites which has by far the biggest collection of videos. Each video on YouTube is labeled by one or more tags. Therefore, we use YouTube as our labeled database.

Given a keyword query, the text-based video search engine (powered by Google) in YouTube can return rather good results, hence we can use YouTube search to find the semantically related videos. For the target vlog, we submit each word w in $W_{intrinsic}$ as a query to YouTube searcher, and get the corresponding search results R_w (for simplicity, only the top-ranked 20 results are included). For each result r in R_w, we extract the video's representative frame f_r (which is usually the first frame of the video) and the corresponding tags.

Then, among the semantically related videos, visually related ones are selected through content-based similarity between the vlog video and the result videos found on YouTube. We define the visual similarity between a result video r and the vlog video v as the maximum image similarity between the representative frame f_r of r and the keyframe f_v of v:

$$\mathrm{Sim}_{video}(r, v) = \max_{f_v \in F_{key}(v)} \mathrm{Sim}_{image}(f_r, f_v)\,, \tag{3}$$

where $F_{key}(v)$ is the keyframe set of the vlog video v.

After the above two search stages, we have obtained a batch of videos which are relevant to the vlog both semantically and visually with regard to the intrinsic annotation word w. We then gather the tags of all the reserved videos into a tag set $T(w)$, which is adopted as the external annotation candidates for the vlog.

This process is applied for each intrinsic annotation word w in $W_{intrinsic}$. Finally, we obtain the word set $W_{external}$ for external annotation candidates:

$$W_{external} = \bigcup_{w \in W_{intrinsic}} T(w)\,. \tag{4}$$

2.2.2. Context-based annotation refinement

Although the videos used for annotation expansion are all semantically and visually relevant to the target vlog, it dose not follow that all the tags of the videos are also relevant to the vlog. In

[2] http://www.google.com

[3] http://video.google.com

[4] http://www.youtube.com

the process of annotation expansion, we have to deal with the serious problem of semantic drift. Therefore, we should refine the expanded annotation candidates and delete the irrelevant words. We calculate the relevance between an annotation candidate c and the vlog by comparing c with the words in $W_{intrinsic}$.

As we know, when comparing two words, we should consider not only the semantics in them but also the specific contexts they are in. In this paper, we propose a novel *context histogram* to depict the semantics of a word in a specific context. For a word w, its context is substantially a set of words which confines its specific semantics. We first calculate the one-to-one correlation between w and each of the words in its context $W_{context}$. Then, we organize all the correlation values into a histogram and get the context histogram for w with respect to $W_{context}$ (as illustrated in Figure 1). The problem of context comparison is now reduced to histogram comparison. Here we simply use histogram intersection as a metric of the context histogram similarity.

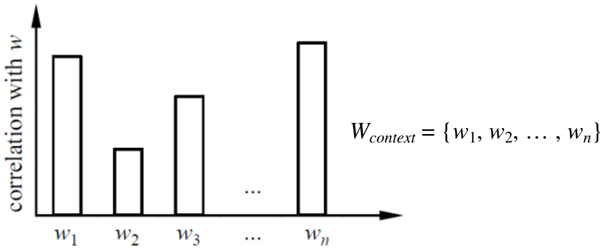

Figure 1. Context histogram of *w* with respect to *W*

We perform the context-based external annotation refinement as follows: For an intrinsic annotation word w of $W_{intrinsic}$, we create its context histogram with respect to $W_{context} = W_{intrinsic} - \{w\}$; while for an annotation candidate c in $W_{external}$, we also build its context histogram with respect to the same $W_{context} = W_{intrinsic} - \{w\}$. In order to compare c with w, we calculate both their one-to-one word correlation Sim_{word} and their contextual similarity $\mathrm{Sim}_{context}$. The total correlation between c and w is defined as:

$$\mathrm{Sim}_{total}(c,w) = \alpha \mathrm{Sim}_{word}(c,w) + \beta \mathrm{Sim}_{context}(c,w)\,, \qquad (5)$$

where α and β are adjustable parameters. Then, we calculate the relevance between c and $W_{intrinsic}$ as:

$$\mathrm{Rel}(c, W_{intrinsic}) = \max_{w \in W_{intrinsic}} \mathrm{Sim}_{total}(c,w)\,. \qquad (6)$$

Only those annotation candidates with high relevance to $W_{intrinsic}$ are kept in $W_{external}$.

After the refinement, we merge $W_{intrinsic}$ and $W_{external}$ to get the final annotation for the target vlog:

$$W_{final} = W_{intrinsic} \cup W_{external}\,. \qquad (7)$$

3. EVALUATION CRITERIA

Since no ground truth is available, effective criteria are needed to evaluate the annotation results. The search-based criteria are widely adopted, which use the performance in the search stage to evaluate the quality of an annotation. There are also score-based criteria, which make judgment directly on the annotation. For example, in [2], three types of predicted annotation words, “perfect”, “correct”, and “wrong”, were defined. Each type was assigned with a corresponding score (1, 0.5, and -1). Then, the evaluation score of a given annotation was calculated as:

$$E = (p + 0.5r - w)/N\,, \qquad (8)$$

where N denotes the number of words in the annotation, and p, r, w are the number of “perfect”, “correct”, and “wrong” words respectively. This score-based criterion is widely used for the evaluation of multi-label annotations.

However, we believe that for a high-quality annotation, two key aspects should both be placed emphasis on, which we call *accuracy* and *completeness*. The accuracy measures the relevance between the annotation words and the vlog’s semantics, that is, how accurately the words reflect the vlog’s content; while the completeness reflects how completely the annotation depicts the volg. For example, for a video clip from the movie “Titanic” which depicts Titanic's sinking, the annotation W_1 = “Titanic” is very accurate but not complete; while another annotation W_2 = “Titanic, crash, sink, iceberg, USA” seems to be better, although the word “USA” is only a “correct” one. The evaluation score mentioned above is biased to accuracy, and is inclined to attach higher scores to those annotations with fewer words even if they cover only a small part of the whole semantics of the vlog. So it will evaluate 1 (the highest score) for W_1 (p = 1, r = 0, w = 0, N = 1), but 0.9 for W_2 (p = 4, r = 1, w = 0, N = 5).

In this paper, we propose a new score-based evaluation criterion which takes both accuracy and completeness into account. We assume that the presence of a “perfect” word in an annotation will contribute 1 score; while the absence of a “perfect” word will degrade the annotation, and thus we punish the missing “perfect” word by attaching score -1. Similarly, a present “wrong” word gets -1, while an absent “wrong” word 1. For a “correct” word, since it is only partially relevant to the vlog, we would rather it not be included in the annotation lest it should lead to semantic drift. Thus we assign 0.5 to a “correct” word’s presence, and 1 for its absence. The scoring strategy designed for both the presence and absence of the “perfect”, “correct”, and “wrong” words is summarized in Table 1.

Table 1. The scoring strategy for annotation word.

	perfect	correct	wrong
present	1	0.5	-1
absent	-1	1	1

Using this improved word scoring strategy, we create our annotation evaluation score. Suppose we use K different annotation methods to annotate the same vlog and get K corresponding annotations: W_1, W_2,…, W_K. In order to compare the K methods, we first merge all the annotations into a single word set W_{all}:

$$W_{all} = W_1 \cup W_2 \cup \ldots \cup W_K\,. \qquad (9)$$

Thus W_{all} includes annotation words from all the methods. Then we identify the “perfect”, “correct”, and “wrong” words in W_{all}, and calculate the evaluation score $E_{a\&c}$ for each annotation as follows:

$$E_{a\&c} = \frac{p_{present} - p_{absent} + 0.5r_{present} + r_{absent} - w_{present} + w_{absent}}{|W_{all}|}\,, \qquad (10)$$

where $|W_{all}|$ is the total number of annotation words in W_{all}. $p_{present}$, $r_{present}$, and $w_{present}$ are the number of “perfect”, “correct”, and “wrong” words existing in the annotation; while p_{absent}, r_{absent}, and w_{absent} are the number of the three types of words in W_{all} but not included in the annotation. Again, take the Titanic video clip mentioned above for example, under our evaluation criterion, W_1

($p_{present} = 1$, $p_{absent} = 3$, $r_{absent} = 1$, $r_{present} = w_{present} = w_{absent} = 0$, $|W_{all}| = 5$) gets -0.2, while W_2 ($p_{present} = 4$, $r_{present} = 1$, $p_{absent} = r_{absent} = w_{present} = w_{absent} = 0$, $|W_{all}| = 5$) gets 0.9.

Compared with the previous evaluation criterion, ours is more reasonable in that it offers a good balance between accuracy and completeness. It is a generic criterion for evaluating multi-label annotation, and can also be used to evaluate image and general video annotation besides vlog annotation.

4. EXPERIMENTS

We have built a vlog database, which contains 1000 vlogs obtained from the web or submitted by users. To add a vlog to the database, the user only has to provide a video clip and the corresponding textual content (title and description). Automatic vlog annotation will then be conducted off line.

After the automatic annotation, each word will be judged manually as “perfect”, “correct”, or “wrong”. Then, according to the evaluation criterion introduced in Section 3, a score $E_{a\&c}$ is calculated. We compare the average evaluation scores after each stage of our annotation model, which are listed in Table 2 (where $W_{candidate}$ stands for the word set $W_{external}$ before refinement).

Table 2. The average evaluation scores after each stage.

	$W_{intrinsic}$	$W_{intrinsic} \cup W_{candidate}$	W_{final}
$E_{a\&c}$	0.41	0.45	0.72

As indicated in Table 2, the intrinsic annotations are, unsurprisingly, of low quality. By external annotation extraction, the scores can be improved slightly. This is because the external annotation candidates do bring in some “perfect” words, but there are also “correct” and “wrong” ones, which incur punishment in the scores. After the context-based refinement, the evaluation scores increase remarkably, as most of the irrelevant and partially irrelevant words are filtered out.

We also test the effectiveness of our annotation model through search. Given a query word, the system returns all the vlogs that are annotated with the query word, and ranks them in descending order according to their evaluation score $E_{a\&c}$. “P@m” [7], the precision for the top-ranked m search results, is calculated for 50 randomly selected queries, and then the average P@m values (m = 5, 10, 15, 20) based-on the annotations after each stage of our annotation model are illustrated Figure 2.

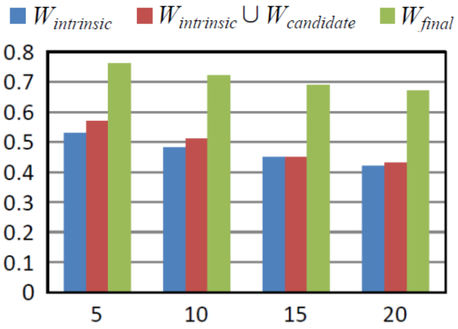

Figure 2. Average P@*m* over 50 queries.

From Figure 2, we can see that the search results based-on intrinsic annotations are not quite satisfactory. With the external annotation candidates appended, the results are improved slightly; while the refinement in annotation boosts the search performance greatly. The result in Figure 2 accords with that in Table 2, which implicitly proves the validity of our score-based evaluation criterion.

5. CONCLUSION

In this paper, we have proposed an automatic annotation model for video blogs. In order to guarantee high quality of a vlog’s annotation, we extract informative keywords not only from the textual content of the target vlog but also from external recourses which are semantically and visually relevant to it, then context-based refinement is performed to further improve the annotation. We have also defined a new score-based evaluation criterion, which reflects both the accuracy and completeness of an annotation. Experimental results demonstrate the effectiveness of the proposed vlog annotation model and evaluation criterion.

In the future, we will try to make better use of the visual contents of vlogs to further improve the annotation performance, since in our current model more emphasis is placed on the text aspect. Some advanced techniques in the field of natural language processing and data mining can also be applied to obtain higher-quality keywords. Furthermore, besides vlog annotation, we are interested in other fields of vlog management, such as automatic vlog categorization and personalized vlog search.

6. ACKNOWLEDGMENT

This work is supported by the National Hi-Tech Research and Development Program of China (863 Program, Grant No. 2006AA01Z315), and the National Natural Science Foundation of China (Grant No. 60475010).

7. REFERENCES

[1] G.J. Qi, X.S. Hua, Y. Rui, J. Tang, T. Mei, H.J. Zhang, “Correlative Multi-Label Video Annotation”, Proceedings of the 15th International Conference on Multimedia, ACM, USA, pp. 17-26, 2007.

[2] X.J. Wang, L. Zhang, F. Jing, W.Y. Ma, “AnnoSearch: Image Auto-Annotation by Search”, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, USA, pp. 1483-1490, 2006.

[3] C. Wang, F. Jing, L. Zhang, H.J. Zhang, “Scalable Search-Based Image Annotation of Personal Images”, Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, ACM, USA, pp. 269-278, 2006.

[4] X. Rui, M. Li, Z. Li, W.Y. Ma, N. Yu, “Bipartite Graph Reinforcement Model for Web Image Annotation”, Proceedings of the 15th International Conference on Multimedia, ACM, USA, pp. 585-594, 2007.

[5] A. Natsev, A. Haubold, J. Tesic, L. Xie, R. Yan, “Semantic Concept-Based Query Expansion and Re-ranking for Multimedia Retrieval”, Proceedings of the 15th International Conference on Multimedia, ACM, USA, pp. 991-1000, 2007.

[6] R. Cilibrasi, P. Vitanyi, “Automatic Extraction of Meaning from the Web”, Proceedings of IEEE International Symposium on Information Theory, USA, pp. 2309-2313, 2006.

[7] J. Liu, B. Wang, M. Li, Z. Li, W.Y. Ma, H. Lu, S. Ma, “Dual Cross-Media Relevance Model for Image Annotation”, Proceedings of the 15th International Conference on Multimedia, ACM, USA, pp. 605-614, 2007.