LATENT SEMANTIC RETRIEVAL OF PERSONAL PHOTOS WITH SPARSE USER ANNOTATION BY FUSED IMAGE/SPEECH/TEXT FEATURES

Yi-sheng Fu[1], *Chia-yu Wan*[2] *and Lin-shan Lee*[1,2]

[1]Graduate Institute of Computer Science and Information Engineering, National Taiwan University
[2]Graduate Institute of Communication Engineering, National Taiwan University

mayaplus@speech.ee.ntu.edu.tw, d94942018@ntu.edu.tw,lslee@gate.sinica.edu.tw

ABSTRACT

While users prefer high-level semantic photo descriptions (e.g., who, what, when, where), we wish to minimize the need to annotate photos using such descriptions by the user. We propose a latent semantic personal photo retrieval approach using fused image/speech/text features. We use low-level image features to derive relatoionships among sparsely annotated photos, and probabilistic latent semantic analysis (PLSA) models based on fused image/speech/text features to analyze photo "topics". We then retrieve the photos using text or speech queries of simple high-level semantic words only. In preliminary experiments, while only 10% of the photos were manually annotated, the photos could be well retrieved with very encouraging results.

Index Terms: image retrieval, semantic analysis, latent topics, fused features

1. INTRODUCTION AND PROBLEM OVERVIEW

With the growing popularity of digital cameras, many people have saved huge collections of personal photos. A resulting challenge is how to browse across the huge collection and exactly find a desired photo. This calls for an efficient photo retrieval approach.

Content-based image retrieval has been an active research area for years, with many successful approaches based on low-level image features implemented with "query by example" [1,2] or similar. However, this is not very attractive in practice, because it requires that the user provides an example photo as the query. In fact, most users prefer high-level semantic descriptions of photos that use words as indices or queries, such as who, where, when, what (objects/events) and so on, but again, this is not an attractive solution if it requires manual annotation of each individual photo. This observation has led to the idea of annotating photos with speech [3,4]. When such a spoken photo annotation is taken as a spoken document, the problem becomes one of spoken document retrieval.

Many spoken document retrieval approaches have been successful[5,6], but these approaches usually suffer from the problem of word usage diversity, i.e., the query and its relevant documents may use different sets of words. This problem is especially serious for photos, because the annotation may describe location (where), but the user may look for a person (who), i.e., both annotation and query are typically free-form and vary significantly. Semantic matching strategies have been developed to solve this problem by discovering latent topics inherent in the query and documents with latent semantic indexing (LSI) and probabilistic latent semantic analysis (PLSA) as two typical examples [7,8]. In both cases the relevance score between a query term and the spoken documents can be obtained via a set of latent topics, and relevant documents can be retrieved even using query terms that are completely different from those used in the documents [9]. This is because common topics are usually found in sets of documents that each includes a set of similar terms, or in sets of terms that each appears in a set of similar documents, and such topical information is used in retrieval. Recent image retrieval works also adopted the idea of semantic topics [10].

The above semantic matching methods have not solved the photo retrieval problem described here either. Assume that photo annotation can be formulated into six categories: *who*, what (*object* or *event*), *when*, *where*, and *others*. When labeling a photo, users typically select only one or two categories. As such, related photos may not be labeled by similar terms (e.g., some by *where* and some by *who*), and the relationships among terms in different categories cannot be trained using latent topics. For example, given a *where* query or category, many photos taken at that location may not be collected if they are annotated with words in other categories. In other words, the above six categories of labels are orthogonal, but user annotations are usually very sparse, and personal photos users generally annotate far too few photos to train such topic models. Thus the problem is quite different from the spoken document retrieval problem, even if photos have spoken annotations.

Here we propose a user-friendly latent semantic retrieval approach for personal photos with sparse annotation using fused image/speech/text features. We use low-level image features to derive the relationships among photos, since these features are really the universal language describing photos. However, we train semantic models with PLSA using fused image/speech/text features to analyze the topics of these photos. In PLSA, "terms" are discrete, while low-level image features are continuous. Thus for each given photo we use low-level image features to select as its "image terms" those groups of "cohort photos" with similar image characteristics. We then fuse these image characteristics with the speech/text features of available user annotations. Speech/text annotations can be very sparse, that is, only a few words regarding semantics (e.g., *who* or *where*) are needed for only a small subset of the photos. The sparse text/speech annotations serve as user interface for the whole photo archive, since other photos that have not been annotated are automatically related by fused feature semantics using PLSA.

2. THE PROPOSED APPROACH

2.1. Overview of the proposed approach

As shown in Fig. 1, the proposed approach includes a preparation phase (left part) and a retrieval phase (right part). The low level image features are first extracted and used to select the "cohort photos" (Blocks (B) and (C), lower left to middle of the figure) for each

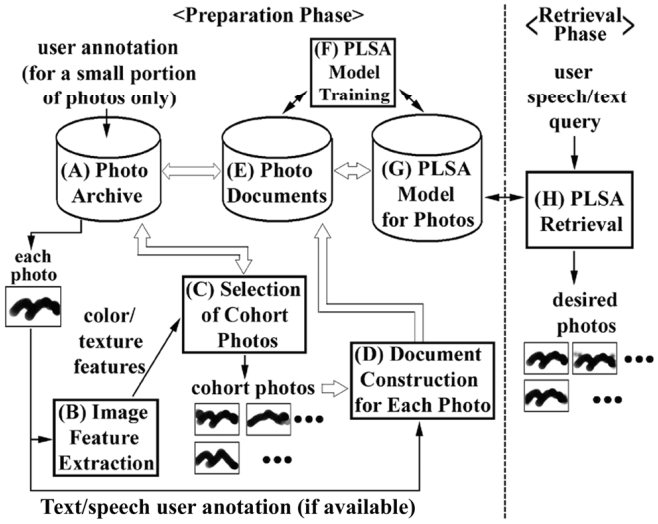

Fig. 1. The proposed approach: preparation phase includes document construction for each photo and PLSA model training for photo documents, while retrieval phase is based on PLSA.

photo in the photo archive (Block (A), upper left corner). The cohort photos, used as discrete "image terms", together with the speech/text annotation by the user as "text/speech terms", if available, are then fused to construct a "document" for each photo (Block (D), lower middle). These "documents" and their "terms" are then used to train the PLSA topic model (Blocks (E)(F)(G), upper middle). The user query then includes only very few semantic words, in either speech or text form. PLSA finally gives the desired photos.

2.2. Probabilistic latent semantic analysis (PLSA)

Probabilistic latent semantic analysis (PLSA) uses a set of latent topic variables, $\{z_k, k = 1, 2, \ldots, K\}$, to characterize the term-document co-occurrence relationships [6] given a set of terms, $\{t_j, j = 1, 2, \ldots, M\}$, and a set of documents $d_i, \{d_i, i = 1, 2, \ldots, N\}$, assuming document d_i and term t_j are both independetly conditioned on an associated latent topic z_k. The joint probability of this observed pair (d_i,t_j) is then expressed by the following equations:

$$P(t_j, d_i) = P(d_i)P(t_j|d_i), \tag{1}$$

$$P(t_j|d_i) = \sum_{k=1}^{K} P(t_j|z_k)P(z_k|d_i). \tag{2}$$

All PLSA model parameters can be trained using EM algorithm by maximizing a total likelihood function. With the latent topic variables, retrieval can then be based on topics rather than terms, so topically relevant documents can be retrieved even using different sets of terms. For photos, however, topics clearly have to do with the scene, but image features are represented using real numbers, while the terms in PLSA have to be discrete. That is why we use image features to select cohort photos with similar image characteristics, and use them as discrete "image terms" in PLSA, as we given below.

2.3. Color features from the images

Color histogram popularly used in image retrieval is adopted here [11]. Each photo n can be represented by a color histogram H_n, in which each entry $H_n(i)$ is the number of pixels belonging to the color bin i. The HSV color space is quantized into 166 colors, including 18 levels of hues (H)* 3 levels of saturation (S)* 3 levels of values (V) + 4 levels of grays[11]. The distance $d_{n,l}$ between two photos n and l is then defined by the L2 distance measure,

$$d_{k,l} = \sum_{i=0}^{N-1} (H_n(i) - H_l(i))^2, \tag{3}$$

where N=166 here.

2.4. Texture features from images

The Gabor texture features previously proposed and frequently used for image analysis, produced by a bank of Gabor filters at multiple scales and orientation [12], are adopted here including four scales and six orientations.

2.5. Cohort photos selected using image features

For PLSA modeling, we need "discrete terms" describing the scene characteristics of each photo, but the above color and texture features do not directly translate to such "discrete terms". So we use each individual photo in the archive as a discrete "image term", and use the above color and texture features to select photos with similar scene characteristics for each photo, referred to as "Cohort Photos", to be used as discrete "image terms" for that photo, as explained in the next section.

Here we discuss how these "Cohort Photos" are selected. The L2 distance in Eq. (3) is used as the distance measure not only for color features in Section 2.3 but also for texture features in Section 2.4. We use a total of three methods to select cohort photos. The first method is based simply on the combination of ranks (i.e., the closest top photos) with respect to color and texture features. In the second and third methods, we use one set of features (color or texture) to select the top 10% photos as the candidates, and then use the other set of features to re-score (or re-rank) the selected photos. These three methods are actually complementary to each other, so they jointly generate three sets of cohort photos for each given photo, to be used to construct the photo documents as presented below.

2.6. Construction of photo documents with fused features

Each photo in the archive is represented as a document consisting of discrete terms for PLSA modeling. We first define every photo in the archive as a discrete "image term", and then we further represent each photo as a document composed of the "image terms" for all of its cohort photos selected as described above using color and texture features. These terms jointly describe the image and scene characteristics of each photo. We use three methods to extract cohort photos based on image similarity as in Section 2.5. For each method, the "image terms" for the top most similar photos are included in the document for the given photo. When the same photo appears in more than one of the three top lists, the counts are simply used as term frequencies for the "image terms". These are shown in the right half of Fig. 2.

On the other hand, the speech/text annotation for a given photo (if any) is also included in its document. This is straightforward: we simply define word, character, syllable and bi-syllabic patterns

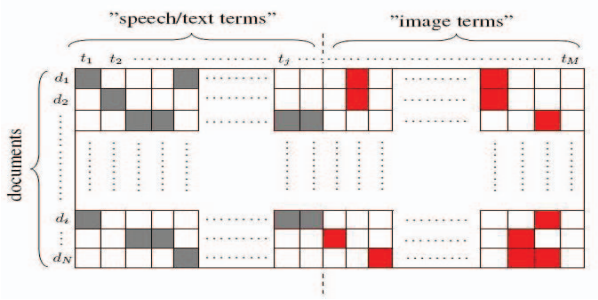

Fig. 2. The co-occurrence matrix used in PLSA model training. The image/speech/text information are all represented by discrete terms

as "speech/text terms" (the annotation is in Mandarin Chinese) for word- and subword-level indexing as in conventional spoken document retrieval. The subword units (character and syllable) are used to handle out-of-vocabulary (OOV) words as usual in spoken document retrieval. Since OOV words are not in the vocabulary and cannot be correctly recognized, they cause serious problems in retrieval. Use of subword units (characters and syllables here) can offer some help to this problem. For speech annotation, utterances are represented in word- and subword-based lattices and all arcs of the lattices are included as "speech terms". These word and subword terms in the lattices are given less weight in PLSA training, in order to reduce interference from noisy word/subwords, but still add indexing functionality if these terms appear in the lattices. These are shown in the left half of 2. In this way, we construct documents for all photos with fused image/speech/text features.

2.7. Latent semantic retrieval with fused image/speech/text features

The PLSA model is then trained with the constructed documents based on fused image/speech/text features. Because few photos are annotated, the obtained latent topics are based primarily on image semantics, i.e., photos of the same latent topic look similar. The input query can be in either speech or text form, represented as a sequence of observed word- or subword-based terms, and the relevance score with respect to each photo is then calculated as usual, without using any image term. Note that there are four types of terms in each photo document: image terms, word terms, character terms, and syllable-based terms. For unannotated photos, the latter three types of terms are simply blank. The central idea of PLSA-based latent semantic retrieval is that a query and a document may have a high relevance score even if they do not share any terms in common, as long as they share the same latent topic.

More precisely, a speech/text query Q is treated as a sequence of n observed terms, $Q = \hat{t_1}, \hat{t_2}, \cdots \hat{t_j} \cdots \hat{t_n}$. The photos or documents are then sorted by the relevant score $P(Q|d_i)$,

$$P(Q|d_i) = P(\hat{t_1}|d_i)P(\hat{t_2}|d_i)..P(\hat{t_j}|d_i)..P(\hat{t_n}|d_i), \quad (4)$$

$$P(\hat{t_j}|d_i) = \sum_{k=1}^{K} P(\hat{t_j}|z_k)P(z_k|d_i), \quad (5)$$

where the probabilities $P(\hat{t_j}|z_k)$ and $P(z_k|d_i)$ are obtained from the PLSA model. In this way, unannotated photos that have no terms in common with the text/speech query (since the query contains only word/character/syllable terms) can also be retrieved, because the matching is not based on term co-occurrences but on latent topics.

Fig. 3. Retrieved photos by the text query "Place de la Concorde" (in Chinese) for 10% speech annotation case. Only the photo ranked 9-th has the speech annotation, and only the photos ranked 3rd and 5-th are incorrect.

3. PRELIMINARY EXPERIMENTAL RESULTS

3.1. Experimental setup

In the preliminary experiments, an archive of 1429 photos for trips of several students to several locations in America and Europe was used. Only 50% of these photos were annotated by the users with text labels and 20% with speech labels. In such labels each photo was annotated by only one of the six categories: *who*, what (*object* or *event*), *when*, *where*, and *others*. Each annotation includes 1 to 3 Chinese words or 2 to 6 Chinese characters (or syllables, since in Mandarin Chinese all characters are produced as a monosyllable). The recognition accuracies for the speech annotation are relatively low, 77.2%, 72.9% and 65.7% for syllables, characters and words, respectively, apparently because of the OOV problem (since 40.1% of the speech annotation include OOV words) and the spontaneous nature of the speech annotation. In the experiments we assume 10%, 20%, 30%, 40% and 50% of photos are annotated by text labels, or 10%, 20% by speech labels by randomly deleting parts of the annotations. We assume text and speech annotations never exist jointly. Two users who contributed the photos participated in the test. They together generated 28 text queries (including 13 "*where*" queries, 11 "*who*" queries and 4 "*object*" queries). Each query includes 1 to 3 Chinese words, or 2 to 6 Chinese characters (or syllables). 8 out of the 28 queries (28.5%) include OOV words, too. For each query, the system displayed a ranked list of the retrieved photos. The users were asked to identify relevant photos he or she recognized in the top n photos along the given list, in which n ranged from 10 to 50 with interval of 10. The precision rates were then averaged.

Top N photos	Averaged precision						
	Speech annotation		Text annotation				
	(a)10%	(b)20%	(c)10%	(d)20%	(e)30%	(f)40%	(g)50%
(1)top 10 photos	0.467	0.445	0.500	0.481	0.461	0.500	0.514
(2)top 20 photos	0.403	0.410	0.432	0.429	0.429	0.452	0.455
(3)top 30 photos	0.386	0.376	0.413	0.387	0.411	0.414	0.419
(4)top 40 photos	0.365	0.358	0.391	0.370	0.387	0.394	0.403
(5)top 50 photos	0.361	0.363	0.386	0.350	0.378	0.379	0.389

Table 1. Averaged precision of top N photos

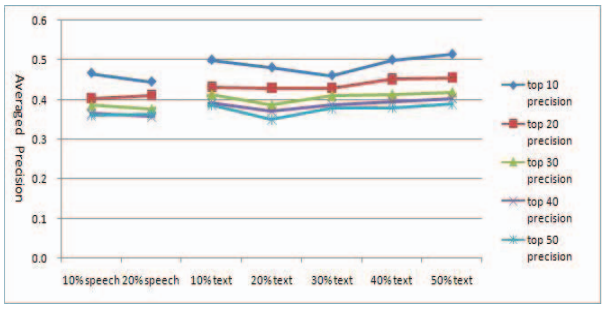

Fig. 4. averaged precision of top n photos retrieved

3.2. Photo retrieval results

Fig. 3 shows one example of the first 9 photos retrieved by the text query "Place de la Concorde (in Chinese)" in the experiment of 10% of speech annotation. Only the two photos retrieved at ranks 3 and 5 are incorrect, all others are correct. In fact here only the photo ranked 9 was annotated with a speech label "Place de la Concorde (4 syllables in Mandarin Chinese)." This is an OOV word and cannot be correctly recognized. As a result, the word and the 4 corresponding characters are all incorrectly recognized, although 3 out of the 4 syllables are correctly recognized. The 3 correctly recognized syllables explained why many related photos were actually correctly retrieved, definitely because of the fused image/speech/text features and the semantic topic of PLSA. The two incorrectly retrieved photos of ranks 3 and 5 were actually taken at a location different from "Place de la Concorde", but probably only the user himself can identify such difference. This is why the performance of semantic retrieval of personal photos is difficult to evaluate, because very often only the users themselves can determine whether a photo is relevant or not. As another example, the query "sun rise" may retrieve many photos of "sun set," while only the user knows which one is which. This is different from the task of "query by example" retrieval system, in which the relevant images are simply those close to the query example.

Table 1 summarizes the complete average precision results. The left part (columns (a)(b)) are for 10% and 20% of speech annotation/ while the right part (columns (c)-(g)) are for 10%-50% of text annotation. The five rows (1)-(5) are respectively for the top n photos, $n = 10, 20, 30, 40, 50$. For example, for 10% speech annotation (column (a)) the averaged precision of top 20 photos is 0.403, but for 10% text annotation (column (c)), the averaged precision is 0.432. The results in Table 1 are also plotted in Fig. 4 for different percentages of speech/text annotations. From Table 1 and Fig. 4, a very important observation is that the retrieval accuracy is only slightly dependent on the percentage of user annotation. For example, in Fig. 4 the degradation in precision for text annotation percentage reduced from 50% to 10% is only very limited. In fact, the precision for 10% and 20% of text annotation is very close. This verifies the power of latent semantic retrieval with fused image/speech/text features. The strong latent topic relationships and the integration of different types of features make the user annotation less critical.

On the other hand, the performance degraded with speech annotation as compared to text annotation. But it is interesting that the degradation was even more serious for 20% annotation in most cases. Clearly this had to do with the relatively low recognition accuracies. More annotation very probably introduced more noise than more information in both PLSA model training and retrieval. But this also indicated that very sparse annotation is fine for the approach proposed here. This may not be ture if the recognition accuracy is higher.

4. CONCLUSION

In this paper, we propose a new approach of latent semantic retrieval of personal photos with sparse user annotation using fused image/speech/text features. The approach is user friendly, because only very simple annotation is needed for small portion of photos, while all photos can be retrieved based on high level semantics.

5. REFERENCES

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and B. Dom, "Query by image and video content: the QBIC system," IEEE Computer, Sep. 1995.

[2] John R. Smith, Shih-Fu Chang, "VisualSEEk: a fully automated content-based image query system," ACM Multimedia 1996.

[3] J. Chen, T. Tan, P. Mulhem, and M. Kankanhalli, "An improved method for image retrieval using speech annotation," Proceedings of the 9th International Conference on Multi-Media Modeling 2003.

[4] Timothy J. Hazen, Brennan Sherry and Mark Adler, "Speech-based annotation and retrieval of digital photographs," Interspeech 2007.

[5] C. Chelba, J. Silva and A. Acero," Soft indexing of speech content for speech in spoken documents," Computer Speech and Language, vol. 21, no. 3, pp.458-478, July 2007.

[6] Yi-chen Pan, Hung-lin Chang and Lin-shan Lee, "Analytical comparison between position specific posterior lattices and confusion networks based on words and subword units for spoken document indexing," Automatic Speech Recognition & Understanding, pp.677-682 , Dec 2007.

[7] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. Harshman, L.A. Streeter, and K.E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure," Proc. ACM SIGIR Conf. R&D in Information Retrieval, 1988.

[8] T. Hofmann, "Probabilistic latent semantic indexing," Proc. ACM SIGIR Conf. R&D in Informational Retrieval, 1999.

[9] Ya-chao Hsieh, Yu-tsun Huang, Chien-chih Wang and Lin-shan Lee, "Improved spoken document retrieval with dynamic key term lexicon and probabilistic latent semantic analysis(PLSA)," ICASSP 2006, vol. 1, May 2006.

[10] Yi-Hsuan Yang, Po-Tun Wu, Ching-Wei Lee, Kuan-Hung Lin, Winston H. Hsu, "ContextSeer: Context Search and Recommendation at Query Time for Shared Consumer Photos," ACM Multimedia 2008(full paper), Vancouver, Canada.

[11] M. J. Swain and D. H. Ballard, "Color indexing," Int. Journal of Computer Vision, 1991.

[12] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," IEEE T-PAMI, Aug. 1996.