SPEECH-BASED VISUAL CONCEPT LEARNING USING WORDNET

Xiaodan Song,[*] *Ching-Yung Lin,*[**] *Ming-Ting Sun*[*]

*: Dept. of Electrical Engineering, Univ. of Washington, Seattle, WA 98195 {song, sun}@ee.washington.edu

**: IBM T. J. Watson Research Center, Hawthorne, NY 10532, chingyung@us.ibm.com

ABSTRACT

Modeling visual concepts using supervised or unsupervised machine learning approaches are becoming increasing important for video semantic indexing, retrieval, and filtering applications. Naturally, videos include multimodality data such as audio, speech, visual and text, which are combined to infer therein the overall semantic concepts. However, in the literature, most researches were conducted within only one single domain. In this paper we propose an unsupervised technique that builds context-independent keyword lists for desired visual concept modeling using WordNet. Furthermore, we propose an Extended Speech-based Visual Concept (ESVC) model to reorder and extend the above keyword lists by supervised learning based on multimodality annotation. Experimental results show that the context-independent models can achieve comparable performance compared to conventional supervised learning algorithms, and the ESVC model achieves about 53% and 28.4% improvement in two testing subsets of the TRECVID 2003 corpus over a state-of-the-art speech-based video concept detection algorithm.

1. INTRODUCTION

As the amount of multimedia data increases, content-based image/video indexing, filtering, and retrieval are becoming increasingly important. Supervised machine learning methods have shown its effectiveness on modeling generic models to address these issues. Although they have the best performance in the NIST TREC concept detection benchmarking (2002-2004), a huge amount of work is required to manually label the training contents [1]. Even with an enormous labeling effort, the assumption of the similarity between the training and the testing data made by these supervised learning schemes limits the capability of generalization for the system. When the dataset changes, we have to reassign the training data and redo the tedious labeling work. In many practical applications, it is desirable if we have an autonomous learning scheme without any supervision.

The correspondence between the video and the speech data provides possibility to achieve this goal. In the automatic image indexing area, [10] attempts to discover the statistical links between the visual features and the words by estimating the joint distribution of the words and the regional image features, and posing annotation as statistical inference in a graphical model. Many people have been following that direction and obtained some promising results [11]. However, training data with manual labeling is still needed to learn the joint models. In our previous work, we showed the possibility of using cross-modality data to achieve autonomous learning [2]-[4]. We developed an autonomous concept learning approach which uses the "imperfect" association between the visual content and audio/text data in video sequences or images to automatically train the concept models. In order to achieve the above goal of autonomous learning, we used an unsupervised learning mechanism to first use speech information to detect visual semantic, and then use Generalized Multiple Instance Learning (GMIL) to refine the concept models explicitly. In the previous systems, we used existing tools for the speech-based concept detection.

In this paper, we develop techniques for an unsupervised keyword expansion, and extend the supervised keyword expansion, for the speech-based visual concept detection. We learn the speech-based visual concept models using a dictionary, e.g., WordNet, which has been used for knowledge discovery, keyword expansion, and disambiguate word senses in supervised learning schemes [12]. We show our techniques can achieve successful performance.

It has been found that speech cues are helpful for detecting visual-based concepts [5][6]. In TRECVID 2003, H. Nock *et al*. [5][6] proposed a procedure for semantic concepts retrieval using speech information alone. For nine of the concepts, the speech-based modeling can achieve better performance than visual-based approaches. They use supervised learning method to train visual concept models using the annotated terms as the concept labels, and the words in a window of shots as the keywords. Models are represented by keyword terms and their frequencies. Then, manual work is done to select terms to generate the keyword list. Retrieval of the shots containing the interested concept then proceeds by ranking the shots against the keyword list according to their

0-7803-9332-5/05/$20.00 ©2005 IEEE

OKAPI [7] scores. This is a conventional supervised learning framework, where the manual labeling work is necessary for obtaining the annotation of each shot. The manual selection for refining the keyword list is another obstacle for the scalability of this scheme.

In this paper, inspired by learning from dictionaries as people usually do, we propose to automatically generate the speech-based visual concept keyword list by WordNet [8], a large scale semantic database for lexical organization. Paradigmatic relationships between entries such as hyponymy, antonymy, and polysemy are covered in WordNet. In some scenarios, this method generates reasonable results. In some cases, however, this method may be too general to adapt to different video sequences, and the performance is not as good as the one learnt by supervised learning from similar context. Therefore, we also experiment some methods to combine this unsupervised approach with prior supervised methods. We propose an Extended Speech-based Visual Concept (ESVC) model which utilizes both keyword lists generated by supervised learning and WordNet (Figure 1).

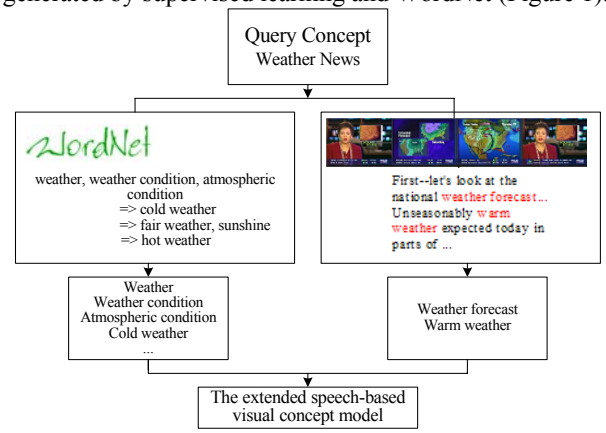

Figure 1. Illustration of combining word lists generated from WordNet and supervised learning to generate an extended speech-based visual concept model

The main contributions of this paper are summarized as follows.

1) We build the keyword lists used for modeling speech-based visual concepts by the hyponyms generated by WordNet.

2) We propose an ESVC model to boost the performance by reordering and enlarging the keyword list learnt from supervised learning by the keyword list generated by WordNet.

3) We propose "common information gain" to measure the information we obtain from the keyword list generated by WordNet to the one generated by supervised learning.

The rest of the paper is organized as follows. In Section 2, we present the process to generate the keyword list by WordNet and supervised learning, and show how to combine them to obtain a new keyword list. We then discuss how to apply the learnt speech-based visual concept models for retrieval in Section 3. In Section 4, we compare the experimental results generated by different keyword lists. Finally, conclusions are shown in Section 5.

2. GENERATING THE KEYWORD LIST

In this section, we discuss how to (1) extract the keyword list from WordNet, (2) utilize the video context, and (3) combine them to get a new keyword list for modeling speech-based visual concepts.

2.1 Unsupervised Learning from WordNet

In WordNet, there are two types of meaning provided: lexical and semantic. Synonymy and antonymy provide lexical relations between word forms, while hyponymy is a semantic relation between word meanings, which represents "is one kind of". For example, {scorcher} is a hyponym of {hot weather}, and {hot weather} is a hyponym of {weather}.

Table 1 shows the results for Hyponyms search of the noun "weather" from WordNet. All the words and phrases compose of our keyword list generated from WordNet.

Table 1. Hyponyms search of noun "weather" from WordNet

weather, weather condition, atmospheric condition
=> cold weather
=> freeze, frost
=> fair weather, sunshine, temperateness
=> hot weather
=> scorcher
=> sultriness
=> thaw, thawing, warming
=> precipitation, downfall
=> rain, rainfall

In Table 2, we show part of the keyword lists generated by WordNet for several concepts. Some of the keywords are commonly used in many situations; nevertheless, some of them are seldom used in our daily life.

Table 2. Keyword lists generated by WordNet

Airplane	Animal	Building	Weather news
airplane	animal	building	weather
aeroplane	beast	edifice	atmospheric
plane	brute	walk-up	cold weather
airline	critter	butchery	freeze
airbus	darter	apartment	frost
twin-aisle	peeper	tenement	sunshine
airplane	creature	architecture	temper
biplane	Fauna	call center	scorcher
passenger	microorganism	sanctuary	thaw
aircraft	poikilotherm	Bathhouse	Rain

2.2 Supervised Learning from the Video Context

Speech cues may be derived from one of two sources: manual transcriptions such as closed caption, or the results of automatic speech recognition (ASR) on the speech segments of the audio. Given transcriptions of either type, the transcripts are split into documents for learning the concept models. Documents are defined as the words occurring symmetrically around the center of a shot (±2

surrounding shots). These documents mapped from the shots, which are manually annotated as containing a particular concept, are selected for that concept. Then, for each document in the selected set, stop-words are removed by using a standard stop-words list and stemming are processed for all the words. Finally, information gain (IG), which measures the number of bits of information by knowing the presence or absence of a term in a document, is calculated to rank and select the keywords [9]. Those terms with information gains less than a predetermined threshold (0.02 in this paper) are removed from the keyword list.

Table 3 illustrates the keyword lists generated from the video context by the above approach. The terms are ordered by the decreasing of the information gain. Compare two keyword lists in Table 2 and Table 3, it is interesting to see that part of the words in these two keyword lists are the same, while some keywords in Table 3 represent the context relationship instead of the lexical or semantic relationship. For example, from the keyword list generated by WordNet, we can see that "Africa" has nothing to do with Animal. Also, from the keyword list generated by WordNet, we can see that "Africa" has nothing to do with "Animal". However, in the keyword list generated from the video context, "Africa" is an important word related to "Animal", since it has a high value of information gain. The same thing happens in "Airplane" and "Iraq". Thus, we can see it may be hard for a context-free keyword list generated from WordNet to adapt to different situations.

Table 3. Keyword lists generated from the video context

Airplane	Animal	Building	Weather news
Plane	animal	house	rain
fly	Africa	president	temperature
ground	park	school	weather
military	safari	damage	forecast
land	nation	tornado	shower
weapon	land	court	storm
pilot	wild	city	thunderstorm
generator	gorilla	destroy	snow
airplane	wildlife	town,	lake
hospital	elephant	police	southeast
war	extinct	residence	warm
Iraq	breed	building	meteorologist

2.3 The Extended Speech-based Visual Concept (ESVC) Model

The intuition of merging two keyword lists is that if both keyword lists include this word, we will know this word is important for the retrieval, and we will keep its rank at the original position or ahead of it. Otherwise, it should be kept at the original position or put behind it. Thus we define a common information gain of word w as:

$$CG(w) = IG(w) + IG(w)P(q \mid w) \quad (1)$$

where $P(q \mid w)$ indicates whether the word w is included in WordNet generated keyword list ($P(q \mid w) = 1$) or not ($P(q \mid w) = 0$). The new keyword list is ordered by the values of *CG* computed by equation (1).

3. RANKING SPEECH-BASED VISUAL CONCEPT DETECTION RESULTS

After obtaining the keyword lists, standard Okapi [7] formula is applied in ranking the video shots. Each unigram and bigram term in the intersection of the query and document term lists contributes a score of

$$s = \frac{tf}{c_1 + c_2 \times \frac{dl}{avdl} + tf} \times w^{(1)} \times qtf \quad (2)$$

where *tf* and *qtf* are the document and query counts for a given term, *dl* is the length of the document, *avdl* is the average length of the documents in the corpus, $w^{(1)}$ is the inverse document frequency, computed as

$$w^{(1)} = \log\left(\frac{N - n + 0.5}{n + 0.5}\right) \quad (3)$$

where N is the total number of documents in the corpus, and n is the number of documents containing a given term. We used $c_1 = 0.5$, $c_2 = 1.5$ for unigram scoring and $c_1 = 0.05$, $c_2 = 0.05$ for the bigrams.

4. EXPERIMENTAL RESULTS

We demonstrate the performance of our algorithm by applying the model for a concept detection task upon the NIST Video TRECVID 2003 dataset. In our experiments, the development set, which was manually annotated [1], is further divided into four parts: ConceptTraining (CR, 38 hours), ConceptValidate (CV, 6 hours), ConceptFusion1 (CF1, 6 hours) and ConceptFusion2 (CF2, 12 hours). Similar to [5][6], our goal is to use only speech cue for retrieving video shots that include specific visual objects or scenes. The performance is measured by the Average Precision which is widely adopted in NIST TREC evaluations [6].

The retrieval performances of applying different keyword lists upon 11 concepts in TREC-2003 video benchmark are shown in Table 4. CR is used as the training set, and CF1 and CF2 are used as the testing set. "HJN" represents the algorithm proposed by Nock *et al.* [5][6], which is a supervised learning algorithm from annotated video sequences with additional manual selection on the keyword list; "WordNet" represents using the keyword list generated by WordNet; and "Extension" means using the ESVC model. Combining HJN models with visual detectors, the IBM TREC 2003 concept detection system performs best in terms of the mean average precision. The HJN models perform better than the visual-only detectors in half of the submitted detectors.

The results in Table 4 show that the unsupervised learning by WordNet can achieve comparable

performance comparing to HJN for some concepts, such as Airplane, Building, Madeleine Albright, and People. The Mean Average Precisions (MAPs) are 87.3% and 88.5% over the original HJN models in CF1 and CF2, respectively. For the ESVC model, the performance is better than that of HJN. The relative improvements of the MAPs of the extension model are 53.1% and 28.4% better than the original HJN models in CF1 and CF2, respectively.

Table 4. Comparison of retrieval performances obtained by different approaches, with the best performance marked by yellow

Concept Name	Approaches	Testing Set	
		CF1	CF2
Airplane	HJN	0.258	0.2087
	WordNet	0.2898	0.19
	Extension	0.3179	0.3081
Animal	HJN	0.1146	0.0223
	WordNet	0.0145	0.0387
	Extension	0.1329	0.0519
Building	HJN	0.0162	0.0349
	WordNet	0.0743	0.0159
	Extension	0.0823	0.0234
Face	HJN	0.237	0.3194
	WordNet	0.1663	0.2988
	Extension	0.3114	0.4724
Madeleine Albright	HJN	0.0528	0.3009
	WordNet	0.2529	0.419
	Extension	0.5566	0.4219
Nature Vegetation	HJN	0.0733	0.0777
	WordNet	0.012	0.0129
	Extension	0.0906	0.043
Outdoors	HJN	0.0825	0.2093
	WordNet	0.0394	0.1566
	Extension	0.1618	0.1653
People	HJN	0.0436	0.0364
	WordNet	0.0526	0.0413
	Extension	0.1693	0.0578
Physical Violence	HJN	0.002	0.0065
	WordNet	0	0.0138
	Extension	0.0933	0.0108
Road	HJN	0.0285	0.0352
	WordNet	0.0247	0.0149
	Extension	0.0546	0.0135
Weather News	HJN	0.7818	0.3917
	WordNet	0.6874	0.4335
	Extension	0.7915	0.6276
Mean Average Precision	HJN	0.1597	0.1610
	WordNet	0.1394	0.1424
	Extension	0.2444	0.2067

5. CONCLUSIONS

We presented an unsupervised learning algorithm by hyponyms search from WordNet for learning speech-based visual concept models. We also propose an Extended Speech-based Visual Concept (ESVC) model to extend the keyword list learnt from supervised video context and the WordNet. Experimental results upon a detection task on 11 concepts in TREC-2003 video benchmark show that comparing to the conventional supervised learning algorithm, the unsupervised learning scheme achieves comparable performance and the ESVC model achieves better performance. In the future, we plan to further extend the ESVC model based on Latent Semantic Analysis and Latent Dirichlet Allocation models, and utilize the speech-based methods in conjunction with visual models using random graph methods.

6. ACKNOWLEDGEMENT

We would like to thank Dr. Belle L. Tseng for her code for calculating average precision values, and Dr. Harriet Nock for kindly providing us the data used for performance comparison.

7. REFERENCES

[1]. C.-Y. Lin, B. L. Tseng, and J. R. Smith, "Video Collaborative Annotation Forum: Labels on Large Multimedia Dataset," Proc. of TRECVID 2003, Gaithersburg, Nov. 2003.

[2]. X. Song and C.-Y. Lin and M.-T. Sun, "Cross-modality automatic face model training from large video databases," Proc. of FPIV, Washington DC, June 28, 2004.

[3]. X. Song, C.-Y. Lin and M.-T. Sun, "Autonomous Visual Model Building Based on Image Crawling through Internet Search Engines," Proc. of MIR, New York, NY, October 15-16, 2004.

[4]. X. Song, C.-Y. Lin and M.-T. Sun, "Autonomous Learning of visual concept models," Proc. of ISCAS, Kobe, Japan, May 23-26, 2005.

[5]. H. J. Nock, G. Iyengar, C. Neti, "Issues in Speech-based Retrieval of Video," Proc. on Multilingual Spoken Document Retrieval, 2003.

[6]. H. J. Nock, W. Adams, G. Iyengar, C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, J. R. Smith, and B. Tseng, "User-trainable Video Annotation Using Multimodal Cues," Proc. of SIGIR, 2003.

[7]. S. E. Robertson, S. Walker, and M. Beaulieu, "Okapi at TREC-7," TREC-7, Gaithersburg, November 1998.

[8]. C. Fellbaum, "WordNet: An Electronic Lexical Database," MIT Press, Cambridge, MA, USA, 1998.

[9]. Y. Yang, and J. P. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. of the ICML, pp. 412-420, 1997.

[10]. K. Barnard, P. Duygulu, N. D. Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching Words and Pictures", *Journal of Machine Learning Research*, Vol 3, pp 1107-1135, 2003.

[11]. J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Trans. on PAMI*, vol. 25, no. 9, pp. 1075-1088, 2003.

[12]. A. B. Benitez and S.-F. Chang, "Image Classification Using Multimedia Knowledge Networks," Proc. of ICIP, Barcelona, Spain, Sep 14-17, 2003.