EXPLOITING CONCEPT ASSOCIATION TO BOOST MULTIMEDIA SEMANTIC CONCEPT DETECTION

Sheng Gao, Xinglei Zhu and Qibin Sun

Institute for Infocomm Research, Singapore 119613
{gaosheng, xzhu, qibin}@i2r.a-star.edu.sg

ABSTRACT

In the paper we study the efficiency of semantic concept association in multimedia semantic concept detection. We present an approach to automatically learn from the corpus the association strength between pair-wise semantic concepts. We discuss two usages of association strength: 1) applying positive concepts with high association strength for selecting expressive component in the model-based fusion and 2) applying negative concepts with low association strength as filters. We evaluate its efficiency on the task of semantic concept detection on the large-scale news video dataset from TRECVID 2005 development set. Our experimental results demonstrate that exploiting positive association reduces the size of feature dimension in the model-based fusion and significantly improves the rank performance of system. The mean average precision is increased to 0. 215 on the validation set and 0.206 on the evaluation set. Compared to the traditional model-based fusion, the improvement is about 9.1% and 3.5%, respectively. The average feature dimension is reduced to 43 from 312.

Index Terms – multimedia semantic concept detection, concept association strength, feature reduction.

1. INTRODUCTION

Advanced by annual TREC video retrieval evaluation, large-scale multimedia semantic concept detection and search have been extensively studied in recent years. Not only huge volume of annotated news video corpus (i.e. video shot segmentation, concept annotation at key-frame, moderate size of predefined semantic concepts; see [3] for details) is accessible for research, TRECVID also provides a platform to evaluate the state-of-art technology in multimedia information retrieval. One of the lessons learned from the systems developed by participants is that fusing multi-modality and various features are critical. In the top systems built by IBM, Columbia University, CMU, etc., multi-modality features as well as various feature extractors are integrated to enhance the system performance.

Multimedia data such as videos have three modalities: audio, visual and textual (e.g. text transcribed from speech signal using automatic speech recognition and text captured by video OCR). For each modality, many feature extractors are suitable for the content representation. For instance, many types of visual features (e.g. color histogram, texture, motion, edge, shape) are extracted to describe visual content in the keyframe images. Any single modality and extractor alone is not powerful enough to capture the rich content in multimedia data. In general, two ways are used for fusion. One way is to concatenate all features in one vector, which generates undesirable high index dimension. Thus, the issue of curse of dimensionality must be addressed. Furthermore, forming one feature vector is not always realistic in practice because sometimes some modalities are lost. It is a natural phenomenon in multimedia data. For example, sometimes there is no speech signal and thus there is no text from speech recognition. Therefore, another fusion way, i.e. model-based transformation (*MBT*), is preferred [1, 2, 7, 8, 9].

In model-based transformation, a classifier for each feature is firstly trained for each semantic concept. Thus, a set of classifiers is collected for each concept. Suppose we have *M* types of features and *N* concepts, *M*N* classifiers should be trained. These classifiers are treated as the bases to map a training sample into *M*N*-dimensional model score space [1, 2, 7]. Based on the model score space representation, another classifier is trained for each concept to reach the final classification decision. In the classification stage, a test sample is first mapped into a *M*N*-dimensional model space vector, and then the final decision is made using the classifier trained on the model score space. This scheme has been proven successful by all systems developed for multimedia semantic concept detection and search in TRECVID[1] [1, 2] (see TRECVID workshop papers for details).

The model-based transformation in the above is still facing the curse of dimensionality. For a concept lexicon [3] of moderate size in which 101 concepts are annotated, the dimension of model space will reach 1,010 if 10 types of features are used (note: 10 types are not much by studying the top systems [1]). In the MBT method, it is not easy to answer which modalities or features are important and which concepts are more critical to boost the specific concept detection. Wu & Chang in [9] studied the first issue. They applied principle component analysis (PCA) to reduce the feature dimension and independent component analysis (ICA) to identify the important modalities. In the paper, we will address the second issue.

Relation between semantic concepts is well studied in natural language processing (e.g. *WordNet* is built to describe relation among words). Detecting one concept will boost chances of detection of another concept. For example, the concept *airplane* has a high association with *sky* but has never co-occurred with *sports*. Thus, if the detectors of *sky* and *sports* have high accuracy, they can be used to enhance the *airplane* detector. In [4], Naphade & Huang developed multinet (i.e. Bayesian network) to integrate concept relation into semantic indexing and retrieval in video. However, training and inference in Bayesian network are high computation consuming, especially for large-scale semantic concept detection.

In the paper, we will study 1) automatically extracting the semantic concept association from the corpus and 2) exploiting the knowledge to boost semantic concept detection. We will evaluate the efficiency for semantic concept detection based on TRECVID 2005 development set.

In the next section, we will introduce extraction of semantic concept association. Then the experimental evaluation results and

[1] http://www-nlpir.nist.gov/projects/trecvid/

analysis are shown in Section 3. Finally, we will conclude our findings.

2. EXTRACTING CONCEPT ASSOCIATION

Although *Wordnet* describes concept relations, it cannot indicate the strength of association between two concepts. It is also too general to reflect domain knowledge, which has been proven to play an important role in many applications such as machine translation, speech recognition, language modeling, etc. Exploiting domain knowledge often significantly improves the system performance. Here we study extracting domain-specific semantic concept association from the news video corpus. Thanks to LSCOM (i.e. Large Scale Concept Ontology for Multimedia), we have large-scale annotated video corpus for semantic concept lexicon of moderate size. Based on the annotated corpus, we can automatically extract the association between any two concepts and their strength.

In the recent version LSCOM, 449 concepts are annotated, of which 39 concepts are used here. The subset is used in the evaluation of high-level feature extraction task in TRECVID 2006[2]. The names and the identity number (ID) of the semantic concepts are listed in Table 1.

Table 1 Names and ID of the semantic concepts in TRECVID'06

ID	Concept	ID	Concept	ID	Concept
1	Sports	14	Sky	27	Computer_TV-screen
2	Entertainment	15	Snow	28	Flag-US
3	Weather	16	Urban	29	Airplane
4	Court	17	Waterscape_Waterfront	30	Car
5	Office	18	Crowd	31	Bus
6	Meeting	19	Face	32	Truck
7	Studio	20	Person	33	Boat_Ship
8	Outdoor	21	Government-Leader	34	Walking_Running
9	Building	22	Corporate-Leader	35	People-Marching
10	Desert	23	Police_Security	36	Explosion_Fire
11	Vegetation	24	Military	37	Natural-Disaster
12	Mountain	25	Prisoner	38	Maps
13	Road	26	Animal	39	Charts

2.1 Learning Concept Association from Corpus

To learn concept association, following the terms in [4], we notate the concept pairs in which two concepts co-occur with each other at least once as *positive* association and *negative* association otherwise. In the following, we will note a concept as *target* concept when we calculate its co-occurrence with other concepts, which are noted as *associated* concept.

The annotated corpus is the development set of TRECVID 2005. Each news video file is segmented into the shots, which are represented by a few keyframes. For each keyframe and concept, the annotator labels it as positive if it is relevant with the concept and negative otherwise. Annotation of the shot can be derived from the keyframes. A shot is positive for a concept if at least one keyframe in the shot is relevant with it. Because we are interested in the shot level annotation in the TRECVID, we will count the statistical association on the shots for the pair-wise concepts.

Given the target concept *A* and a concept *B*, we measure the strength of their association, *Str*, as,

$$Str = \frac{\#(A,B)}{\#(A)} \quad (1),$$

where #(A, B) is the number of shots relevant with both *A* and *B*, and #(A) is the number of shot relevant with *A*. It is the measurement of conditional probability of *B* on *A*. Higher value of *Str* means concept *B* has a stronger relation with *A*. The measurement in Eq. (1) is asymmetric, i.e. the conditional probability of *B* on *A* is not equal to that of *A* on *B*. It is easy to be verified from Eq. (1). Although their numerators are same, their denominators may be different. For example, in the TRECVID 2005 corpus, *outdoor* ranks the first in terms of the association strength for the concept *airplane*, however, airplane ranks the 25th for *outdoor* (the first is *person*).

The pair-wise association strengths among 39 concepts are shown in an *association map* in Figure 1. Y-axis is the target concept ID while X-axis is the associated concept ID. The brightness of the blocks corresponds to the association strength. The association strength between the concept and itself is defined as one. It is the brightest diagonal line in the figure. The darkest area is the negative association. The association map clearly shows most of the concepts have a higher association strength with *outdoor* (see vertical line at ID=8, X-axis) and *person* (see vertical line at ID=20, X-axis). The concept *person* associates with all concepts. It is in accordance with our intuition that person is central in news video. It means that it has little information to discriminate the target concept from others.

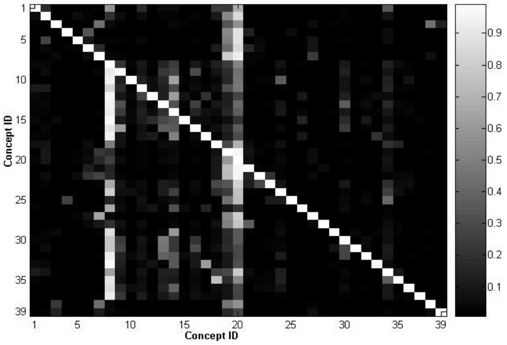

Figure 1 Association map for pair-wise concept association strength (X-axis, Y-axis: the concept ID listed in Table 1)

2.2 Exploiting Concept Association for Indexing

The associated concepts with high association strength boost the detection of the target concept. Sometimes, the target concept may have poor detection performance, but some of its associated concepts can be detected with high accuracy. Then we can use the detection outputs of its associated concepts to help detecting the target concept. Here we introduce an application of association map to select the most efficient feature components for indexing in the MBT fusion.

For simplicity, we assume one type of low-level feature and *N* semantic concepts. Thus, *N* classifier are trained, each for one concept. Following the terms in [2], each classifier is treated as a basis model which plays a similar role as the eigenvector in the eigenspace, and it maps the low-level feature into one component in model score space. Traditionally, the *N* basis models are equally treated, and the feature dimension in the MBT fusion will be *N*. In Figure 1, the association of a target concept with other concepts varies in a large range from one concept to another. The components corresponding to concepts with strong association strength are much more important than others. We exploit the association strength to select them.

[2] http://www-nlpir.nist.gov/projects/tv2006/tv2006.html

For a specific target concept, we sort the concept (or basis model) according to their association strength defined in Eq. (1). Then a predefined threshold is used to prune concepts with unqualified association strength. Thus the size of basis models is reduced as well as the dimension of the fused feature vector. Because the model bases are chosen based on the domain knowledge extracted from the corpus, it is not affected by the type of low-level features. If K ($K<N$) basis models are chosen and there are M types of features, then the indexing dimension in MBT fusion will be $K*M$. It is clear that indexing dimension may be different across target concepts.

2.3 Exploiting Negative Association

In Section 2.1, we have introduced the negative association, i.e. concepts have never co-occurred with the target concept. The basis models corresponding to these concepts will not be used for indexing. Here we discuss using these concept detectors as a filter to prune the most impossible ranking document for a specific target concept.

Given the target concept *A*, we assume another two concepts, *B* and *C*, having negative association with it. In semantic concept detection task, the ranking lists on test database *A*, *B*, and *C*, are *RA*, *RB*, and *RC* respectively. Each element in the ranking list is a shot. If *B* and *C* detectors have high accuracy, we should expect most of their top-*N* shots will be strongly relevant with *B* and *C*. Due to their negative association with *A*, the union of their top-*N* shots should have the smallest chance to occur among the top-*N* for ranking *RA*. If the shots in the union occur in the top-*N* of *RA*, we can safely remove them from *RA* to improve ranking performance of concept *A*.

Of course, the operation has risk, especially when the concept detectors as the filters that have low performance. In practice, we choose the concepts having high performance, i.e. more than a threshold, as the filter. We will experimentally study its effect on the ranking performance in next section.

Using filter operation to improve ranking performance is not new. But, traditionally, the filter detector needs extra resources, e.g. manual annotation beside the common annotation set. For example, to use commercial detector, the developer must label their own training samples because commercial concept is not in the semantic concept lexicon. The novelty of using the negative association concept as filter is in that 1) extra labeling and training are not required and 2) it provides an efficient way to use available resources to achieve their best ability.

3. EVALUATION AND ANALYSIS

We have discussed the method to exploit the strength of pair-wise concept association for indexing and filtering in the above. Now we evaluate its efficiency on semantic concept detection on the TRECVID 2005 development set. The set has 137 mpeg news videos. We randomly split the videos into three sets, i.e. 70% (96 videos, ~40,000 keyframes) for training, 15% (20 videos, ~10,000 keyframes) for validation, and 15% (21 videos, ~10,000 keyframes) for evaluation. The 39 semantic concepts have been listed in Table 1.

In the experiment, we use the MBT to fuse multiple visual features as shown below:

- Global color correlogram (*GCC*) in HSV space: 324-dimension [11].
- Co-occurrence texture extracted from global gray-level co-occurrence matrix (*GLCM*): 64-dimension [12].
- 3-D global color histogram in HSV (*HSV*): 162-dimension.
- 3-D global color histogram in RGB (*RGB*): 125-dimension.
- 3-D global color histogram in LAB (*LAB*): 125-dimension.

For each type of feature and concept, one SVM classifier (*SVM*) or linear discriminative function (*LDF*) classifier is trained. The details are shown in Table 2. If a classifier is trained for a feature, a mark '+' is given. Otherwise, it is left empty. Thus there are 8 classifiers trained for each concept.

Table 2 Description of trained classifiers (+: the classifier is trained for the feature)

	SVM	LDF
GCC	+	+
GLCM	+	+
HSV	+	+
RGB	+	
LAB	+	

3.1 Tuning Classifiers

We train SVM classifier using SVM-light [10] tool package and LDF using our developed ROC optimized learning algorithm [6]. Now we will discuss automatically setting the configuration of SVM to get the optimal average precision. The tuned parameters are kernel type and kernel parameters. For example, for linear kernel its parameter is a coefficient controlling the trade-off between training error and margin. For polynomial and RBF kernel, the kernel parameters are the polynomial order and gamma, respectively, besides the trade-off coefficient. For each type of kernel, we linearly search in a range to find its combination of trade-off coefficient and polynomial order or gamma. Then, the configuration which gives best average precision on validation set is chosen. Finally, the SVM model is trained on the best configuration and evaluated on evaluation set. All other parameters in SVM-light tool are assigned as the default setting.

In ROC optimized learning algorithm [6], the tuned parameter is *alpha* which controls the smoothness of sigmoid function. Similar to tuning SVM, we linearly search *alpha* in a range to get the optimal one giving the best average precision. Then, the selected configuration is used for training.

3.2 Effect of Positive Concept Association

To study the efficiency of exploiting the positive concept association on indexing and semantic concept detection, we build the semantic concept detection system using the strength of the positive concept association. It is compared with a benchmark system, which is trained using the traditional MBT fusion. Hereafter, the two systems are named as *AssociationMap* and *Benchmark*.

As introduced above, 8 classifiers are trained for each semantic concept. Thus, the feature dimension in the model score space is 312 (8*39). Using the 312-dimensional feature, a SVM classifier is trained for each concept using the tune method in section 3.1. The MBT classifier is then used to rank the shots in the validation set and evaluation set. The official NIST evaluation metric, i.e. average precision (AP) at the top-2000 retrieved shots, is reported here.

A threshold (0.2 in our system) is set to get the top-*N* concepts having high association strength for each semantic concept. Here the minimal K is set equal to 5. As K associated concepts are chosen, we construct $K*8$-dimensional model space feature vectors. In our experiment, the average K on 39 concepts is ~5.4 and the maximal value is 8. Thus the maximal dimension is 64. Comparing with 312, there is a significant reduction of the feature dimension. Each component in the association strength has a clear meaning, while in other feature reduction methods such as PCA, it is difficult to explain its component in the eigenspace.

The AP values of the two systems over 39 semantic concepts are shown in Figure 2 for the validation set and in Figure 3 the evaluation set. From these two figures, it is obvious that exploiting concept association significantly improves the average precision over most concepts. The mean average precision (MAP) of the *AssociationMap* system reaches 0.215 on the validation set and 0.206 on the evaluation set. For comparison, the benchmark system has 0.197 on the validation set and 0.199 on the evaluation set. A little improvement is observed. To further illustrate that the *AssociationMap* boosts ranking, we compare their precision at top-100 shots. The precision for the benchmark is 0.333 on the validation set and 0.326 on the evaluation set while they are 0.334 and 0.323 for the *AssociationMap*, respectively. Although the *AssociationMap* uses a much lower indexing dimension, its ranking performance is competitive with the benchmark using the full dimension indexing.

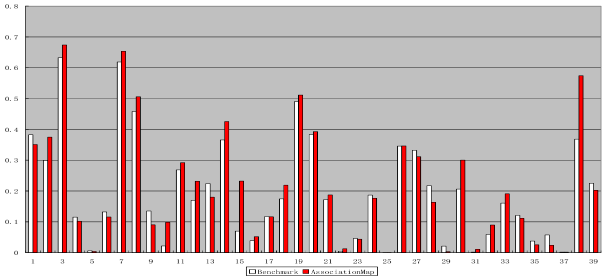

Figure 2 Comparison of the average precision on the validation set between the *AssociationMap* (red bar) and benchmark (white bar) (X-axis: concept ID. Y-axis: AP value.)

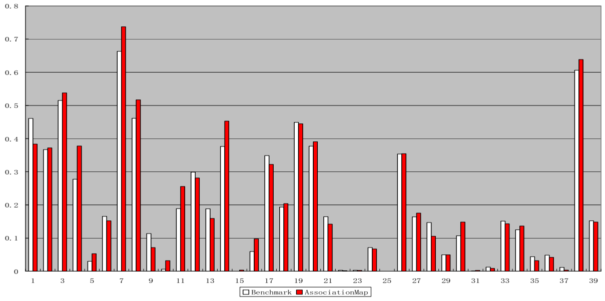

Figure 3 Comparison of the average precision on the evaluation set between the *AssociationMap* (red bar) and benchmark (white bar) (X-axis: concept ID. Y-axis: AP value.)

3.3 Effect of Negative Concept as Filter

In this section we experimentally study how the negative concepts as the filters affect the ranking performance. We use the filter concepts to prune the ranking shots of target concept in the *AssociationMap* system to explore whether the filter concepts improve the ranking performance.

Besides the negative concepts of each target concept, positive concepts with association strength less than a threshold (0.1 in our system) are also considered as the candidates of filter concepts. We remove the concepts whose AP values on the validation set are less than a threshold (0.3 in our system). Thus, the final filter concepts are determined for each target concepts. As discussed in Section 2.3, we use the ranking shots in the filter concepts to prune the target concepts. Due to limited space, we only show the overall performance on 39 concepts. After filtering, the MAPs are 0.211 on the validation set and 0.205 on the evaluation set, and the precisions at top-100 are 0.331 and 0.322, respectively. Compared with the *AssociationMap* system without filter, there is a little reduction in ranking performance. Our preliminary analysis on each concept reveals that filter operation seems to help the concepts with poor performance and deteriorate those with good performance. For example, filter operation improves AP of *corporate-Leader* to 0.0023 from 0.0016 on the evaluation set. But it reduces AP of *sky* from 0.453 to 0.430. However, it cannot yet get a general conclusion about it. Considering the gain from the positive concept association, the negative association should have more benefit than that has been exploited here. In future, we will study how to efficiently integrate the negative association knowledge into semantic concept detection.

4. CONCLUSION

In the paper we presented an approach to automatically learn the association strength among semantic concepts from the corpus and study the usage of association strength in indexing and semantic concept detection. Its efficiency is evaluated on semantic concept detection based on the large-scale news video dataset from TRECVID 2005. Our experimental results demonstrate that exploiting positive association concepts significantly improve the system performance. The mean average precision is increased to 0.215 on the validation set and 0.206 on the evaluation set. Compared to the traditional model-based fusion, the improvement of MAP is about 9.1% and 3.5% respectively. The average feature dimension is significantly reduced to 43 from 312.

5. REFERENCES

[1] A. Amir, et al., "IBM research TRECVID-2005 video retrieval system", *Proc. of TRECVID'05 Workshop*.

[2] G. Iyengar, et al., "Discriminative model fusion for semantic concept detection and annotation in video", *Proc. of ACM MM'03*.

[3] L. Kennedy, et al. "LSCOM Lexicon Definitions and Annotations Version 1.0", *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3*, 2006.

[4] M. R. Naphade, et al., "Probabilistic semantic video indexing", *Proc. of NIPS'00*.

[5] M. R. Naphade, et al., "A light scale concept ontology for multimedia understanding for TRECVID 2005," IBM Research Technical Report, 2005.

[6] S. Gao & Q. B. Sun, "Classifier optimization for multimedia semantic concept detection", *Proc. of ICME'06*.

[7] S. Gao, et al., "Automatic image annotation through multi-topic text categorization", *Proc. of ICASSP'06*.

[8] D. H., Wang, et al., "Discriminative fusion approach for automatic image annotation", *Proc. of MMSP*'05.

[9] Y. Wu & E.-Y. Chang, "Optimal multimodal fusion for multimedia data analysis", *Proc. of ACM MM'04*.

[10] T. Joachims, *Learning to Classify Text Using Support Vector Machines*. Dissertation, Kluwer, 2002.

[11] J. Huang, et al., "Image indexing using color correlograms", *Proc. of CVPR*'97.

[12] R. M. Haralick, et al. "Textural features for image classification", IEEE Trans. Systems, Man and Cybernetics, Vol. 3, No.6, pp.610-621, 1973.