IMPROVING SEMANTIC CONCEPT DETECTION AND RETRIEVAL USING CONTEXTUAL ESTIMATES

Yusuf Aytar, O. Bilal Orhan and Mubarak Shah

Department of Computer Science, University of Central Florida, Orlando, FL 32816
E-mail: {yaytar, oborhan, shah} @cs.ucf.edu

ABSTRACT

In this paper we introduce a novel contextual fusion method to improve the detection scores of semantic concepts in images and videos. Our method consists of three phases. For each individual concept, the prior probability of the concept is incorporated with detection score of an individual SVM detector. Then probabilistic estimates of the target concept are computed using all of the individual SVM detectors. Finally, these estimates are linearly combined using weights learned from the training set. This procedure is applied to each target concept individually. We show significant improvements to our detection scores on the TRECVID 2005 development set and LSCOM-Lite annotation set. We achieved on average +3.9% improvements in 29 out of 39 concepts.

1. INTRODUCTION

Detecting high level concepts in image/video domain is an important step in achieving semantic search and retrieval [1]. The main trend in understanding semantic concepts is computing low-level features using texture, color, motion and shape on an annotated data set, and then ranking and retrieving data using the models trained for each concept (e.g. support vector machine (SVM) classification [2]). However such models are generally built independent of each other, lacking the relationships among semantic concepts. It is obvious that the occurrence of some concepts increases the probability of the occurrence of other concepts. Similarly, some concepts do not occur together.

There have been a few attempts to exploit the relationships between semantic concepts over the past few years. The most common approach in semantic level fusion is to train individual SVM detectors for each concept at the early stage and employ their scores as features to train another SVM classifier. This idea was first introduced by Iyengar et. al. [3], where they attempted to put the individual detector scores in a vector, called model vector, and used an SVM to train a classifier using this vector. Similarly Snoek et. al [4] fused scores of individual detectors into a context vector and input this vector to a stacked classifier. Jiang et. al. [5] experimented with a scheme similar to [3] except a linear fusion step was used to combine the context-based SVM classifiers' results with the individual detector output.

There are also some graphical models that exploit those relationships in probabilistic structures. Naphade et. al. [6] used an explicit concept linkage in a Bayesian network to obtain an inference between concepts. In a later study Naphade et. al. [7] used a factor graph framework and sum-product algorithm to perform learning and inference. Yan et. al. [8] experimented with different directed and undirected graphical models to explore the concept relationships in a unified probabilistic framework.

The relationships between concepts can be better represented by directed models rather than undirected co-occurrence relations. For instance, when a car concept exists in an image, it is very likely that the outdoor concept will also exist. However, it is not as likely to see a car in an image when the outdoor concept is known to exist. In our approach, we use the conditional probability values, which can be seen as directed relationship representations between concepts acquired from the training data.

Given a target concept we attempt to improve its detection score. First the detection scores from individual SVM detectors are computed. Then we incorporate the concept priors to refine the detection scores of each individual detector. We employ a probabilistic prediction rule using all individual concept detectors to estimate detection probabilities for the target concept. Finally, we apply a weighted linear combination to aggregate the estimated probabilities into a final detection score. We apply this procedure to all concepts individually on the images. One of the main contributions of our approach is the fact that we bring individual detector scores into the same concept domain by computing an estimation of that concept, and then try to explore the relationships in the concepts domain.

Despite the significant efforts in the past, we believe that relationships between concepts were not adequately exploited. Our experiments show that for some concepts using only the contextual information provided by other concepts may yield results comparable to or sometimes even better than those of the baseline SVM results of the target

concept. These results highlight the importance and the capability of the relationships between concepts.

As mentioned earlier in the previous approaches [5, 8], not all concepts gain a performance increase by context based fusion for various reasons. With our system, over 70% of all 39 concepts achieve a performance increase.

The paper is organized as follows. In section 2, we present our contextual fusion approach. In section 3 we present improvements on our baseline SVM results for TRECVID'05 video collections. And finally the paper concludes with future work and discussions.

2. CFPE APPROACH

This section presents our contextual fusion method using probabilistic estimates (CFPE) in three phases. For each individual concept, the prior probability of the concept is incorporated with the detection score from the individual SVM detector. Then probabilistic estimates for the existence of the target concept are computed using all of the individual SVM detectors. Finally, these estimates are linearly combined using weights learned from the training set using least squares method. This procedure is applied for each target concept individually. The overview for our method is shown in Fig. 1.

There are two main differences between our approach and the approaches in [5, 3]. Firstly, most of the individual concept detectors have low accuracy. Therefore, we can not directly rely on SVM scores. We incorporate the prior probabilities of concepts with the SVM scores. It is important to indicate that this step itself has no effect on the retrieval, since we apply the same transformation to all the scores of that individual SVM detector. However, with this approach we bring the different detection scores together on a more realistic and comparable basis. Secondly, we do not directly fuse the individual detection scores, but instead we compute the probabilistic estimates from the individual SVM detectors and combine them. Using our approach, we put the different SVM scores in the same concept domain to combine them more meaningfully.

2.1 Incorporating Priors

Given an image X and a concept C_i, assume that the true label is y_i, where

$$y_i = \begin{cases} 1, & C_i \quad occurs \quad in \quad X \\ 0, & otherwise \end{cases}. \qquad (1)$$

Each individual detector produces a detection score, $P_I(y_i=1)$, which defines the probability of the occurrence of concept C_i in the image. The accuracy of each individual detector, λ_i, is computed using a validation set. Also, each concept has a prior probability, $P(y_i=1)$, which is the probability of the occurrence of the concept in a random image and computed using training images.

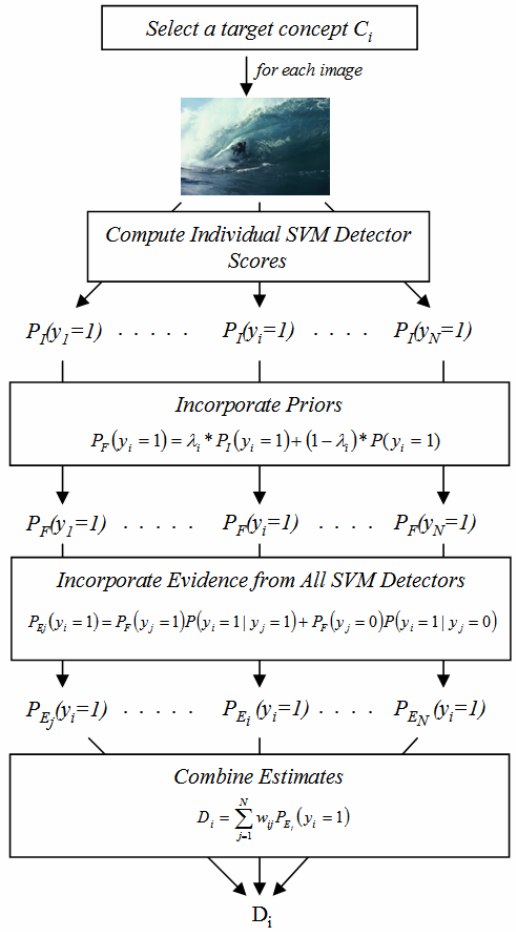

Fig. 1. The steps of the proposed approach for improving detection score for concept C_i.

In order to get a realistic final detection score, denoted as $P_F(y_i=1)$, we linearly combine the prior probability and SVM detection score using a weight which is proportional to the accuracy of the SVM detector.

$$P_F(y_i=1)=\lambda_i * P_I(y_i=1)+(1-\lambda_i)*P(y_i=1). \qquad (2)$$

2.2 Incorporating Evidence from Other SVM Detectors

The main assumption in this step is that each individual SVM can make a meaningful estimate, not only for its original concept, but also for other concepts. During the computation of the estimates, we consider the conditional probabilities between each concept and the individual detection scores. The estimates of all SVMs on the occurrence of concept C_i are computed by the following probabilistic rule:

$$P_{E_j}(y_i=1)=P_F(y_j=1)P(y_i=1|y_j=1)+P_F(y_j=0)P(y_i=1|y_j=0), \qquad (3)$$

where $P_{Ej}(y_i=1)$ is the estimate of the j^{th} SVM for C_i. $P(y_i=1|y_j=1)$ and $P(y_i=1|y_j=0)$ are the conditional probabilities extracted from the training data:

$$P(y_i = 1 | y_j = a) = \frac{P(y_i = 1, y_j = a)}{P(y_i = 1)}, \quad (4)$$

where $P(y_i=1|y_j=a)$ is the ratio of images satisfying this condition to all the images in the training set, and $P(y_i=1)$ is the prior probability of C_i which is the ratio of images that contain C_i to all the images in the training set. Eq. (3) implies that, the estimation score of an SVM for its own concept is directly the score of that SVM.

2.3 Combining Estimates

After the estimates from all the SVMs are computed, they should be combined to produce a final detection score, D_i. A weighted linear combination is applied for combining all the estimates as shown in the following equation:

$$D_i = \sum_{j=1}^{N} w_{ij} P_{E_j}(y_i = 1), \quad (5)$$

where w_{ij} denotes the weight of the estimate of the j^{th} detector for C_i. This weight matrix w is learned from the training data using least squares method. For each concept we have a set of linear equations which is the implementation of Eq.(5) for each image in the training set, and then we computed the weight vector for that concept by applying least squares method. Combination of these weight vectors constructs the weight matrix w.

3. EXPERIMENTS AND RESULTS

The data set consists of a series of broadcast news videos from TRECVID 2005 [9]. The selected concept set is the LSCOM-Lite set of 39 concepts [10]. 74523 video shots are grouped into 3 sets, as 50% training, 25% validation and 25% testing. The individual SVM detectors are trained using color moment features of 5x5 grid image patches and their accuracies are computed from the validation data set. Conditional probabilities between concepts and the prior probabilities of concepts are computed from the training set.

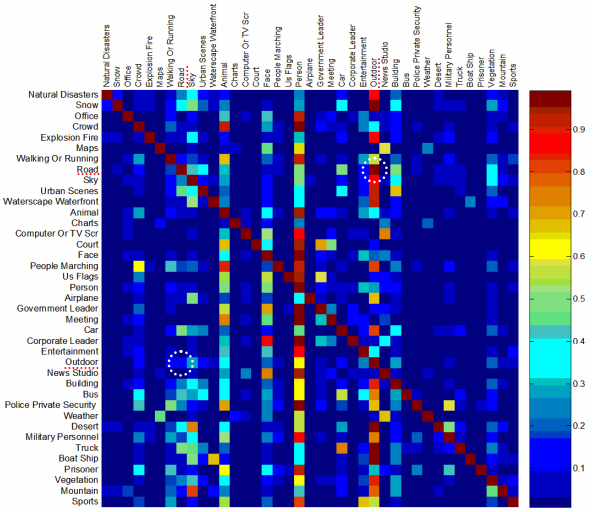

Fig. 2. The conditional probabilities between concepts

The conditional probability relations, as shown in Fig. 2, demonstrate the probability of observing i^{th} column when j^{th} row is observed, $P(y_i=1|y_j=1)$. By focusing on 'Outdoor' and 'Road' concepts it is obvious that the relationship matrix is not symmetric. As can be observed in Fig. 2 the intersection of the *Road* row and the *Outdoor* column demonstrates a high probability of existence of the *Outdoor* concept given the *Road* concept. On the other hand probability of the *Road* concept given the *Outdoor* concept is not as probable.

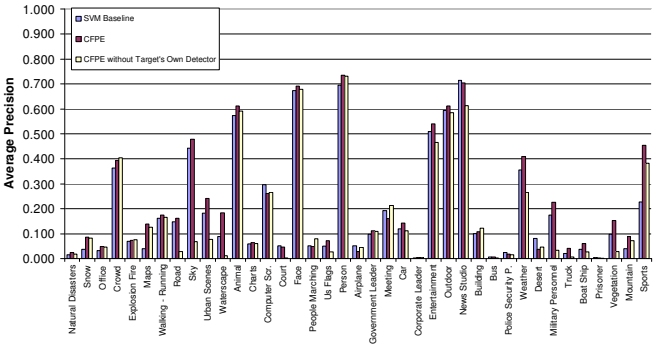

Fig. 3. Average Precision plots of all 39 concepts for baseline and two variations of the proposed approach.

Fig. 3 shows the SVM baseline retrieval results compared to the result of CFPE with and without the estimate of target concept's own detector. These results are also given numerically in Table 1. The error metric used is the non-interpolated average precision (AP) calculation used in the TRECVID challenge. We achieved on the average +3.9% improvements in 29 out of 39 concepts and -1.6% degradation in the remaining concepts. Yan et. al. [8] observed that the best multi-concept modeling approaches can bring 2-3% improvement in terms of mean average precision (MAP). The mean average precision improvement in our approach is +2.5% over all concepts. Specifically concepts like {*Maps, Urban Scenes, Waterscape Waterfront, Weather, Military Personnel, Vegetation*} resulted in over +5% increase in detection performance.

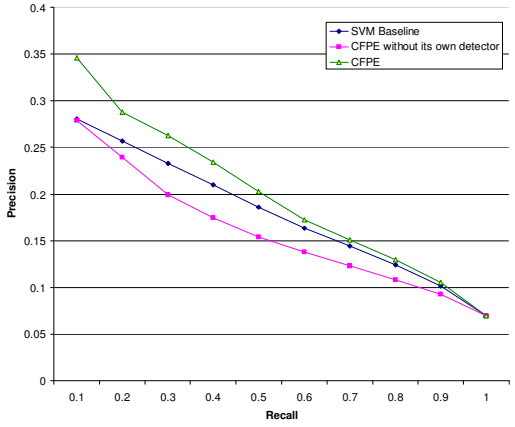

Fig. 4. Precision-Recall curve for baseline and two variations of the proposed approach.

Fig. 4 shows the Precision-Recall curves in which the precisions are averaged over all concepts. It is evident that our CFPE approach gives better performance than the SVM baseline in terms of precision-recall. Furthermore it should be noted that even without the estimate of target concept's own detector CFPE performs close to the individual SVM baseline.

Table 1. The Average Presicion results of concept retrieval using our methods compared to detection baseline

Concept Name	SVM	CFPE	CFPE*
Natural Disasters	0.016	0.025	0.018
Snow	0.039	0.087	0.083
Office	0.032	0.048	0.048
Crowd	0.362	0.393	0.404
Explosion Fire	0.070	0.074	0.075
Maps	0.040	0.138	0.126
Walking Or Running	0.162	0.176	0.167
Road	0.149	0.162	0.029
Sky	0.442	0.478	0.068
Urban Scenes	0.182	0.242	0.077
Waterscape	0.088	0.184	0.012
Animal	0.574	0.612	0.592
Charts	0.060	0.065	0.061
Computer Scr.	0.296	0.261	0.265
Court	0.052	0.047	0.004
Face	0.674	0.691	0.678
People Marching	0.053	0.049	0.080
Us Flags	0.050	0.072	0.027
Person	0.694	0.735	0.731
Airplane	0.053	0.029	0.046
Government Leader	0.098	0.112	0.110
Meeting	0.193	0.161	0.213
Car	0.119	0.142	0.112
Corporate Leader	0.004	0.005	0.006
Entertainment	0.508	0.539	0.465
Outdoor	0.594	0.611	0.585
News Studio	0.715	0.705	0.615
Building	0.101	0.109	0.123
Bus	0.008	0.007	0.003
Police Security P.	0.025	0.019	0.016
Weather	0.355	0.410	0.265
Desert	0.082	0.038	0.047
Military Personnel	0.174	0.225	0.035
Truck	0.022	0.041	0.007
Boat Ship	0.037	0.061	0.026
Prisoner	0.006	0.004	0.002
Vegetation	0.097	0.154	0.029
Mountain	0.040	0.089	0.073
Sports	0.228	0.455	0.383

* CFPE without target's own SVM detector

3. CONCLUSION AND FUTURE WORK

We have developed a new probabilistic contextual fusion method for improving the performance of semantic concept detection in images and videos. Our method considers the reliability of individual detectors and refines the detection scores. Using the refined scores each detector computes a probabilistic estimate for the existence of each concept. These estimates are then linearly combined with the weights that are learned from the training set.

Compared to the most recent approaches our CFPE method achieves promising results. Although the latest works such as [5] has been able to improve 18 out of 39 concepts, we achieved 29 improvements out of 39 concepts. Furthermore without the detection knowledge of the target concept we could detect 18 of the concepts better than our individual baseline SVM detectors. These results show that the contextual relations provide valuable information and should be properly exploited.

We intend to extend our experiments to the whole LSCOM concept set of 449 concepts. By extending the concept size we may be able to find more informative semantic relationships and improve our results.

4. ACKNOWLEDGEMENTS

This work was funded by the Disruptive Technology Office. The views and conclusions are those of the authors, not of the US Government or its agencies.

5. REFERENCES

[1] A. G. Hauptmann, et al.,"Informedia at TRECVID 2003: Analyzing and searching broadcast news video," in Proc. of TRECVID, 2003.

[2] M.R. Naphade and J.R. Smith, "Learning visual models of semantic concepts," in Proceedings of international Conference on Image Processing (ICIP 2003), Barcelona, Spain, September 2003, vol. 2, pp. 531–534.

[3] G. Iyengar, H. Nock, and C. Neti, "Discriminative model fusion for semantic concept detection and annotation in video," ACM Multimedia, pp. 255-258, Berkeley, USA, 2003.

[4] C.G.M. Snoek, et al., "The mediamill TRECVID 2004 semantic video search engine," in Proc. of TRECVID, 2004.

[5] Wei Jiang, et. al., "Active Context-based concept fusion with partial user labels," In IEEE International Conference on Image Processing (ICIP 06), Atlanta, GA, USA, 2006.

[6] M. R. Naphade, et al., "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems," in Proc. of ICIP, 1998.

[7] M. Naphade and J. R. Smith, "A Hybrid Framework for Detecting the Semantics of Concepts and Context," Conf. on Image and Video Retrieval, pp. 196-205, Urbana, IL, June 2003

[8] Yan, R., et. al., "Mining Relationship between Video Concepts Using Probabilistic Graphical Model," IEEE International Conference on Multimedia and Expo (ICME'06), July 9-12 2006.

[9] TRECVID, "Trec video retrieval evaluation," in *http://www-nlpir.nist.gov/projects/trecvid/.*

[10] M. R. Naphade, et. al. , "A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005," IBM Research Technical Report, 2005.