ACTIVE CONTEXT-BASED CONCEPT FUSION WITH PARTIAL USER LABELS

Wei Jiang[1] *Shih-Fu Chang*[1] *Alexander C. Loui* [2]

[1] Department of Electrical Engineering, Columbia University, New York, NY
[2] Kodak Research Labs, Eastman Kodak Company, Rochester, NY

ABSTRACT

In this paper we propose a new framework, called active context-based concept fusion, for effectively improving the accuracy of semantic concept detection in images and videos. Our approach solicits user annotations for a small number of concepts, which are used to refine the detection of the rest of concepts. In contrast with conventional methods, our approach is active, by using information theoretic criteria to automatically determine the optimal concepts for user annotation. Our experiments over TRECVID 2005 development set (about 80 hours) show significant performance gains. In addition, we have developed an effective method to predict concepts that may benefit from context-based fusion.

Index Terms— Active content-based concept fusion, Semantic concept detection

1. INTRODUCTION

Recognition of semantic information from visual content has been an important goal for research in image/video indexing. In recent years, NIST TRECVID video retrieval evaluation has included a task in detecting high-level features, such as locations, objects, people, and events from videos. Such high-level features, termed *concepts* in this paper, have been found to be very useful in improving quality of retrieval results in searching broadcast news videos [6].

This paper addresses the problem of enhancing concept detection accuracy by exploring useful context and active user input. Semantic concepts usually do not occur in isolation - knowing the contextual information (e.g., outdoor) of an image is expected to help detection of other concepts (e.g., cars). Based on this idea, several *Context-Based Concept Fusion* (*CBCF*) methods have been proposed. The Multinet approach [3] models the correlation between concepts with a factor graph and uses loopy probability propagation to modify the detection of each concept based on the detection confidence of other concepts. In [4], models based on Bayesian Networks are used to capture the statistical interdependence among concepts present in consumer photographs. The Discriminative Model Fusion (*DMF*) method [5] generates a model vector based on the detection score of individual detectors, and a SVM is then trained to refine the detection of original concepts. However, the results reported so far have indicated that not all concepts benefit from the CBCF strategy. As reported in [1], no more than 8 out of 17 concepts gain performance improvement by using CBCF. The lack of consistent performance gain could be attributed to several reasons: (1) insufficient data for learning reliable relations among concepts, (2) unreliable detectors, and (3) scales and complexity of the concept relations. Interestingly, results in [4] suggests that user-provided labels are much more effective in helping inferring other concepts compared to automatically detected labels.

In this paper, we propose a new fusion paradigm, called *Active CBCF*, to effectively exploit the contextual relations among concepts. We submit that in several applications, users may be willing to be involved in the process and annotate a few concepts in an image. The human annotated concepts are then used to help infer and improve detection of other concepts (up to hundreds). Human assisted annotation is not new in the literature. But a new, interesting question arises in the active paradigm - if a user is to annotate only a very small number of concepts e.g., (1-3), which concepts should we ask him/her to annotate? We propose an active system that adaptively selects the right “key” concepts, different for each image, for user annotation. In contrast, conventional methods are passive. Users are asked to annotate all concepts or a subset of arbitrarily chosen concepts.

Based on the statistical principles, our method considers the mutual information between concepts, the estimated performance of individual concept detectors, and the confidence of detection for each concept for a given image. Experiments over TRECVID 2005 data (80 hours 61000 subshots) show our active CBCF approach achieves significant performance gains – 11.8% improvement in terms of mean average precision over individual concept detectors without CBCF.

In implementing a baseline CBCF system, we also develop a simple but effective method to predict concepts that may benefit from the use of CBCF. This is motivated by the observation mentioned earlier that not all concepts benefit from context-driven statistical fusion - a prudent step in selective application of the strategy is necessary. We first describe the baseline CBCF system and the prediction process in Sec. 2. The core algorithm of active CBCF system is presented in Sec. 3, with experiment results shown in Sec. 4.

2. A BASELINE CBCF APPROACH

This section introduces our baseline CBCF method followed by the criterion to determine which concepts to use CBCF learning. The scheme of our CBCF baseline method is shown in Fig.1. Given an image x, for concept S_i, assume that the true label is y_i, where $y_i = 1$ if x contains S_i, and $y_i = 0$ otherwise. The individual detector produces a detection score $P_I(y_i = 1)$. A context-based SVM is built to utilize the relationships among concepts to refine the detection result as follows. The concepts are treated as features of images, i.e., x is represented in an $N-1$-dim feature space as $[y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_N]^t$. Based on this feature an SVM classifier is trained. Then in the testing stage, for a test image the estimated $[P_I(y_1 = 1), \dots, P_I(y_{i-1} = 1), P_I(y_{i+1} = 1), \dots, P_I(y_N = 1)]^t$ is used to approximate the true labels (since the true labels are unknown for a test image) and is provided to the context-based SVM to be classified. The classification result is $P_C(y_i{=}1)$, which is the prediction of existence of S_i inferred from concepts other than S_i, based on the relationships among concepts and detection of other concepts. Finally, $P_C(y_i{=}1)$ and $P_I(y_i{=}1)$ are linearly combined into $P_F(y_i{=}1)$ as the refined detection result:

$$P_F(y_i{=}1) = \lambda_i P_C(y_i{=}1) + (1 - \lambda_i) P_I(y_i{=}1) \qquad (1)$$

The scheme described above is similar to the DMF scheme in [5], except we use a linear fusion step at the end to combine the SVM fusion results with original detector output. In addition, in the training stage we used ground truth labels, rather than the individual detector outputs (as in [5]), to train context-based SVMs. By this, we are able to use the same data set to train the individual detectors and context-based SVMs, alleviating the problem of not having enough data for separate training and validation. Our experiments have indicated such a training process is better than the alternative that uses detector outputs for training the context-based SVM.

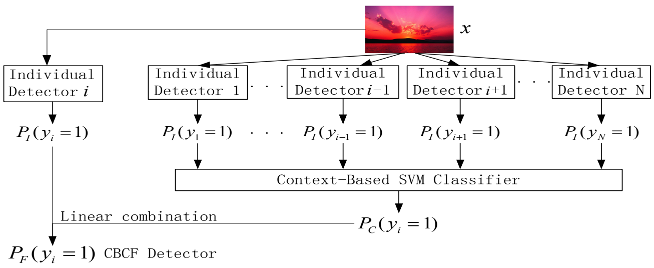


Fig. 1. Context-based SVM for S_i in our multi-level CBCF method.

As discussed above, not all concepts benefit from CBCF learning. Intuitively, two reasons may cause performance deterioration on a concept using CBCF: first the concept has weak relationships with other concepts; second the related concepts have poor individual detectors. This also suggests an intuitive criterion: concept S_i will use CBCF learning when (1) S_i is strongly related to some other concepts, and the average performance of individual detectors of the related concepts are robust; and (2) the performance of its own individual detector is not perfect so that there is space for improvement. Specifically, the relationship between two concepts S_i and S_j can be measured by their mutual information $I(S_i; S_j)$. By using a validation set, we get validation error rate $E_I(S_i)$ that estimates the robustness of individual detector for S_i. Our criterion for applying the CBCF strategy to concept S_i is:

$$E_I(S_i) > \theta, \text{ and } \mathrm{Avg}_{\{S_j: I(S_j;S_i)>\gamma\}} E_I(S_j) < \eta \qquad (2)$$

Note the first part favors concepts whose detection performance is “improvable”, while the second part requires a concept has correlated concepts (in terms of mutual information) and they have adequate strength to help improve accuracy of the given concept. Our experiment results (Sec. 4) show this criterion is indeed effective, making 33 correct predictions for the 39 concepts from TRECVID 2005.

3. THE ACTIVE CBCF APPROACH

We present the proposed active CBCF method (ACBCF) in this section. The work flow of the ACBCF method is illustrated in Fig.2. In ACBCF the user is willing to label a small number of concepts, and the labelled information is utilized to help detection of other concepts as follows. For a test image x, the user labels a set of concepts $\mathbf{S}^* = \{S_1^*, \dots, S_n^*\}$. The labelled ground truth $y_1^*, \dots, y_n^*$ are used to replace the corresponding hypotheses $P_I(y_1^*{=}1), \dots, P_I(y_n^*{=}1)$ from individual detectors to generate a mixed feature vector. In the mixed feature vector, the entries corresponding to the labelled concepts are the labelled ground truth from user, and the other entries are still the estimated hypotheses from individual detectors. This mixed feature vector is provided to the context-based SVM to get a new classification result $\tilde{P}_C(y_i{=}1)$. Since the labelled ground truth provides accurate inputs instead of the inaccurate machine-predicted hypotheses, $\tilde{P}_C(y_i = 1)$ is expected to be better than original $P_C(y_i{=}1)$. And the combined estimation (by using the ACBCF detector) $\tilde{P}_F(y_i = 1)$ should outperform $P_F(y_i{=}1)$.

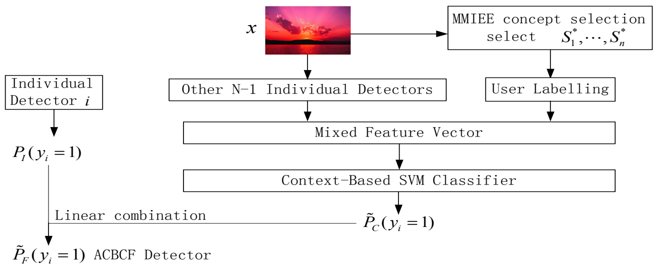


Fig. 2. ACBCF estimation for concept S_i.

Different from passive labelling that randomly selects concepts, our ACBCF actively selects concepts for annotation. In following subsections, we present a statistical criterion to select concepts based on mutual information between concepts, and the error rate and entropy of individual detection results.

3.1. Maximizing mutual information

The average Bayes classification error for concept detection is $\overline{\mathcal{E}} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{E}(S_i)$, where $\mathcal{E}(S_i)$ is the error rate for concept S_i. The problem is to find an optimal $\mathbf{S}^*$ with a small size n so that the labelled information can minimize $\overline{\mathcal{E}}$ of the rest of concepts. Directly searching for $\mathbf{S}^*$ to minimize $\overline{\mathcal{E}}$ is practically infeasible. Instead, we try to minimize the upper bound of $\overline{\mathcal{E}}$, which can be derived as [2]: $\overline{\mathcal{E}} \leq \frac{1}{2N}\sum_i H(S_i) - \frac{1}{2N}\sum_i I(S_i; S^*_{1:n})$, where $S^*_{1:n} = \{S^*_1, \ldots, S^*_n\}$. Minimizing the upper bound equals the maximizing mutual information $\sum_i I(S_i; S^*_{1:n})$. $I(S_i; S^*_{1:n})$ can be written as [7]:

$$I(S_i;S^*_{1:n}) = \sum_{k=1}^{n} I(S_i; S^*_k) - \sum_{k=2}^{n}\left[I(S^*_k;S^*_{1:k-1}) - I(S^*_k;S^*_{1:k-1}|S_i)\right]$$

The above expression can be further simplified as follows:

$$I(S^*_k; S^*_{1:k-1}) - I(S^*_k; S^*_{1:k-1}|S_i) = H(S_i) + H(S^*_k)$$

So maximizing $\sum_i I(S_i; S^*_{1:n})$ equals maximizing:

$$\sum_i \sum_{k=1}^{n} I(S_i; S^*_k) - \sum_i \sum_{k=2}^{n} H(S^*_k) \tag{3}$$

With the objective function of Eq.(3), we can implement a sequential concept selection process as follows:

1. Select the 1st optimal S^*_1 with largest $g^*_1 = \sum_i I(S_i; S^*_1)$
2. Select the kth optimal S^*_k, $k > 1$, with largest $g^*_k = \sum_i [I(S_i; S^*_k) - H(S^*_k)] = -\sum_i H(S^*_k|S_i)$

The selection criteria listed above are actually quite intuitive - the best concept has the highest total mutual information with the rest of concepts, and the next optimal concepts are those having low conditional entropy (i.e., high correlations with the rest of concepts).

3.2. Further consideration

When g^*_k is large, accurate labels for S^*_k is desired to better estimate rest of the concepts. However if S^*_k can be classified accurately by the individual detector, i.e., user's labelling won't make big difference. In this case, it will be more beneficial to ask user to label a different concept that has low detection accuracy. Thus we need to take into consideration the robustness of the individual detectors. Furthermore, for a test image, if an individual detector has very high confidence on its hypothesis, we may risk assuming that the estimated label of this detector is close to the true label. Then user's labelling will not make big difference either. Those informative concepts whose individual detectors are not confident about their estimations should get higher priorities for user annotation.

From validation, we get the validation error rate $E_I(S_i)$ of the individual detector for every concept S_i, which estimates the average robustness of the individual detector. The confidence of an individual detector about its hypothesis can be measured by the entropy of its output: $H_I(S_i|x) = -P_I(y_i{=}1)\log P_I(y_i{=}1) - P_I(y_i{=}0)\log P_I(y_i{=}0)$. Note this measure is computed for each image, thus it is image dependent. Let $\tilde{g}^*_k$ be the normalized version of g^*_k (normalized to [0,1]). To take into consideration the mutual information among concepts as well as the robustness and confidence of individual detectors, our criterion selects "key" concepts as follows: the optimal S^*_k are those having the largest G^*_k, where:

$$G^*_k = \beta\tilde{g}^*_k + (1-\beta)E_I(S^*_k)H_I(S^*_k|x)\,. \tag{4}$$

Note the inclusion of the last term makes the selection criterion image-dependent. This is consistent with intuition - we should ask users different questions for different images.

4. EXPERIMENTAL RESULTS

The data set contains 137 international broadcast news videos from TRECVID 2005 [6] and contains 39 concepts. It is separated into 4 data sets for training detectors, different stages of fusion methods, and final testing (shown in Table 1). Pairwise co-occurrences of concepts in subshots are counted based on the training set, validation set, selection set 1 and selection set 2, based on which the mutual information in Eq.(2) is calculated. Our individual concept detectors are SVM-based classifiers over simple image features such as grid color features and texture, extracted from key frames of a video subshot. Such classifiers have been shown to be effective for detecting generic concepts [1]. The outputs of the SVM classifiers are transformed into probabilities through a sigmoid function.

Table 1. The data sets for experiments

Name	Size	Usage
Training Set	41837 subshots	train SVMs
Validation Set	4515 subshots	learn RBF kernel for SVM
Selection Set 1	6021 subshots	learn θ, η, γ, λ_i in Eq.(1, 2)
Selection Set 2	3011 subshots	learn β in Eq.(4)
Test Set	6506 subshots	performance evaluation

4.1. Evaluation of baseline CBCF

In this experiment, we evaluate the baseline CBCF approach in Fig.1 and the criterion in Sec.2 to select concepts for updating. After validation and the selection process 1, 16 concepts are predicted to benefit from the use of CBCF. These concepts are: airplane, building, charts, corporate-leader, desert, explosion-fire, government-leader, map, natural-disaster, office, people-marching, road, snow, truck, urban, and vegetation. Fig.3 shows the Average Precision (AP) for the 39 concepts of the baseline CBCF and the individual detectors on the test set. AP is the official TRECVID performance metric. It is related to the multi-point average precision value of a precision-recall curve. From the figure we can see that CBCF actually improves the detection of 18 concepts, and 14 of them are automatically picked up by the proposed criterion in Sec.2. Among the 16 selected concepts, 2 do not have performance improvements, i.e., the precision of the prediction method is 14/16 and the recall is 14/18. It is interesting to note that "government leader" and "office" are two concepts that benefit most from CBCF. Furthermore, most of the concepts not selected are rejected by η and γ in Eq.(2), which indicates that the context relation and the robustness of related context is important to qualify a concept for benefiting from CBCF.

4.2. Evaluation of the ACBCF approach

To show the effectiveness of our ACBCF method, it is compared with three other methods: the passive labelling (randomly select concepts for user to label), the baseline CBCF without active labelling, and the individual detector without concept fusion. Note that with user's interaction, some subshots are labelled for each concept. If we calculate precision or recall based on all the test subshots, the comparison will

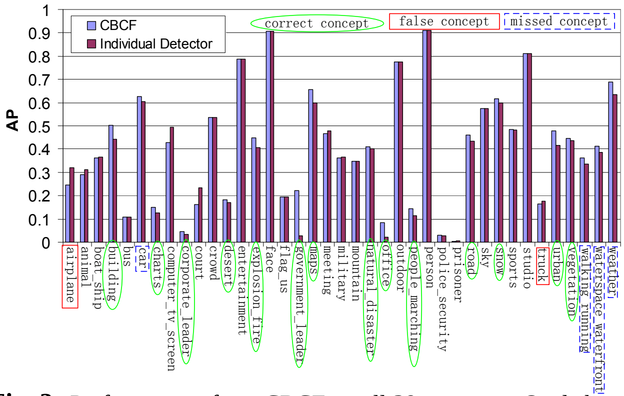


Fig. 3. Performance of our CBCF on all 39 concepts. Oval shapes indicate that CBCF improvement is correctly predicted for the given concept. Red boxes for false prediction and dashed boxes for misses.

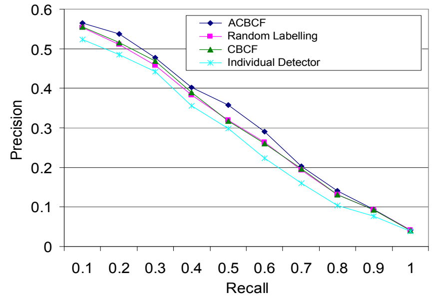
be unfair to the algorithms without user labelling. To avoid this problem, we compare the algorithms as follows: for each concept S_i, those subshots selected to be labelled by either our ACBCF or random labelling are taken out, and the rest of subshots are included for calculating precision or recall.

Fig.4 (a) shows the Precision-Recall curves (the precision is averaged on the selected 16 concepts) when the user labels only 1 concept for each subshot. Fig.4 (b) gives the corresponding AP on each of the selected 16 concepts. The figure shows that out of the selected 16 concepts, 14 has obvious performance improvements using our ACBCF. Furthermore the improvements for most concepts are significant, e.g. 650% for government-leader, 14% for map, 357% for office, 34% for people-marching, 13% for road, and 25% for urban. The Mean AP (MAP) of ACBCF on the entire 16 concepts is 0.322, which outperforms random labelling by 4.1%, outperforms baseline CBCF by 3.1%, and outperforms the individual detector by 11.8%. Moreover the ACBCF yields more stable results than both random labelling and CBCF. For example, as reported in Sec.4.1, on 2 of the selected concepts (airplane and truck) the CBCF performance actually deteriorates. The deterioration is especially severe for airplane. On the contrary, the proposed ACBCF remains more effective consistently across all selected concepts.

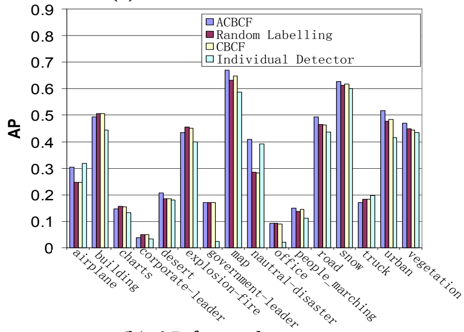
In addition, we evaluate the performance when different numbers of concepts are labelled by the user for each subshot. The MAPs on the selected 16 concepts are given in Fig.5. Compared with random labelling, when more concepts are labeled, the advantage of ACBCF is more significant. Also, on test set, when 1 concept is labelled for each subshot, outdoor is selected to be labeled for 70% of subshots. When 2 concepts are labelled, besides outdoor, airplane is the second most popular concept selected for 43% of subshots. When 3 concepts are labeled, court is selected for 36% of subshots as the third most popular concept besides outdoor and airplane.

5. CONCLUSION

We have developed a new statistical framework, active context-based concept fusion, for improving the performance of detecting semantic concepts in images and videos. The framework incorporates user's interaction to annotate a small number of key concepts per image, which are then used to improve detection of other concepts. In addition, we propose a simple yet accurate method in predicting concepts that may benefit from context-based fusion. Our experiments over TRECVID05 data have confirmed the effectiveness of the proposed paradigm.



(a) Recall-Precision curve



(b) AP for each concept

Fig. 4. Performance with 1 concept labelled for each subshot.

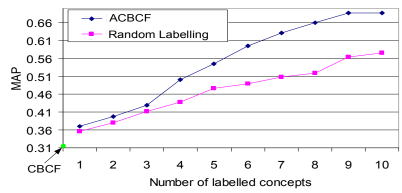


Fig. 5. MAP with different numbers of labelled concepts.

6. REFERENCES

[1] A. Amir, et al., "IBM research TRECVID-2003 video retrieval system", *Proc. NIST Text Retrieval Conf. (TREC)*, 2003.

[2] M.E. Hellman and J.Raviv, "Probability of error, equivocation, and the chernoff bound", *IEEE Trans. on Information Theory*, 16(1970), no.4, 368–372.

[3] M.R. Naphade, et al., "A factor graph framework for semantic video indexing", *IEEE Trans. on CSVT*, (12)1:40-52, 2002.

[4] S. Paek and S.-F. Chang. "Experiments in Constructing Belief Networks for Image Classification Systems", *Proc. ICIP*, Vancouver, Canada, Sept. 2000.

[5] J. Smith, et al., "Multimedia semantic indexing using model vectors", *Proc. ICME*, vol.2, pp. 445-448, 2003.

[6] TRECVID, "Trec video retrieval evaluation", *in http://wwwnlpir. nist.gov/projects/trecvid/*.

[7] N. Vasconcelos, "Feature selection by maximum marginal diversity: optimality and implications for visual recognition", *Proc. CVPR*, vol.1, pp.762-769, 2003.