DATA MODELING STRATEGIES FOR IMBALANCED LEARNING IN VISUAL SEARCH

Jelena Tešić *Apostol Natsev* *Lexing Xie* *John R. Smith*

IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY
{jtesic, natsev, xlx, jsmith}@us.ibm.com

ABSTRACT

In this paper we examine a novel approach to the difficult problem of querying video databases using visual topics with few examples. Typically with visual topics, the examples are not sufficiently diverse to create a robust model of the user's need. As a result, direct modeling using the provided topic examples as training data is inadequate. Otherwise, systems resort to multiple content-based searches using each example in turn, which typically provides poor results. We propose a new technique of leveraging unlabeled data to expand the diversity of the topic examples as well as provide a robust set of negative examples that allow direct modeling. The approach intelligently models a pseudo-negative space using unbiased and biased methods for data sampling and data selection. We apply the proposed method in a fusion framework to improve discriminative support vector machine modeling, and improve the overall system performance. The result is an enhanced performance over any of the baseline models, as well as improved robustness with respect to training examples, visual features, and visual support of video topics in TRECVID. The proposed method outperforms a baseline retrieval approach by more than 18% on the TRECVID 2006 video collection and query topics.

1. INTRODUCTION

In this paper we investigate a novel approach for improving the performance of the visual-based component of a video search system. In particular, we address the problem of querying using a small number of visual examples. There is extensive prior work on query analysis in the traditional fields of text-based Information Retrieval. The NIST TRECVID [1] benchmark search task motivated researchers to examine the value of semantic and context modeling in real-life scenarios. However, query modeling is much more difficult when confined to visual processing alone, and usefulness of the visual topic examples has been largely undermined. Meanwhile the need for composing such topics or attempting to detect complex information with relatively few visual examples is increasing across domains like broadcast, content sharing sites, and security. We alleviate the problem of automatically formulating complex models of queries given only a (small) set of positive examples per topic of interest. The major contributions of this paper can be summarized as follows:

(i) We investigate the effects of pseudo negative data sampling and selection techniques for imbalanced learning. Specifically, we use a more sophisticated modeling of the descriptor space by means of smart data sampling and selection. This results in a significant improvement of the underlying support vector machine model, and ultimately, of the system performance.

This material is based upon work funded in part by the U. S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government.

(ii) We estimate the overall impact of the visual-based search component on the search performance over a series of topics within the scope of the TRECVID benchmark's Search Task.

We find that applying multiple biased sampling and selection methods across a variety of features results in enhanced performance over all of the baseline models, as well as in improved robustness with respect to training examples, visual features, and visual support.

The experiments in this paper were performed using the 2003, 2005 and 2006 TRECVID corpora and 73 query topics. We represent video shots and query examples with a data point in four descriptor spaces: global color (166), local color (225), global texture (96) and local texture (108). The number in the brackets note the dimensionality of the descriptors space. We use standardized TRECVID average precision (AP) as a performance measure. AP emphasizes returning more relevant documents earlier. Further details are provided in [2]. Section 2 summarizes our search-by-visual-example framework, Section 3 describes the key improvement elements of the visual component that made a significance contribution to our search system, while Section 4 presents our experimental results, followed by concluding remarks in Section 5.

2. VISUAL-BASED SEARCH

The problem of search-by-visual-example takes one or more images as input, and tries to answer the intuitive questions of "retrieve more that look like these". More recent development in large-scale multimodal retrieval shows that search-by-visual-example is a effective component of a retrieval system and it filters and re-ranks text-based retrieval results so that the precision of the final result is largely improved [1].

The visual-based component of the IBM TRECVID search system uses the unique approach of formulating the topic answering problem as a discriminant modeling one. The combination hypothesis modeling method, proposed initially in [3], fuses the selective MECBR (multi-example content based retrieval) approach with the discriminant SVM (support vector machines) one. Detailed design is presented in [2, 3]. Figure 1 illustrates the main idea. Circles show a single content-based retrieval (CBR) query and associated nearest neighbors. The full MECBR baseline is then achieved by fusing the individual CBR query results using OR logic (e.g., overall distance of a candidate point is computed as the closest distance to any of the query points). The SVM approach with nonlinear kernels, on the other hand, allows us to learn nonlinear decision boundaries even when descriptors are high dimensional. We fix the kernel type to Radial Basis Kernels, and select global SVM kernel parameters for each descriptor to avoid over-fitting. Since there are no negative examples provided, we generate pseudo-negative examples by randomly sampling data points. We build a set of primitive SVM classifiers, whereby the positive examples are used commonly across all classifiers but the pseudo-negative data points come from different

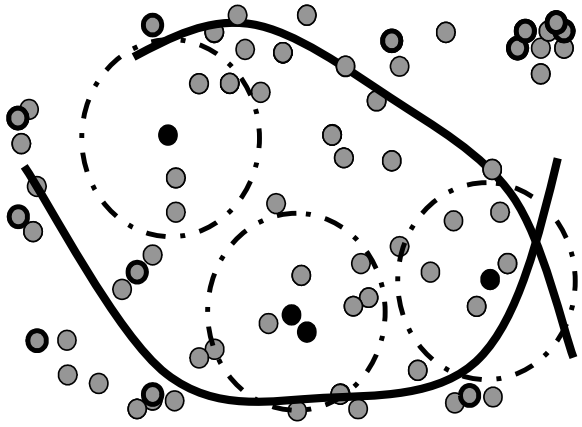

Fig. 1. Learning a Combination Hypothesis. Each line represents a primitive SVM hyperplane between the same set of positive examples (black filled) and a randomly sampled bag of pseudo-negative examples (black edge). Each dash-dot circle represents a single CBR query and associated nearest neighbors within the circle. The intersection of the SVM hyperplanes with the union of the CBR circles forms the final set of results.

sample sets. The SVM scores corresponding to each primitive SVM model are then fused using AND logic (e.g., MAX distance score aggregation) to obtain a final discriminative model, as illustrated by the dividing lines in Figure 1. This combination hypothesis approach was shown to significantly improve retrieval results over either of the individual approaches on the TRECVID 2003 video corpus and query topics [3].

3. DESCRIPTOR SPACE MODELING

One primary limiting factor for visual search is that there is only a *very* small number of distinct positive examples, and no negative examples. We propose two strategies to address this challenge: (a) fusing a number of primitive SVM predictions trained on the same set of positives and different views of pseudo-negative data points so that the final SVM model corresponds to the intersection of several hyper-spaces, and (b) sampling pseudo-negative data points so that they model the test space well. The objective here is to carefully select the pseudo-negatives to model the input space well, and to balance the number of pseudo-negative data points with the number of positive examples to avoid the imbalanced learning problem [4]. The descriptor space is high-dimensional, and its structure can run counter to intuition based on Euclidean spaces of small dimensionality [5]. In statistics, the "dimensionality curse" refers to the fact that a large number of observations is needed to obtain an acceptable estimate in high dimensions. Samples quickly become "lost" in the wealth of the space, while simultaneously, the required sample size increases exponentially with dimensions [6]. Since the descriptor dimensionality is typically large, there is only a small probability that any data point will be inside the nearest-neighbor sphere. Moreover, results in [5] demonstrate that all points converge to the same distance from the topic point in high dimensions, and the concept of nearest neighbors becomes meaningless if there is no inherent smaller dimensionality. Therefore, the inherited objective is to maximize the number of selected pseudo-negative data points in the descriptor space. We propose to:

- *maximize* the number of pseudo-negative data points under constraints of imbalanced learning and complexity, and
- *carefully* select data points so that the descriptor space is well represented.

3.1. Increasing the Data Sample

In the SVM fusion framework of primitive models, we select $K * N$ pseudo-negative points for training in a bagging framework, where K is the number of primitive SVM models to be fused (i.e., number of bags), and N is the number of pseudo-negative data points selected for each primitive model (i.e., bag size). We can increase the overall number of selected pseudo-negatives in two ways: (a) maximize the number of primitive SVM models, K, and (b) maximize the number of data points selected for each primitive model, N.

Number of primitive SVM models. Due to the inherent nature of the high-dimensional descriptor space and the over-fitting tendency of the SVM classifier with small training samples, it can be expected that the performance of a fused outcome of a number of primitive SVM classifiers would be better than any single run. In [3] we confirmed that fusing multiple primitive models outperforms the single-model (no bagging) approach. However, the time to perform a query is proportional to the number of primitive models chosen so K cannot be too large. To be realistic, we limit K to 10, as the number of views to be fused. Experiments in global color space using TRECVID 2003 dataset show that increasing K from 1 to 10 improves MAP from 0.05 to 0.0784. Further increase does not result in additional performance gains as the performance has already saturated at $K = 10$: for $K = 20$, MAP is 0.779, for $K = 30$, MAP is 0.0781, and for $K = 40$ MAP is 0.0787.

Imbalanced learning. In selecting the number of pseudo-negative points N for each primitive SVM model, the objective is to minimize the under-sampling rate of negative examples while avoiding the imbalance problem in the learning process [4], and therefore we need to balance the ratio of negatives and positives rather than maximize the number of negatives alone. We adopted $N = 50$ as a fixed pseudo-negatives bag size in [3]. We revisit this decision here, and make N a function of P for every topic, where P is the number of visual examples for the topic. As reported in [7], maximum ratios should be less than 10 ($max\{N/P\} < 10$) so that SVM classifiers perform correctly. In order to investigate the method sensitivity to the number of pseudo negative data samples used, we compared the performance of our baseline approach of [3] with the balanced ratio approach for the 2003 and 2005 TRECVID search topics. We consider four visual descriptor spaces and take the fusion over all four spaces, as that proved to be more robust than any of the individual descriptors [2]. In both cases (2003 and 2005 topics), the overall mean AP of the visual component exhibited an improvement, as shown in Table 1. The 2003 and 2005 TRECVID evaluations show that the average precision measure in video retrieval is strongly influenced by the number of pseudo-negative samples, especially for the queries with low average precision. Maximizing the total number of pseudo-negatives $K \times N$ improved the modeling of the topic space without compromising the learning by over-sampling negatives. These experiments show that the dependency on modeling is higher for the more difficult topics, and that modeling can have a significant impact on them.

3.2. Smart Sampling and Selection

In the machine learning community, a number of solutions to the class-imbalance problem have been proposed both at the data and al-

Topics (year)	N=50	N=10*P	% change
2005	0.085781	0.087707	2.25%
2003	0.076184	0.078634	3.22%
AP range (#topics)			
$AP < 0.01$ (11)	0.004767	0.006904	44.84%
$AP \in (0.01, 0.04)$ (9)	0.022979	0.028490	23.98%
$AP > 0.15$ (5)	0.364495	0.360221	-1.17%

Table 1. Retrieval performance measured using mean AP and gain percentage w.r.t. number of primitive models used: 2003 and 2005 visual topics, and over 3 groups of 2003 topics grouped by baseline visual AP

gorithmic level, and it has been shown that smart sampling can help preserve the class distribution [7]. Smart sampling needs to provide a diverse set of pseudo-negative data points that well capture the diversity in the descriptor space for discriminant learning purposes. We have investigated sampling pseudo-negative points from the bottom of the ranked list obtained by running MECBR. As can be observed in Figure 2, checkered points far from positive examples do not make good training candidates as the learned model is not discriminative enough and it includes a large portion of the descriptors space.

Descriptor space modeling using pseudo-negative data selection involves two stages: (a) sampling of the data points (b) selection of the data points for each primitive SVM. We have investigated two approaches for sampling and selection of $N \times K$ points from the dataset: (i) **random sampling**—randomly sample $K \times N$ points and, for each bag, select N random points from the sample, and (ii) **cluster-based sampling**—cluster the descriptor space using k-means clustering so that the resulting number of clusters be up to $2 \times N \times K$, and for each bag, randomly select N of the centroids of the formed clusters as pseudo-negative data samples.

The probability of finding near-duplicate positive examples in the test set is low since the query examples usually are from outside of the targeted test set. In any case, we have established a low distance threshold of $\epsilon = 0.01$. If a selected data point falls within ϵ distance of a positive example, we treat it as a *pseudo-positive* example, since increasing the number of positives may further help the training process.

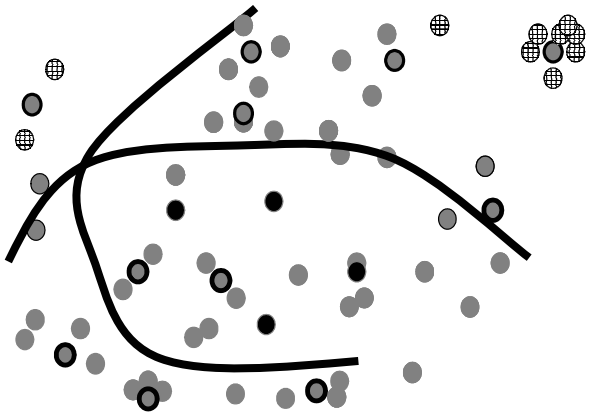

Fig. 2. Data modeling considerations in selecting pseudo-negatives: (i) checkered points—distant negative examples, which are not very useful for discriminant learning (ii) gray points with black edge—centroids of data clusters, which are more representative and useful.

Based on the data sampling and selection, we propose five different domain modeling approaches: (1) *SVM-RANDOM* approach randomizes both data sampling and data selection, and it is used as a baseline model, (2) *SVM-CLUSTER* approach uses centroids for data sampling. Training is further boosted by assigning a positive label to a set of clusters closest to the positive data points, (3) *SVM-CLUSTER-OUT* approach uses the same approach as *CLUSTER*, but the pseudo-negatives are sampled both from within the test set as well as from an outside set from the same domain. Benefits come from the fact that two different data sources are used for modeling, and thus the SVM models learned can be complementary.

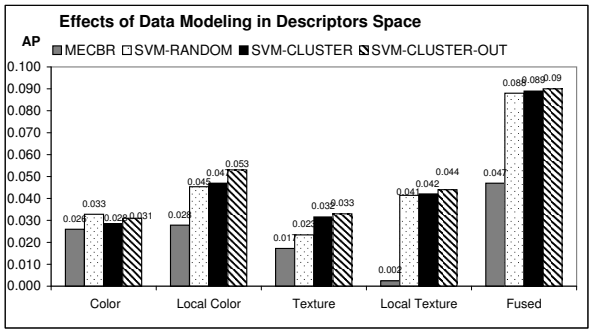

Fig. 3. Retrieval performance of data modeling approaches measured using MAP over 2005 TRECVID query topic and four descriptor spaces.

4. EXPERIMENTS

We evaluate the impact of data modeling techniques to visual search, and the overall improvement over a text-based retrieval baseline using TRECVID 2005 and 2006 datasets. The two collections contain 401 broadcast news video programs from six U.S., Arabic, and Chinese channels, recorded in the fourth quarter of 2004 and 2005, respectively, and are segmented into total of 125,249 shots. Each video comes with a speech transcript obtained through automatic speech recognition, as well as machine translation for the non-English sources. Each collection has 24 NIST-distributed queries with pooled ground-truth. The dataset used for outside collection sampling in the *CLUSTER-OUT* approach is the TRECVID 2005 development set.

In the first two experiments, we evaluate the proposed data modeling methods in Sec. 3.2 for visual search using four descriptors. All the approaches were tested for K=10, and N=10*P, as described in Sec. 3.1. Due to the space limitation, we will not present evaluation per topic, but report the average evaluation over 24 queries per descriptors space on TRECVID 2005 only, shown in Figure 3. The overall improvement of the cluster methods over the *MECBR* baseline, and over the *SVM-RANDOM* baseline are shown in Figure 3. Although the overall performance improvement after fusion across all four descriptors is modest (3%), the gain on a per-descriptor basis can be significant. For example, the performance gain over *MECBR* is 100% in the local color space, and the gain over the *SVM-RANDOM* approach is nearly 40% in the global texture space. This confirms our findings in Table 1 that better query modeling and sampling approaches are relevant in some scenarios but not in others. In particular, depending on the discriminative power of the low-level descriptors used, and the diversity of the query topic examples, data modeling and smart sampling can lead to substantial improvements (as in the case of local color and global texture), or can lead to very small improvement or even a loss (as in the global color case). Further work is needed to determine good criteria when to apply biased pseudo-negative sampling and selection approaches.

Next, we evaluate the impact of the visual-based component on the overall video search performance by fusing it with a text-based

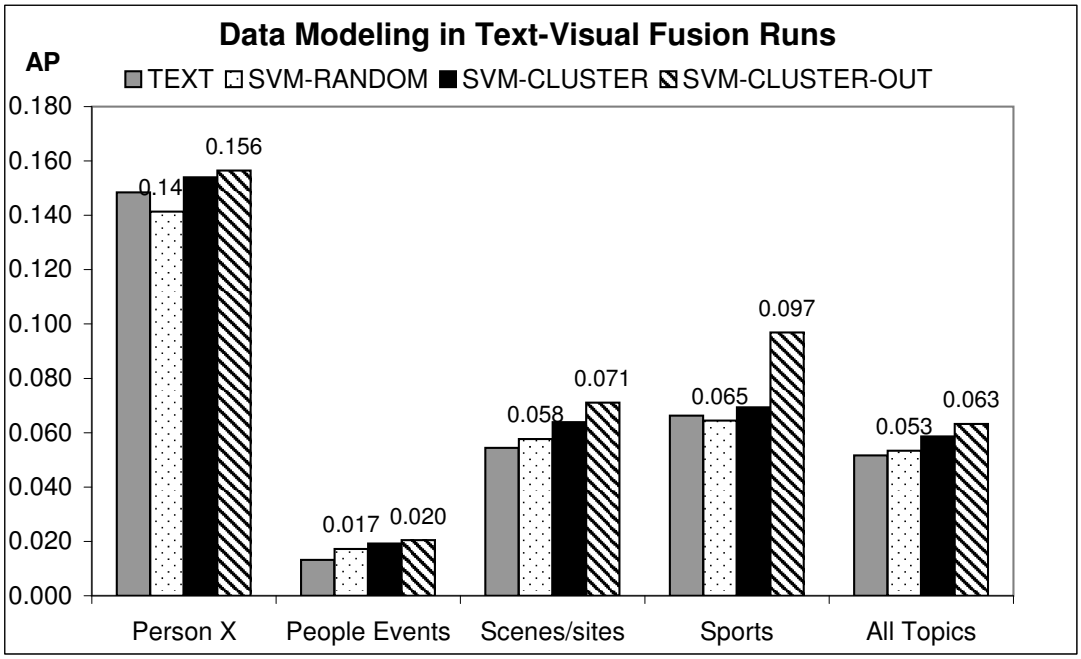

Fig. 4. Retrieval performance of data modeling approaches measured using mean AP over query topic classes in 2006.

retrieval component (based on speech transcripts). Our speech-based retrieval baseline is based on the JuruXML semantic search engine using story boundary and topic refinement approaches [8]. Visual and speech-based retrieval components are fused using topic dependent weights [9], where we learn the fusion weights on one corpus and apply the weights to the other. The results are shown in Figure 4.

From the results, it is evident that the *SVM-RANDOM* visual component is not that strong and attributes to only 3% improvement over the text baseline, thereby justifying the need for a more robust visual-based search. The *SVM-CLUSTER* approach offers a more significant 12% improvement over the text baseline, and it is 10% better than the *SVM-RANDOM* baseline. The *SVM-CLUSTER-OUT* run gives the best performance, resulting in 21% improvement over the text baseline. Our conjecture that sampling from an outside set can help prevent over-fitting on the time period and diversify the pseudo-negative sample set was also validated, as the *SVM-CLUSTER-OUT* run improved over the *SVM-RANDOM* run by 18%, and outperformed the *SVM-CLUSTER* approach by 8%. In conclusion, modeling of the descriptor space and different data inputs are helpful tools in visual-based search. Regardless of the visual relevance of the topic samples, we have demonstrated that visual-based search improves the existing text (speech) retrieval baseline by more than 20% on the difficult topics from the TRECVID 2006 benchmark.

5. CONCLUSION

In this work we proposed a robust framework for the visual-based component of video search, and demonstrated the need for this component in answering queries with few examples. The proposed modeling strategies of the training space were designed to overcome the imbalanced learning problem and high-dimensionality of descriptors using smart data sampling and selection. We find that applying multiple biased sampling and selection methods across a variety of features results in enhanced performance over any of the baseline models. More importantly, we have proved that the sophisticated approach to modeling of the training samples improves visual search and consistently improves the text baseline over a range of visual examples and a range of visual support of the diverse topics in TRECVID benchmark: up to 53.43% for 2005 and 21.54% for 2006 TRECVID topics. We are working on context-based modeling of negative samples for each primitive model, and on further up-sampling of positive examples.

6. REFERENCES

[1] P. Over, T. Ianeva, W. Kraaij, and A.F. Smeaton, "Trecvid 2006 an introduction," in *NIST TRECVID-2006 Workshop*.

[2] M. Campbell et al, "IBM research TRECVID-2006 video retrieval system," in *NIST TRECVID 2006 Workshop*.

[3] A. Natsev, M. Naphade, and J. Tešić, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *ACM Multimedia*, 2005.

[4] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *International Conference on Artificial Intelligence*, Las Vegas, Nevada, 2000.

[5] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is "Nearest Neighbor" Meaningful?," in *International Conference on Database Theory (ICDB)*, 1999.

[6] B.W. Silverman, *Density Estimation for statistics and data analysis*, Chapman and Hall, January 1986.

[7] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *15th European Conference on Machine Learning (ECML)*, 2004.

[8] T. Volkmer and A. Natsev, "Exploring automatic query refinement for text-based retrieval," in *International Conference on Multimedia and Expo(ICME)*, 2006.

[9] L. Xie, A. Natsev, and J. Tešić, "Dynamic multimodal fusion in video search," in *International Conference on Multimedia and Expo(ICME)*, 2007.