IMPACT OF NOVEL SOURCES ON CONTENT-BASED IMAGE AND VIDEO RETRIEVAL

Arnab Ghoshal, Sanjeev Khudanpur

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, USA

Dietrich Klakow

Spoken Language Systems
Saarland University, Saarbrücken, Germany

ABSTRACT

The problem of content-based image and video retrieval with textual queries is often posed as that of visual concept classification, where classifiers for a set of predetermined visual concepts are trained using a set of manually annotated images. Such a formulation implicitly assumes that the training data has similar distributional characteristics as that of the data which need to be indexed. In this paper we demonstrate empirically that even within the relatively narrow domain of news videos collected from a variety of news programs and broadcasters, the assumption of distributional similarity of visual features does not hold across programs from different broadcasters. This is manifested in considerable degradation of ranked retrieval performance on novel sources. We observe that concepts whose spatial locations remain relatively fixed between various sources are also more robust to source mismatches, and vice versa. We also show that a simple averaging of multiple visual detectors is more robust than any of the individual detectors. Furthermore, we show that for certain sources using only 20% of the available annotated data can bridge roughly 80% of the performance drop, while others can require larger amounts of annotated data.

Index Terms— Multimedia systems, Information retrieval, Robustness

1. INTRODUCTION

In recent years, content-based video retrieval has been cast as a visual concept classification problem [1, 2]. While these methods have been shown to be effective to various degrees on different datasets, very little work has been done to analyze their robustness to mismatched training-test conditions. The NIST TRECVID retrieval evaluations provide an excellent opportunity for doing such analysis. Over the last few years, the data collected for these evaluations have come from a variety of sources, often quite different from the manually annotated data that the participating systems are trained on. A study of the trend in performance of all state-of-the-art systems on the "High-level feature extraction" task of recent TRECVID evaluations [3, 4, 5], clearly show that mismatches in the sources of training and test data lead to severe degradations in retrieval performance.

We have previously proposed adaptation of visual concept detectors to the different video sources, and showed significant improvements in retrieval performance through source adaptation [6]. One can argue that the efficacy of source adaptation demonstrates that the visual features are not robust to source mismatches. In other words, this indicates the presence of source dependent covariates in the visual data, and demonstrates that modeling such covariates can improve prediction of visual concepts. Depending on the significance that the covariates play in prediction, one would also expect the prediction accuracy to be adversely affected for a particular source when the visual detectors are not trained using data from that source.

In this paper we investigate the effects of source mismatch on video retrieval using the TRECVID 2005 development dataset. We show that programs from different broadcasting sources differ noticeably in the distributions of visual features, which indicate that the visual data from different sources may not be very good predictors of each other. In section 3 we simulate the effect of novel sources by leaving each source out of the training, one at a time. Finally, in section 3.4 we adapt the baseline models to the novel sources by using different percentages of the available annotated data.

2. HIDDEN MARKOV MODELS FOR IMAGE RETRIEVAL

Let a collection $\mathcal{L} \equiv \{(I_1, C_1), \ldots, (I_L, C_L)\}$ of image+caption pairs be provided, where the caption $C_l = \{c_1^l, \ldots, c_{N_l}^l\}$ denotes the set of visual concepts present in image I_l. Let $I \equiv \langle i_1, \ldots, i_T \rangle$ denote segments (regions) in image I, and let $x_t \in \mathbb{R}^d$ represent the visual features (like color, texture, edges) of each image region i_t.

Following the formulation presented in [6], we train a pair of HMMs for each concept c: an HMM $\mathcal{H}_c^+$ trained on all images labeled with the concept c, and an HMM $\mathcal{H}_c^-$ trained on rest of the images. The models are trained using the standard maximum-likelihood (ML) training procedure.

For an unlabeled image collection $\{I\}$, we calculate the likelihoods $\ell(I|\mathcal{H}_c^+)$ and $\ell(I|\mathcal{H}_c^-)$ for each concept c:

$$\ell(I|\mathcal{H}_c^+) = \sum_{s_1^T} \prod_{t=1}^{T} f_{\mathcal{H}_c^+}(x_t|s_t)\, p_{\mathcal{H}_c^+}(s_t|s_{t-1}),$$

where $f_{\mathcal{H}}(\cdot|\cdot)$ denote the *emission densities*, modeled as mixtures of Gaussians, and $p_{\mathcal{H}}(\cdot|\cdot)$ denote the *transition probabilities* for HMM $\mathcal{H}$. We compute the posterior probability of a concept given an image as:

$$\texttt{score}(I, c) = \frac{\ell(I|\mathcal{H}_c^+)\, p(c)}{\ell(I|\mathcal{H}_c^+)\, p(c) + \ell(I|\mathcal{H}_c^-)\, (1 - p(c))}, \qquad (1)$$

which is used to rank-order the images for each concept c.

The likelihood of individual image blocks can be calculated as

$$\ell(i_t|\mathcal{H}) \equiv \sum_{s_1^T \in \mathcal{S}^T} f_{\mathcal{H}}(x_t|s)\, I_{(s_t=s)}\, p(s_1^T),$$

where $I_{(s_t=s)} = 1$ iff s is the state reached at position t, and $I_{(s_t=s)} = 0$ otherwise; $p(s_1^T) = \prod_{t=1}^{T} p(s_t|s_{t-1})$ is the probability of the state sequence. Using $\ell(i_t|\mathcal{H})$ in equation 1, we can similarly compute $p(c|i_t)$, the posterior probability of a concept c, given each image block.

This work was sponsored by the NSF PIRE Grant No OISE-0530118.

3. EFFECT OF NOVEL SOURCES

The TRECVID 2005 dataset consists of about 170 hours of video from 13 different news programs from the following 6 broadcasters – CCTV and NTDTV in Mandarin, LBC in Arabic, and CNN, MSNBC, and NBC in English [3]. The collection is divided by NIST into a 74K-keyframe (137 videos) development (DEV) set and 78K-keyframe (140 videos) evaluation (EVAL05) set. Keyframes in DEV are manually marked for the presence of 39 selected concepts from the LSCOM-lite [7] concept vocabulary. We further divide DEV into a training (TRN) set $\mathcal{L}$ of 57K keyframes from 107 videos, and a validation (VALID) set of 17K keyframes from 30 videos.

Each keyframe is segmented into a 5×7 rectangular grid of 50×50 pixel blocks, from which we extract low-level visual features like color moments, oriented edge strength features, and gray-level co-occurrence features for texture. The features are used in two different setups – 1) all features are stacked together in a single 96-dimensional vector and PCA is performed to decorrelate the dimensions, whiten the data, and reduce the feature vector dimension to 80 (we refer to this as the 'Fusion' feature); 2) in the second setup PCA decorrelation, whitening and dimensionality reduction is performed on each feature type separately to yield 9-dim color, 63-dim edge, and 14-dim texture features. Note that the 'Fusion' features are not simply a concatenation of these 3 vectors, and hence have a different dimensionality.

3.1. Source-specific differences in video

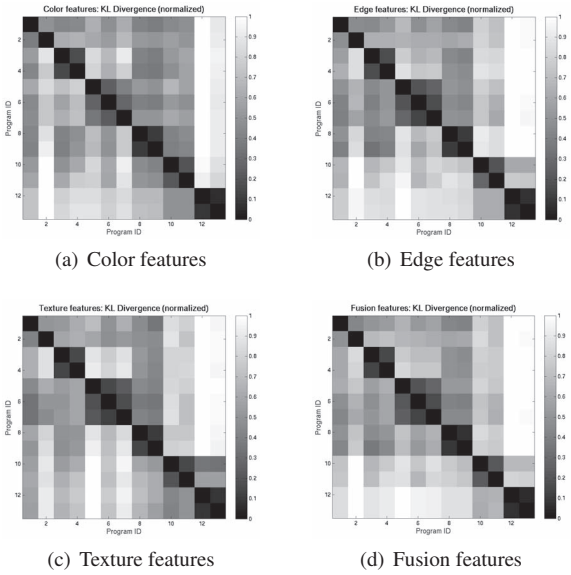
Video data collected from different television programs tend to "look" different. The differences can be attributed to, among others, different production settings and guidelines like studio design, on-screen graphics, recording equipments etc. One way of quantifying the differences between the programs is using the Kullback-Leibler Divergence (KLD) between the distributions of various visual features extracted from the videos of each program.

Let $\mathcal{L}_1, \ldots, \mathcal{L}_K$ be the K disjoint source-specific subsets, corresponding to the 13 different news programs in the TRECVID 2005 data. Further, let $\{\bar{x}_1, \ldots, \bar{x}_T\}$ be the quantized visual features for an image I, and let $p_k(\bar{x})$ be the distribution of the quantized visual features for source k. We compute the KLD between each pair of sources, and the resulting 13×13 matrix of $D(p_{k_1}||p_{k_2})$ values for various feature types are plotted in Figure 1. The diagonals correspond to the divergence of the feature distribution of a source from itself, and are 0, by definition.

The low block-diagonal values show that programs from same broadcasters have similarly distributed color, edge and texture features. We can argue, based on the noticeable distributional differences among visual features from different broadcasters, that models trained on data from one source are expected to be relatively poor predictors of the data from a different source. Such an assertion is indeed validated by the retrieval results on novel sources.

3.2. Retrieval performance on novel sources

To simulate the effect of video data from novel sources we performed a leave-one-out experiment, by removing each of the 6 broadcasters from the training data, one at a time, and using the resulting models to index the validation set data from the broadcaster that was left out of training. We perform ranked-retrieval for each of the concepts and compute the mean average precision [8] (mAP) for all the visual concepts. The performance is compared with that of the situation where data from all the groups are present in training, which is the standard training setup. The results are shown in Table 1.



(a) Color features (b) Edge features

(c) Texture features (d) Fusion features

Fig. 1. KL Divergence between visual feature distributions of various news programs. The program IDs are from the following broadcasters: 1-2) CCTV4, 3-4) CNN, 5-7) LBC, 8-9) MSNBC, 10-11) NBC, and 12-13) NTDTV. We can see that programs from the same broadcaster have similarly distributed visual features.

From the results we see that for all the sources and all types of visual features in consideration, we see large degradations in retrieval performance. The average degradation for various visual features is plotted in Figure 2. Furthermore, we see that by using the average of the scores for different visual features, and average of the block-posteriors $p(c|i_t)$, not only do we get better performance than the individual features, the average also has the most robust performance overall. The robustness of the average score is consistent with several earlier results, most notably that of bagging predictors [9].

3.3. Spatial distribution of concept posterior probabilities

For each concept c, we compute a location-specific discriminant information $d(c,t)$, which is a measure of the average significance of the image region t in distinguishing between the sets of relevant and non-relevant images for the concept c. Let $p_c^+(t)$ (and $p_c^-(t)$) denote the mean posterior probability $p(c|i_t)$ at any given image location t, over the set of relevant (and non-relevant) images for the concept. Let $\sigma_c^+(t)$ and $\sigma_c^-(t)$ be the respective standard deviations. We define the location-specific discriminant information as:

$$d(c,t) \triangleq \frac{p_c^+(t) - p_c^-(t)}{\sqrt{\sigma_c^+(t)\sigma_c^-(t)}}. \quad (2)$$

As before, using our standard and leave-one-out setups, we compute the discriminant information for the *seen* and *unseen* cases of Table 1. We denote them as two vectors $\mathbf{d}_{seen}(c)$ and $\mathbf{d}_{unseen}(c)$, indexed by the location t. For a concept whose spatial location remains relatively fixed across various sources, $\mathbf{d}_{seen}(c)$ and $\mathbf{d}_{unseen}(c)$ are expected to be close together, whereas they are expected to be farther apart for concepts whose spatial location tends to vary. In other words, the distance between the vectors $\mathbf{d}_{seen}(c)$ and $\mathbf{d}_{unseen}(c)$ is a measure of how invariant the spatial location of a particular concept is across various sources.

Table 1. Effect of novel sources – Retrieval performance on TV05 VALID set.

Feature type		CCTV	CNN	LBC	MSNBC	NBC	NTDTV
Color	Seen	0.266	0.217	0.318	0.168	0.217	0.203
	Unseen	0.173	0.131	0.155	0.117	0.113	0.113
Edge	Seen	0.305	0.228	0.355	0.205	0.265	0.242
	Unseen	0.184	0.142	0.170	0.130	0.128	0.156
Texture	Seen	0.290	0.213	0.329	0.211	0.238	0.225
	Unseen	0.174	0.123	0.172	0.133	0.117	0.137
Fusion	Seen	0.322	0.245	0.372	0.224	0.280	0.255
	Unseen	0.199	0.159	0.195	0.154	0.147	0.169
Avg. of feats	Seen	0.336	0.265	0.404	0.253	0.300	0.270
	Unseen	0.228	0.176	0.237	0.179	0.169	0.175
Avg. of blocks & feats.	Seen	0.352	0.267	0.409	0.280	0.313	0.291
	Unseen	0.248	0.187	0.265	0.192	0.183	0.214

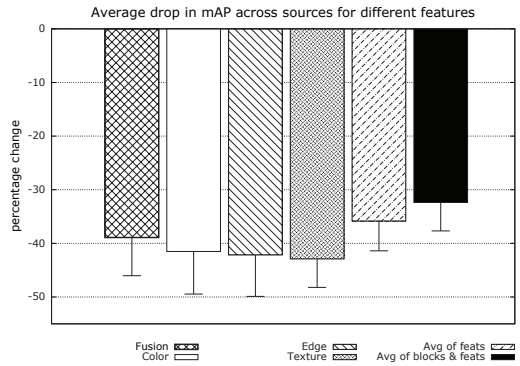


Fig. 2. Effect of novel sources – drop in retrieval performance on TV05 VALID set.

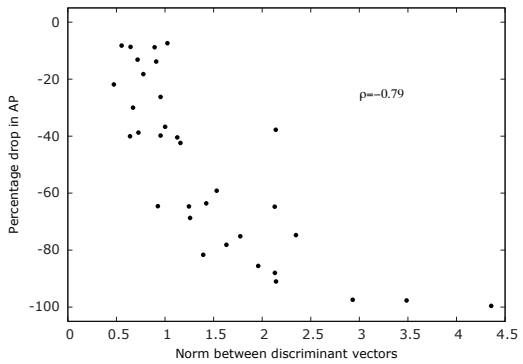


Fig. 3. Effect of novel sources – location specific concepts are more robust to source mismatches. Each dot corresponds to a concept.

We plot $\| \mathbf{d}_{seen}(c) - \mathbf{d}_{unseen}(c) \|_2$ against the drop in average precision for concept c from the *seen* source to the *unseen* source setup (Figure 3). The strong correlation seen in the plot ($\rho = -0.79$, p-value < 0.001) clearly indicates that concepts which tend to appear in fixed locations in video frames are more robustly retrieved.

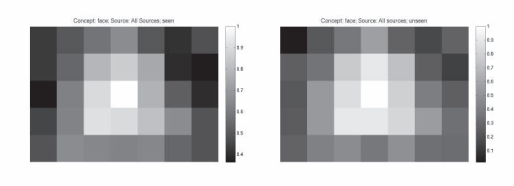
3.4. Using data from novel sources

We first proposed using source-dependent (SD) models, obtained by *adapting* source-independent (SI) models (i.e. those trained on data from all sources), to improve video retrieval in [6]. Based on that work we expect that a good way of utilizing annotated data from a novel source is to *adapt* the existing models to that source, using maximum *a posteriori* (MAP) adaptation [10]. More recently, [11] and [12] have also applied this idea of model adaptation to new video sources in a support vector machine (SVM) framework.[1] While [11] also uses the TRECVID 2005 dataset, important differences with the current work are: 1) the experimental setup presented here is more stringent since we remove *all programs* of a broadcaster from the training data to simulate the effect of novel source, because models trained on data from a program are often found to be good predictors for other programs of the same broadcaster; and 2) we present the learning rate characteristics for different sources across a much broader spectrum of labeled dataset sizes.

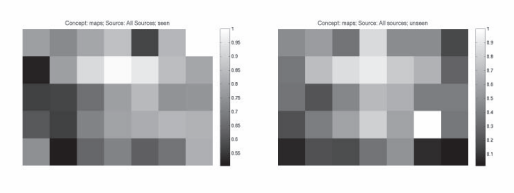
To estimate the amount of adaptation data required for a source, we adapt the models using random subsets containing 20%, 40%, 60%, 80%, and 100% of the available data from each source. For each of the 6 sources, we compare the results of three different systems — the baseline is when no source-specific data is used in training (the *unseen* results in Table 1), the second is the SI model which is trained on data from all sources (the *seen* results in Table 1), and the third system is the SD model obtained by adapting the SI model to various amounts of the source-specific data. For each source, we repeat the above experiment for 10 different random samplings of the training data, and plot the mean of the mean average precisions (mAP) and the standard deviation in Figure 5.

The plots reconfirm the expected result that model adaptation is a good way of utilizing annotated data from a novel source. The results also show that while certain sources like CCTV and NBC require relatively little annotated data to achieve reasonable performance (only 20% of available annotated data bridges roughly 80% of the performance drop between the matched and mismatched conditions), other sources like NTDTV or MSNBC can require larger amounts of annotated data to achieve the same performance gain. These are just some initial observations and more detailed analysis is needed to understand why the different sources exhibit different

[1] Similarities between our observations regarding effects of novel data sources on video retrieval, and those presented in [11, 12], together with the efficacy of model adaptation as an ideal method for utilizing small amounts of labeled in-domain data, should convince the reader that choice of classifier is not the overarching factor in content-based video retrieval.



(a) Face; seen AP = 0.92, unseen AP = 0.85



(b) Maps; seen AP = 0.20, unseen AP = 0.07

Fig. 4. Location-specific discriminant information for various concepts. Concepts like `face`, with little change in the distribution of $d(c,t)$ are more robustly retrieved, while less robust concepts like `maps` show noticeable differences in distribution of $d(c,t)$.

learning rate characteristics.

4. DISCUSSION AND CONCLUSIONS

In this paper we investigated the effect of novel sources of video data on the performance of content-based retrieval systems. We showed that one source of mismatch is in the commonly used visual features themselves. While we do not expect color, edge or texture features to be robust to changes in photometric conditions, more work is needed in identifying robust visual features. With the non-robust features, one can still contemplate learning a set of transforms for matching features from one source to another – a solution which can lead to gains over baseline conditions.

We have also demonstrated that concepts whose spatial location changes in novel sources are far less robustly detected than those whose location remains relatively fixed. While on one hand it has been observed that using location information significantly improves retrieval performance on the TRECVID tasks [13, 6], on the other hand using location information may make the models less robust. The model proposed in [14] disregards the location information, and generally performs worse. We are, however, investigating whether the model is more robust for certain visual concepts.

5. REFERENCES

[1] A. Amir *et al.*, "IBM Research TRECVID-2005 Video Retrieval System," in *Proc. TRECVID Workshop*, 2005.

[2] C. Snoek *et al.*, "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia," in *Proc. ACM Multimedia*, 2006, pp. 421–430.

[3] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton, "TRECVID 2005 - an overview," in *Proc. TRECVID Workshop*, 2005.

[4] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton, "TRECVID 2006 - an overview," in *Proc. TRECVID Workshop*, 2006.

[5] P. Over, G. Awad, W. Kraaij, and A. Smeaton, "TRECVID 2007 - overview," in *Proc. TRECVID Workshop*, 2007.

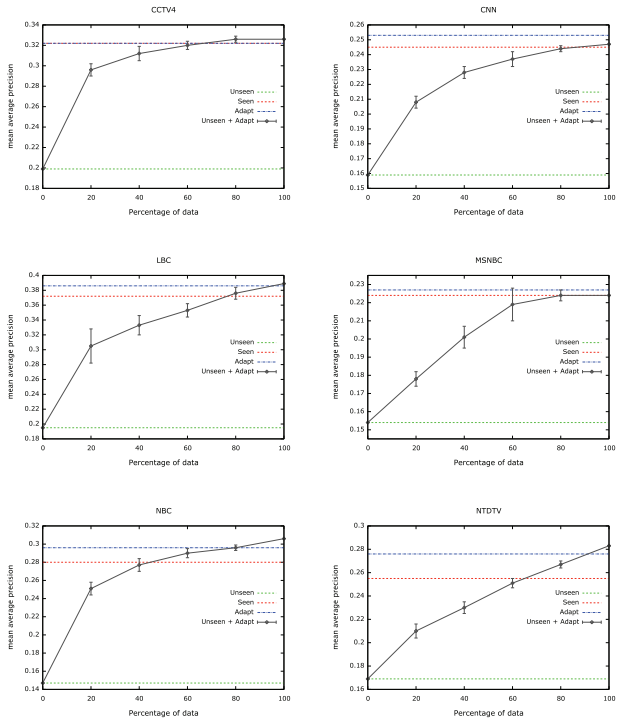


Fig. 5. Learning rates for various sources. While for sources like CCTV4 and NBC, very little data is sufficient to achieve good performance, for a source like NTDTV the learning rate is almost linear.

[6] A. Ghoshal and S. Khudanpur, "Source Adaptation for Improved Content-Based Video Retrieval," in *Proc. IEEE ICASSP*, 2006, pp. II–133–II–136.

[7] M. Naphade *et al.*, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, July 2006.

[8] E. M. Voorhees and D. K. Harman, Eds., *NIST Special Publication 500-250: The Tenth Text REtrieval Conference*. Department of Commerce, NIST, 2001.

[9] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, August 1996.

[10] J-L Gauvain and C-H Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[11] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proc. ACM Int. Conf. Multimedia*, September 2007, pp. 188–197.

[12] W. Jiang, E. Zavesky, S-F Chang, and A. Loui, "Cross-domain learning methods for high-level visual concept classification," in *Proc. IEEE Int. Conf. Image Processing*, October 2008.

[13] D. Klakow, "Using Regional Information in Language Model Based Automatic Concept Annotation and Retrieval Of Video," in *Proc. IEEE ICASSP*, 2006, pp. II–129–II–132.

[14] A. Ghoshal, P. Ircing, and S. Khudanpur, "Hidden Markov Models for Automatic Annotation and Content-Based Retrieval of Images and Video," in *Proc. ACM SIGIR*, 2005.