A Novel Statistical Approach for Image and Video Retrieval and Its Adaption for Active Learning

Moitreya Chatterjee
USC Institute for Creative Technologies
Playa Vista, CA, USA
metro.smiles@gmail.com

Anton Leuski
USC Institute for Creative Technologies
Playa Vista, CA, USA
leuski@ict.usc.edu

ABSTRACT

The ever expanding multimedia content (such as images and videos), especially on the web, necessitates effective text query-based search (or retrieval) systems. Popular approaches for addressing this issue, use the query-likelihood model which fails to capture the user's information needs. In this work therefore, we explore a new ranking approach in the context of image and video retrieval from text queries. Our approach assumes two separate underlying distributions for query and the document respectively. We then, determine the extent of similarity between these two statistical distributions for the task of ranking. Furthermore we extend our approach, using Active Learning techniques, to address the question of obtaining a good performance without requiring a fully labeled training dataset. This is done by taking Sample Uncertainty, Density and Diversity into account. Our experiments on the popular TRECVID corpus and the open, relatively small-sized USC SmartBody corpus show that we are almost at-par or sometimes better than multiple state-of-the-art baselines.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering, Retrieval Models; H.5.1 [**Multimedia Information Systems**]: Video (e.g., tape, disk, DVI)

Keywords

KL-Divergence; Active Learning; Clustering; Uncertainty

1. INTRODUCTION

The ever increasing volume of multimedia content (such as images, videos, etc.), thanks to the web, has necessitated effective multimedia search systems. The goal of these search (or retrieval) systems, is to be able to rank the content (multimedia samples) in order of relevance to a given text query by the user. The popular web-based image/video search systems, e.g. Google, Yahoo, etc. are common examples.

MM '15, October 26-30, 2015, Brisbane, Australia.

http://dx.doi.org/10.1145/2733373.2806368.

Conventional search systems typically, first automatically annotate the samples with multiple text labels (each typically one word long), specifying different objects, events, etc. called *concepts*. They then, treat every sample as simply a text document containing the labels and rank them based on the similarity of these labels with the user's query [11]. However the first step of automated annotation, tends to be noisy and introduces errors. Thus, these approaches are often not effective. Direct-retrieval approaches, inspired from cross-language retrieval techniques, bypass the task of annotation [5]. Such techniques treat the query to be coming from one language and the feature representations of the documents to be ranked, to be coming from some other.

For the task of ranking, many of these direct-retrieval systems [6, 9, 2], including the Continuous Relevance Model (CRM), uses query-likelihood as a ranking function [5]. The query-likelihood models assume a language model over the documents and the query to have been sampled out from that distribution. It then ranks the documents by computing the conditional probability of the query given a document. This style of modeling ignores the user's information need, since the query is assumed to be modeled by the document distribution [3]. This raises a central research question: Can there be a new way to rank the documents, without making such an assumption?

In this work, we address this fundamental issue by exploring a ranking function that captures the notion of relevance of the query and the document as a measure of distance.

Another major concern with these retrieval systems is that they require a large fully annotated training corpus for a good performance. The human-effort costs for doing this, is prohibitively high. This begs an answer to another research question: Can we achieve a good retrieval performance without requiring a fully annotated training data (i.e. would a partially annotated training dataset suffice)?

Active Learning addresses this issue by outputting an order of labeling the unlabeled training samples such that a good retrieval performance is achieved before all unlabeled data is queried for their labels. Typical active learning systems consists of a retrieval/classification engine and a sample selection engine, which ranks the unlabeled samples, typically by a measure of *informativeness* [11]. In our work, we determine this informativeness by combining measures of Sample Density and Diversity [2] with that of Sample Uncertainty, while the ranking function that we talked about is used for the retrieval engine. We call this system KLActive. Our experiments show that KLActive performs better than

the state-of-the-art on the popular, large TRECVID corpus and almost at-par on the smaller USC SmartBody corpus.

2. BACKGROUND

2.1 Query-Likelihood and CRM

The research community has already explored retrieval of multimedia content using Query-Likelihood as a ranking function [6, 9, 4]. The idea behind query-likelihood is to compute the quantity $P(w|\mathbf{r_i})$, the conditional probability of the text query word w given a M-dimensional feature vector $\mathbf{r_i}$, of an image/video sample i. Assuming we have t samples to rank, we compute $P(w|\mathbf{r_i})$ for each $i = 1, 2, ..., t$ and then rank them based on their conditional probability scores. CRM, one such approach, proposes to compute this probability as follows: $P(w|\mathbf{r_i}) = P(w, \mathbf{r_i})/P(\mathbf{r_i})$ [4]. The joint-distribution is then estimated from the training data ($Train$) by

$$P(w, \mathbf{r_i}) = \sum_{J \in Train} (P(J)P(w|J) \prod_{s_k \in \mathbf{r_i}, k=1}^{M} P(s_k|J)) \quad (1)$$

2.2 CBIR Systems and KL-Divergence

Content-Based IR (CBIR) systems seek to rank the image/video samples just like conventional multimedia search systems, except here the query is in the form of an image/video rather than text. The research community has addressed this question by utilizing distance measures to compute similarity between the query and the sample to be ranked. One such popular distance measure is the Kullback-Leibler Divergence (KL-Divergence) [7, 10]. KL-Divergence between two continuous distributions P and Q (defined over the variable $\mathbf{r}$) is defined as :

$$D(P||Q) = \int_{\mathbf{r}} P(\mathbf{r}) \log(P(\mathbf{r})/Q(\mathbf{r}))d\mathbf{r} \quad (2)$$

CBIR systems, typically compute a language model (a Bag-of-Words) representation for both the query and the sample to be ranked and are thus able to compute a KL-Divergence based similarity between the two distributions.

2.3 Active Learning

Active Learning is a technique for determining an optimal order of labeling the unlabeled samples in the training data, such that a system achieves a good performance at the task of retrieval/classification on unseen test data, even before it has been trained with the fully annotated training dataset. This is typically done by ranking the unlabeled training samples by a measure of *informativeness* and labeling the samples with a higher score first and using them to train the retrieval/classification engine [11, 2]. This process is repeated in batches. CRMActive, a promising recent approach, computes this *informativeness* measure by combining quantities measuring Sample Uncertainty (how sure is the retrieval engine about the relevance of a query to the sample in question), Sample Density (does the sample chosen for labeling represent the population or is it an outlier) and Sample Diversity (how much do the samples chosen for labeling resemble each other) [2].

For computing Sample Uncertainty, CRMActive first uses Normalized CRM [4] for annotating the unlabeled samples with concepts from the vocabulary. Assuming a vocabulary $\mathcal{V}$ consisting of k concepts, it computes $P(w_i|\mathbf{r})$, $w_i \in \mathcal{V}$, $i = 1, 2, ..., k$ for the image/video described by the M-dimensional feature vector $\mathbf{r}$ and picks the top-n concepts, where $0 < n < k$, as relevant for the particular image/video (in decreasing order of relevance). Now, Sample Uncertainty of an unlabeled sample $\mathbf{x}$, $unct(\mathbf{x})$ is obtained by computing the difference between the conditional probabilities of the most relevant label, i.e. $i = 1$ and the first irrelevant one $i = n + 1$ [2].

For computing Sample Density and Diversity, CRMActive clusters the samples in the feature space, taking into account the agreement between the labels of the samples of a cluster.

Sample Density of an unlabeled sample $\mathbf{x}$ in cluster C is then computed by

$$den(\mathbf{x}) = \frac{p(\mathbf{x})}{\max_{\mathbf{x}_i \in Train} p(\mathbf{x}_i)},$$

where $Train$ is the training set and $p(\mathbf{x})$ is the standard Gaussian Kernel Density Estimate:

$$p(\mathbf{x}) = \frac{1}{|C|} \sum_{\mathbf{x}_i \in C} K_{Gauss}(\mathbf{x}, \mathbf{x}_i)$$

and $|C|$ is the total number of samples in cluster C. For two M-dimensional vectors $\mathbf{x}$, $\mathbf{y}$, $K_{Gauss}(\mathbf{x}, \mathbf{y})$ is defined by:

$$K_{Gauss}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{M} (\frac{1}{\sqrt{2\pi\sigma}} \exp(-(x_i - y_i)^2/2\sigma^2)) \quad (3)$$

where σ controls the spread of the kernel.

Assuming there are T clusters, each "represented" by their respective centroids in set $\mathcal{S}$, $\mathcal{S} = \{rep(C_1), ..., rep(C_T)\}$, Sample Diversity of an unlabeled sample $\mathbf{x}$ is defined as:

$$div(\mathbf{x}) = 1 - \max_{\mathbf{x}_i \in \mathcal{S}} \frac{K_{Gauss}(\mathbf{x}, \mathbf{x}_i)}{\sqrt{K_{Gauss}(\mathbf{x}, \mathbf{x}) \times K_{Gauss}(\mathbf{x}_i, \mathbf{x}_i)}},$$

Finally, *informativeness* of an unlabeled sample $\mathbf{x}$ is defined as :

$$Info(\mathbf{x}) = \lambda_1 \times unct(\mathbf{x}) + \lambda_2 \times den(\mathbf{x}) + \lambda_3 \times div(\mathbf{x}), \quad (4)$$

where $\sum_{i=1}^{3} \lambda_i = 1$; $\forall i$, $\lambda_i \geq 0$.

3. PROPOSED APPROACH

CRM-based approaches use query-likelihood for ranking and assume the query to be sampled out from the distribution governing the documents. The query is not modeled directly, thereby ignoring the user's information needs. In order to mitigate this weakness, drawing inspiration from cross-language retrieval [3], we explore a model comparison based approach for ranking. Our approach assumes that there is a separate probability distribution governing both the query and the samples to be ranked. We now compare the similarity between the expected value of the query distribution and that of the sample distribution, using KL-Divergence. This permits the ranking of the samples.

Let θ_d be the distribution governing the samples to be ranked, $E_q\theta$ be the expected value of the query distribution and $\mathbf{r}$ be the random variable over which the distributions are defined. Then the KL-Divergence between these two distributions according to Equation 2 is proportional to:

$$-D(E_q\theta||\theta_d) \propto \int_{\mathbf{r}} E_q\theta(\mathbf{r}) \log \theta_d(\mathbf{r})d\mathbf{r}, \quad (5)$$

where both the query and the document distributions are approximated by kernels. θ_d is approximated by the Gaussian kernel (Equation 3) $K_{Gauss}(\mathbf{r}, \mathbf{r}_d)$, with $\mathbf{r_d}$ being the feature vector representing the document(d) to be ranked. However, different from CBIR approaches, our query distribution is not directly representable in terms of $\mathbf{r}$, since our queries are texts. Taking a cue from cross-language retrieval [3], we address this issue by the following mapping:

$$E_q\theta(\mathbf{r}) = \frac{\sum_{J \in Train} K_{Gauss}(\mathbf{r}, \mathbf{r}_J).\beta_J(q)}{\sum_{J \in Train} \beta_J(q)}, \quad (6)$$

where $\mathbf{r}_J$ is the feature vector of the current training image J, $Train$ is the training set, q is the query term and $K_{Gauss}(.,.)$ is the Gaussian kernel of Equation 3 and the query distribution, $\beta_J(q)$, has the following multinomial form:

$$\beta_J(q) = \lambda \frac{N_{q,J}}{N_J} + (1 - \lambda)\frac{N_q}{N} \quad (7)$$

where $N_{q,J}$ is the number of times q occurred in the annotation of image J, N_J is the number of annotation labels for image J, N_q is the total number of times q occurred in the training set, and N is the total length of all annotations across the training data. λ denotes a parameter that controls the degree of smoothing. Equation 6 computes the mapping statistic over the entire training set and is thus, in effect an expected value of the query distribution.

Note that the integral in Equation 5 is actually an integration over all M-dimensions of the features. Also, due to the limitation of expressing continuous integrals in computers, we approximate this representation by a sum over the corresponding feature values which occur in the dataset. This version of the Equation 5 constitutes our ranking function.

In order to extend this model for active learning, we compute measures of Sample Density and Sample Diversity following the cluster-refinement based approach proposed in CRMActive [2]. However, we determine Sample Uncertainty of an unlabeled sample $\mathbf{x}$ as follows:

$$unct(\mathbf{x}) = \frac{1}{-D(q_1||\mathbf{x}) - (-D(q_{k+1}||\mathbf{x}))}, \quad (8)$$

where $q_1, ..., q_k$ (in decreasing order of relevance) are the top-k most relevant labels assigned to $\mathbf{x}$ by our model. The denominator in Equation 8 gives a measure of the gap (distance) between the KL-Divergence measures of the most relevant label and the first irrelevant one with respect to the sample $\mathbf{x}$. This therefore points to the level of uncertainty the model has about the labels it assigns to $\mathbf{x}$. Now, we combine the three measures of Density, Diversity and Uncertainty using Equation 4. Thus, we have in place both the ranking system as well as the sample selection(order determining) system. We call this combined system KLActive.

4. EXPERIMENTS

4.1 Methodology

We conduct experiments to test the effectiveness of an algorithm at the task of retrieving relevant images/videos for a query concept and see how its performance evolves under an active learning setting. In our experiments, the goal of an algorithm is to rank the test samples by their similarity to a single word query without annotating the test samples, i.e. direct-retrieval. The algorithm starts with the initial training dataset, a small section of which is labeled and the rest being unlabeled. For each concept label in the vocabulary, the algorithm ranks the test samples by their similarity to the concept. It then selects a batch of K unlabeled training samples, we reveal the labels for these selected samples, and the algorithm repeats the ranking task. For each round, we compute precision scores of the algorithm for the top 5 retrieved images/videos per concept and report their average. We call this score (AP for the top 5 documents): P@5.

4.2 Datasets

TRECVID 2007: The portion of the TRECVID 2007 video corpus, which is annotated, has 110 short video clips, with a total of about 21,500 frames [1]. Each frame in every video is annotated with at most 16 concept labels selected from a set of 36 concepts such as "crowd", "building", "airplane", etc. that constitute the whole vocabulary. This corpus has been used extensively in video retrieval/annotation experiments [11, 2]. For every frame, a 225-dimensional feature vector (color moment, edge orientation histogram, wavelet PWTTWT texture) is computed, as described in Zha et al. [11]. The frames from 13 randomly selected videos constitute our test set for our experiments, while we use the rest of the data (frames from 97 videos) for training. We selected 4000 frames from the training data as the initial set of labeled samples, containing at least 1 positive example of every concept. We set the batch size (i.e. the number of samples which are annotated in every iteration), to be 2400.

USC SmartBody: SmartBody is an open virtual character animation platform. It ships with a library of 274 animation clips such as walking, hand beat gesture, pointing, etc. [8]. The animations are defined on a 3D skeleton consisting of 119 individual joints, the 3D coordinates of which are available from the SmartBody API. A dataset created with this software was recently made public [2]. Here, each video is annotated (not at the frame-level) using at most 6 concept labels from a vocabulary of 30 labels such as "Legs", "Arms", "Face", etc. The dataset along with the features - the differences between the minimum and the maximum values for the skeleton angles at 9 joints (neck, left(L)/right(R) shoulders, L/R elbows, L/R hip joints, and L/R knees)- is open to the research community. Using these features, we encode every video as a 9-dimensional feature vector. We randomly selected 24 videos for testing and used the rest of the data (250 animations) for training. We selected 40 animations from the training data as the initial set of labeled samples, containing at least one positive example of each concept. We now set, batch size (i.e. the number of samples which are annotated in every iteration), to be 23.

4.3 Baseline Systems

For the task of direct retrieval in an active learning context, we compare KLActive with three different approaches. The first and the second one being an active learning system that uses KL-Divergence based model comparison (KL-Random) and Normalized CRM (CRMRandom) as the retrieval engines respectively, while the samples are selected randomly in each batch of active learning. The results are averaged over 3 runs with different random seeds for both cases. The only prior approach that is known to us, that selects the samples non-randomly in this setting, is CRMActive [2]. We therefore, also compare our approach with theirs. For our experiments, the values of the parameters λ, σ and the validated parameters of the baseline approaches

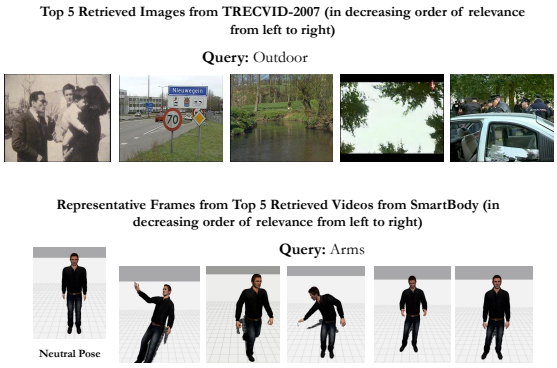

Figure 1: Example retrieval results by KLActive after Round 0 for TRECVID and SmartBody Datasets.

are all selected by performing 10-fold cross-validation on the first annotation batch. The values of all the parameters for the active-learning setup are reused from past work [2].

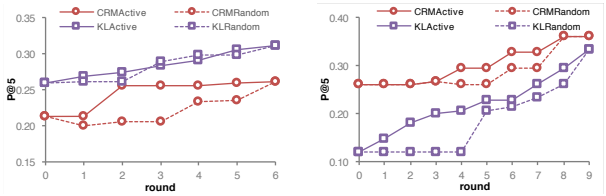

Table 1: AP scores for top-5 videos/images (on Y-axis) for retrieval on TRECVID (a), SmartBody (b).

4.4 Results and Discussion

Table 1 presents the retrieval results for the TRECVID Corpus and for the USC SmartBody corpus over different rounds of active learning. The first observation that stands out from the results is that all the models tend to improve their performance as successive rounds of active learning progress. This leads us to conclude that supplying more training data helps improve the performance. The next noteworthy observation is that for both the datasets the Active Learning version of both the CRM and KL-Divergence algorithms generally outperform their random counterparts. Thus it seems prudent, in the context of active learning, to label unlabeled samples based on a measure of sample *informativeness* rather than random sample selection.

However, the more interesting observation is a comparison of the two ranking schemes: KL-Divergence and query-likelihood based CRM. For TRECVID, clearly the KL- Divergence based approaches win while this is not the case for SmartBody. We hypothesize that this is primarily due to difference in dataset size. TRECVID is a much larger dataset compared to SmartBody. Further for the experiments, we start with much more samples in TRECVID (4000 as opposed to 40) and then in every pass we add more labeled samples (2400 as opposed to 23). Now recall that the idea in query-likelihood is to estimate the parameters of a distribution over the documents(θ_d) and the query is treated to be sampled out from this distribution. On the other hand in KL-divergence based model comparison, we assume separate distributions for the query(θ_q) and the documents(θ_d) and compare the two. This results in two fundamentally different ranking functions, Equation 1 for CRMActive and Equation 5 for KLActive. Due to the paucity of samples in SmartBody, KLActive fails to capture the true query distribution θ_q. This problem however, is less acute for TRECVID. Thus, the KL-Divergence based models do not perform as well on SmartBody as they do in TRECVID. This is also borne out by the fact that in SmartBody, the difference between the two approaches in terms of precision values at Round-0 (when there is less labeled training data) is much larger as compared to Round-9.

Figure 1 shows sample retrieval results for both SmartBody and TRECVID for sample queries in the first round using KLActive. The relevance of the retrieved results to the queries is very apparent, showing the model's effectiveness.

5. CONCLUSIONS

This work explores a novel ranking scheme for retrieval of multimedia content based on a statistical model comparison-based approach of query and document distributions. Furthermore, we extend the model in order to adapt it to an active learning setting. Experimental results show the effectiveness of our approach viz-a-viz other baselines. KLActive is a promising avenue to explore for future research.

Acknowledgements

The project or effort described here has been sponsored in part by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

6. REFERENCES

[1] Trecvid 2007: Trec video retrieval evaluation. link: http://www-nlpir.nist.gov/projects/tv2007/tv2007.html.

[2] M. Chatterjee and A. Leuski. An active learning based approach for effective video annotation and retrieval. *arXiv preprint arXiv:1504.07004*, 2015.

[3] V. Lavrenko. *A generative theory of relevance*. Springer Science & Business Media, 2008.

[4] V. Lavrenko, S. Feng, and R. Manmatha. Statistical models for automatic video annotation and retrieval. In *Proceedings of ICASSP*, volume 3, 2004.

[5] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[6] A. Llorente, R. Manmatha, and S. Rüger. Image retrieval using markov random fields and global image features. In *Proceedings of the ACM CIVR*, 2010.

[7] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *NIPS*, 2003.

[8] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann. Smartbody: Behavior realization for embodied conversational agents. In *AAMAS*, 2008.

[9] M. Wang, X.-S. Hua, Y. Song, X. Yuan, S. Li, and H.-J. Zhang. Automatic video annotation by semi-supervised learning with kernel density estimation. In *Proceedings of the 14th ACM Multimedia*, 2006.

[10] X. Wu, W.-L. Zhao, and C.-W. Ngo. Near-duplicate keyframe retrieval with visual keywords and semantic context. In *Proceedings of the 6th ACM CIVR*, 2007.

[11] Z.-J. Zha et al. Interactive video indexing with statistical active learning. *Multimedia, IEEE Transactions on*, 14(1), 2012.