Application of Discriminative Models for Interactive Query Refinement in Video Retrieval

Amit Srivastava, Saurabh Khanwalkar, Anoop Kumar
Raytheon BBN Technologies, Cambridge, MA, USA
Email: {asrivast, skhanwal, akumar}@bbn.com

ABSTRACT

The ability to quickly search for large volumes of videos for specific actions or events can provide a dramatic new capability to intelligence agencies. Example-based queries from video are a form of content-based information retrieval (CBIR) where the objective is to retrieve clips from a video corpus, or stream, using a representative query sample to find more like this. Often, the accuracy of video retrieval is largely limited by the gap between the available video descriptors and the underlying query concept, and such exemplar queries return many irrelevant results with relevant ones. In this paper, we present an Interactive Query Refinement (IQR) system which acts as a powerful tool to leverage human feedback and allow intelligence analyst to iteratively refine search queries for improved precision in the retrieved results. In our approach to IQR, we leverage discriminative models that operate on high dimensional features derived from low-level video descriptors in an iterative framework. Our IQR model solicits relevance feedback on examples selected from the region of uncertainty and updates the discriminating boundary to produce a relevance ranked results list. We achieved 358% relative improvement in Mean Average Precision (MAP) over initial retrieval list at a rank cutoff of 100 over 4 iterations. We compare our discriminative IQR model approach to a naïve IQR and show our model-based approach yields 49% relative improvement over the no model naïve system..

Keywords: Video Application, Video Retrieval, Relevance Feedback, Interactive Query Refinement, IQR.

1. INTRODUCTION

With the advent of technologies that simplify capturing and storing large amounts of video, retrieval methods are becoming essential to comb through large archives to find the videos of interest. Unmanned aerial vehicles (UAVs) are able to provide large amounts of video data over terrain of interest to analysts at defense and intelligence agencies, who scour through the data to find relevant information. An approach to video retrieval is centered on the idea of a "user query concept", in which the user of the system defines the kind of video clips that are of interest. The aim of Content based video retrieval (CBVR) system is to then is to learn this query concept and deliver appropriate results to the user. The query concept is typically semantic (e.g. person exiting a vehicle), and the system learns the query concept and deliver appropriate results to the analyst [1] [2].

Current research in the areas of relevance feedback and results re-ranking has demonstrated potential to improve CBVR. A relevance feedback (RF) based approach where an user provides feedback about the initial results in the hope of getting better results on the basis of this feedback [2] have been demonstrated to be extremely useful in text retrieval applications [3], and are now being applied to image and video retrieval [1]. In the video retrieval domain, re-ranking techniques based on sematic similarity have demonstrated improvement [4]. In addition, using feedback provided by users has demonstrated significantly superior retrieval [5][6][7]. By combining the feedback and re-ranking there can be significant improvements in video retrieval [8].

In this paper, we describe an Interactive Query Refinement (IQR) system which will allow intelligence analyst to iteratively refine search queries on UAV video archives. The IQR system applies novel feature transformations on kinematic and content descriptors from video tracker and learns a statistical query model via interactive user feedback. The IQR system leverages human-computer interaction in an efficient manner learning complex query concepts to achieve rapid improvement in video retrieval. We present the IQR algorithm flowchart and highlight the key conceptual steps in the next sections followed by experimental results and observations.

2. APPROACH

IQR is essentially supervised relevance feedback, which is a form of interactive machine learning that leverages user input to improve the effectiveness of information retrieval systems in an iterative manner [9]. Supervised relevance

Sixth International Conference on Machine Vision (ICMV 2013), edited by Antanas Verikas,
Branislav Vuksanovic, Jianhong Zhou, Proc. of SPIE Vol. 9067, 90671N · © 2013
SPIE · CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2051887

feedback can be entirely human-driven or can be machine assisted. The choice of the design depends on many factors: the complexity of the retrieval task, the quality of the corpus and the features used in the retrieval problem, and the amount of user feedback available, among others. In human-driven techniques, the user chooses the segments for relevance feedback to the system and also chooses the amount and balance (positive vs. negative) of feedback at each iteration to guide the development of the query model for best retrieval. Machine-assisted IQR denotes the ability of the system to choose the samples that would maximize the information in relevance feedback. In machine learning paradigm, this can also be considered as a special case of pool-based active learning, where the learner or the system has access to a pool of unlabeled elements and can request labeling of a certain number of elements from the user. In this paper, we focus only on machine-assisted IQR systems.

2.1 Naïve IQR

An IQR system that iteratively presents sample results for user relevance feedback and simply ranks the relevant results at the top of the results list and the irrelevant results at the bottom of the list is considered as a Naïve IQR system. Such a system does not have capability of learning the query concept and also does not produce a query model that can improve the ranking of other relevant results not necessarily judged through the user feedback. We use this naïve IQR approach as a baseline to compare to our model-based approach in a machine-assisted iterative feedback framework.

The cross-correlation technique is a method to get the displacement information of two consecutive images by comparing the similarity of a pair of image signals [6,7,8]. The traditional cross-correlation method was showed in Fig.1 and the cross-correlation coefficient was defined as formula

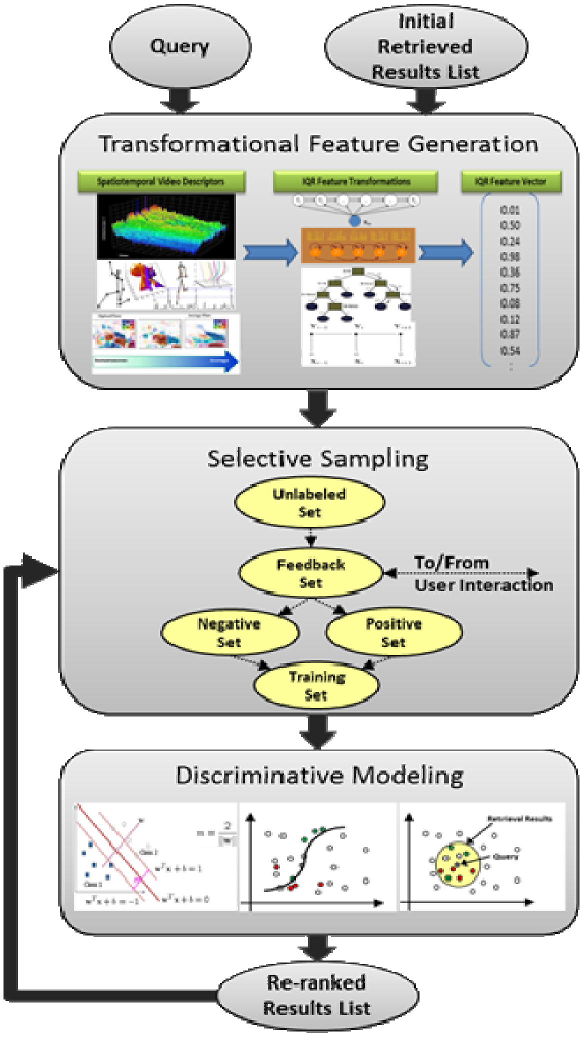

Figure 1. Model-based IQR Algorithm Flowchart

2.2 Model-based IQR

Figure 1 shows the algorithmic flowchart of the model-based approach to IQR, designed as a re-ranking tool that applied to the retrieval results list generated by a video descriptor vector matching algorithm in response to an example-based query. Every retrieved video segment is characterized by sequences of video descriptors and other low-level features extracted by the video tracking and event descriptor extraction components that are a part of a CBVR system. The query video clip is similar in characteristics to the retrieved video segments and is also input to the IQR algorithm. e output of the IQR process is the same list of video segments results re-ranked according to their relevance to the query.

2.2.1 Support Vector Machines (SVM)

The linear SVM is a simple linear discriminative model that corresponds to the hyper-plane separating the relevant examples for a query from the non-relevant examples. The SVM corresponds to this optimal hyper-plane that minimizes the training classification error while maximizing the margin, which is a measure of generality – capacity of the classifier to perform robustly on unseen data. Let $\{x_1, \cdots, x_n\}$ be the set of n labeled examples for a query and let $y_i \in \{1, -1\}$ be the class label of x_i, which is a k-dimensional feature vector. Note that the features in this vector can be continuous, binary, or categorical. The optimal separating hyper-plane, represented as the weight vector w, can then be found by solving the following constrained optimization problem:

$$\text{minimize } \frac{1}{2}\|w\|^2 \quad \text{subject to } y_i(w^T x_i + b) \geq 1 \quad \forall i \tag{1}$$

To solve this quadratic programming problem, we need to use Lagrange multipliers to convert this problem into its dual form, also a quadratic programming problem as shown in equation 2. We used the Kernel Adatron algorithm which is a conjugate gradient descent-based technique to solve for α_i in the dual QP problem above. We experimented with polynomial and radial basis function (RBF) kernels and compared the performance achieved with linear models [10].

$$\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \qquad \text{subject to } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i\, y_i = 0 \tag{2}$$

2.2.2 Logistic Regression

We also investigated logistic regression (Logit) model, which have been extensively used for binary classification problems where the posterior probability of the relevant class is a desired quantity [11]. Our primary objective was to combine the outputs from these models to produce a better estimate of the posterior probability of the relevant class for effective re-ranking. Using the same notations as above, the parameters of a Logit model are encoded as the weight vector w such that the probability of the relevant class given the feature vector x_i is given by:

$$P\left(y = 1/x_i\right) = \frac{1}{\left(1+e^{-w^T x_i}\right)} \tag{3}$$

In our model-based approaches we employed transformational feature generation. The video tracking and descriptor extraction modules generate a number of distinct video features and descriptors that vary along dimensions such as:

1. Feature type (binary, categorical, continuous)
2. Time resolution (per-frame, per-segment, end-frames)
3. Feature dimension (single-valued scalar, multi-dimensional vector)

Table 1 shows the different types of feature transformation function templates. The columns entitled "Feature Constraints" correspond to the constraints on the features used by each transformation function; e.g., discrete sequence statistics can only be applied to categorical or binary, scalar video features on the frame-level resolution. The advantage of this template-based framework is that new transformation functions can be easily integrated into the system with minimal changes.

Table 1. Types of Feature Transformation Functions

		Feature Constraints			
	Transformation Function	Type	Dimension	Resolution	Examples
Simple	Simple Statistics	Continuous	Scalar	Frame/Tracklet	Min., Max., Mean, Median crossing rates, number of peaks/valleys …
	Higher-Order Statistics	Continuous	Scalar	Frame/Tracklet	Variance, Kurtosis, Skewness, 1st order derivatives statistics …
	Discrete Sequence Statistics	Categorical, Binary	Scalar	Frame/Tracklet	first element, last element, #unique elements, element unigram, element bigram, predominant element, non-consecutive element bigrams, element trigrams …
	Identity Function	Binary, Categorical, Continuous	Scalar	Frame, Segment, End-frames	Feature name/value
Complex	Vector-Distance Statistics	Continuous	Vector	Tracklet	Pair-wise Kullback-Leibler distance between decimated vectors
	Vector-Max Sequence Statistics	Continuous	Vector	Tracklet	Maximum-valued dimension ID sequence

3. EXPERIMENTAL SETUP

3.1 Dataset

We used a 4-hour dataset from the A. P. Hill corpus to test the IQR system [11]. The A.P Hill data set was recorded at an altitude of 7,500 feet with a slant range of 2.0-3.75km. Table 2 presents the statistics of the dataset. The queries we processed have the following events: Carrying, Decelerating, Digging, Entering a Facility, Gesturing, Getting into a Vehicle, Maintaining Distance, Standing, Throwing, Turning, U-Turn, and Walking. The total length of the video is 4 hours, with 50 queries covering 12 events. The size of the result set was 1500.

3.2 Evaluation Measure

Precision, recall, and a plot showing precision vs. recall are commonly used methods to evaluate the performance of retrieval system. Precision is defined as the number of relevant videos retrieved divided by total number of relevant videos. Its complementary measure, recall is defined as the number of relevant videos retrieved divided by the total number of videos retrieved. A plot between precision and recall can be used to compare systems, but it is desirable to have a measure. Therefore we use average Precision (AP) computes the average value of precision over the interval of recall between 0 and 100%.

$$\mathrm{AP} = \sum_{k=1}^{n} \mathrm{P(k)}\Delta \mathrm{r(k)} \tag{4}$$

Where P(k) is precision at cut-off k, Δr(k) is the change in recall from results k-1 to k, and n is the number of retrieved results. In all our evaluations we empirically set n to 100.

The mean of AP over all the queries is defined as the Mean Average Precision (MAP) [12]. MAP is given by the equation below.

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AP(q)}}{\mathrm{Q}} \tag{5}$$

Where q is the q^th query, AP(q) is the AP at rank 100 for the q^th query, and Q is the total number of queries. MAP_i is defined as the MAP computed after ith iteration. We use MAP to evaluate the performance of the systems.

4. RESULTS

In this section we present and analyze the results of various IQR approaches. We particularly compare the model based approaches to naïve, improvement due to the models and multiple iterations of IQR.

4.1 Comparison of Model-based IQR to Naïve System

We compute the MAP before and after 4 iterations of IQR for all the reviewed approaches. We observe that SVM based discriminative approaches outperform the Naïve methods. MAP0, or MAP on the initial retrieval (without any IQR), is 2.4 and the best MAP after 4 iterations is 11.0 obtained by the SVM with linear kernel.

4.2 Performance of SVM based model

We plot precision vs. recall to further analyze the performance of the models. From the plot in Figure 2a it is clear that the precision for SVM is higher than that of naïve IQR and initial retrieval without IQR

4.3 Transformational Feature Generation

One of the key contributions of this work is development of methods for feature transformation. MAP for SVM with feature transformation is 10.4 and without feature transformation is 8.6.

4.4 IQR Improvement over Iterations

Finally, we evaluate the performance of IQR over 4 rounds. From Figure 2b, e note that the MAP increases after each round of IQR. Model based approaches result in higher MAP not only after iteration 4, but also after every iteration.

5. CONCLUSIONS

This paper introduces the concept of Interactive Query Refinement (IQR) in video retrieval for leveraging human feedback to improve the retrieval precision in an iterative manner. More specifically, we propose a model-based approach to IQR and compare it with a naïve no model approach. We introduce a novel technique to transform raw video features to create a larger augmented set for training models.

Through our experiments, we show that IQR yields a significant 358% improvement in MAP as compared to initial retrieval result ranking (without any IQR). Additionally, we show that our model-based approach is far superior to a naïve approach with significant gains of 49% relative to naïve IQR. Finally, we also show that our novel feature transformation technique gives better performance as compared to using raw low-level video features directly with discriminative models such as SVM. While we have not seen significant gains using the non-linear kernels we reviewed, there are several others that can be evaluated for further improvement. In addition, multiple kernel leaning approach can be reviewed and compared to feature transformations.

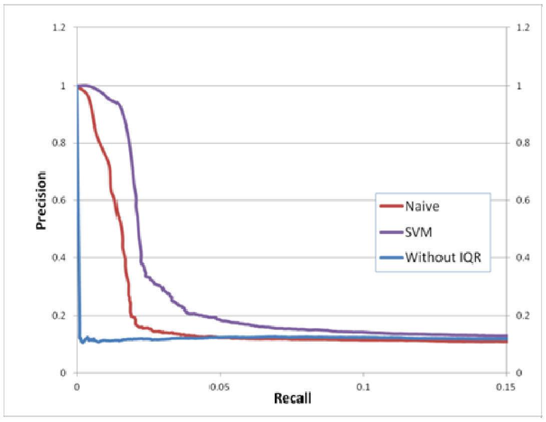

Figure 2a. Comparing Precision vs. Recall curves of SVM based model with Naïve model and without IQR

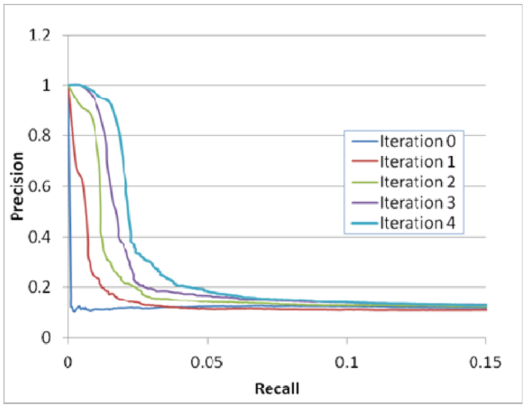

Figure 2b. Precision vs. Recall curves showing improvements with number of iterations

REFERENCES

[1] Y. Rui, T. Huang, M. Ortega and S. Mehrotra, "Relevance feedback: A power tool in interactive content-based image retrieval," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 8, no. 5, 1998.

[2] W. Smeulders and e. al, "Content-Based Image Retrieval at the End of the Early Years," *TRAMI,* vol. 22, no. 12, 2000.

[3] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," *Journal of the American Society for Information Science,* vol. 41, no. 4, 1990.

[4] H. H. Winston, K. L. S and C. Shi-Fu, "Video Search Reranking through Random Walk over," in *Proceedings of the 15th international conference on Multimedia*, 2007.

[5] A. Hauptmann and J. Christel M, "Successful Approaches in the TREC Video Retrieval Evaluations," in *ACM Multimedia*, 2004. A. K. A., M. A. H., S. A. E and K. M.S., "Multimodal fusion for multimedia analysis: a survey," Multimedia Systems, vol. 16, pp. 345-379, 2010.

[6] Su, J. H., Huang, Y. T., Yeh, H. H., & Tseng, V. S. (2010). Effective content-based video retrieval using pattern-indexing and matching techniques. Expert Systems with Applications, 37(7), 5068-5085.

[7] Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 41(6), 797-819.

[8] Jones, S., Shao, L., Zhang, J., & Liu, Y. (2012). Relevance feedback for real-world human action retrieval. Pattern Recognition Letters, 33(4

[9] Yang, Y., Nie, F., Xu, D., Luo, J., Zhuang, Y., & Pan, Y. (2012). A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(4), 723-742.

[10] Frie, T. T., Cristianini, N., & Campbell, C. (1998). The kernel-adatron algorithm: a fast and simple learning procedure for support vector machines. In Machine Learning: Proceedings of the Fifteenth International Conference (ICML'98)

[11] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J.K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xiaoyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai, “A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video” in Proceedings of IEEE Comptuer Vision and Pattern Recognition (CVPR), 2011.

[12] A. Turpin and F. Schole, "User performance versus precision measures for simple search tasks," in SIGIR, 2006.