Efficient Search: The Informedia Video Retrieval System

Alexander Hauptmann Jonathan J. Wang Wei-Hao Lin Jun Yang Michael Christel
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
+1 412-268-1448

alex+@cmu.edu, jjwang+@cmu.edu, whlin@cs.cmu.edu, juny@cs.cmu.edu, christel@cs.cmu.edu

ABSTRACT
We introduce an interface for efficient video search that exploits the human ability to quickly scan visual content, after an automatic system has done its best to arrange the images in order of relevance. While extreme video retrieval is taxing to the human, it has also been shown to be very effective. The system will demonstrate several ways to rapidly scan images, change search queries, and employ different types of relevance feedback.

Categories and Subject Descriptors
H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *evaluation, video*

General Terms
Experimentation, Human Factors

Keywords
Video retrieval, relevance feedback, multi-modality integration.

1. System Approach
Our approach builds on existing work in the Informedia Video Search system [1]. This includes search based on text keywords, and image similarity. While traditional video retrieval has been based on shot keyframes as the core unit of analysis and search, we have expanded this to use multiple keyframes for both analysis and display, enabling retrieval of complex activities and scenes which are not easily represented by a single keyframe.

As search can take on a variety of forms, the system can efficiently exploit image similarity based on typical color features as well as SIFT feature points clustered into visual words [2]. SIFT features around points of interest can also be used fast region-based similarity comparisons, even when the regions to be compared are of different sizes. Thus the system searches video for similar regions to ones specified by the user in a sample query image.

Relevance feedback (RF) is a key technique to improving retrieval performance, RF proceeds by leveraging users annotations of a small number of selected video documents from the initial retrieval results and then feeding them back to update the retrieval models. Formally, we denote the relevance information as $y_1,...,y_F$ associated with the feedback documents $D_1,...,D_F$. This can be viewed as a learning component in a retrieval system, where the system learns from a small amount of relevant examples to adjust the ranking function accordingly. In this system we mainly consider using the additional annotated data to adjust the combination parameters λ in the probabilistic retrieval models.

Given relevance judgments, our model-based relevance feedback approach computes the maximum a posteriori estimation for the updated combination parameters of different retrieval components each of which has returned its own ranked list of results,

$$\begin{aligned}\lambda^* &= \arg\max_{\bar{\lambda}} P(\bar{\lambda} \mid y, D, Q, \lambda) \\ &= \arg\max_{\bar{\lambda}} P(\bar{\lambda} \mid \lambda) \prod_t P(y_t \mid D_t, Q, \bar{\lambda}) \\ &= \arg\max_{\bar{\lambda}} \left[\log P(\bar{\lambda} \mid \lambda) + \sum_t \log P(y_t \mid D_t, Q, \bar{\lambda}) \right]\end{aligned}$$

where λ are the initial parameters for combination and λ* are the updated parameters after relevance feedback. The prior probability can be defined in many ways and we particularly define it as a Gaussian distribution with mean λ and a pre-defined variance. This can be interpreted as a maximum likelihood formulation, which finds a compromise between two factors: minimizing the distance between the updated model parameters and the initial model parameters, and on the other hand maximizing the likelihood for the feedback data.

We also utilize an additional selection strategy that is computationally less complex. The crucial insights come from an analysis of temporal sequences in video concepts. After noticing that the semantic concept in a keyframe of a shot is the single best predictor for the concept in the next shot, we found this would also hold true for query results. In other words, if we find a relevant shot, we predict that the same 'query concept' is likely to be relevant in the adjacent shot. This gives a framework for re-ranking. When a shot is marked as relevant by the user, we simply insert the neighbors of this shot at the top of the shots to be presented to the user on the next display page.

2. ACKNOWLEDGMENTS
This work was supported in part by the National Science Foundation under Grants No. IIS-0205219 and CNS-0751185

3. REFERENCES
[1] A. Hauptmann, W.-H. Lin, R. Yan, J. Yang, M.-Y. Chen Extreme Video Retrieval: Joint Maximization of Human and Computer Performance, In Proc. of 8th ACM Multimedia 2006, Oct. 23-27, 2006, Santa Barbara, CA

[2] Yang, J., Jiang Y.G., Hauptmann, A. and Ngo, C.W. 2007, Evaluating bag-of-visual-word representation in scene classification, MIR'07 ACMMM, September 2007