Etter Solutions Research Group
TRECVID 2007

David Etter

October 19, 2007

Abstract

Etter Solutions Research Group participated in the TRECVID conference for the first time in 2007. We submitted five runs in the area of fully automatic search.

Fully Automatic Search Runs

- **F_A_1_ESRG_A1_1**

  The first run is a required baseline, using only ASR/MT features. A sliding window of three shots is used to create a "bag of words" representation of the shot.

- **F_A_1_ESRG_KN2_2**

  The second run is a required baseline, using no text from the ASR/MT output. This run uses visual keypoint features with shot description expansion using a semantic network.

- **F_A_2_ESRG_AN3_3**

  The third run is an optional run combining ASR/MT features with shot term expansion using a semantic network.

- **F_A_2_ESRG_AKN4_4**

  The fourth run is an optional run using a weighted fusion model with the ASR/MT and visual keypoint features. Shot description expansion using a semantic network is applied to both feature sets.

- **F_A_2_ESRG_AKNF5_5**

  The fifth run is an optional run using a weighted fusion model with the ASR and visual keypoint features, shot description expansion, and automatic relevance feedback.

Our primary focus in TRECVID 2007 is the evaluation of retrieval utilities on a video search system. We apply both automatic relevance feedback and shot (document) expansion using a semantic network. These utilities have proven to be very effective in the text retrieval domain. Our top MAP was found in the second run, consisting of visual keypoint features with shot description expansion. Runs using only the ASR/MT features performed poorly, but showed slight improvement in MAP when combined with shot description expansion.

1. Fully Automatic Search Task

Etter Solutions Research Group participated in the fully automatic search task for TRECVID 2007. Our primary research focus this year was on the application of retrieval utilities [1]. Specifically, we are interested in the application of shot description expansion, semantic networks, and automatic relevance feedback. These utilities have proven effective in the area of text retrieval and we analyze their application to the TRECIVD test collection.

1.1 ASR/MT Search

The TRECVID 2007 guidelines [2] require two baseline runs for every automatic search system. The first baseline run is a text only run, which consists of the ASR/MT [3] output of the audio translation from the video collection. Our approach to this task is similar to that applied by many past TRECVID participants [4] [5] . A sliding window of three shots is used to build a "bag of words" representation of the corresponding shot. We then apply term frequency inverse document frequency weighting with a cosine similarity measure. The mean average precision for the required run, F_A_1_ESRG_A1_1**,** was .004.

1.2 Semantic Search

Another important utility found in the text retrieval domain is that of query expansion or document expansion [6] and the application of a semantic network [7]. Our implementation of this utility is through shot description expansion where we attempt to boost recall by including additional concept descriptions, which are related through a semantic network. We apply a modified weighting scheme where the original descriptions are given greater weight compared to their semantic counterparts.

This technique is first applied to the required baseline run using no text from the ASR/MT output. The visual only features used in this run are based on the keypoint features donated by City University of Hong Kong [8]. These keypoint features use a DoG detector and SIFT descriptor with keyframes represented by a 500 dimensional vector of visual words. The detectors were trained on the TRECVID 2005 – 2006 collections using the 374 LSCOM concepts [9] [10]. Our baseline run applies shot expansion with a semantic network in the attempt to expand the original vocabulary from the 374 concepts to include all semantically related concepts. The mean average precision for the required run, F_A_1_ESRG_KN2_2**,** was .017.

The shot expansion utility is also applied to the original baseline ASR/MT run. This run uses the sliding window of three keyframes and expands the original shot to include additional terms from the semantic network [11]. We again apply a modified weighting scheme were higher weights are applied to the original ASR terms. The mean average precision for the optional run, F_A_2_ESRG_AN3_3, was .005.

1.3 Automatic Relevance Feedback

Runs F_A_2_ESRG_AKN4_4 and F_A_2_ESRG_AKNF5_5 combine the text features of the ASR/MT with the visual keypoint features in a fusion model. In our model, results are obtained separately from the two steams of data and then fused [12] with a weighting scheme that gives higher weights to the visual features. Each run includes the shot expansion utility applied separately to the two data streams. The final run, F_A_2_ESRG_AKNF5_5, incorporates an automatic relevance feedback utility. This application of relevance feedback [1] analyzes the top n shots returned from each data stream and reformulates the original query, based on the assumption that these keyframes are the most relevant. The mean average precision for the optional run, F_A_2_ESRG_AKNF5_5, was .015.

2. Analysis and Conclusions

Our first year of participation in the TRECVID conference focused on the application of retrieval utilities. We submitted five runs in the automatic search task, which analyze the effects of shot description expansion, semantic networks, and automatic relevance feedback to our system recall. The best performing run was based on visual keypoint features only, with shot description expansion using a semantic network. ASR/MT feature runs did not perform as well as the visual feature runs and had a negative impact on runs were a fusion model was applied to both visual and ASR streams. In future research, we hope to boost the recall of our system through an expanded semantic network and by increasing the number of concepts in our visual vocabulary. We would like to thank City University of Hong Kong for their donation of visual keypoint features.

3. References

[1]. **Grossman, D.A., and Frieder, O.** *Information Retrieval: Algorithms and Heuristics.* s.l. : 2nd edition, Springer, 2004.
[2]. TREC Video Retrieval Evaluation (TRECVID). *http://www-nlpir.nist.gov/projects/tv2007/.*
[3]. **Marijn Huijbregts, Roeland Ordelman and Franciska de Jong.** Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition. *To appear in proceedings of SAMT, December 5-7 2007, Genova, Italy.* 2007.
[4]. **Smeaton, A. F.** Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Inf. Syst. 32, 4 June.* 2007, pp. 545-559.
[5]. **Shih-Fu Chang, Wei-Ying Ma, Arnold Smeulders.** Recent Advances and Challenges of Semantic Image/Video Search. *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on.* April 2007.
[6]. **Zobel, Bodo Billerbeck and Justin.** Document Expansion versus Query Expansion for Ad-hoc Retrieval. *Proceedings of the 10th Australasian Document Computing Symposium, Melbourne, Australia.* 2005.
[7]. **Fellbaum, Christiane, editor.** *WordNet: An Electronic Lexical Database.* s.l. : The MIT Press, Cambridge, MA, 1998.
[8]. **Yu-Gang Jiang, Chong-Wah Ngo, Jun Yang.** Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. *ACM International Conference on Image and Video Retrieval (CIVR'07), Amsterdam, The Netherlands.* 2007.
[9]. **A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu.** Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. *Columbia University ADVENT Technical Report #222-2006-8.* March 2007.
[10]. **Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, Jon Curtis.** Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia, vol. 13, no. 3, pp. 86-91, Jul-Sept.* 2006, pp. pp. 86-91.
[11]. **Pereira, Amit Singhal and Fernando C. N.** Document Expansion for Speech Retrieval. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* . 1999, pp. 34-41.
[12]. **Snoek, C. G., Worring, M., and Smeulders, A. W.** Early versus late fusion in semantic video analysis. *Proceedings of the 13th annual ACM international conference on Multimedia* . 2005, pp. 399-402.