Florida International University - University of Miami TRECVID 2016

Yilin Yan[1], Samira Pouyanfar[2], Sheng Guan[2], Haiman Tian[2], Hsin-Yu Ha[2], Mei-Ling Shyu[1], Shu-Ching Chen[2], Winnie Chen[3], Tiffany Chen[3] and Jonathan Chen[4]

[1]Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33146, USA
[2]School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
[3]School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907, USA
[4]Miami Palmetto Senior High School
Miami, FL 33156, USA
y.yan4@umiami.edu, spouy001@cs.fiu.edu, sguan005@cs.fiu.edu, htian005@cs.fiu.edu, hha001@cs.fiu.edu, shyu@miami.edu, chens@cs.fiu.edu, chen1219@purdue.edu, chen1791@purdue.edu

Abstract

This paper demonstrates the framework and results from the team "Florida International University - University of Miami (FIU-UM)" in TRECVID 2016 [1] Ad-hoc Video Search (AVS) task [2]. The following two runs were submitted:

- M_D_FIU_UM.16_1*: CNN features + linear SVM + concept scores combination type I*
- M_D_FIU_UM.16_2*: CNN features + linear SVM + concept scores combination type II*

In both runs, the features are first extracted by the CNN (Convolutional Neural Network) structure of AlexNet [3]. Then, using the linear SVM (Support Vector Machine) classifiers, the scores of each concept for the key frames are generated. For run 1 and run 2, the scores from the aforementioned model are combined in different ways for different queries. From the submission results, run 2 outperforms run 1. The submission details are listed as follows.

- *Class*: M (Manually-assisted runs)
- *Training type*: D (IACC & non-IACC non-TRECVID data)
- *Team ID*: FIU-UM (Florida International University - University of Miami)
- *Year*: 2016

1 Introduction

Previously in year 2015, the TRECVID project [4] includes a semantic indexing (SIN) task which aims to recognize the semantic concept contained within a video shot. It has been well-acknowledged that there are several challenges in the SIN task, such as data imbalance, scalability, and semantic gap [5, 6, 7, 8, 9, 10, 11, 12].

In this year, the task was changed to the Ad-hoc video search (AVS) task which is to model the end user search use-cases. Comparing to the SIN task, the new AVS task looks for not only the video segments that contain persons, objects, activities, locations, etc. but also the video segments of their combinations.

The automatic annotation of semantic concepts in video shots can be an essential technology for retrieval, categorization, and other video exploitations [13, 14, 15, 16, 17, 18, 19, 20]. The semantic concept retrieval research directions include (1) developing robust learning approaches that adjust to the increasing size and the diversity of the videos, (2) fusing the information from other sources such as audio and text, (3) detecting the low-level and mid-level features that have a high discriminating capability, etc. [21, 22, 23, 24, 25, 26, 27, 28].

The size of the high-level semantic concepts provided by IACC remains the same as the SIN task of the previous year, which has 346 concepts in total. For each of the 346 semantic concepts, a list of ground truth labels is provided for training. Given the test collection (IACC.3), master shot reference, 30 Ad-hoc queries were released by NIST for testing. Each query can be a combination of some of the 346 concepts and/or some other concepts not included in the training set. Each participated group is allowed to submit a maximum of $1,000$ possible shots from the test collection for each query, which are ranked according to their likelihood of containing the target query. The submission result is rated by using the mean inferred average precision (mean xinfAP) [29] based on the assessment of a 2-tiered random sampling (1-200@100% and 201-1000@11.1%).

This paper is organized as follows. Section 2 describes our proposed framework and the specific approaches utilized for each run. Section 3 shows the submission results in details. Section 4 summarizes the whole paper and proposes some future directions to pursue and plan for next year.

2 The Proposed Framework

Our framework of the TRECVID 2016 AVS task is shown in Figure 1. In this year, key frames are already extracted from the videos in the IACC.3 collection and are provided to the participants. Thus, the key frame extraction step is skipped; while the key frames extracted in the previous years are used for training.

2.1 CNN Feature Extraction

Ten kinds of low-level key frame features were extracted from each frame in the training and testing data last year. In this year, we use a pre-trained deep learning model, Alexnet [3], a Convolutional Neural Network (CNN) structure trained on the ImageNet database for object feature extraction. AlexNet contains five convolutional layers and three fully-connected layers as shown in Figure 1. CNN features are extracted from all the training and testing key frames from the $8th$ layer, i.e., the output layer with $1,000$-dimensions. The Alexnet structure is well trained and proven with great performance.

2.2 Classification

After feature extraction, feature vectors are normalized for each key frame. Support Vector Machine (SVM), one of the state-of-the-art algorithms in the data mining area [30, 31, 32, 33] including multimedia classification, is used for classification and concept score generation. The general idea of SVM is to build a separating hyperplane to classify the data instances so that the geometric margin is maximized. For efficiency, linear kernels are adopted in our proposed framework.

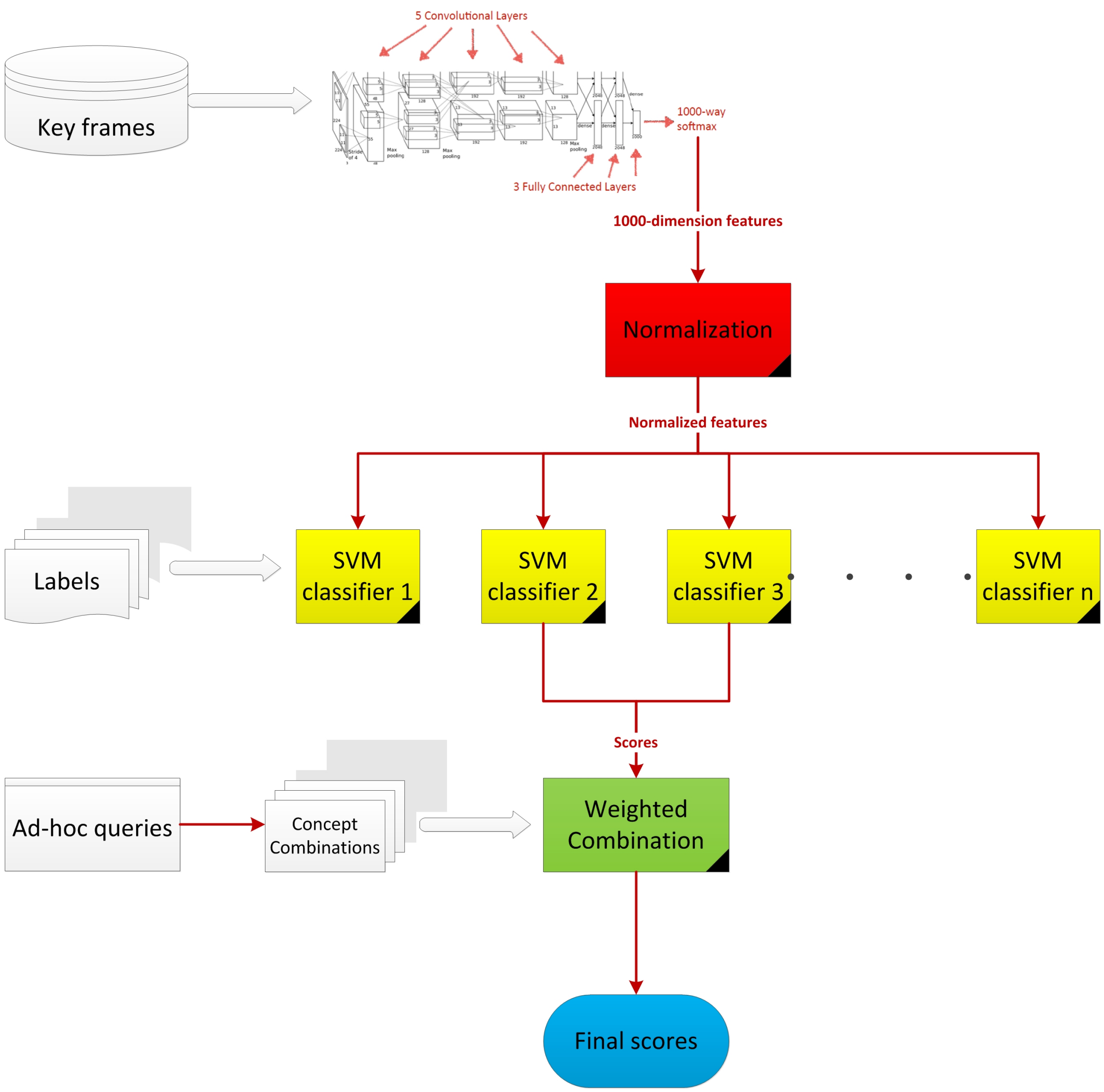

Figure 1. The proposed framework for the Ad-hoc video search task

2.3 Query Formulation and Score Combination

Our two runs are both manually-assisted runs. For an initial ad-hoc query, a group member formulates it into a combination of concepts based on its topic and query interface without the knowledge of the collection or the search results. For queries containing the concepts not included in the 346 concepts from the training set (e.g., the concept "guitar" in "501 Find shots of a person playing guitar outdoors"), a similar concept like "music instruments" is selected. For run 1 and run 2, different weights are used for the score combination based on our empirical studies.

3 Experimental Results

3.1 Data

Given the information (including the IACC.3 test collection, master shot reference, and 30 Ad-hoc queries) released by NIST and the concept definitions, a list of at most 1000 shot IDs from the test collection for each ad-hoc query was returned and ranked according to their likelihood of containing the target query. TRECVID 2016 test data set (IACC.3) contains $4,593$ Internet Archive videos with the durations between 6.5 and 9.5 minutes (144GB, 600 hours in total). The train data set combines the development and test data sets from the 2010 to 2015 issues of the SIN task, namely the IACC.1.tv10.training, IACC.1.A-C, and IACC.2.A-C data sets. Each contains about 200 hours of videos drawn from the IACC.1 and IACC.2 collections using videos with durations ranging from 10 seconds to (3.5 to 6.4) minutes, respectively.

The overall framework of the TRECVID 2016 AVS task contains three stages:

1. Model training: using TRECVID 2010-2012 training videos as the training data.
2. Model evaluation: using TRECVID 2013-2015 training videos as the testing data to evaluate the framework and tune the parameters of the models.
3. Model testing: using TRECVID 2010-2015 training videos as the TRECVID 2016 training data, and TRECVID 2016 testing videos as the testing data to generate the ranking results for the submission.

3.2 Evaluation

A subset of the submitted ad-hoc query results (20) announced after the submission date were evaluated by the assessors at NIST pooling and sampling. Measures (indexing) are shown as follows [34].

1. Mean extended inferred average precision (mean xinfAP) [29] which allows the sampling density to vary so that it can be 100% in the top strata. This is the most important one for average precision.
2. As in the past years, other detailed measures based on recall and precision are generated and given by the sample_eval software provided by the TRECVID team.

3.3 Performance

All of the measures below were based on the assessment of a 2-tiered random sampling (1-200@100% and 201-1000@11.1%) of the full submission pools and the sample_eval software was used to infer the measures.

Figure 2 and Figure 3 present the performance of our ad-hoc query video search results. The x-axis is the concept number; while the y-axis is the inferred average precision. More clearly, Table 1 shows the inferred mean average precision (MAP) values of the first 5, 10, 15, 20, 30, 100, 200, 500 and 1000 shots. The inferred true shots and mean xinfAP are shown in Table 2.

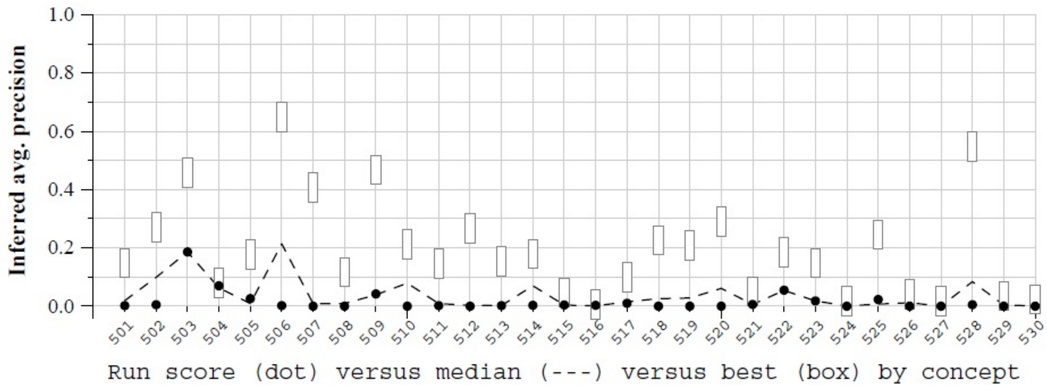

Figure 2. Run scores (dot) versus median (—) versus best (box) for *M_D_FIU_UM.16_1*

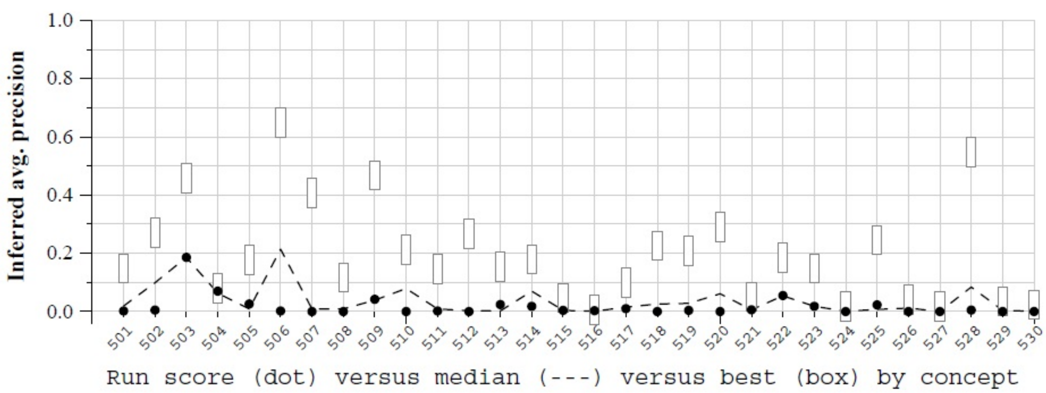

Figure 3. Run scores (dot) versus median (—) versus best (box) for *M_D_FIU_UM.16_2*

Framework	5	10	15	20	30	100	200	500	1000
M_D_FIU_UM.16_1	0.220	0.180	0.149	0.132	0.111	0.070	0.056	0.046	0.032
M_D_FIU_UM.16_2	0.220	0.183	0.151	0.132	0.112	0.076	0.062	0.053	0.039

Table 1: The MAP values at first *n* shots for all 2 runs

Framework	Inferred true shots returned	Mean xinfAP
M_D_FIU_UM.16_1	949	0.015
M_D_FIU_UM.16_2	1177	0.017

Table 2: Inferred true shots returned and Mean xinfAP

4 Conclusion and Future Work

In this notebook paper, the framework and results of team FIU-UM in TRECVID 2016 AVS task are summarized. It can be seen that there are a lot of improvements that need to be done based on the results. The following directions will be investigated.

- In our framework, only global features are utilized. The object-level features can also be explored by R-CNN.
- SVM classifiers need to be adopted to address the data imbalance issue.
- Some other advanced CNN structures can be integrated to reach a better performance.
- More training data should be collected by a general purpose search engine like Google using the query definition to further improve the retrieval accuracy.

The second stage of the AVS task, namely concept combination and score fusion make this task a totally new one comparing to the previous SIN task. It will be helpful to exchange ideas and thoughts with other groups so that novel approaches can be developed for further performance improvement.

References

[1] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Qunot, Maria Eskevich, Robin Aly, and Roeland Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[2] Alan F. Smeaton, Paul Over, and Wessel Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Ajay Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[4] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, Georges Quenot, and Roeland Ordelman. Trecvid 2015 - an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[5] Shu-Ching Chen, Arif Ghafoor, and R. L. Kashyap. *Semantic Models for Multimedia Database Searching and Browsing*. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[6] Xiuqi Li, Shu-Ching Chen, Mei-Ling Shyu, and Borko Furht. An effective content-based visual image retrieval system. In *Proceedings of the IEEE International Computer Software and Applications Conference*, pages 914–919, August 2002.

[7] Lin Lin, Guy Ravitz, Mei-Ling Shyu, and Shu-Ching Chen. Effective feature space reduction with imbalanced data for semantic concept detection. In *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, pages 262–269, 2008.

[8] Yilin Yan, Mei-Ling Shyu, and Qiusha Zhu. Supporting semantic concept retrieval with negative correlations in a multimedia big data mining system. *International Journal of Semantic Computing*, 10:247–268, 2016.

[9] Mei-Ling Shyu, Thiago Quirino, Zongxing Xie, Shu-Ching Chen, and Liwu Chang. Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems. *ACM Trans. Auton. Adapt. Syst.*, 2(3), September 2007.

[10] Mei-Ling Shyu, Choochart Haruechaiyasak, and Shu-Ching Chen. Category cluster discovery from distributed www directories. *Information Sciences*, 155(3):181–197, 2003.

[11] Shu-Ching Chen and R.L. Kashyap. Temporal and spatial semantic models for multimedia presentations. In *Proceedings of the 1997 International Symposium on Multimedia Information Processing*, pages 441–446, 1997.

[12] Yilin Yan, Min Chen, Mei-Ling Shyu, and Shu-Ching Chen. Deep learning for imbalanced multimedia data classification. In *2015 IEEE International Symposium on Multimedia (ISM)*, pages 483–488, December 2015.

[13] Shu-Ching Chen, Srinivas Sista, Mei-Ling Shyu, and R. Kashyap. Augmented transition networks as video browsing models for multimedia databases and multimedia information systems. In *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, pages 175–182, November 1999.

[14] Xiuqi Li, Shu-Ching Chen, Mei-Ling Shyu, and Borko Furht. Image retrieval by color, texture, and spatial information. In *Proceedings of the 8th International Conference on Distributed Multimedia Systems*, pages 152–159, September 2002.

[15] Xin Chen, Chengcui Zhang, Shu-Ching Chen, and Min Chen. A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval. In *Proceedings of the Seventh IEEE International Symposium on Multimedia*, pages 37–44, December 2005.

[16] Yilin Yan, Yang Liu, Mei-Ling Shyu, and Min Chen. Utilizing concept correlations for effective imbalanced data classification. In *Proceedings of the IEEE 15th International Conference on Information Reuse and Integration*, pages 561–568, August 2014.

[17] Shu-Ching Chen, S.H. Rubin, Mei-Ling Shyu, and Chengcui Zhang. A dynamic user concept pattern learning framework for content-based image retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 36(6):772–783, November 2006.

[18] Mei-Ling Shyu, C. Haruechaiyasak, Shu-Ching Chen, and Na Zhao. Collaborative filtering by mining association rules from user access sequences. In *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*, pages 128–135, April 2005.

[19] Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, and R. L. Kashyap. Identifying overlapped objects for video indexing and modeling in multimedia database systems. *International Journal on Artificial Intelligence Tools*, 10(4):715–734, 2001.

[20] Yilin Yan, Mei-Ling Shyu, and Qiusha Zhu. Negative correlation discovery for big multimedia data semantic concept mining and retrieval. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 55–62, February 2016.

[21] Shu-Ching Chen, Mei-Ling Shyu, and R.L. Kashyap. Augmented transition network as a semantic model for video data. *International Journal of Networking and Information Systems*, 3(1):9–25, 2000.

[22] Mei-Ling Shyu, Shu-Ching Chen, and Rangasami L Kashyap. Generalized affinity-based association rule mining for multimedia database queries. *Knowledge and Information Systems*, 3(3):319–337, 2001.

[23] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. Effective supervised discretization for classification based on correlation maximization. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, pages 390–395, 2011.

[24] Dianting Liu, Yilin Yan, Mei-Ling Shyu, Guiru Zhao, and Min Chen. Spatio-temporal analysis for human action detection and recognition in uncontrolled environments. *International Journal of Multimedia Data Engineering and Management*, 6(1):1–18, January 2015.

[25] Shu-Ching Chen, Mei-Ling Shyu, and Chengcui Zhang. Innovative shot boundary detection for video indexing. In Sagarmay Deb, editor, *Video Data Management and Information Retrieval*, pages 217–236. Idea Group Publishing, 2005.

[26] Lin Lin and Mei-Ling Shyu. Weighted association rule mining for video semantic detection. *Int. J. Multimed. Data Eng. Manag.*, 1(1):37–54, January 2010.

[27] Shu-Ching Chen, Mei-Ling Shyu, and Chengcui Zhang. An intelligent framework for spatio-temporal vehicle tracking. In *Proceedings of the 4th IEEE International Conference on Intelligent Transportation Systems*, pages 213–218, August 2001.

[28] Tao Meng, Yang Liu, Mei-Ling Shyu, Yilin Yan, and Chi-Min Shu. Enhancing multimedia semantic concept mining and retrieval by incorporating negative correlations. In *Proceedings of the IEEE International Conference on Semantic Computing*, pages 28–35, June 2014.

[29] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, New York, NY, USA, 2008. ACM.

[30] Chao Chen, Qiusha Zhu, Lin Lin, and Mei-Ling Shyu. Web media semantic concept retrieval via tag removal and model fusion. *ACM Transactions on Intelligent Systems and Technology*, 4(4):61:1–61:22, October 2013.

[31] Yilin Yan, Jun-Wei Hsieh, Hui-Fen Chiang, S.-C. Cheng, and Duan-Yu Chen. Plsa-based sparse representation for object classification. In *Proceedings of the 22nd International Conference on Pattern Recognition*, pages 1295–1300, August 2014.

[32] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[33] Tao Meng and Mei-Ling Shyu. Leveraging concept association network for multimedia rare concept mining and retrieval. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, pages 860–865, July 2012.

[34] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.