Florida International University - University of Miami TRECVID 2015

Yilin Yan[1], Miguel Gavidia[2], Tarek Sayed[1], Hsin-Yu Ha[3], Mei-Ling Shyu[1], Shu-Ching Chen[3], Winnie Chen[4] and Tiffany Chen[4]

[1]Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33146, USA
[2]Department of Computer Science
University of Miami, Coral Gables, FL 33124, USA
[3]School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
[4]School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907, USA
y.yan4@umiami.edu, m.gavidia@miami.edu, t.sayed@miami.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 2015 Semantic Indexing (SIN) task [1]. Four runs were submitted, and the summary of these four runs is given as follows:

- 2C_M_A_FIU_UM.15_1*: MCA late fusion - Multiple Correspondence Analysis (MCA) based ranking using the MCA scores of all ten key frame (KF) features.*
- 2C_M_A_FIU_UM.15_2*: MCA early fusion - MCA based ranking using the selected five KF features.*
- 2C_M_A_FIU_UM.15_3*: Run* 1 + *Time information - MCA late fusion combined with MCA scores from frames other than key frames.*
- 2C_M_A_FIU_UM.15_4*: MCA early fusion - MCA based ranking using the selected four KF features.*

In Run 1, the MCA scores from ten KF features are combined and re-ranked. For Run 3, the result from the aforementioned run (i.e., run 1) and the time information extracted from frames other than key frames are fused. In this way, we wanted to test whether the time information could help improve the results. In Run 2 and Run 4, different feature sets are combined to feed to the same baseline MCA-based model. As a result, from the submission results, Run 2 outperforms the other three runs.

- *Processing type*: Automatic
- *Class*: M - main, single concepts
- *Training type*: A (only the IACC data)

1 Introduction

In the TRECVID 2015 project [2], the semantic indexing (SIN) task aims to recognize the semantic concept contained within a video shot. This task has several challenges such as data imbalance, scalability, and semantic gap [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. The automatic annotation of semantic concepts in video shots can be an essential technology for retrieval, categorization, and other video exploitations. 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 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. [16, 17, 18, 19].

The size of the high-level semantic concepts remains the same as the SIN task of the previous year, which has 60 concepts. For each of the 60 semantic concepts, the participants are allowed to submit a maximum of $2,000$ possible shots, and the submission result is rated by using the mean inferred average precision (mean xinfAP) [20].

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.

2 The Proposed Framework

The proposed framework of the TRECVID 2015 SIN task is shown in Figure 1. The key frame level features (KF) are extracted and normalized. In Run 1, the MCA late fusion model is applied; while for Run 2 and Run 4, the MCA early fusion model is applied. Run 3 is based on Run 1 and also includes the time information. The xinfAP values are calculated from models trained on TRECVID 2015 training data and evaluated on TRECVID 2015 testing data.

2.1 Data Pre-processing and Feature Extraction

A key frame for each shot is provided to the SIN task participates in both training and testing videos. Ten kinds of KF features are extracted from each frame in training and testing data, including CEDD [21], Cooccur [22], Gabor [23], Haar [24], LBP [25], Sobel [26], HoG [27, 28], Canny edge histogram [29], and color histograms in HSV and YCbCr spaces. Before extracting the features, histogram equalization is employed to regulate the contrast of frames [30, 19]. Then, these features are extracted and normalized for each key frame.

2.2 Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) [31, 32, 33, 34, 35] is a data analysis technique for nominal/categorical data, which has been used to detect and represent the underlying structures in a data set. The procedure therefore appears to be the counterpart of principle component analysis for categorical data. MCA extends correspondence analysis (CA) by providing the ability to analyze tables containing some measure of correspondence between the rows and columns with more than two variables. To the best of our knowledge, our team is the first one to apply the MCA technique in the area of multimedia information retrieval.

MCA first projects the features and classes into a 2D space spanned by the first principle component (i.e., the eigenvector with the largest eigenvalue) and the second principle component (i.e., the eigenvector with the second largest eigenvalue). The correlation between different feature-value pairs and different classes can be used as an indication of the similarity between them. If one feature has two feature-value pairs in a binary classification application, the angle of one feature-value pair and one class definitely equals the angle of the other feature-value pair and the other class. MCA can be applied for semantic concept detection. One can use the similarity value from the angle file as the criterion to evaluate the distance of the testing data instance to the semantic concept.

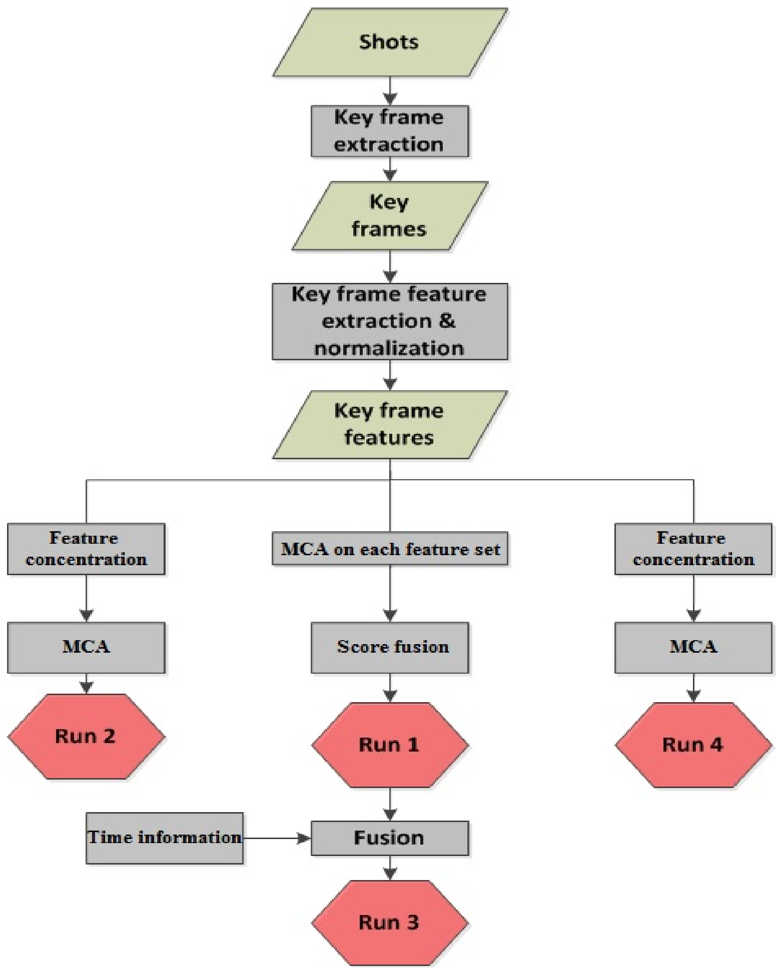

Figure 1. The proposed framework for semantic indexing

2.3 Information Fusion

There are two main categories of the fusion methods, namely early fusion and late fusion. For early fusion, we concentrated four (Run 4) / five (Run 2) low-level features as the feature super vectors. The MCA model is then applied for concept detection. On the other hand, late fusion concentrates on the scores generated by the MCA model for each low-level feature. Another MCA model is used to fuse these scores to have a final score for each shot.

Time information is also very important in multimedia data. Motivated by this fact, we extracted the features from frames other than key frames in order to get the relative time information. A concept may not appear or be clearly shown in the key frame of a shot, while the identical concept may be detected in the other frames in the same shot. Therefore, in Run 3, we fused the results from Run 1 with the MCA scores generated by the other frames that carry the time information.

3 Experimental Results

3.1 Data

Given the test collection (IACC.2.C), master shot reference, and concept definitions, for each target concept, a list of at most 2000 shot IDs from the test collection was returned and ranked according to their likelihood of containing the target concept. TRECVID 2015 test data set (IACC.2.C) contains the 200 hours of videos drawn from the IACC.2 collection using videos with durations between 10 seconds and 6 minutes. The train data set combines the development and test data sets from the 2010 and 2011 issues of the SIN Task, namely the IACC.1.tv10.training, IACC.1.A, IACC.1.B, and IACC.1.C data sets. Each contains about 200 hours of videos drawn from the IACC.1 collection using videos with durations ranging from 10 seconds to a little bit longer than 3.5 minutes.

The overall framework of TRECVID 2015 SIN task contains three stages:

1. Model training: using TRECVID 2014 training videos as the training data.

2. Model evaluation: using TRECVID 2014 testing videos as the testing data to evaluate the framework and tune the parameters of the models.

3. Model testing: using TRECVID 2014 training + TRECVID 2014 testing videos as the TRECVID 2015 training data, and TRECVID 2015 testing videos as the testing data to generate the ranking results for the submission.

3.2 Evaluation

A subset of the submitted concept results (20) announced after the submission date were evaluated by the assessors at NIST pooling and sampling. Measures (indexing) are shown as follows [36].

1. Mean extended inferred average precision (mean xinfAP) [37], 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 201-2000@11.1%) of the full submission pools and the sample_eval software was used to infer the measures.

Figure 2 to Figure 5 present the performance of our semantic indexing 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 10, 100, 1000 and 2000 shots. The inferred true shots and mean xinfAP are shown in Table 2.

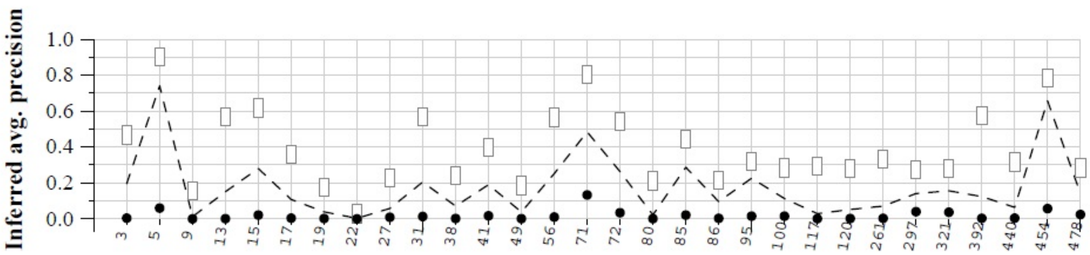

Figure 2. Run scores (dot) versus median (—) versus best (box) for *2C_M_A_FIU_UM.15_1*

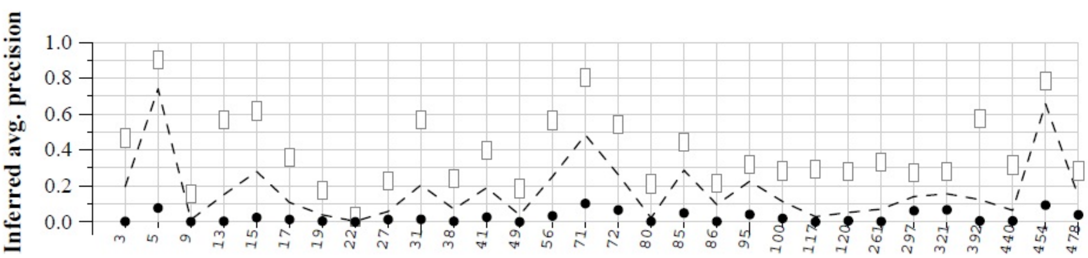

Figure 3. Run scores (dot) versus median (—) versus best (box) for *2C_M_A_FIU_UM.15_2*

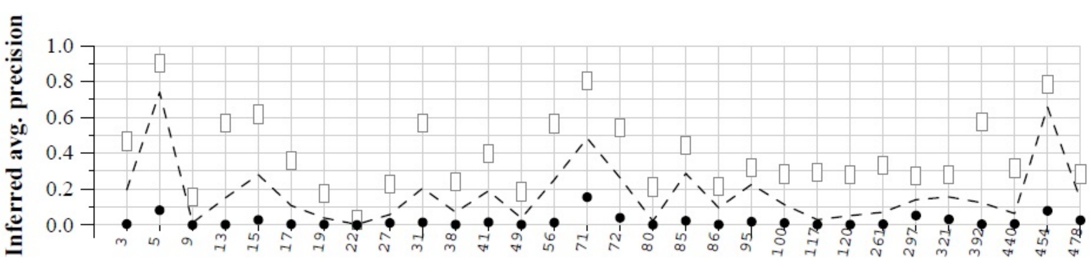

Figure 4. Run scores (dot) versus median (—) versus best (box) for *2C_M_A_FIU_UM.15_3*

4 Conclusion and Future Work

In this notebook paper, the framework and results of team FIU-UM in TRECVID 2015 SIN task are summarized. We can tell there are still a lot of improvements need to be done based on the results. Some important

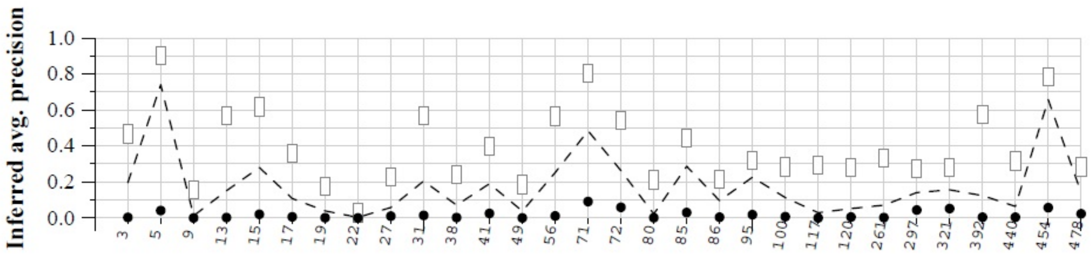

Figure 5. Run scores (dot) versus median (—) versus best (box) for *2C_M_A_FIU_UM.15_4*

Framework	10	100	1000	2000
2C_M_A_FIU_UM.15_1	0.180	0.114	0.064	0.049
2C_M_A_FIU_UM.15_2	0.250	0.149	0.072	0.055
2C_M_A_FIU_UM.15_3	0.187	0.132	0.071	0.053
2C_M_A_FIU_UM.15_4	0.200	0.113	0.061	0.048

Table 1: The MAP values at first *n* shots for all 4 runs

Framework	Inferred true shots returned	Mean xinfAP
2C_M_A_FIU_UM.15_1	2960	0.018
2C_M_A_FIU_UM.15_2	3314	0.026
2C_M_A_FIU_UM.15_3	3207	0.021
2C_M_A_FIU_UM.15_4	2865	0.018

Table 2: Inferred true shots returned and Mean xinfAP

directions are desired to be investigated:

- In our framework, only global features are utilized. Object-level and mid-level features need to be explored.
- The proper re-ranking strategy needs to be explored in depth to further improve the retrieval accuracy.
- The proper filtering strategy needs to be adopted to address the data imbalance issue.
- Deep learning techniques should be integrated to reach a better performance.

It is also necessary to exchange ideas and thoughts with other groups to come up with novel approaches to further improve the performance.

References

[1] Alan F. Smeaton, Paul Over, and Wessel Kraaij. *High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements*. Springer US, first edition, 2009.

[2] 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.

[3] S.-C. Chen, S. Sista, M.-L. Shyu, and R.L. Kashyap. Augmented transition networks as video browsing models for multimedia databases and multimedia information systems. In *The 11th IEEE International Conference on Tools with Artificial Intelligence*, pages 175–182, Nov. 1999.

[4] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht. Image retrieval by color, texture, and spatial information. In *The 8th International Conference on Distributed Multimedia Systems*, pages 152–159, Sept. 2002.

[5] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht. An effective content- based visual image retrieval system. In *The 26th IEEE International Computer Software and Applications Conference*, pages 914–919, Aug. 2002.

[6] X. Huang, S.-C. Chen, M.-L. Shyu, and C. Zhang. User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval. In *Third International Workshop on Multimedia Data Mining (MDM/KDD'2002), in conjunction with the 8th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 100–108, July 2002.

[7] M-L. Shyu, S-C. Chen, Q. Sun, and H. Yu. Overview and future trends of multimedia research of content access and distribution. *International Journal of Semantic Computing*, 1(1):29–66, March 2007.

[8] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Video semantic concept discovery using multimodal-based association classification. In *IEEE International Conference on Multimedia & Expo*, pages 859–862, July 2007.

[9] L. Lin and M.-L. Shyu. Mining high-level features from video using associations and correlations. In *IEEE International Conference on Semantic Computing (ICSC09)*, pages 137–144, September 2009.

[10] C. Chen, M.-L. Shyu, and S.-C. Chen. Supervised multi-class classification with adaptive and automatic parameter tuning. In *IEEE International Conference on Information Reuse and Integration (IRI09)*, pages 433–434, August 2009.

[11] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen. Weighted subspace filtering and ranking algorithms for video concept retrieval. *IEEE Multimedia*, 18(3):32–43, 2011.

[12] M.-L. Shu, C. Chen, and S.-C. Chen. Multi-class classification via subspace modeling. *International Journal of Semantic Computing*, 5(1):55–78, 2011.

[13] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Dianting Liu. Utilizing context information to enhance content-based image classification. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 2(3):34–51, 2011.

[14] Qiusha Zhu, Mei-Ling Shyu, and Shu-Ching Chen. Discriminative learning assisted video semantic concept classification. In Frank Y. Shih, editor, *Multimedia Security and Steganography*. CRC Press, 2012.

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

[16] 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.

[17] Q. Zhu and M.-L. Shyu. Sparse linear integration of content and context modalities for semantic concept retrieval. *IEEE Transactions on Emerging Topics in Computing*, 3(2):152–160, June 2015.

[18] Dianting Liu, Yilin Yan, Mei-Ling Shyu, Guiru Zhao, and Min Chen. Spatio-temporal analysis for human action detection and recognition in uncontrolled environments. *Int. J. Multimed. Data Eng. Manag.*, 6(1):1–18, January 2015.

[19] 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 *Semantic Computing (ICSC), 2014 IEEE International Conference on*, pages 28–35, June 2014.

[20] 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*, SIGIR '08, pages 603–610, New York, NY, USA, 2008. ACM.

[21] Savvas A. Chatzichristofis and Yiannis S. Boutalis. Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Proceedings of the 6th International Conference on Computer Vision Systems*, ICVS'08, pages 312–322, Berlin, Heidelberg, 2008. Springer-Verlag.

[22] Kenneth L. Critchfield, John F. Clarkin, Kenneth N. Levy, and Otto F. Kernberg. Organization of co-occurring axis ii features in borderline personality disorder. *British Journal of Clinical Psychology*, 47(2):185–200, 2008.

[23] S.E. Grigorescu, N. Petkov, and P. Kruizinga. Comparison of texture features based on gabor filters. *Image Processing, IEEE Transactions on*, 11(10):1160–1167, Oct 2002.

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511–I–518 vol.1, 2001.

[25] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 582–585 vol.1, Oct 1994.

[26] Hany Farid and EeroP. Simoncelli. Optimally rotation-equivariant directional derivative kernels. In Gerald Sommer, Kostas Daniilidis, and Josef Pauli, editors, *Computer Analysis of Images and Patterns*, volume 1296 of *Lecture Notes in Computer Science*, pages 207–214. Springer Berlin Heidelberg, 1997.

[27] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[28] Kai-Ting Chuang, Jun-Wei Hsieh, and Yilin Yan. Modeling and recognizing action contexts in persons using sparse representation. In Jeng-Shyang Pan, Ching-Nung Yang, and Chia-Chen Lin, editors, *Advances in Intelligent Systems and Applications - Volume 2*, volume 21 of *Smart Innovation, Systems and Technologies*, pages 531–541. Springer Berlin Heidelberg, 2013.

[29] John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, Nov 1986.

[30] Li-Chih Chen, Jun-Wei Hsieh, Yilin Yan, and Duan-Yu Chen. Vehicle make and model recognition using sparse representation and symmetrical surfs. *Pattern Recognition*, 48(6):19791998, June 2015.

[31] Lin Lin, G. Ravitz, Mei-Ling Shyu, and Shu-Ching Chen. Correlation-based video semantic concept detection using multiple correspondence analysis. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 316–321, Dec 2008.

[32] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. Enhancing concept detection by pruning data with mca-based transaction weights. In *Multimedia, 2009. ISM '09. 11th IEEE International Symposium on*, pages 304–311, Dec 2009.

[33] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. Feature selection using correlation and reliability based scoring metric for video semantic detection. In *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*, pages 462–469, 2010.

[34] Qiusha Zhu, Lin Lin, and Mei-Ling Shyu. Correlation maximisation-based discretisation for supervised classification. *International Journal of Business Intelligence and Data Mining*, 7(1/2):40–59, August 2012.

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

[36] 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.

[37] 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*, SIGIR '08, pages 603–610, New York, NY, USA, 2008. ACM.