# Quaero at TRECVID 2010: Semantic Indexing

Bahjat Safadi[1], Yubing Tong[1], Franck Thollard[1], Georges Quénot[1], Tobias Gehrig[2], Hazim Kemal Ekenel[2], and Rainer Stifelhagen[2]

[1]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
[2]Karlsruhe Institute of Technology, P.O. Box 3640, 76021 Karlsruhe, Germany

## Abstract

The Quaero group is a consortium of French organization working on Multimedia Indexing and Retrieval[1]. UJF-LIG and KIT participated to the semantic indexing task and UJF-LIG participated to the organization of this task. This paper describes these participations. For the semantic indexing task, a classical approach based on feature extraction, classification and hierarchical late fusion was used. Four runs were submitted corresponding to the use or not of genetic algorithm-based fusion and of two distinct fusion optimization methods. Both led to a small performance improvement and our best run has an infAP of 0.0485 (33/101).

UJF-LIG also co-organized the task with the support of the Quaero programme while taking care of avoiding conflict between participation and organization. We defined a new version of the previous HLF detection task, now called semantic indexing. Two versions of the task were proposed with different numbers of concept to detect: 10 for the "light" version and 130 for the "full" version. We organized as in 2007-2009 the collaborative annotation for this task using an active learning approach. An improvement was made this year by the use of relations between concepts during the annotation process. We also assessed 30 additional concepts for the evaluations. 10 of them were included in the official TRECVID 2010 evaluation and 20 more were delivered later.

## 1 Participation to the semantic indexing task

### 1.1 Introduction

The classical approach for concept classification in images or video shots is based on a three-stage pipeline: descriptors extraction, classification and fusion. In the first stage, descriptors are extracted from the raw data (video, image or audio signal). Descriptors can be extracted in different ways and from different modalities. In the second stage, a classification score is generated from each descriptor and, for each image or shot, and for each concept. In the third stage, a fusion of the classification scores obtained from the different descriptors is performed in order to produce a global score for each image or shot and for each concept. This score is generally used for producing a ranked list of images or shots that are the most likely to contain a target concept. In this work we have tried to improve the performance af a generic classification system with the use of features obtained from a face detection and categorization system. The original system uses a combination of low level features, including color, texture, SIFTs [3] and audio, and intermediate level features [2].

[1]http://www.quaero.org

### 1.2 Use of faces features

KIT has run its face detection and classification system on TRECVID 2010 video shots. It classified the number of faces visible in an image using a Modified Census Transformation (MCT) based face detector [8]. Additionally it classified gender (male or female), ethnicity (Asian, black or white) and age (child or adult) for detected faces using one common framework for feature extraction and classification using block-based Discrete Cosine Transform (DCT) [9] and Support Vector Machines (SVM) [10] respectively. Training was performed on datasets composed of images from the Color Feret Face Database [11, 12], MORPH-II Database [13], the FGNET Aging Database [14], some images collected from the web [15] and a small subset of the TRECVID development set.

We made a number of attempts for making use of the information extracted by KIT in order to improve the performance of our global system. Since many of the 130 concepts are not obviously related to face or person detection or classification and the task requires

Table 1: One-fold cross-validation result of the fusion process

| Run | MAP | last fusion level method |
|---|---|---|
| Multimodal_map | 0.1339 | Average Precision weighting |
| Multimodal_opt | 0.1403 | Direct optimization weighting |
| Multimodal_faces_map | 0.1356 | Average Precision weighting |
| Multimodal_faces_opt | 0.1422 | Direct optimization weighting |
| Multimodal_faces_ga_map | 0.1368 | Average Precision weighting |
| Multimodal_faces_ga_opt | 0.1432 | Direct optimization weighting |

submitting results on all the 130 concepts, we used our best system for concept recognition without the use of face processing as a baseline and then tried various approaches for improving over this baseline using KIT's output, many of which didn't work, possibly because the baseline, using a number of other cues, was already quite good even for concepts directly related to face recognition.

Our first attempt was to substitute the scores of KIT's detector when their output exactly matches or was very close to one of the target concepts, e.g. Female_Face_Close-Up or Asian_People. This did not lead to any improvement. We then tried some fusion between the score of the baseline system and the KIT scores and we could not get any improvement either.

Our second attempt was to combine all of the ten outputs of the KIT's detectors into a single 10-component vector in order to use it just as another intermediate level descriptor. This is very similar to the percept descriptors used in the baseline system. Classifiers were then trained with this descriptor and predictions were made directly from it. Though the individual performance of this descriptor is a bit low compared to other visual descriptors, it is able to produce a significant increase of the total system performance when combined with the other descriptors.

Table 1 shows the results obtained for various combinations of descriptors and various fusion strategies. On development data, the gain obtained by the use of the face detection and classification information can be seen between Multimodal_map (0.1339) and Multimodal_faces_map (0.1356, +1.3%) or between Multimodal_opt (0.1403) and Multimodal_faces_opt (0.1422, +1.4%). This gain is small but it is statistically significant. All fusion optimizations have been made by concept at this stage; therefore, the contribution of the new descriptor was automatically used where it is useful and with the appropriate weight. Two different optimization methods were used for the final fusion stage: a weighting by the average precision evaluated by cross-validation (*_map) or a direct search of the otpimal weights, also by cross-validation (*_opt). The latter is known to be better for the last stages of the hierarchical fusion but more unstable otherwise. An additional gain was obtained by the use of Genetic Algorithms (GA, *_ga_*) for the intermediate levels of fusion. This also leads to a small performance gain.

On test data and on all the 30 concepts, the gain obtained by the use of the face detection and classification information is a bit higher. It can be seen between Multimodal_map (0.0466) and Multimodal_faces_map (0.0476, +2.1%) or between Multimodal_opt (0.0471) and Multimodal_faces_opt (0.0485, +3.0%).

On test data and on the 4 concepts related to people categories, the gain obtained by the use of the face detection and classification information is significantly higher. It can be seen between Multimodal_map (0.0523) and Multimodal_faces_map (0.0598, +14.3%) or between Multimodal_opt (0.0553) and Multimodal_faces_opt (0.0635, +14.8%).

## 1.3 Performances on the semantic indexing task

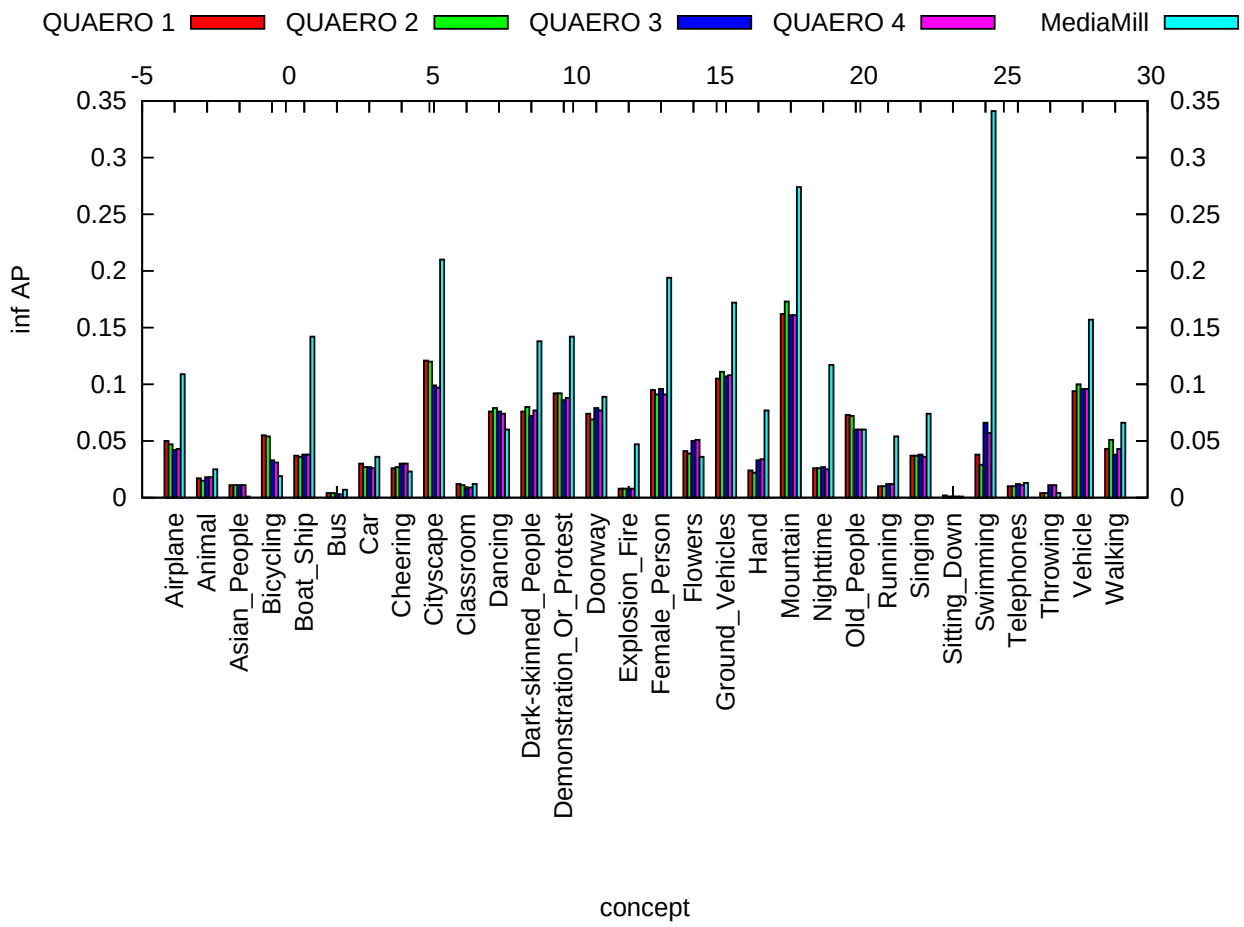
The four system variants including face features were used for the official TRECVID 2010 SIN submissions. They were prioritized according to the predictions on the development set. Table 2 and figure 1 present the result obtained by the four runs submitted. Although the absolute performances are quite different between the one obtained during the cross validation step, the ranking of the run is almost the same. Our best run is in the first third of the submissions. The use of Genetic Algorithms for the intermediate levels of fusion led to a smaller improvement than on the development set and even to a slight loss in the case of direct weight optimization.

Figure 2 provides insight of the performances of the Quaero best run at the concept level. The error bars provides the min/mean/max values of the systems where the line plots the performance of the Quaero run. As can be seen on the figure, our run is most of the time better than the average performances. On some concepts it is close to the best system on these concepts

Table 2: InfAP result and rank on the test set for all the 30 TRECVID 2010 concepts

| System/run | MAP | rank |
|---|---|---|
| Best submission | 0.0900 | 1 |
| F_A_Quaero_RUN02_2 (Multimodal_faces_opt) | 0.0485 | 32 |
| F_A_Quaero_RUN01_1 (Multimodal_faces_ga_opt) | 0.0484 | 33 |
| F_A_Quaero_RUN03_3 (Multimodal_faces_ga_map) | 0.0479 | 34 |
| F_A_Quaero_RUN04_4 (Multimodal_faces_map) | 0.0476 | 36 |
| Median submission | 0.0385 | 51 |

Figure 1: Quaero results at TRECVID 2010



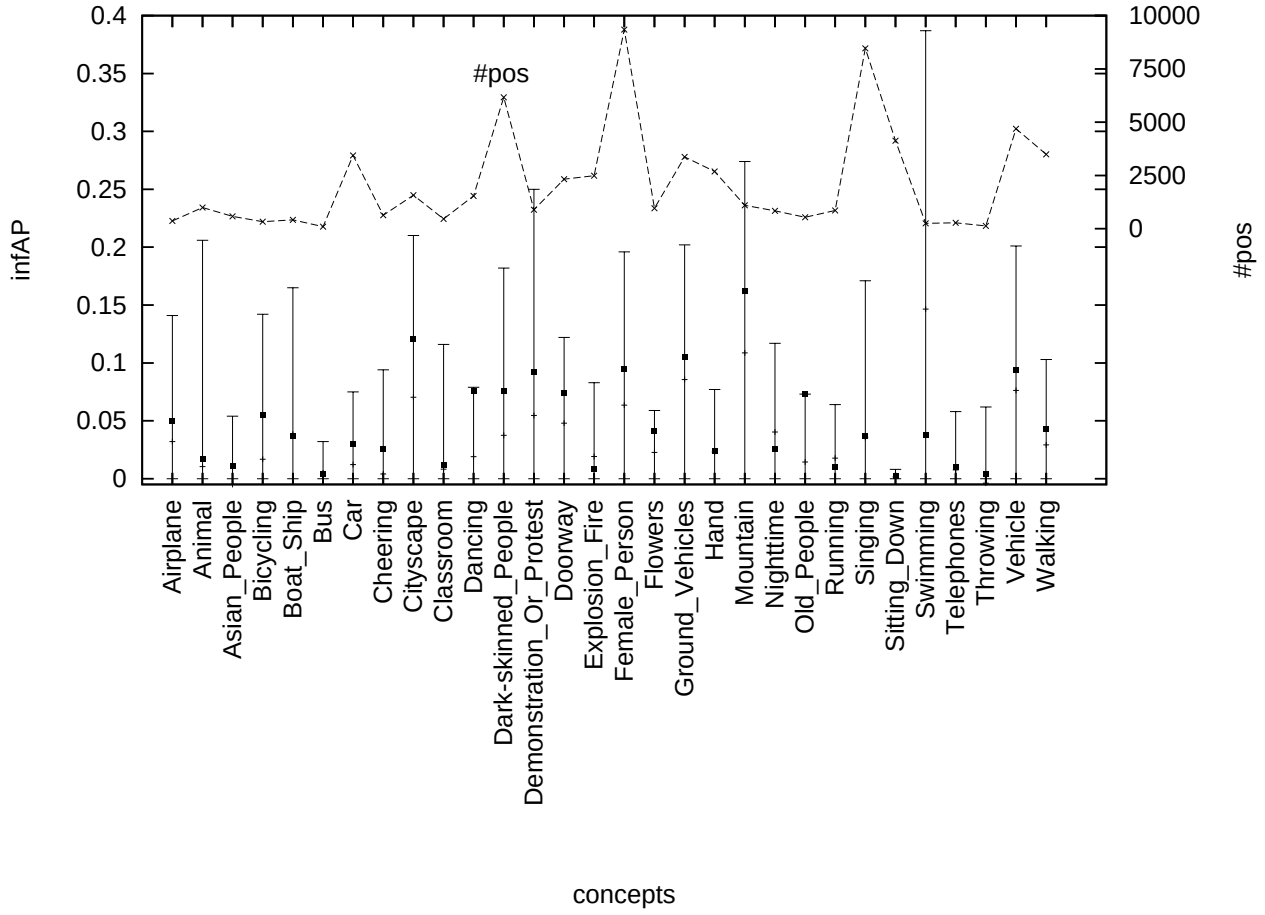

(e.g. Dancing/Old_People).

## 2 Organization of the semantic indexing task

This year, UJF-LIG has co-organized the semantic indexing task at TRECVID with the support of Quaero. A list of 130 target concepts has been produced, 30 of which have been officially evaluated at TRECVID and 20 more have also been assessed after the official evaluation.

The 130 concepts are structured according to the LSCOM hierarchy [4]. They include all the TRECVID "high level features" from 2005 to 2009 plus a selection of LSCOM concepts so that we end up with a number of generic-specific relations among them. We enriched the structure with two relations, namely *implies* and

Figure 2: Relative performance of the Quaero system




*excludes*. The goal was to promote research on methods for indexing many concepts and using ontology relations between them.

TRECVID provides participants with the following material:

- a development set that contains roughly 200 hours of videos;
- a test set that contains roughly 200 hours of videos;
- shot boundaries (for both sets);
- a set of 130 concepts with a set of associated relations;
- elements of ground truth: some shots were collaboratively annotated. For each shot and each concept, four possibilities are available: the shot has been annotated as positive (it contains the concept), the shot has been annotated as negative (it does not contain the concept), the shot has been skipped (the annotator cannot decide), or the shot has not been annotated (no annotator has seen the shot).

The goal of the semantic indexing task is then to provide, for each of the 130 concept, a ranked list of 2000 shots that are the most likely to contain the concept. The test collection contains 146,788 shots. A light version of the task has also been proposed in order to facilitate the asses to small and/or new groups. More information about the organization of this task can be found in the TRECVID 2010 overview paper []

## 2.1 Development and test sets

Data used in TRECVID are free of right for research purposes as it comes from the Internet Archive (http://www.archive.org/index.php). Table 3 provides the main characteristics of the collection set.

Table 3: Collection feature

| Characteristics | TRECVID 2010 |
|---|---|
| #videos | 11640 |
| Duration (total) | ∼400 hours |
| min;max;avg ± sd | 11s;1h;132s±93s |
| # shots | 266,473 |
| # shots (dev) | 119,685 |
| # shots (test) | 146,788 |

The whole set of videos has been split equally into two parts, the development set and the test set. Both sets were automatically split into shots using the LIG shot segmentation tool [5].

### 2.2 The evaluation measure

The evaluation measure used by TRECVID is the MAP (Mean Average Precision). Given the size of the corpus, the inferred MAP is used instead as it saves human efforts and has shown to provide a good estimate of the MAP [6].

### 2.3 Annotations on the development set

Shots in the development set have been collaboratively annotated by TRECVID 2010 participants. As concepts density is low, an active learning strategy has been set up in order to enhance the probability of providing relevant shots to annotators [1]: the active learning algorithm takes advantage of previously done annotations in order to provide shots that will more likely be relevant. Although this strategy introduces a bias, it raises the number of examples available to systems. Moreover, it exhibits some trend in the concept difficulty. As an example, the number of positive examples for the concept *Person* is larger than the number of negative examples. This means that the active learning algorithm was able to provide more positive examples than negative ones to annotators, meaning that *Person* is probably a "too easy" concept.

A total of 3,042,296 single concept × shots annotations were made, of which 819,297 by Quaero and the remainder by the TRECVID participants. Among these, 2,669,775 (87.8%) were done at least once, 267,191 (8.78%) were done at least twice and 105,330 (3.46%) were done three times. The multiple annotations were selected by the active learning tool as those being the more likely to correspond to errors or ambiguities and were made for cleaning as much as possible the annotations made. The resulting 2,669,775 annotations were amplified by the use of relations between concepts to 6,241,010 usable annotations. The relation used included the "implies" and "excludes" relations. These ∼6.2 M annotations represent about 40% of all the possible annotations on the development set. These have been selected by an active learning procedure that makes them almost as efficient as if the whole annotation was performed [1].

### 2.4 Assessments on the test set

Table 4 presents the concepts that were assessed.

The first column indicates who made the assessments and for which sub-task:

**light:** assessments made by NIST for the light and full tasks

**fullT:** assessments made by NIST for the full task

**fullQ:** assessments made by Quaero for the full task

**addQ:** additional assessments made by Quaero; not part of the official evaluation but usable for complementary post-workshop system evaluation.

Besides the two official TRECVID evaluations (light and full) that were intended to be generic, we propose to use these assessments for creating sub-lists of concepts (according to the X marks) and by placing restrictions on the type of data (still image, motion audio) that can be used by the systems. The sub-lists correspond to "static" concepts (Sta.), "dynamic concepts" (Dyn), "person (related)" concepts (Per.) and "multi-modal" concepts (M-M). The systems can be compared either in a generic way of in specific ways according to these sub-lists. Table 4 also indicate the number of positive and negative samples found for each concept in the development set.

## 3 Acknowledgments

This work was partly realized as part of the Quaero Programme funded by OSEO, French State agency for innovation. It was also partly supported by the UJF APIMS project. Some results from the IRIM network were also used in these experiments.

## References

[1] Stéphane Ayache and Georges Quénot. Video Corpus Annotation using Active Learning, In *30th European Conference on Information Retrieval (ECIR'08)*, Glasgow, Scotland, 30th March - 3rd April, 2008.

Table 4: Assessed concepts

| Assess. | Sta. | Dyn. | Per. | M-M | #pos | #neg | TV10 | LSCOM | LSCOM_Name |
|---|---|---|---|---|---|---|---|---|---|
| light | | X | | X | 360 | 17956 | 004 | 125 | Airplane_Flying |
| light | X | | | | 419 | 17836 | 015 | 233 | Boat_Ship |
| light | X | | | | 93 | 17772 | 019 | 227 | Bus |
| light | X | | | X | 1569 | 18195 | 028 | 068 | Cityscape |
| light | X | | | | 458 | 18098 | 029 | 275 | Classroom |
| light | | X | | | 895 | 17999 | 041 | 006 | Demonstration_Or_Protest |
| light | X | | | | 2684 | 17634 | 059 | 156 | Hand |
| light | X | | | | 838 | 20907 | 084 | 352 | Nighttime |
| light | | X | | X | 8466 | 16319 | 105 | 013 | Singing |
| light | X | | | | 281 | 17732 | 117 | 366 | Telephones |
| fullT | X | | | | 3436 | 16591 | 021 | 221 | Car |
| fullT | X | | | | 2327 | 20146 | 044 | 1578 | Doorway |
| fullT | | X | | X | 2484 | 17387 | 049 | 203 | Explosion_Fire |
| fullT | X | | X | X | 3229 | 17254 | 052 | 3155 | Female-Human-Face-Closeup |
| fullT | X | | | | 960 | 17618 | 053 | 307 | Flowers |
| fullT | X | | | | 3366 | 16727 | 058 | 609 | Ground_Vehicles |
| fullT | X | | | | 1096 | 20242 | 081 | 236 | Mountain |
| fullT | | X | | | 4131 | 17733 | 107 | 3149 | Sitting_Down |
| fullT | X | | | X | 4686 | 55028 | 126 | 108 | Vehicle |
| fullT | | X | | | 855 | 18107 | 100 | 003 | Running |
| fullQ | X | | | | 998 | 60101 | 006 | 201 | Animal |
| fullQ | X | | X | | 577 | 18218 | 007 | 246 | Asian_People |
| fullQ | | X | | | 329 | 17997 | 013 | 197 | Bicycling |
| fullQ | | X | | | 628 | 18482 | 027 | 033 | Cheering |
| fullQ | | X | | X | 1540 | 20727 | 038 | 028 | Dancing |
| fullQ | X | | X | | 6167 | 15886 | 039 | 258 | Dark-skinned_People |
| fullQ | | X | | | 250 | 22603 | 115 | 131 | Swimming |
| fullQ | | X | | | 133 | 19289 | 120 | 035 | Throwing |
| fullQ | | X | | | 3492 | 17978 | 127 | 012 | Walking |
| fullQ | X | | X | X | 533 | 18091 | 086 | 359 | Old_People |
| addQ | X | | X | X | 18175 | 11696 | 002 | 181 | Adult |
| addQ | X | | | | 9334 | 16178 | 018 | 226 | Building |
| addQ | X | | | | 8091 | 14388 | 030 | 212 | Computer_Or_Television_Screens |
| addQ | X | | | | 1747 | 18235 | 031 | 281 | Computers |
| addQ | X | | X | X | 9347 | 15073 | 051 | 103 | Female_Person |
| addQ | | X | | X | 57 | 18016 | 062 | 010 | Helicopter_Hovering |
| addQ | X | | | | 12693 | 14396 | 067 | 3156 | Indoor |
| addQ | X | | | X | 622 | 17773 | 068 | 157 | Indoor_Sports_Venue |
| addQ | X | | X | X | 236 | 17859 | 070 | 326 | Infants |
| addQ | | | | X | 3936 | 16218 | 071 | 637 | Instrumental_Musician |
| addQ | X | | | | 6725 | 18430 | 074 | 153 | Landscape |
| addQ | X | | X | X | 16643 | 11763 | 075 | 104 | Male_Person |
| addQ | X | | | | 349 | 19184 | 079 | 077 | Military_Base |
| addQ | X | | | X | 2514 | 17581 | 083 | 225 | News_Studio |
| addQ | X | | | | 6761 | 15912 | 098 | 206 | Road |
| addQ | X | | | | 3939 | 17837 | 101 | 403 | Scene_Text |
| addQ | X | | | X | 541 | 18250 | 112 | 058 | Stadium |
| addQ | | X | | | 1884 | 20595 | 128 | 205 | Walking_Running |
| addQ | X | | | | 3764 | 16663 | 129 | 209 | Waterscape_Waterfront |
| addQ | X | | | | 9461 | 19085 | 091 | 2451 | Plant |

[2] S. Ayache, G. Quénot, J. Gensel, and S. Satoh. Using topic concepts for semantic video shots classification. In Springer, editor, *CIVR – International Conference on Image and Video Retrieval*, 2006.

[3] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. A comparison of color features for visual concept classification. In *ACM International Conference on Image and Video Retrieval*, pages 141–150, 2008.

[4] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13:86–91, 2006.

[5] Georges Quénot, Daniel Moraru, and Laurent Besacier. CLIPS at TRECvid: Shot boundary detection and feature detection. In *TRECVID'2003 Workshop*, Gaithersburg, MD, USA, 2003.

[6] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR*. ACM 978-1-60558-164-4/08/07, July 2008.

[7] Paul Over, George Awad, Jon Fiscus, Brian Antonishek, Martial Michel, Alan F. Smeaton, Wessel Kraaij and Georges Quénot TRECVID 2010 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics, In *Proceedings of the TRECVID 2010 workshop*, Gaithersburg, USA, 15-17 Nov. 2010.

[8] Christian Küblbeck and Andreas Ernst. Face detection and tracking in video sequences using the modified census transformation. *Image and Vision Computing*, 24(6):564–572, June 2006.

[9] Hazim Kemal Ekenel. *A Robust Face Recognition Algorithm for Real-World Applications*. PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, February 2009.

[10] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/\~cjlin/libsvm`.

[11] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[12] P. Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[13] Karl Ricanek Jr. and Tamirat Tesafaye. MORPH: A Longitudinal Image Database of Normal Adult Age-Progression. In *IEEE 7th International Conference on Automatic Face and Gesture Recognition (FGR'06)*, pages 341–345, April 2006.

[14] FG-NET Aging Database.

[15] Clemens Siebler. Gesichtsbasierte Geschlechtserkennung auf Bildsequenzen. Studienarbeit, Universität Karlsruhe (TH), Karlsruhe, Germany, July 2008.