VITALAS at TRECVID-2009

Christos Diou [1,2], George Stephanopoulos [2], Nikos Dimitriou [1,2],
Panagiotis Panagiotopoulos [1,2], Christos Papachristou [1,2], Anastasios Delopoulos [1,2],
Henning Rode [3], Theodora Tsikrika [3], Arjen P. de Vries [3],
Daniel Schneider [4], Jochen Schwenninger [4],
Marie-Luce Viaud [5], Agnès Saulnier [5],
Peter Altendorf [6], Birgit Schröter [6], Matthias Elser [6],
Angel Rego [7], Alex Rodriguez [7], Cristina Martínez [7] Iñaki Etxaniz [7]
Gérard Dupont [8], Bruno Grilhères [8], Nicolas Martin [8],
Nozha Boujemaa [9], Alexis Joly [9], Raffi Enficiaud [9],
Anne Verroust [9], Souheil Selmi [9], Mondher Khadhraoui [9]

[1] Multimedia Understanding Group, ECE Dept., Aristotle University of Thessaloniki, Greece
[2] Informatics and Telematics Institute, Centre for Research and Technology Hellas
[3] CWI, Amsterdam, The Netherlands
[4] Fraunhofer IAIS
[5] French Audiovisual Institute
[6] Institut für Rundfunktechnik
[7] Robotiker-Tecnalia
[8] EADS Val de Reuil
[9] INRIA Rocquencourt, France

ABSTRACT

This paper describes the participation of VITALAS in the TRECVID-2009 evaluation where we submitted runs for the High-Level Feature Extraction (HLFE) and Interactive Search tasks.

For the HLFE task, we focus on the evaluation of low-level feature sets and fusion methods. The runs employ multiple low-level features based on all available modalities (visual, audio and text) and the results show that use of such features improves the retrieval effectiveness significantly. We also use a concept score fusion approach that achieves good results with reduced low-level feature vector dimensionality. Furthermore, a weighting scheme is introduced for cluster assignment in the “bag-of-words” approach. Our runs achieved good performance compared to a baseline run and the submissions of other TRECVID-2009 participants.

For the Interactive Search task, we focus on the evaluation of the integrated VITALAS system in order to gain insights into the use and effectiveness of the system’s search functionalities on (the combination of) multiple modalities and study the behavior of two user groups: professional archivists and non-professional users. Our analysis indicates that both user groups submit about the same total number of queries and use the search functionalities in a similar way, but professional users save twice as many shots and examine shots deeper in the ranked retrieved list.The agreement between the TRECVID assessors and our users was quite low. In terms of the effectiveness of the different search modalities, similarity searches retrieve on average twice as many relevant shots as keyword searches, fused searches three times as many, while concept searches retrieve even up to five times as many relevant shots, indicating the benefits of the use of robust concept detectors in multimodal video retrieval.

High-Level Feature Extraction Runs

1. A_VITALAS.CERTH-ITI_1: Early fusion of all available low-level features.
2. A_VITALAS.CERTH-ITI_2: Concept score fusion for five low-level features and 100 concepts, text features and bag-of-words with color SIFT descriptor based on dense sampling.
3. A_VITALAS.CERTH-ITI_3: Concept score fusion for five low-level features and 100 concepts combined with text features.
4. A_VITALAS.CERTH-ITI_4: Weighting scheme for bag-of-words based on dense sampling of the color SIFT descriptor.
5. A_VITALAS.CERTH-ITI_5: Baseline run, bag-of-words based on dense sampling of the color SIFT descriptor.

Interactive Search Runs

1. vitalas_1: Interactive run by professional archivists
2. vitalas_2: Interactive run by professional archivists
3. vitalas_3: Interactive run by non-professional users
4. vitalas_4: Interactive run by non-professional users

1. INTRODUCTION

VITALAS, now in its final year, is an EU-funded Integrated Project that aims at the development of a system capable of large-scale indexing and retrieval of video and images, specifically targeted towards multimedia professionals and archivists [1]. One of the key features of the system is the ability to perform "Concept search", i.e., retrieve multimedia documents using High-Level Features or *Concepts*. VITALAS participated in the High-Level Feature Extraction (HLFE) with the aim to evaluate (i) the effectiveness of a set of low-level features extracted from multiple modalities, (ii) the use of multiple external concept models at the fusion stage (concept score fusion), and (iii) the use of weighted cluster assignments when constructing frequency histograms from local descriptors in the bag-of-words model. Overall, the submitted HLFE runs achieved good results, with our best run ranked 28^{th} in a total of 222 submissions and 7^{th} in a total of 42 best runs. Details about the system components, submitted HLFE runs and results are presented in Section 2.

We also participated in the Interactive Search task in order to evaluate the integrated VITALAS system (including its user interface) from a user perspective.Since the VITALAS system supports search on (the combination of) multiple modalities, the aim of our experiments was to analyze and compare the use and effectiveness of the different search functionalities and also examine how different types of users interact with the system by studying the behavior of professional archivists and non-professional users. Although there was a low agreement between the TREVID assessors and our users, our experiments provided us with valuable insights into the search behavior of different user groups. Section 3 gives a brief overview of the VITALAS system, describes our experimental set-up, and presents and discusses the results of our analysis.

2. HIGH-LEVEL FEATURE EXTRACTION

The HLFE system was developed by the Multimedia Understanding Group (affiliated with both Aristotle University of Thessaloniki and CERTH-ITI). Some of the system components have already been used in the VITALAS system successfully, while others are candidates for inclusion and are under evaluation. The general architecture of the system is illustrated in Figure 1.

It is assumed that the temporal segmentation of video into shots has already been performed. For TRECVID-2009 participants, this segmentation was performed by C. Petersohn using the method described in [21]. Features that don't use the temporal dimension are extracted from keyframes of the video shot. The multiple low-level features that are extracted from each video shot are subsequently combined in the fusion module and then employed by the classifier in order to produce the final ranking.

2.1 Low-Level Features

Low-level features were extracted from three modalities, visual, audio and text. For the visual modality the multiple extracted features can be categorized into "Global", if the feature is directly computed from the entire keyframe, "Regional" if the features are computed on the basis of keyframe regions or "Local" if the features are based on specific points in the keyframe/shot.

2.1.1 *Global and Regional Visual Features*

The global and regional features used are a variation of the MPEG-7 Color Structure Descriptor [19] (CSD), a feature extracted using the keyframe Dominant Color (DCOLOR), one based on the lines detected in various angles by the Hough transform (HOUGH), a feature based on the Integrated Weibull distribution extracted from keyframe regions [26] (WBL), and a feature based on the output of a face detector (FACE).

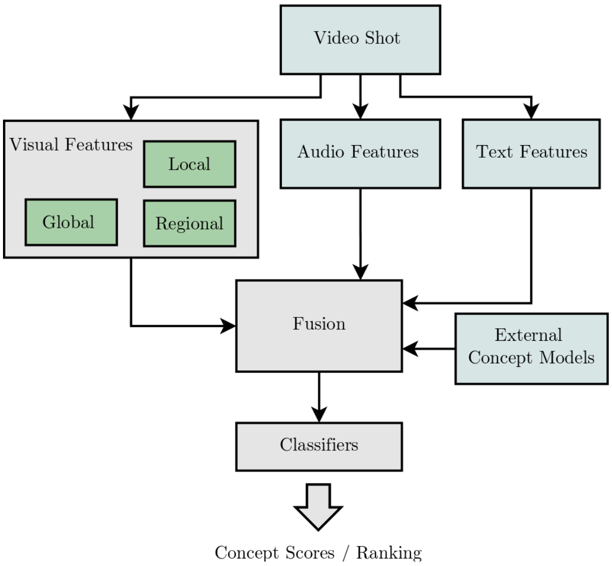


Figure 1: The HLFE system architecture. Features of all modalities are extracted from each shot and then combined using early fusion or concept score fusion. A classifier uses the produced feature vector to assign scores to each concept/shot combination and produce the final ranking.

For CSD, the color structure histogram is computed in a manner similar to the simple color histogram, with a significant difference in the accumulation process: At each image point, a region covered by a structuring element (a 3×3 box in our case) is examined and the colors present in the element are incremented. Note that only a color's presence or absence is important and not the amount present in the structuring element.

For DCOLOR, *octrees* [10] are used in order to achieve effective and fast color reduction in the keyframe (see also [6]). More specifically, a color signature is extracted from each keyframe I,

$$CS_I = \{(c_0, p_0), \dots, (c_N, p_N)\} \tag{1}$$

where each c_i is one of the dominant colors and p_i is the corresponding percentage of c_i in the image, after reduction with octrees. Since the number of colors N depends on the image, the Earth Mover's Distance (EMD) [22] is employed in order to compare two color signatures. To produce the final feature vector, a method similar to the "contexture histograms" used in [26] is applied. Since this method is also used for the extraction of the WBL feature, it is presented here for completeness.

We consider the manually labeled image regions corresponding to 15 concepts[1] that form 15 sets $\mathbf{P}_i, i = 1, \dots, 15$

[1] The concepts are: *building*, *car*, *charts*, *crowd*, *desert*, *fire*, *maps*, *mountain*, *road*, *sky*, *smoke*, *snow*, *US-flag*, *vegetation*, *water* and the corresponding image regions have been manually extracted from images in the TRECVID-2005 de-

of reference images. For a keyframe I and a set $\mathbf{P}_i$ two values can be computed, namely the average and best distance of I to the images in $\mathbf{P}_i$,

$$H^i_{avg} = \frac{1}{|\mathbf{P}_i|}\sum_{j=0}^{|\mathbf{P}_i|} EMD(CS_I, CS_{P_i(j)}) \tag{2}$$

$$H^i_{best} = \min_j(EMD(CS_I, CS_{P_i(j)})) \tag{3}$$

This approach leads to a 30-dimensional feature vector for dominant color.

For the description of keyframes based on the direction of lines detected by the Hough transform, we use a simple histogram of angles that was extracted from the Hough accumulator matrix. An accumulator matrix with 24 angles is used and after computing the values for the matrix cells, only values that exceed a predefined threshold and are local maxima are kept (otherwise they are set to zero). Adding all rows of the accumulator matrix leads to a 24-bin angle histogram, that is normalized to a unit vector with its L^2 norm to produce the final feature vector.

The WBL feature is extracted using the method described in [26]. Initially, the colorspace of each keyframe is transformed to the Gaussian color model [11] and each color channel is filtered using one Gaussian derivative filter for each image direction (horizontal and vertical). This process is repeated for two filter scales (i.e., two values of σ). The edges of a region in any of the 12 resulting images is assumed to follow an Integrated Weibull distribution,

$$\frac{\gamma}{2\gamma^{\frac{1}{\gamma}}\beta\Gamma(\frac{1}{\gamma})}\exp\left\{-\frac{1}{\gamma}\left|\frac{r-\mu}{\beta}\right|^{\frac{1}{\gamma}}\right\} \tag{4}$$

where β and γ are the distribution parameters and μ is assumed to be zero (this is ensured by pre-processing of the region values). Computation of the β and γ values is performed numerically, while comparison between two distributions is achieved using a metric derived in [26] from the Cramér von Mises statistic $C = \frac{\min(\beta_1,\beta_2)}{\max(\beta_1,\beta_2)}\frac{\min(\gamma_1,\gamma_2)}{\max(\gamma_1,\gamma_2)}$. A set of overlapping keyframe regions of two different sizes are considered for each keyframe and using the reference concepts in a manner similar to DCOLOR, the following values are computed

$$H^{i,s_R}_{avg} = \frac{1}{|\mathbf{P}_i||R|}\sum_{r\in R}\sum_{j=0}^{|\mathbf{P}_i|} C(S_r, S_{P_i(j)}) \tag{5}$$

$$H^{i,s_R}_{best} = \max_{r\in R}(\sum_{j=0} C(S_r, S_{P_i(j)})) \tag{6}$$

where s_R is the region size corresponding to the set of regions R (of the same size), $S_r, S_{P_i(j)}$ are the signatures of the region r and image $P_i(j)$ of the i-th reference concept respectively. There are two region sizes, two scales for the Gaussian derivative filters and two values for each region size and scale leading to 8 values for each reference concept and a 120-d feature vector based on the Integrated Weibull distribution.

One more regional low-level feature was implemented using the output of a Viola-Jones [28] face detector. The produced feature vector has 20 dimensions and each dimension provides different type of information about the detected faces. Faces are assigned to three categories depending on the percentage of keyframe area they occupy ($[0, 0.05)$, $[0.05, 0.1)$ and $[0.1, 1]$). Three dimensions contain the percentage of frames in the shot that have faces of each category and depending on whether this percentage exceeds 20%, another three dimensions are 0 or 1. Four dimensions for each category (total 12) also contain the percentage of frames in the shot where 0, 1, 2 or more than 2 faces of that category appear. One additional dimension contains the average area percentage that the biggest face of each frame covers, while one last entry is binary, indicating whether this percentage is greater than 20%.

2.1.2 Local Visual Features

Local features are extracted for both color (C-LOCAL) and motion (M-LOCAL) information. The methods presented in [25] and [18] and the corresponding binaries that have been made publicly available by K. van de Sande and I. Laptev were used respectively.

For color local features dense uniform sampling of image points is used (instead of a keypoint detector) and for each point, the Opponent-SIFT descriptor is extracted. A sampling of such descriptors extracted from the development set was used to compute a set of 4000 cluster centers using k-means. Based on those centers, feature vectors were constructed using a typical bag-of-words approach, i.e., the feature vector is a histogram of cluster frequencies in the keyframe.

The same approach and number of clusters were also used for local motion information. In this case, space-time interest points are detected in the entire shot and the “Histograms of Oriented Gradients” and “Histograms of Optical Flow” descriptors are used for each point.

2.1.3 Weighted Bag-of-Words Features

We computed one more low-level feature based on local descriptors that uses a form of soft assignment of descriptors to clusters (W-LOCAL). Apart from modeling the uncertainty associated with the assignment of descriptors to clusters, its goal is to “normalize” the assignment with respect to the statistical behavior of each cluster’s samples.

More specifically, we associated a covariance matrix Σ_c with each cluster center c in the local descriptor space and for simplicity, we assumed that this matrix is diagonal (i.e., the descriptor space dimensions are uncorrelated). Using the Mahalanobis distance gives

$$d^2_w(c, x) = \sum_{i=1}^{N}\frac{1}{\sigma^2_i}(x_i - c_i)^2$$

where N is the number of dimensions in the descriptor space.

For each data point x, all distances $d_w(c, x)$ from the cluster centers c are computed and only the top k distances are kept. The point x contributes $1/d_w(c, x)$ to each of the k dimensions corresponding to c in the feature vector.

2.1.4 Audio Features

Low-level features based on audio (AUDIO) are extracted using the same principles as the local features for visual input. MFCCs extracted from small overlapping audio segments of the video shot act as the local descriptor. A sample of the development set is used to produce a set of 500 clusters in the MFCC descriptor space. For the construction of

velopment set.

the final $500-d$ feature vector, soft weighting that takes into account the top k cluster centers is applied.

2.1.5 Text Features

Extraction of text features (TEXT) is based on the output of an ASR system [14] available to all TRECVID participants. Again, a bag-of-words model is followed, using a fixed vocabulary for each concept. The vocabulary is constructed (similar to [23]) by the words that appear in the shots that are positive examples for the concept after stemming and stop-word removal. Each dimension in the final feature vector signifies the number of occurrences of the corresponding vocabulary word in the video shot.

2.2 High-Level Features

The low-level features extracted from video shots are combined in a single feature vector prior to the final classification stage using two approaches, early fusion and concept score fusion. Early fusion is the simple concatenation of the feature vectors, possibly followed by normalization to a predefined value range separately for each dimension. Concept score fusion uses a set of intermediate classifiers to produce the final feature vector.

2.2.1 Concept Score Fusion

Concept score fusion is based on the assumption that detectors trained for a set of concepts (the *base* concepts) can assist in the detection of other concepts (the *target* concepts). Similar approaches have also been used in [7], [29] and [20]. Instead of directly using the available feature vectors, the base concept detectors are applied separately for each feature. The resulting scores are then concatenated to form a single feature vector consisting of concept scores only.

In our implementation 100 base concept detectors were trained on the TRECVID-2005 dataset using the LSCOM annotations [17]. Selection of the base concepts was based on criteria related to the detector's "consistency", rather than effectiveness: (i) The training set of the base concept detector must have a number of positive examples that exceeds a predefined threshold. (ii) Rather than examining the absolute effectiveness of the detector, the ratio of average precision to the prior of the concept within an evaluation set was used. The base concept scores are computed for each low-level feature separately and the resulting scores are concatenated to form the final feature vector.

2.2.2 The Classifier

The concept detectors used are SVM [27, 4] classifiers based on the LibSVM [13, 5] implementation. Two kernels are used, the radial basis function kernel and the χ^2 kernel $K(\mathbf{u},\mathbf{v}) = e^{-\frac{1}{A}D(\mathbf{u},\mathbf{v})}$, with

$$D(\mathbf{u},\mathbf{v}) = \frac{1}{2}\sum_{1}^{N}\frac{(u_i-v_i)^2}{u_i+v_i} \tag{7}$$

The results of [31] indicate that the χ^2 is more suitable for histogram/vocabulary features, such as the local features of section 2.1.2. In addition, our experience shows that the χ^2 is less affected by increased dimensionality, compared to the RBF kernel whose performance deteriorates for a large number of dimensions.

Two additional steps were taken to ease the computational requirements of classifier training: (i) Instead of using the entire development set for training, all positive and a subset of 10000 randomly selected negative examples were used. (ii) In order to avoid the heavy computational cost of cross-validation without significant loss in performance, heuristics were used to select the training parameters so that they approximate the optimal ones. The training set priors of each class were used to set the penalty factors C_+ and C_- for each class, while a predefined value was used for the other kernel parameters ($\gamma = 1/\text{N}$ for RBF, and the mean of the distances from cluster centers for parameter A of the χ^2 kernel).

2.3 HLFE Runs

A total of 5 runs were submitted aiming at the evaluation of the multi-modal features used, the soft weighting scheme for the local features (Section 2.1.3) as well as the concept score fusion method (Section 2.2.1) in the TRECVID-2009 dataset.

The training data used to construct the concept detectors of the submitted runs are the result of the collaborative participants' effort organized and coordinated by the LIG and LIF groups [2]. The annotation unit is the video shot, while keyframes were extracted using a simple periodic selection policy.

Bag-of-words using dense sampling and color SIFT descriptor (Run 5, baseline run). This is the baseline run, using the C-LOCAL feature that achieved very good results in TRECVID-2008.

Weighting scheme for bag-of-words features (Run 4). The same local descriptor, but with soft cluster assignment (W-LOCAL).

Concept score fusion for five low-level features combined with text features (Run 3). Concept score fusion for WBL, DCOLOR, HOUGH, CSD and C-LOCAL combined with TEXT.

Concept score fusion for five low-level features combined with text and local features (Run 2). Same as Run 3, with the addition of C-LOCAL.

Early fusion of all low-level features (Run 1). Early fusion of all features presented in Section 2.1.

2.4 HLFE Results

The HLFE submissions were evaluated by the TRECVID organizers using a reduced ground truth sample and the results are shown in Table 1 in terms of inferred Average Precision (infAP) [30].

The most obvious observation to be made is that the simple concatenation of all available features leads to the best results in most cases and achieves significantly higher mean infAP overall. This result can be explained by the fact the this is the only run that includes all available information and multimodal low-level features: Concepts *Person-playing-a-musical-instrument*, *People-dancing* and *Singing* present a huge improvement compared to other runs (400%, 919% and 261% respectively, compared to the baseline), and that can only be attributed to the presence of audio features. Another noteworthy case is the concept *Female-*

No.	Concept	Run 1	Run 2	Run 3	Run 4	Run 5
1	Classroom	0.0560	0.0290	0.0270	**0.0740**	0.0310
2	Chair	**0.0840**	0.0660	0.0600	0.0460	0.0570
3	Infant	**0.0070**	0.0010	0.0030	0.0040	0.0050
4	Traffic-intersection	0.1380	0.1320	**0.1550**	0.0810	0.0860
5	Doorway	**0.2100**	0.1540	0.1950	0.1030	0.1510
6	Airplane_flying	0.1000	**0.1060**	0.0780	0.0490	0.0350
7	Person-playing-a-musical-instrument	**0.2000**	0.0390	0.0300	0.0150	0.0410
8	Bus	0.0370	0.0330	**0.0490**	0.0250	0.0190
9	Person-playing-soccer	**0.4440**	0.4040	0.3720	0.3000	0.3050
10	Cityscape	0.1900	**0.1920**	0.1670	0.1200	0.1830
11	Person-riding-a-bicycle	0.0300	0.0270	**0.0570**	0.0260	0.0110
12	Telephone	**0.0280**	0.0120	0.0090	0.0130	0.0100
13	Person-eating	**0.0050**	0.0020	0.0020	0.0020	0.0010
14	Demonstration_Or_Protest	**0.0770**	0.0300	0.0370	0.0240	0.0240
15	Hand	**0.2280**	0.1680	0.2160	0.1030	0.1420
16	People-dancing	**0.2140**	0.1070	0.1120	0.0250	0.0210
17	Nighttime	**0.2670**	0.2080	0.1580	0.1830	0.1920
18	Boat_Ship	0.1820	**0.1910**	0.1510	0.1620	0.1860
19	Female-human-face-closeup	**0.2920**	0.1740	0.1470	0.1870	0.1900
20	Singing	**0.1480**	0.0290	0.0230	0.0370	0.0410
Mean infAP		0.1468	0.1052	0.1024	0.0789	0.0865

Table 1: Overview of the inferred Average Precision (infAP) results of the HLFE task.

human-face-closeup (54%, possibly due to the FACE feature) while *Person-playing-soccer* and *Doorway* are also improved (46% and 40%), possibly due to the M-LOCAL motion feature. Overall, all indications show that using a large set of multi-modal features leads to significant improvement of results.

Run 3 provides an initial evaluation of the concept score fusion effectiveness. Text is also used directly in this run. The reason for this configuration is that this is one of the operation modes of the VITALAS system being developed. The feature vectors used in this run have a relatively small number of dimensions (5 features ×100 concepts and the sparse text feature vector, compared to the thousands of dimensions used by the other runs), while the RBF kernel was used in the classifier. The results indicate a significantly better performance than the baseline run. Moreover, concept score fusion achieves the best results in three cases.

In order to examine how much information is “lost” by using the concept score fusion method instead of directly applying the features, Run 2 uses the same features with Run 3, with the addition of the C-LOCAL feature of the baseline run. This modification does not lead to any significant improvement overall, increasing our confidence in the concept score fusion approach. Note, however, that Run 4 leads to the best results for three concepts, but with a small difference from the second best.

Runs 1, 2 and 3 all perform better than the baseline, however the same is not true for Run 4 that uses only the W-LOCAL feature. Overall, this run performs slightly worse than the baseline, although it performs better for some individual concepts. Note also that for concept *Classroom* this run achieves the best result overall. These observations show that there is still research that needs to be done on the weighted assignment of local descriptors to clusters, but also that this family of methods has the potential to improve the retrieval results.

Figure 2 depicts a comparison of the VITALAS HLFE runs with the runs submitted by the other TRECVID-2009 participants while Figure 3 restricts the same display to the best run of each participant only.

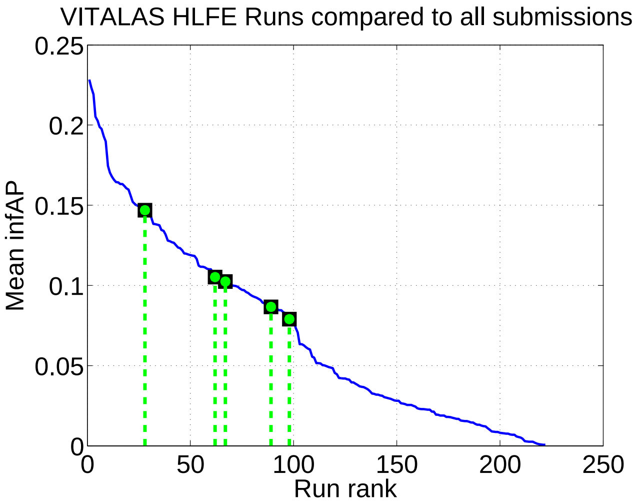


Figure 2: Comparison between VITALAS runs and the runs submitted by each of the other TRECVID-2009 participants. The VITALAS runs are indicated with a square marker.

3. INTERACTIVE SEARCH

We participated in the Interactive Search task with the aim to evaluate the integrated VITALAS system (including its user interface) from a user perspective. This system has been built by the partners of the VITALAS EU-funded research project which has been developing a video and image retrieval system for large collections that integrates different search methods into a single advanced user interface. All

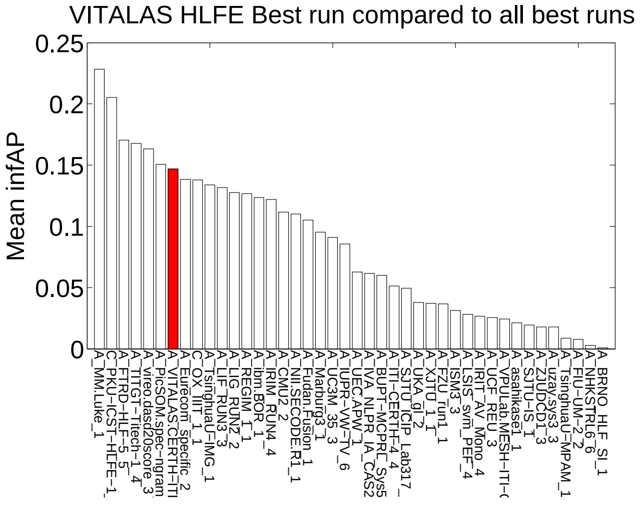


Figure 3: Comparison between the VITALAS best run (shaded bar) and the best run submitted by each of the other TRECVID-2009 participants.

search services have been developed as independent components and integrated by wrapping them into a common web-service architecture. Since the VITALAS system supports search on (the combination of) multiple modalities, the aim of our experiments was to analyze and compare the use and effectiveness of the different search functionalities from a user point of view. We also examined how different types of users interact with the system by studying the behavior of professional archivists and non-professional users.

The remainder of this section is structured as follows. Section 3.1 introduces the VITALAS system, its components, and user interface, while Section 3.2 describes the data preparation and indexing steps. The set-up and execution of the user tests is detailed in Section 3.3 and all evaluation results are presented in Section 3.4.

3.1 The VITALAS System

The TRECVID prototype of the VITALAS system allows users to make use of the following search functionalities: (i) keyword search, (ii) concept search, (iii) visual similarity search, (iv) fused search (combining the above methods), and (v) concept suggestions.

The text search component allows to search the text output coming from automatic speech recognition (ASR) on the video material. It provides common full-text search functionalities, such as keyword search and phrase search, and returns a ranked list of shots. The concept search retrieves shots based on automatically detected concepts. Similarly to keyword queries, the user can search for any combination of concepts and the system ranks the shots in the collection based on the estimated combined relevance. A visual similarity search completes the set of search possibilities. Once users have found one or more relevant examples, they can use them to search for shots in order to find visually similar keyframes. The similarity search therefore enables the retrieval of shots based on visual features, without being bound to the predefined set of concepts. Users are further supported in their search tasks by a concept suggestion service. Whenever a user issues a text search, the service returns concept suggestions related to the submitted query. In this way, users are made aware of automatically detected concepts that might be useful for refining or expanding their text queries.

All search methods are reachable from an integrated user interface. Figure 4 shows the main query interface and result view. The top text field allows users to enter keyword and concept queries. The retrieved shots are shown by a thumbnail keyframe in the mosaic result view below. Each thumbnail can be added to a light-box that is used for gathering possibly relevant shots for the search topic. Furthermore, the thumbnails can be clicked; this opens a zoom view showing the keyframe in a higher resolution, and also enables to enter a detailed view in order to play the shot. The thumbnails can also be used to issue a visual similarity search. The detailed view allows users to play the shot and to jump to any other shot in the video. Unfortunately, the test prototype of the system did not allow to add other shots from the detailed view to the light box. All GUI functionalities are also reachable via hotkeys for a more efficient user handling. The concept suggestions are shown in a special suggestion bar, which also foresees term suggestions that were not available for the TRECVID prototype. Suggested concepts can be added to the last query by clicking them.

3.2 Dataset Preparation

For indexing the TRECVID collection with the VITALAS system, the provided video material, master shot segmentation, and ASR output had to be preprocessed, transformed, and enriched by generated metadata.

3.2.1 ASR Translation, Shot-Alignment, Indexing

We used the provided ASR output from the LIMSI system [9]. It contains the recognized Dutch text associated with corresponding time stamps. Since not all of our test users speak Dutch and are thus unable to issue their queries in Dutch, we employed the Google machine translation services[2] for an automatic translation of the text to English. We translated the ASR text sentence by sentence using the provided Java API.

In a second processing step, the translated text was joined and clustered according to the master shot segmentation. ASR snippets that overlap multiple shots with respect to their time stamps were added to all corresponding shots. Furthermore, we smoothed the ASR shot alignment by including to each shot the text of neighboring shots with exponential fading weights according to the shot distance: each shot contains 8 times its own text, 4 times the text of direct neighboring shots, 2 times the text of second degree neighbors, and 1 time the text of third degree neighbors. This shot smoothing roughly follows the approach proposed by Huurnink and de Rijke [15].

These textual data were indexed by the Lucene retrieval engine[3], which was also used later for text querying.

3.2.2 Keyframes and Visual Feature Extraction

For the visual similarity search, all keyframes were analyzed and indexed by the Maestro system [3]. The creation of the index consists of two complementary phases: the extraction of visual features and the construction of an efficient indexing structure. The latter employs a special encoding of

[2] see `http://code.google.com/p/google-api-translate-java`

[3] see `http://lucene.apache.org`

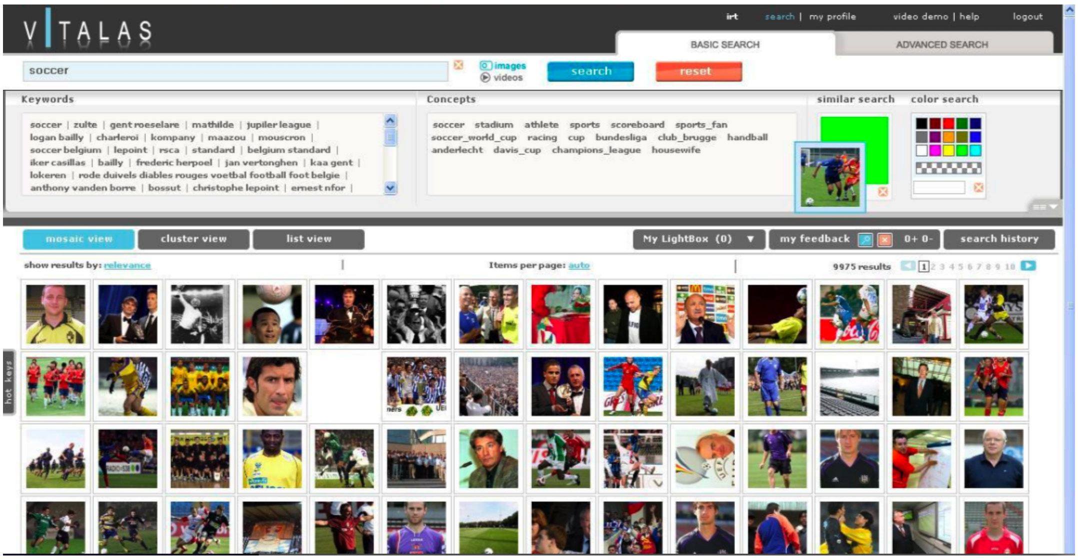

V|TALAS
irt
search | my profile
video demo | help
logout
BASIC SEARCH
ADVANCED SEARCH
soccer
images
videos
search
reset
Keywords
soccer | zulte | gent roeselare | mathilde | jupiler league |
logan bailly | charleroi | kompany | maazou | mouscron |
soccer belgium | lepoint | rsca | standard | belgium standard |
iker casillas | bailly | frederic herpoel | jan vertonghen | kaa gent |
lokeren | rode duivels diables rouges voetbal football foot belgie |
anthony vanden borre | bossut | christophe lepoint | ernest nfor |
Concepts
soccer stadium athlete sports scoreboard sports_fan
soccer_world_cup racing cup bundesliga club_brugge handball
anderlecht davis_cup champions_league housewife
similar search
color search
mosaic view
cluster view
list view
My LightBox (0)
my feedback
0+ 0-
search history
show results by: relevance
Items per page: auto
9975 results
1 2 3 4 5 6 7 8 9 10
hot keys

Figure 4: VITALAS user interface: result view

the set of features at the cost of a slight accuracy loss. Using this technique the implemented indexing structure allows a repository size of nearly 20 millions images to be queried in real-time using a single server with 20GB of main memory [16].

3.2.3 Concept Detection and Concept Indexing

The concept detection of the VITALAS system is explained in detail in the first part of this paper which describes our submission to the HLFE task. We used all concept detector output coming from our own system that was calculated for the HLFE task. In order to extend the number of concepts and offer a more usable concept search, we added the publicly provided MediaMill[4] concept detector output [24] for those concepts that were not part of the HLFE task.

In our experiments in last year's TRECVID, we compared different techniques for concept pruning [7]. Although our retrieval system changed considerably since then, e.g., it now allows to directly index concepts with corresponding scores, it is still necessary to prune the dense shot concept matrix in order to meet the efficiency constraints of an interactive system. Internal evaluations showed that it is better to keep the concept scores of the top rated shots per concept rather than the top rated concepts per shot. Hence, we indexed only the concept detector scores of the top 5000 shots for each concept.

The concepts were indexed by the PF/Tijah retrieval system [12], which is also used for the concept retrieval.

3.3 User Test Setup

We recruited a total of 10 users to participate in our interactive experiments: 4 users are professional archivists employed in institutes hosting large archives of public broadcasting and 6 are non-professional users. None of the users have been involved in the design or implementation of the VITALAS system; therefore, in order to gain some familiarity with the system interface and supported functionalities, all users completed a training session prior to their main search sessions. Each user was then required to complete 12 of the 24 TRECVID 2009 topics, assigned to them based on a latin squares arrangement as illustrated in Figure 5. The order of the topics was not randomized so that the learning effect across user groups could be observed.

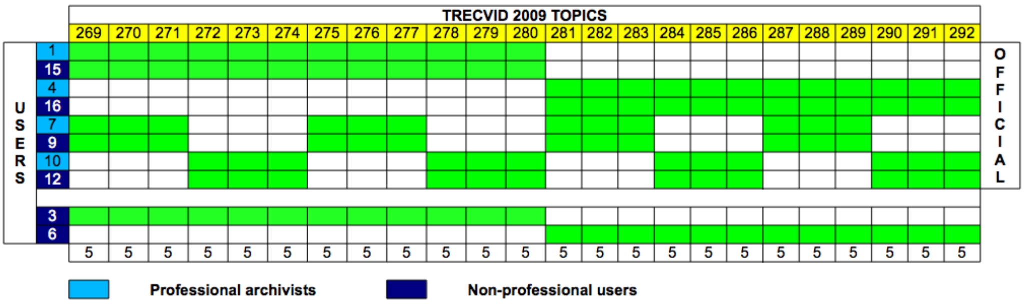
Each user could spend a maximum of 10 minutes on each topic before proceeding to the next. Users were instructed to save those shots that they considered to be relevant to the topic in question. However, the instructions did not emphasize that they should find as many relevant shots as possible, which led all users to only save few shots per topic (about 9 on average), indicating that they possibly focused on those shots they considered to be highly relevant.

Users were asked to fill in a questionnaire consisting of three parts: (i) an *entry questionnaire* provided prior to the training session for collecting background information on the individual and their search experience, (ii) a *search questionnaire* provided after each topic (including the topic in the training session) which asked users to provide their assessment of the topic, the system's performance for that topic, and their perception of their search performance, and (iii) an *exit questionnaire* provided at the end of the search session which asked for an overall evaluation of the VITALAS system and the functionalities it offers. The questionnaires were based on those used by K-Space in TRECVID 2008 [8]. In addition, all user interactions were logged by the system, including the submitted queries, the shots viewed, and the shots saved (i.e., added to the lightbox).

3.4 Results

We submitted four runs, two by professional archivists and two by non-professionals. Each run combined the results of two users that worked on complementary sets of topics (see Figure 5). The analysis presented in this section also includes the results of an additional run by two further novice users. Our analysis is based on the collected search logs. We had foreseen to accompany the quantitative search log analysis with a qualitative analysis based on the questionnaires, but the latter will be included in the final version of

[4] http://www.science.uva.nl/research/mediamill/


TRECVID 2009 TOPICS
USERS
OFFICIAL
Professional archivists
Non-professional users

	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292
1	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩												
15	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩												
4													🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
16													🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
7	🟩	🟩	🟩				🟩	🟩	🟩				🟩	🟩	🟩				🟩	🟩	🟩			
9	🟩	🟩	🟩				🟩	🟩	🟩				🟩	🟩	🟩				🟩	🟩	🟩			
10				🟩	🟩	🟩				🟩	🟩	🟩				🟩	🟩	🟩				🟩	🟩	🟩
12				🟩	🟩	🟩				🟩	🟩	🟩				🟩	🟩	🟩				🟩	🟩	🟩
3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩												
6													🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

Figure 5: Topic assignment based on a latin squares arrangement

the paper.

We first studied the general search behavior of our users, in particular, which search types they used and how often. Table 2 presents the results of this analysis. A *text* type search refers to a query that contains a textual component, irrespective of whether other modalities are also employed. The same applies for the *concept* and *similarity* type searches. The *fused* search type denotes any search that employs two or more modalities. An *entry* search refers to the type of the first query issued by a user for a topic. The table shows the number of queries issued over all topics averaged over all users and over users in each of the user groups. The results indicate that both user groups submitted about the same number of queries. Also the distribution of search types stays similar among the user groups, and a more detailed analysis revealed that the distribution remains roughly similar even among individual users. Furthermore, entry searches were most often keyword queries; in rare cases concepts were also used. The system prototype did not allow the upload of example images and therefore, it was not possible to start a search task with a similarity query.

Table 2: User querying behavior

search	all users		archivists		novices	
type	all	entry	all	entry	all	entry
text	57.5	10.1	60.25	9.5	55.66	10.5
concept	19.5	2.5	22.75	3.25	17.33	2
similarity	13.4	0	13	0	13.66	0
fused	14.5	1.4	14.5	1.75	14.5	1.16
all	81.2	12	85.5	12	78.33	12

A second analysis examined the result investigation behavior of the users. To clarify some terminology, an *add* action refers to a user saving (i.e., adding to the lightbox) a shot he considers to be (potentially) relevant; this is different from shots being judged as *relevant* by the assessors. Furthermore, we refer to all clicks that open a zoom view or a detailed view as *zoom* actions. We examined how often professional and novice users use the zoom action to check whether a thumbnail keyframe is relevant and how far down they investigate the returned ranked list. Our results indicate that both user groups perform almost the same number of zoom actions, professional users though investigate shots deeper in the ranked retrieved list. The median rank of their zoomed and added items is twice as high compared to novice users. On average, the total number of zoom actions of users (irrespective of the user group they belong to) is twice as high as the number of add actions, which indicates that the initial result overview showing thumbnail keyframes is often not enough to judge a shot. We also examined whether users who had been assigned the same topic found the same or different shots. Although each topic was assigned to a total of five users, the proportion of common shots found by more than one user within all added shots for a topic is only 17%. Hence, by far most of the added shots for a given topic are unique among our users. This could be due to the fact that our users added on average only 9 shots per topic to the task search results. It could also indicate that the exhaustiveness of a 10-minute user search session with our system remains low.

We also examined the search effectiveness of our users and their relevance agreement with the TRECVID assessors. It is not clear whether we can indeed talk about relevance agreement in this case, since our users marked potentially relevant shots, while the TRECVID assessors thoroughly checked the relevance of submitted shots. Our results show a low assessor agreement with the results of our users. About 50% of the judged added shots are marked by the assessors as irrelevant. It should also be pointed out that only 50% of the shots added by our users were in fact judged by the assessors, the other half remained outside the judging pool. If we also interpret shots that were zoomed by a user but not added to the lightbox as irrelevant from the user perspective, we find again almost the same disagreement. About 40% of the judged zoomed but not added shots are marked as relevant by the assessors. We further investigated how many of the keyframe thumbnails that were retrieved and displayed to the user in response to any of his/her searches but not added by the user to the results of the search task are relevant according to the assessors. When looking only at the judged shots belonging to this set, we found that 33% of the displayed but not added shots are relevant. Hence, the users missed many shots, by looking only at the keyframe thumbnails. It is important to distinguish here between the two user groups. Professional users added twice as many shots to the results of any search topic than non-professionals, leading to a considerably better recall, but similar precision in both groups.

Finally, we studied the effectiveness of the different search types from system and user perspectives. The results are shown in Table 3 which displays the average number of relevant retrieved shots, added shots, and relevant added shots per query of the specified search type. In this case, we consider the *pure* search types that refer to queries containing a

Table 3: Search type effectiveness

search type	rel. retrieved	added	rel. added
text only	2.76	0.98	0.14
concept only	13.47	2.69	0.87
similarity only	4.32	1.07	0.22
fused	7.88	1.79	0.70

single modality and the *fused* search type that contains two or more modalities. We observed that, from a system perspective, similarity searches retrieve twice as many relevant shots than keyword searches, fused searches three times as many, and concept searches retrieve even up to five times as many relevant shots. From a user perspective, the effectiveness of the different search types stays in the same order, but the differences are smaller. Still, the concept search results in 2.5 as many add actions than a keyword search. Similarity search and keyword search are similarly effective with roughly the same number of add actions.

4. CONCLUSIONS

Overall, the HLFE system achieved good results, both in terms of absolute retrieval effectiveness and with respect to the submissions of other TRECVID-2009 participants. The main conclusions can be summarized as follows: (i) The use of multiple feature modalities definitely improves the concept retrieval effectiveness. In the submitted runs, the use of audio features led to significant improvement for certain features. (ii) Concept score fusion achieves good results with reduced number low-level feature vector dimensions. The combination of concept score fusion and direct use of low-level features did not improve the results significantly, indicating that concept score fusion did not lead to significant loss of information. (iii) The W-LOCAL feature does not improve effectiveness with a few exceptions that motivate us to study in more detail this approach in the future.

Our Interactive Search task experiments examined the use and effectiveness of the different search functionalities of the VITALAS integrated system from a user viewpoint and also studied the search behavior of two different types of users: professional archivists and non-professional users. Our (preliminary) analysis indicates that both user groups submit about the same total number of queries and that the distribution of query types stays also similar among them. However, professional users added twice as many shots to the results compared to non-professionals, leading to a considerably better recall, although a similar precision was achieved by both groups. Professional users also investigated shots deeper in the ranked retrieved list. Since all users saved only few shots per topic, a low number of common shots were found by users assigned to the same topics. Our results also indicate a low agreement between the TRECVID assessors and our users, in addition to the fact that many of shots considered by our users as relevant were not judged, a fact that also makes it more difficult for us to perform a comprehensive analysis and reach reliable conclusions. In terms of the effectiveness of the different search modalities, similarity searches retrieve on average twice as many relevant shots as keyword searches, fused searches three times as many, while concept searches retrieve even up to five times as many relevant shots, indicating the benefits of the use of robust concept detectors in multimodal video retrieval.

ACKNOWLEDGMENTS

This work was supported by the EU-funded VITALAS project (FP6-045389). Christos Diou is supported by the Greek State Scholarships Foundation (`http://www.iky.gr`).

5. REFERENCES

[1] VITALAS, Integrated Project funded by the IST 6th Framework Programme of the European Commission, FP6-045389, visit `http://vitalas.ercim.org` for more information.

[2] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *Proc. of the 30th European Conference on IR Research*, pages 187–198, 2008.

[3] N. Boujemaa, J. Fauqueur, M. Ferecatu, F. Fleuret, V. Gouet, B. Le Saux, and H. Sahbi. Ikona: Interactive generic and specific image retrieval. In *Proceedings of the International Workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR 2001)*, 2001.

[4] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[5] C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines.*

[6] C. Diou, N. Batalas, and A. Delopoulos. Indexing and browsing of color images: Design considerations. In *Advances in Semantic Media Adaptation and Personalization*, volume 93 of *Studies in Computational Intelligence*, pages 329–346. Springer, 2008.

[7] C. Diou, C. Papachristou, P. Panagiotopoulos, A. Delopoulos, G. Stephanopoulos, N. Dimitriou, H. Rode, A. P. de Vries, T. Tsikrika, and R. Aly. Vitalas at trecvid-2008. In *Proc. NIST TRECVID-2008*, 2008.

[8] P. W. et al. K-space at trecvid 2008. In *Proceedings of the TRECVID Workshop*, 2008.

[9] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[10] M. Gervautz and W. Purgathofer. A simple method for color quantization: Octree quantization. In *New Trends in Computer Graphics*. Springer Verlag, Berlin, 1988.

[11] J.-M. Geusebroek, R. van den Boomgaard, A. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, December 2001.

[12] D. Hiemstra, H. Rode, R. van Os, and J. Flokstra. PFTijah: text search in an XML database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR 2006)*, pages 12–17, 2006.

[13] h.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2003.

[14] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using

automatic speech recognition. In B. Falcidieno, M. Spagnuolo, Y. S. Avrithis, I. Kompatsiaris, and P. Buitelaar, editors, *SAMT*, volume 4816 of *Lecture Notes in Computer Science*. Springer, 2007.

[15] B. Huurnink and M. de Rijke. Exploiting redundancy in cross-channel video retrieval. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR 2007)*, pages 177–186, 2007.

[16] A. Joly and O. Buisson. A posteriori multi-probe locality sensitive hashing. In *Proceedings of the 16th ACM international conference on Multimedia (MM 2008)*, pages 209–218, 2008.

[17] L. Kennedy and A. Hauptmann. Lscom lexicon definitions and annotations version 1.0. DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report, 217-2006-3, March 2006.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[19] B. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.

[20] A. Natsev, W. Jiang, M. Merler, J. R. Smith, J. Tešić, L. Xie, and R. Yan. IBM Research TRECVID-2008 Video Retrieval System. In *Proc. of TRECVID 2008*, 2008.

[21] C. Petersohn. Fraunhofer hhi at trecvid 2004: Shot boundary detection system. In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID, 2004.

[22] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

[23] C. G. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1678–1689, 2006.

[24] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, and J. R. R. e. a. Uijlings. The MediaMill TRECVID 2008 semantic video search engine. In *Proceedings of the TRECVID Workshop*, 2008.

[25] 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.

[26] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders. Robust scene categorization by learning image statistics in context. In *International Workshop on Semantic Learning Applications in Multimedia*, page 105, 2006.

[27] V. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1989.

[28] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages 511–518. IEEE Computer Society, 2001.

[29] L. Xie, R. Yan, and J. Yang. Multi-concept learning with large-scale multimedia lexicons. In *Proc. ACM Intl. Conf. on Image Processing (ICIP)*, pages 2148–2151, Oct. 2008.

[30] E. Yilmaz and J. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111. ACM, 2006.

[31] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.