The MediaMill TRECVID 2009 Semantic Video Search Engine

C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij, B. Huurnink, J.R.R. Uijlings, M. van Liempt,
M. Bugalho[†], I. Trancoso[†], F. Yan[‡], M.A. Tahir[‡], K. Mikolajczyk[‡], J. Kittler[‡],
M. de Rijke, J.M. Geusebroek, Th. Gevers, M. Worring, D.C. Koelma, A.W.M. Smeulders

ISLA, University of Amsterdam
Amsterdam, The Netherlands

[†]INESC-ID
Lisboa, Portugal

[‡]CVSSP, University of Surrey
Guildford, Surrey, UK

http://www.mediamill.nl

Abstract

In this paper we describe our TRECVID 2009 video retrieval experiments. The MediaMill team participated in three tasks: concept detection, automatic search, and interactive search. The starting point for the MediaMill concept detection approach is our top-performing bag-of-words system of last year, which uses multiple color descriptors, codebooks with soft-assignment, and kernel-based supervised learning. We improve upon this baseline system by exploring two novel research directions. Firstly, we study a multimodal extension by including 20 audio concepts and fusion using two novel multi-kernel supervised learning methods. Secondly, with the help of recently proposed algorithmic refinements of bag-of-word representations, a GPU implementation, and compute clusters, we scale-up the amount of visual information analyzed by an order of magnitude, to a total of 1,000,000 i-frames. Our experiments evaluate the merit of these new components, ultimately leading to 64 robust concept detectors for video retrieval. For retrieval, a robust but limited set of concept detectors justifies the need to rely on as many auxiliary information channels as possible. For automatic search we therefore explore how we can learn to rank various information channels simultaneously to maximize video search results for a given topic. To further improve the video retrieval results, our interactive search experiments investigate the roles of visualizing preview results for a certain browse-dimension and relevance feedback mechanisms that learn to solve complex search topics by analysis from user browsing behavior. The 2009 edition of the TRECVID benchmark has again been a fruitful participation for the MediaMill team, resulting in the top ranking for both concept detection and interactive search. Again a lot has been learned during this year's TRECVID campaign; we highlight the most important lessons at the end of this paper.

1 Introduction

Robust video retrieval is highly relevant in a world that is adapting swiftly to visual communication. Online services like YouTube and Vimeo show that video is no longer the domain of broadcast television only. Video has become the medium of choice for many people communicating via the Internet. Most commercial video search engines provide access to video based on text, as this is still the easiest way for a user to describe an information need. The indices of these search engines are based on the filename, surrounding text, social tagging, closed captions, or a speech transcript. This results in disappointing retrieval performance when the visual content is not mentioned, or properly reflected in the associated text. In addition, when the videos originate from non-English speaking countries, such as China, or the Netherlands, querying the content becomes much harder as robust automatic speech recognition results and their accurate machine translations are difficult to achieve.

To cater for robust video retrieval, the promising solutions from literature are mostly concept-based [34], where detectors are related to objects, like an *airplane flying*, scenes, like a *classroom*, and people, like *female human face closeup*. Any one of those brings an understanding of the current content. The elements in such a lexicon of concept detectors offer users a semantic entry to video by allowing them to query on presence or absence of visual content elements. Last year we presented the *MediaMill 2008* semantic video search engine [32], which aimed for more robustness of concept detectors in the lexicon rather than extending the number of detectors. Our TRECVID 2009 experiments continue this emphasis on robustness for a relatively small set of concept detectors. A robust but limited set of concept detectors justifies the need to rely on as many multimedia information channels as possible for retrieval. To that end, we explore how we can learn to rank various information channels simultaneously to maximize video search results for a given topic. To improve the retrieval results further, we extend our interactive browsers by supplementing them with visualizations for swift inspection, and a relevance feedback mechanism based on passive sampling of user browsing behavior. Taken together, the *MediaMill 2009* semantic video search engine provides users with robust semantic access to video archives.

The remainder of the paper is organized as follows. We

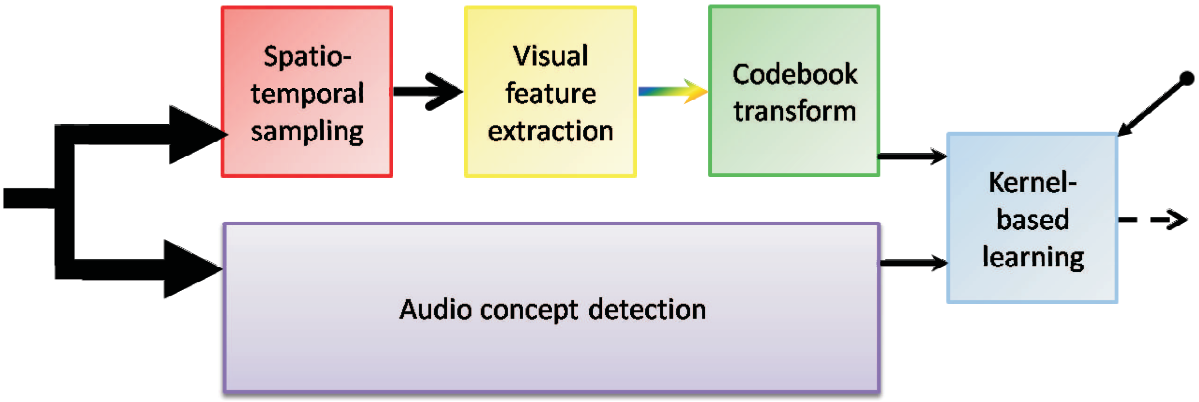

Figure 2: MediaMill TRECVID 2009 concept detection scheme, using the conventions of Figure 1. The scheme serves as the blueprint for the organization of Section 2.

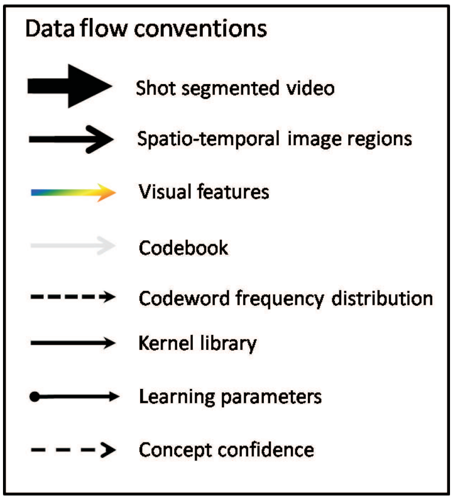

Figure 1: Data flow conventions as used in Section 2. Different arrows indicate difference in data flows.

first define our semantic concept detection scheme in Section 2. Then we highlight our video retrieval framework for automatic search in Section 3. We present the browser innovations of our semantic video search engine in Section 4. We wrap up in Section 5, where we highlight the most important lessons learned.

2 Detecting Concepts in Video

We perceive concept detection in video as a combined multimedia analysis and machine learning problem. Given an n-dimensional multimedia feature vector x_i, part of a shot i [26], the aim is to obtain a measure, which indicates whether semantic concept ω_j is present in shot i. We may choose from various audiovisual feature extraction methods to obtain x_i, and from a variety of supervised machine learning approaches to learn the relation between ω_j and x_i. The supervised machine learning process is composed of two phases: training and testing. In the first phase, the optimal configuration of features is learned from the training data. In the second phase, the classifier assigns a probability $p(\omega_j|x_i)$ to each input feature vector for each semantic concept.

Our TRECVID 2009 concept detection approach builds on previous editions of the MediaMill semantic video search engine [32, 36], which draws inspiration from the bag-of-words approach propagated by Schmid and her associates [19, 24, 51], as well as recent advances in keypoint-based color features [44] and codebook representations [45,47]. We improve upon this baseline system by exploring two novel research directions. Firstly, we study a multimodal extension by inclusding 20 audio concepts [3, 28, 40] and fusion using two novel multi-kernel supervised learning methods [38, 49]. Secondly, with the help of recently proposed algorithmic refinements of the bag-of-words approach [42], a GPU implementation [43], and compute clusters, we scale-up the amount of visual information analyzed by an order of magnitude, to a total of 1,000,000 i-frames. We detail our generic concept detection scheme by presenting a component-wise decomposition. The components exploit a common architecture, with a standardized input-output model, to allow for semantic integration. The graphical conventions to describe the system architecture are indicated in Figure 1. Based on these conventions we follow the video data as it flows through the computational process, as summarized in the general scheme of our TRECVID 2009 concept detection approach in Figure 2, and detailed per component next.

2.1 Spatio-Temporal Sampling

The visual appearance of a semantic concept in video has a strong dependency on the spatio-temporal viewpoint under which it is recorded. Salient point methods [41] introduce robustness against viewpoint changes by selecting points, which can be recovered under different perspectives. Another solution is to simply use many points, which is achieved by dense sampling. Appearance variations caused by temporal effects are addressed by analyzing video beyond

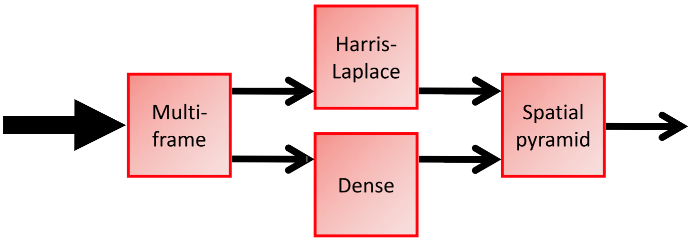

Figure 3: General scheme for spatio-temporal sampling of image regions, including temporal multi-frame selection, Harris-Laplace and dense point selection, and a spatial pyramid. Detail of Figure 2, using the conventions of Figure 1.

the key frame level. By taking more frames into account during analysis, it becomes possible to recognize concepts that are visible during the shot, but not necessarily in a single key frame. We summarize our spatio-temporal sampling approach in Figure 3.

Temporal multi-frame selection In [32, 35] we demonstrated that a concept detection method that considers more video content obtains higher performance over key frame-based methods. We attribute this to the fact that the content of a shot changes due to object motion, camera motion, and imperfect shot segmentation results. Therefore, we employ a multi-frame sampling strategy. To be precise, we sample up to 10 additional i-frames distributed around the (middle) key frame of each shot.

Harris-Laplace point detector In order to determine salient points, Harris-Laplace relies on a Harris corner detector. By applying it on multiple scales, it is possible to select the characteristic scale of a local corner using the Laplacian operator [41]. Hence, for each corner, the Harris-Laplace detector selects a scale-invariant point if the local image structure under a Laplacian operator has a stable maximum.

Dense point detector For concepts with many homogenous areas, like scenes, corners are often rare. Hence, for these concepts relying on a Harris-Laplace detector can be suboptimal. To counter the shortcoming of Harris-Laplace, random and dense sampling strategies have been proposed [10,17]. We employ dense sampling, which samples an image grid in a uniform fashion using a fixed pixel interval between regions. In our experiments we use an interval distance of 6 pixels and sample at multiple scales.

Spatial pyramid weighting Both Harris-Laplace and dense sampling give an equal weight to all keypoints, irrespective of their spatial location in the image frame. In order to overcome this limitation, Lazebnik *et al.* [19] suggest to repeatedly sample fixed subregions of an image, *e.g.*,1x1, 2x2, 4x4, *etc.*, and to aggregate the different resolutions into a so called spatial pyramid, which allows for region-specific weighting. Since every region is an image in itself, the spatial pyramid can be used in combination with both the Harris-Laplace point detector and dense point sampling. Similar to [24,32] we use a spatial pyramid of 1x1, 2x2, and 1x3 regions in our experiments.

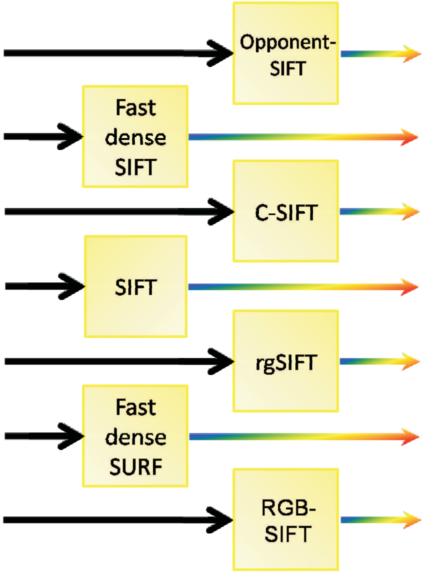

Figure 4: General scheme of the visual feature extraction methods used in our TRECVID 2009 experiments.

2.2 Visual Feature Extraction

In the previous section, we addressed the dependency of the visual appearance of semantic concepts in a video on the spatio-temporal viewpoint under which they are recorded. However, the lighting conditions during filming also play an important role. Burghouts and Geusebroek [4] analyzed the properties of color features under classes of illumination and viewing changes, such as viewpoint changes, light intensity changes, light direction changes, and light color changes. Van de Sande *et al.* [44] analyzed the properties of color features under classes of illumination changes within the diagonal model of illumination change, and specifically for data sets as considered within TRECVID. To speed up the feature extraction process, we adopt the algorithmic refinements of dense sampled bag-of-words proposed by Uijlings *et al.* [42]. We present an overview of the visual features used in Figure 4.

SIFT The SIFT feature proposed by Lowe [23] describes the local shape of a region using edge orientation histograms. The gradient of an image is shift-invariant: taking the derivative cancels out offsets [44]. Under light intensity changes, *i.e.*,a scaling of the intensity channel, the gradient direction and the relative gradient magnitude remain the same. Because the SIFT feature is normalized, the gradient magnitude changes have no effect on the final feature. To compute SIFT features, we use the version described by Lowe [23].

OpponentSIFT OpponentSIFT describes all the channels in the opponent color space using SIFT features. The information in the O_3 channel is equal to the intensity information, while the other channels describe the color information in the image. The feature normalization, as effective in SIFT, cancels out any local changes in light intensity.

C-SIFT In the opponent color space, the O_1 and O_2 channels still contain some intensity information. To add invariance to shadow and shading effects, we have proposed the C-invariant [12] which eliminates the remaining intensity information from these channels. The C-SIFT feature uses the C invariant, which can be intuitively seen as the gradient (or derivative) for the normalized opponent color space O_1/I and O_2/I. The I intensity channel remains unchanged. C-SIFT is known to be scale-invariant with respect to light intensity.

rgSIFT For rgSIFT, features are added for the r and g chromaticity components of the normalized RGB color model, which is already scale-invariant [44]. In addition to the r and g channel, this feature also includes intensity. However, the color part of the feature is not invariant to changes in illumination color.

RGB-SIFT For the RGB-SIFT, the SIFT feature is computed for each RGB channel independently. Due to the normalizations performed within SIFT, it is equal to transformed color SIFT [44]. The feature is scale-invariant, shift-invariant, and invariant to light color changes and shift.

Fast Dense SIFT/SURF We speed up the calculation of densely sampled SIFT [23] and SURF [2] in two ways, described in detail in [42]. First of all we observe that both descriptors are spatial. Both are constructed of 4×4 subregions which are in turn described by the summation of pixel-wise responses over an area. For SIFT the pixel-wise responses are oriented gradient responses, for SURF these are Haar-wavelet responses. By reusing subregions in descriptor creation, we obtain a speed-improvement of a factor 16. To enable this for SIFT we have to make a slight adjustment by removing the Gaussian Weighting around the origin. Experiments showed that this does not influence the final classification accuracy. For the second speed improvement we devised a fast way to do summations of pixel-wise responses over a subregion. Instead of a nested for-loop, we do the summations over a subregion using two matrix multiplications [42]. The use of existing, highly optimized matrix multiplication libraries gives us a speed-improvement of a factor 2 over a naive C++ implementation.

We compute the SIFT [23] and ColorSIFT [44] features around salient points obtained from the Harris-Laplace detector and dense sampling. In addition, we compute SURF [2] features around fast dense sampled points [42]. For all visual features we employ a spatial pyramid of 1x1, 2x2, and 1x3 regions.

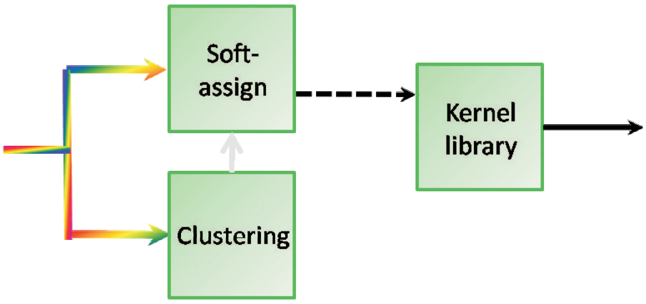

Figure 5: General scheme for transforming visual features into a codebook, where we distinguish between codebook construction using clustering and soft codeword assignment. We combine various codeword frequency distributions into a kernel library.

2.3 Codebook Transform

To avoid using all visual features in an image, while incorporating translation invariance and a robustness to noise, we follow the well known codebook approach, see *e.g.*, [17, 20, 30, 45, 47]. First, we assign visual features to discrete codewords predefined in a codebook. Then, we use the frequency distribution of the codewords as a compact feature vector representing an image frame. By using a vectorized GPU implementation [43], our codebook transform process is an order of magnitude faster for the most expensive feature compared to the standard implementation. Two important variables in the codebook representation are *codebook construction* and *codeword assignment*. Based on last year's experiments we employ codebook construction using k-means clustering in combination with soft codeword assignment and a maximum of 4,096 codewords, following the scheme in Figure 5.

Soft-assignment Given a codebook of codewords, obtained from clustering, the traditional codebook approach describes each feature by the single best representative codeword in the codebook, *i.e.*,hard-assignment. However, in a recent paper [47], we show that the traditional codebook approach may be improved by using soft-assignment through kernel codebooks. A kernel codebook uses a kernel function to smooth the hard-assignment of image features to codewords. Out of the various forms of kernel-codebooks, we selected *codeword uncertainty* based on its empirical performance [47].

Kernel library Each of the possible sampling methods from Section 2.1 coupled with each visual feature extraction method from Section 2.2, a clustering method, and an assignment approach results in a separate visual codebook. An example is a codebook based on dense sampling of rgSIFT features in combination with k-means clustering and soft-assignment. We collect all possible codebook combinations in a (visual) kernel library. By using a GPU implementation [43], this kernel library can be computed efficiently. Naturally, the codebooks can be combined using various configurations. Depending on the kernel-based learning scheme used, we simply employ equal weights in our experiments or learn the optimal weight using cross-validation.

2.4 Audio Concept Detection

The work on extracting audio-related concepts from the audiovisual signal was done by INESC-ID, emphasizing in particular audio segmentation and audio event detection methods [3, 28, 40].

Audio segmentation The audio segmentation module includes six separate components: one for *Acoustic Change Detection*, four components for classification (*Speech/Non-speech*, *Background*, *Gender and Speaker Identification*) and one for *Speaker Clustering*. These components are mostly model-based, making extensive use of feed-forward fully connected Multi-Layer Perceptrons trained with the back-propagation algorithm. All the classifiers share a similar architecture: a Multi-Layer Perceptron with 9 input context frames of 26 coefficients (12th order Perceptual Linear Prediction plus energy and deltas), two hidden layers with 250 sigmoidal units each and the appropriate number of softmax output units (one for each class), which can be viewed as giving a probabilistic estimate of the input frame belonging to that class. The Speaker Clustering component tries to group all segments uttered by the same speaker. The first frames of a new segment are compared with all the same gender clusters found so far. A new speech segment is merged with the cluster with the lowest distance, provided it falls below a predefined threshold. The distance measure for merging clusters is a modified version of the Bayesian Information Criterion. The 4 audio concepts *female-voice*, *child-voice*, *music*, and *dialogue* could potentially be used for detecting the TRECVID video concepts *Infant*, *Classroom*, *Female-close-up*, *Two-People*, *People-Dancing*, *Person-Playing-Music-Instrument*, and *Singing*.

Audio event detection The audio event detection module currently includes more than 70 one-against-all semantic concept classifiers. For each audio event, world and concept examples were chosen from a corpus of sound effects, in order to train models, using a radial basis function support vector machine classifier. Audio features were retrieved using 500 ms window, with 50% overlap: mel-frequency cepstral coefficients and derivatives, zero crossing rate, brightness, and bandwidth. The latter are, respectively, the first and second order statistics of the spectrogram, and they roughly measure the timbre quality. The F-measure results on a separate test corpus of isolated sound effects were generally very good (above 0.8), but the results in real life TRECVID data show the degradation that can be expected from the fact that audio events almost never occur separately, being corrupted by music, speech, background noise and/or other audio events. More sophisticated support vector machine detectors have been built, using new features, different window sizes, different ways of incorporating context, and dimensionality reduction techniques. The time constraints of this evaluation campaign, however, motivated the use of the described baseline approach. The list of 16 audio event adopted in TRECVID includes: *Child-laughter*, *Baby-crying*, *Airplane-propeller*, *Airplane-jet*, *Sirens*, *Traffic-noise*, *Car-engine*, *Bus-engine*, *Dog-barking*, *Telephone-digital*, *Telephone-analog*, *Door-open-close*, *Applause*, *Bite-eat*, *Water* and *Wind*.

2.5 Kernel-based Learning

Learning robust concept detectors from multimedia features is typically achieved by kernel-based learning methods. Similar to previous years, we rely predominantly on the support vector machine framework [48] for supervised learning of semantic concepts. Here we use the LIBSVM implementation [7] with probabilistic output [21, 27]. In order to handle imbalance in the number of positive versus negative training examples, we fix the weights of the positive and negative class by estimation from the class priors on training data. While the radial basis kernel function usually performs better than other kernels, it was recently shown by Zhang *et al.* [51] that in a codebook-approach to concept detection the earth movers distance [29] and χ^2 kernel are to be preferred. In general, we obtain good parameter settings for a support vector machine, by using an iterative search on both C and kernel function $K(\cdot)$ on cross validation data [46]. In addition to the support vector machine framework, we also study the suitability of two novel multi-kernel learning methods for concept detection: *Kernel Discriminant Analysis using Spectral Regression* and *Non-Sparse Multiple Kernel Fisher Discriminant Analysis*.

Multi-Kernel: SR-KDA Linear Discriminant Analysis [11], which is one of the most widely used statistical methods, has been proven successful in many classification problems. Recently, Spectral Regression combined with Kernel Discriminant Analysis (SR-KDA) introduced by Cai et al [5] has been successful in many classification tasks such as multi-class face, text and spoken letter recognition. The method combines the spectral graph analysis and regression for an efficient large matrix decomposition in Kernel Discriminant Analysis. It has been demonstrated in [5] that it can achieve an order of magnitude speedup over the eigen-decomposition while producing smaller error rate compared to state-of-the-art classifiers. In [38], we have shown the effectiveness of SR-KDA for large scale concept detection problem. In addition to superior classification results when compared to existing approaches, it can provide an order of magnitude speed-up over support vector machine. The main computationally intensive operation is Cholesky decomposition, which is actually independent of the number of labels. For more details please refer to [38].

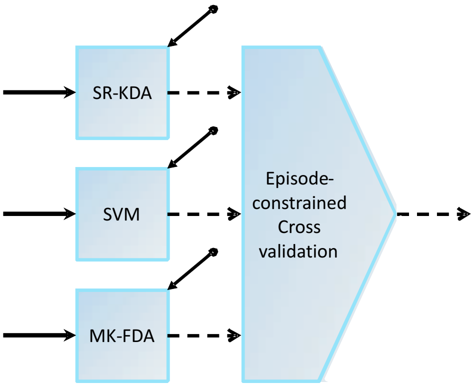

Figure 6: General scheme for kernel-based learning with support vector machines and two novel multi-kernel learning methods, using episode-constrained cross-validation for parameters selection.

Multi-Kernel: MK-FDA Kernel Fisher discriminant analysis has proven to be a very successful classification method in various applications. In many real-world problems, multiple kernels capturing different "views" of the problem are available. In such a situation, one naturally wants to use an "optimal" combination of the kernels. In [50], the authors proposed multiple kernel Fisher discriminant analysis (MK-FDA), where the key idea is to learn the optimal linear combination of kernels by maximizing the ratio of the projected between-class and within-class scatters with respect to the kernel weights. In [50], the kernel weights are regularized with an ℓ_1 norm, which enforces sparsity but may lead to a loss of information. To remedy this, we propose to use an ℓ_2 norm regularization instead. We formulate ℓ_2 MK-FDA as a semi-infinite program, which can be solved efficiently. Experiments show that ℓ_2 regularization tends to produce non-sparse solutions. As a result, less information is lost during the kernel learning process, and the performance is improved over ℓ_1 MK-FDA as well as the uniform weighting scheme. For more details on non-sparse MK-FDA please refer to [49].

Episode-constrained cross-validation From all parameters q we select the combination that yields the best average precision performance, yielding q^*. We measure performance of all parameter combinations and select the combination that yields the best performance. We use a 3-fold cross validation to prevent over-fitting of parameters. Rather than using regular cross-validation for support vector machine parameter optimization, we employ an *episode-constrained* cross-validation method, as this method is known to yield a less biased estimate of classifier performance [46].

The result of the parameter search over q is the improved model $p(\omega_j|x_i, q^*)$, contracted to $p^*(\omega_j|x_i)$, which we use to fuse and to rank concept detection results.

2.6 Submitted Concept Detection Results

We investigated the contribution of each component discussed in Sections 2.1–2.5, emphasizing in particular the role of audio, multi-kernel learning, and scalability by processing 1,000,000 i-frames. In our experimental setup we used the TRECVID 2007 development set as a training set, and the TRECVID 2007 test set as a validation set. The ground truth used for learning and evaluation are a combination of the common annotation effort [1] and the ground truth provided by ICT-CAS [39]. An overview of our submitted concept detection runs is depicted in Figure 7, and detailed next.

Run: Joe The Joe run is our single key frame baseline. It applies the standard sequential forward selection feature selection method on all (visual) kernel libraries computed over key frames only. It obtained a mean infAP of 0.175. This run tends to lag behind our other (multi-frame) runs, especially for dynamic concepts such as *airplane flying*, *people dancing*, *person riding bicycle*, *person playing soccer*, and *person eating*.

Run: William The William run is a cooperation between the University of Amsterdam and the University of Surrey. In this run, each (visual) kernel is trained using SR-KDA with regularization parameter δ [38] which is tuned for each concept using the validation set. Further, instead of using equal weights for each classifier during fusion, weights for individual kernels are learnt for each concept using the classification accuracy i.e. average precision on the validation set. The weighted output from each classifier is then combined using the *SUM* rule [18]. This run has achieved a mean infAP of 0.190. For some concepts (*cityscape*, *people dancing*, *boat/ship*), results are comparable to our top run methods despite the fact that only 1 key frame is processed for every shot in this run while multi-frames per shot are processed in our top runs.

Run: Jack The Jack run is a cooperation between the University of Amsterdam, INESC-ID, and the University of Surrey. In addition to the visual kernels, we also generated an audio kernel using INESC's audio concept detectors. More specifically, the 20 output scores of the 20 audio concept detectors were used as 20 features, and an RBF kernel was build from these features. This audio kernel together with the visual kernels were then used as input to Non-Sparse Multiple Kernel Fisher Discriminant Analysis (MK-FDA) [49], where the optimal kernel weights were learned for each semantic concept. Experiments on the validation set show that by introducing the audio kernel to the kernel set, the mean average precision is improved by 0.01. On the TRECVID 2009 test set this run obtains a mean infAP of 0.193. The concepts that benefit most from the audio kernel are: *person playing musical instrument*, *female human face closeup*, *infant*, *singing*, and *airplane flying*.

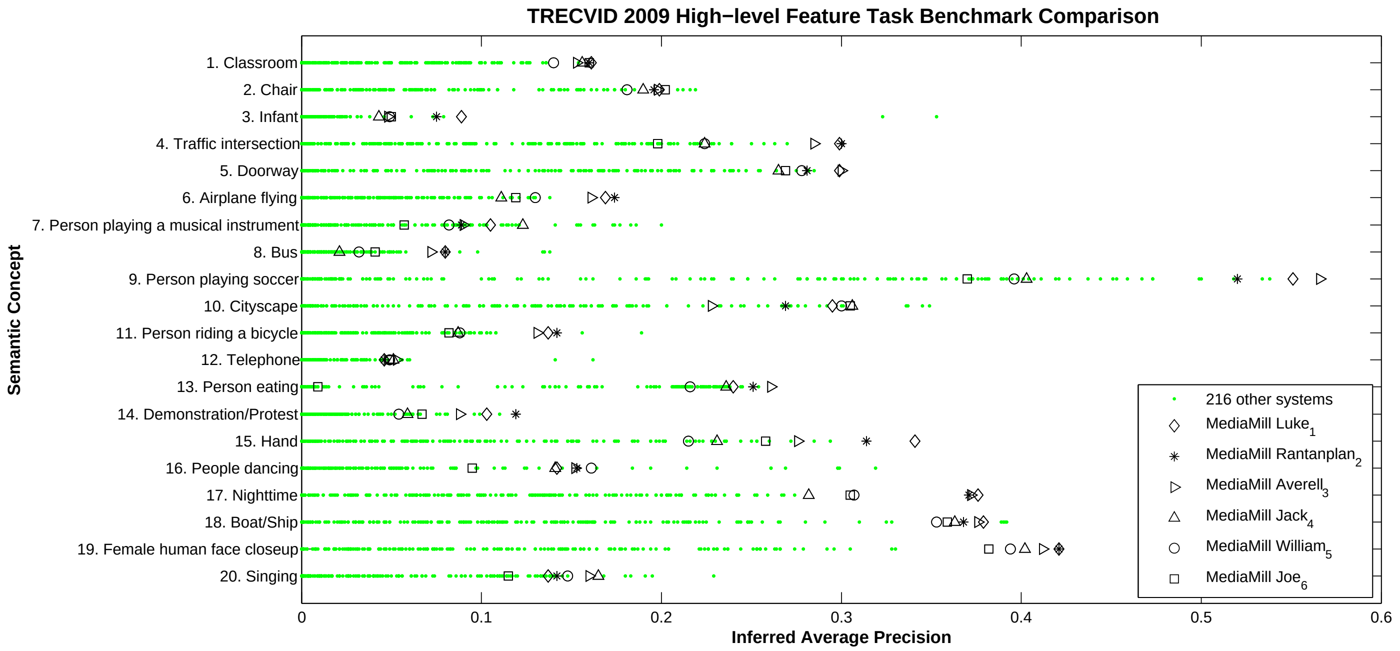

Figure 7: Comparison of MediaMill video concept detection experiments with other concept detection approaches in the TRECVID 2009 High-level Feature Task benchmark.

Run: Averell The Averell run is based on a (visual) kernel library based on SIFT, OpponentSIFT, C-SIFT, and RGB-SIFT only, which have been applied spatio-temporally with up to 5 additional i-frames per shot in combination with a MAX rule combination. This run achieved a mean infAP of 0.219, with the overall highest infAP for 3 concepts: *doorway*, *person playing soccer*, and *person eating*.

Run: Rantanplan The Rantanplan run selects from all the (visual) kernel libraries, all of which have been applied spatio-temporally with up to 10 additional i-frames per shot in combination with AVG and MAX rule combination, the best performer per concept. This run achieved a mean infAP of 0.224, with the overall highest infAP for 4 concepts: *traffic intersection*, *airplane flying*, *demonstration/protest*, and *female human face closeup*.

Run: Luke The Luke run extends upon the Rantanplan run, by applying the standard sequential forward selection feature selection method on all (visual) kernel libraries computed over 1,000,000 i-frames. This run achieved the overall highest mean infAP in the TRECVID2009 benchmark (0.228), with the overall highest infAP for 4 concepts: *classroom*, *nighttime*, *hand*, and *female human face closeup*.

2.7 64 Robust Concept Detectors

Similar to our TRECVID 2008 participation, we again aim for a small but robust lexicon of concept detectors this year. To that end we have employed our Averell run setting on the concept sets of TRECVID 2008 (20 concepts), TRECVID2007 (36 concepts) and an additional black/white detector. All 64 detectors have been donated to the TRECVID community[1] and are included in the 2009 MediaMill semantic video search engine for the retrieval experiments.

3 Automatic Video Retrieval

The MediaMill team continued its effort on automatic search, this year submitting 8 automatic runs. The overall architecture of the search system was based on 3 fundamental search types — transcript-based retrieval, detector-based retrieval, and feature-based retrieval — each of which was submitted individually as a run. In addition we submitted 5 combination runs, consisting of query-dependent and query-independent approaches to video automatic search.

3.1 Baseline Retrieval Approaches

Our baselines correspond to the three information sources of: transcripts, detectors, and low-level features. These are implemented as follows:

Pippin: Transcript-based search Our transcript-based search approach is similar to that of last year, incorporating Dutch automatic speech recognition transcripts and English automatic machine translation transcripts [6]. This year both the University of Twente [13] and LIMSI [9] donated speech recognition transcripts. We evaluated both for retrieval using the 2007 topics, and found that overall retrieval performance could be improved by combining the

[1] Available from: `http://trecvid.nist.gov/trecvid.data.html`

text of both transcripts. This was further confirmed for the 2009 topics with three additional (unsubmitted) runs that we performed using this year's topics. A run using only University of Twente transcripts gained an MAP score of 0.007, a run using only LIMSI transcripts gained an MAP score of 0.009, and a run using combined transcripts gained an MAP score 0.010. We combined the text of both transcripts together with the machine translation for this year's entry, which resulted in a decreased final score of 0.009. At retrieval time, each topic statement was automatically translated into Dutch using the online translation tool `http://translate.google.com`, allowing a search on the machine-translated transcripts with the original (English) topic text, and a search on transcripts from automatic speech recognition using the translated Dutch topic text. The two resulting ranked lists were then combined to form a single list of transcript-based search results. To compensate for the temporal mismatch between the audio and the visual channels, we used our temporal redundancy approach [14]. To summarize this approach, the transcript of each shot is expanded with the transcripts from temporally adjacent shots, where the words of the transcripts are weighted according to their distance from the central shot.

Sam: Detector-based search The detector-based search, using our lexicon of 64 robust concept detectors, consisted of two main steps: 1) concept selection and 2) detector combination. We evaluated a number of concept selection approaches using a benchmark set of query-to-concept mappings, adapted from [15] to the new lexicon. The final concept selection method used for automatic search was to average the score for a concept detector on the provided topic video examples, and select concepts that scored over a threshold. In addition, any detectors with high information content, that were also WordNet synonyms of terms in the topic text, were also selected. As for the combination of multiple selected concepts for a topic, this was done by simply taking the product of the raw selected detector scores for each shot as its retrieval score. No extra normalization or parametrization was done, nor were concepts weighted according to their computed score for the examples. Rather, we used the triangulation of concept detector scores to provide information on the relevance of a shot to a query.

Merry: Feature-based search As we did last year, we treat feature-based search as an on-the-fly concept learning problem, with the provided topic video examples as positive examples, and randomly selected shots from the test collection as pseudo-negative examples. Spatio-temporal sampling of interest regions, visual feature extraction, codebook transform, and kernel-based learning were done as described in Section 2. The resulting model was applied to the shots in the test collection, shots were ranked according to the probabilistic output score of the support vector machine.

3.2 Query-(In)dependent Multimodal Fusion

The final step in our retrieval pipeline is multimodal fusion. Our aim here was to (1) compare query-dependent vs query-independent methods, and (2) investigate the use of the learning to rank framework [22] for video retrieval. In all cases weights and/or models were developed using the TRECVID 2007 and 2008 topics for training. Learning to rank was done according to the SVM-Rank implementation for learning to rank [16].

Gimli: Query-independent fusion Linear combination of the three baseline approaches using weighted combsum fusion.

Legolas: Query-independent learning to rank Learning to rank-based combination of the three baseline approaches.

Aragorn: Query-class based fusion Query-class dependent linear combination of the three baseline approaches using weighted combsum fusion. We utilize the query classes and classification methodology employed by Mei et al. [25].

Gandalf: Predictive reranking Similarly to last year, predict which baseline approach will give the best performance, using various query and result-based features for prediction. Rerank the results of the predicted best baseline with results from the other two baselines.

Frodo: Query-dependent Learning to Rank Learning to rank-based combination of all 6 aforementioned automatic search runs.

3.2.1 Automatic Search Results

Once again this year, the transcript baseline had the lowest overall MAP of all runs with a score of 0.009. At 0.068, detector-based search is the best performing baseline, while feature-based search also does relatively well with a score 0.053. Of the combination approaches, query-dependent learning to rank gives the best retrieval performance of 0.089. Surprisingly, query-independent learning to rank gives the *lowest* performance over all combination strategies. In these experiments, the learning to rank-method is more effective when given both query-dependent and query-independent results as input features.

Figure 8 provides a topic-level summary of the performance of the MediaMill automatic search runs. We see that transcript-based search had consistently low performance, though it did achieve a high AP score relative to other runs for *an airplane or helicopter on the ground, seen from outside*. Feature-based search gave higher performance, doing well for visually distinctive scenes such as *a building entrance* and *printed, typed, or handwritten text, filling more than half of the frame area*. Detector-based search performed best for topics where one or more closely related detectors where available, for instance *something burning*

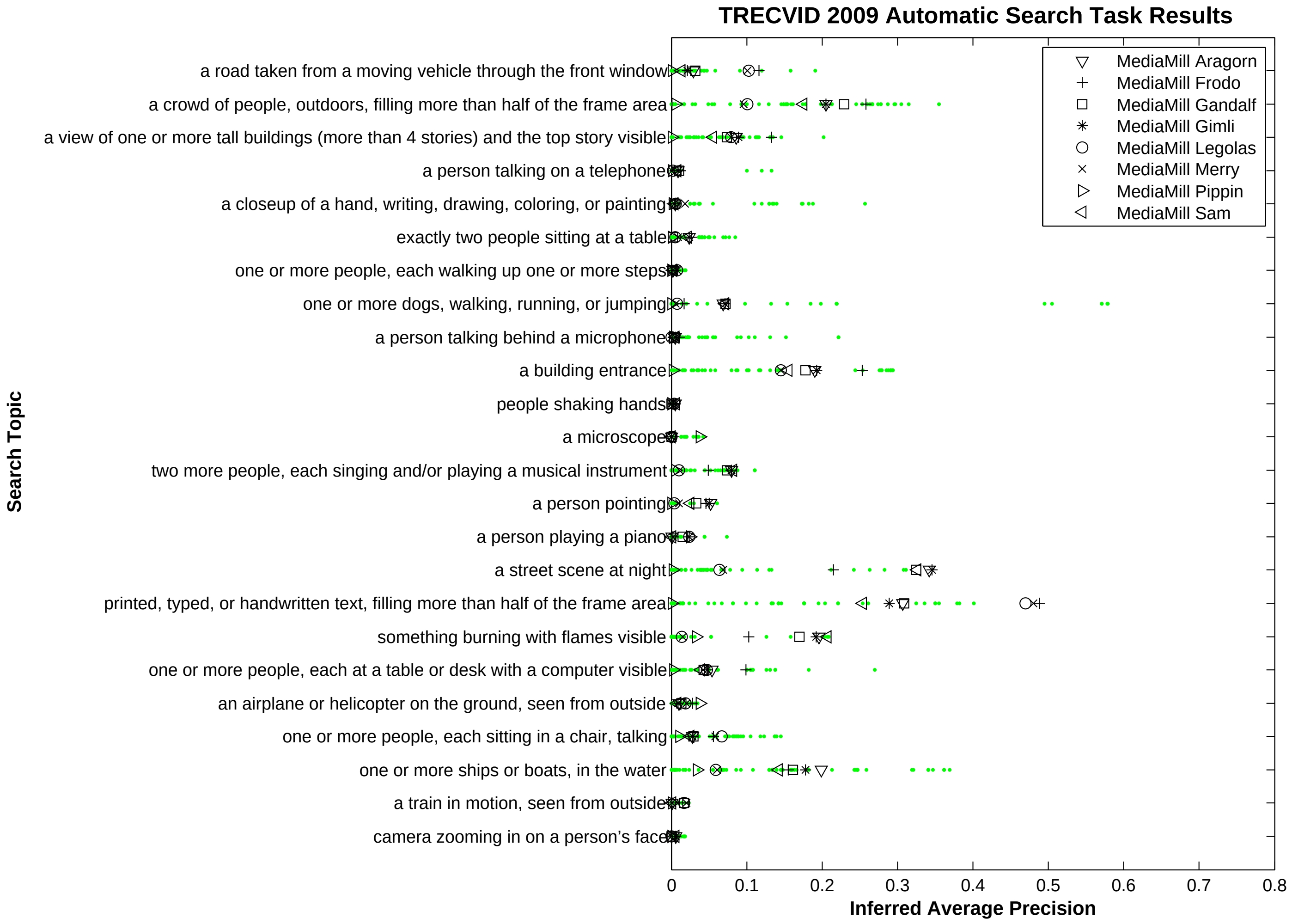

Figure 8: Topic-level comparison of MediaMill automatic video search experiments with other automatic search approaches in the TRECVID 2009 benchmark.

with flames visible, where the *explosion/fire* detector was selected, and *street scene at night* where the *street* and *night* detectors were selected for search. Sometimes results were disappointing: performance for the query for *one or more dogs, walking, running, or jumping*, where the *dog* detectors was selected, was severely degraded by inclusion of scores from the *people walking* detector.

The performance of the query-dependent learning to rank run is 0.089. If we were to select the best performing of the three baselines for each topic, the performance would also be 0.089. This indicates that the fusion approach is capable of performing at least as well as a "best of" approach, at least on an overall level. Performance over individual topics varies, a large boost in performance is obtained for topics where more than one baseline does well, for example for a *a building entrance* AP is increased by 0.098, and for *one or more people, each at a table or desk with a computer visible*, performance more than doubles compared to the highest performing baseline run. Conversely, when a single baseline outperforms the others to a great degree, fusion tends to reduce performance as compared to the best baseline. This is the case for example with the topics *a street scene at night* and *something burning with flames visible.*

4 Interactive Video Retrieval

The performance of interactive video search engines depends on many factors, such as the chosen query method, the used browsing interface with its implied interaction scheme, and the level of expertise of the user. Moreover, when search topics are generic and diverse, it is hard to predict which combination of factors yields optimal performance. Therefore, current video search engines have traditionally offered multiple query methods in an integrated browse environment. This allows the user to choose what is needed. However, while this does offer the user complete control over which strategy to use for which topic, it also allows the user to inadvertently select a sub-optimal strategy.

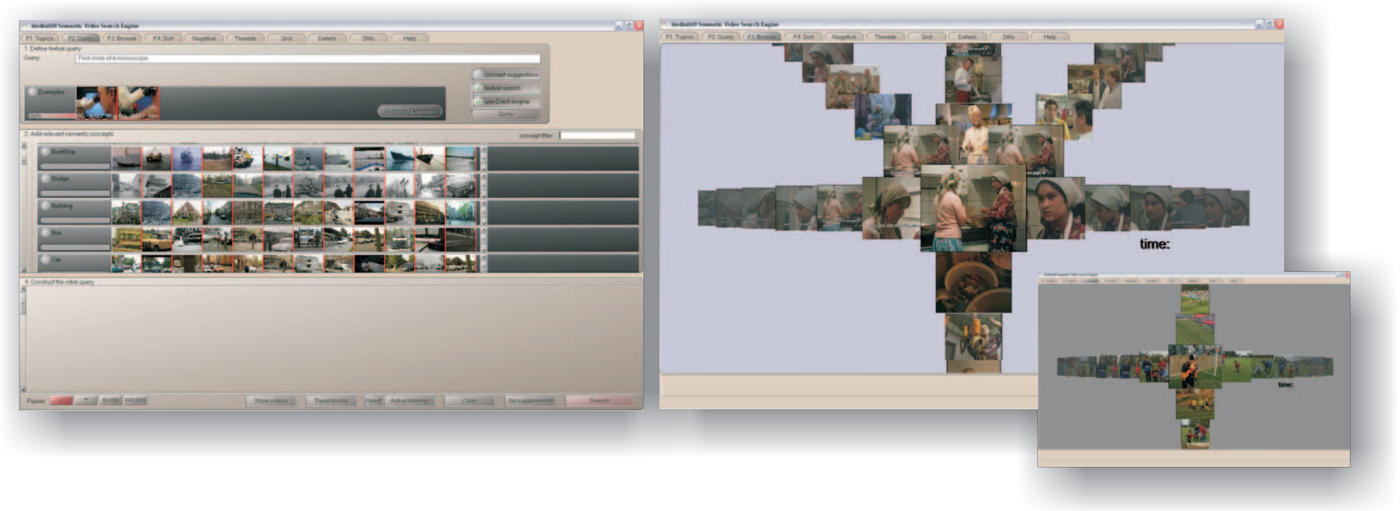

Figure 9: Screenshots of the MediaMill semantic video search engine with its query interface (left), its ForkBrowser [8] (right), and its CrossBrowser [37] (inset).

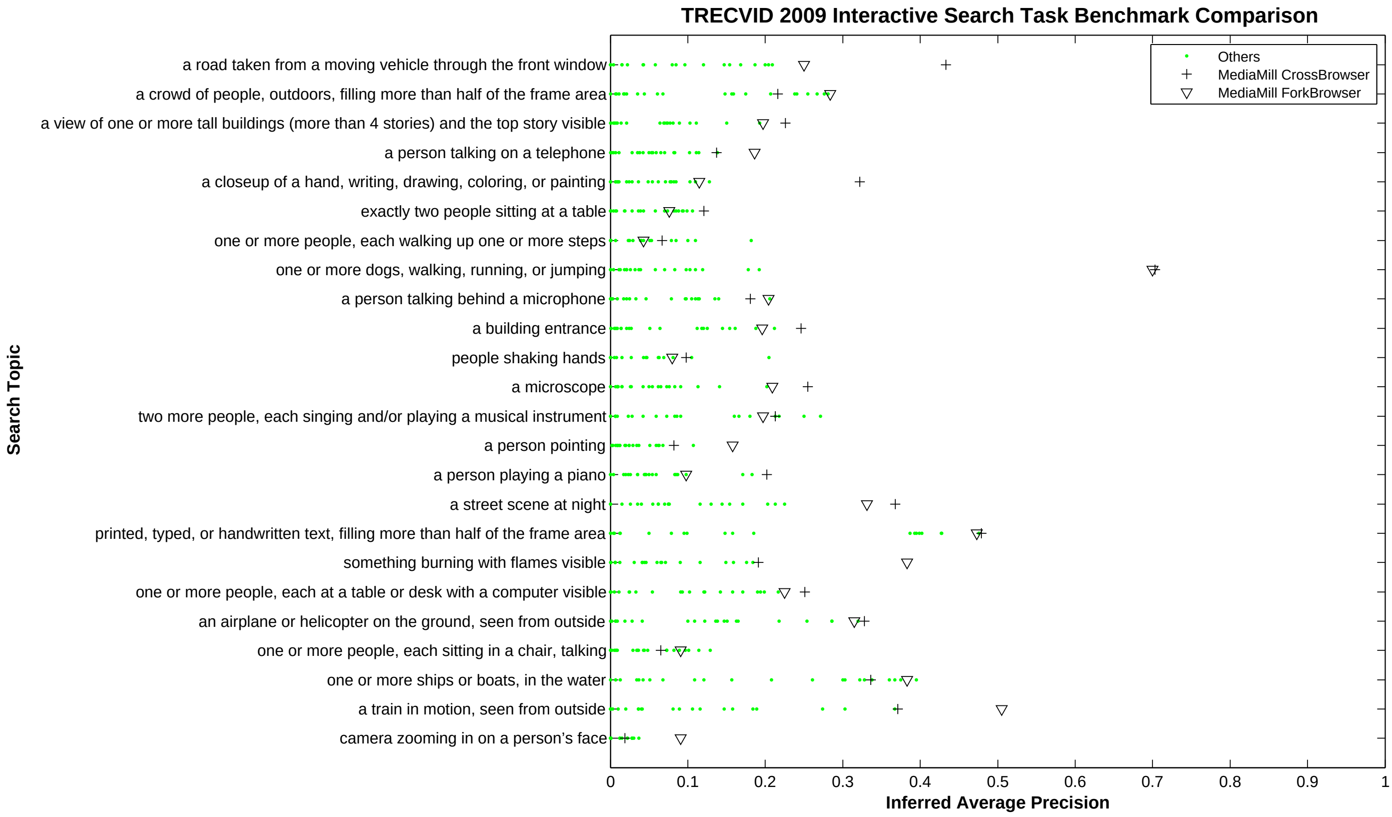

Figure 10: Comparison of MediaMill interactive video search experiments with other interactive video search engines in the TRECVID 2009 benchmark.

4.1 Thread-Based Browsing

The basis for our TRECVID 2009 experiments is the MediaMill Semantic Video Search Engine, see Figure 9. The basic building block behind the browsing interface is the thread; a linked sequence of shots in a specified order, based upon an aspect of their content [8]. These threads span the video archive in several ways. For example, time threads span the temporal similarity between shots, visual threads span the visual similarity between shots, a query thread spans the similarity between a shot and a user-imposed query, and history threads span the navigation path the user follows.

The MediaMill Semantic Video Search Engine allows user to choose between two modes for thread visualization. The first visualization, the CrossBrowser shows the query thread and the time thread in a cross formation. This visualization is most efficient for topics where a single concept query is sufficient for solving a topic [33, 37]. The second visualization, the ForkBrowser, provides the user with two extra diagonal threads, and a history thread. The ForkBrowser is more efficient in handling complex queries where no direct mapping between available concept detectors is possible [8].

4.2 Guiding The User to Results

Our TRECVID Interactive Retrieval experiments focus on helping users to determine the utility of a given retrieval strategy, and on guiding them to a correct set of results. To this end we investigate the benefit of two strategies within the MediaMill Semantic Video Search Engine.

To help users determine the utility of a given retrieval strategy we introduce Active Zooming. This aids users both by helping determine that a subset of visible results is not relevant, and by helping to find a starting point within the selected results. Active Zooming enables the user to quickly and seamlessly visualize a large set of results from a single thread at once. This allows users to make blink-of-an-eye decisions about the contents of a single thread, or, in the case of many relevant results, to quickly select large batches of relevant results at once. The user is then able to either continue browsing the thread, or go back to any other thread.

To help guide users to correct results we introduce a Relevance Feedback strategy based on passive sampling of user browsing behavior in order to guide users to more relevant results. For this, the system continuously monitors user behavior and uses this information on-demand to generate a new set of results. It does so by training a support vector machine model based on positive examples obtained from the user, and negative examples obtained by passive monitoring. By using a pre-computed kernel matrix of inter-shot distances this can be done interactively. The end result is a reranking of the entire collection, which is then available as a thread for visualization.

4.3 Interactive Search Results

We submitted two runs for interactive search. The *Sauron run* was performed by a single expert user. The user was instructed to use the ForkBrowser with Gabor and Wiccest [45] similarity threads. The user was allowed to use Active Zooming and Relevance Feedback techniques on demand. The Saruman run was performed by another single expert user. The user was instructed to use the CrossBrowser together with Active Zooming and Relevance Feedback. We provide a preliminary analysis of the logging data for both runs.

In Figure 10 we show a per-topic overview of interactive video retrieval results. The log-analysis indicates that the users employed a variety of strategies to retrieve results. We highlight a few typical cases. When relevant concept detectors are available for a topic, these are taken as the entry point for search by both users. For example, the users selected the *Hand* detector for the topic *a closeup of a hand, writing, drawing, coloring, or painting.* We found the capability to analyze and view multiple frames from individual shots to be a significant benefit. For example, the results for *one or more dogs...* were largely found by selecting the opening credits of a single television program, in which a dog can be seen running. This was however not apparent in the key frames of these shots. For other topics, such as *train in motion* or *camera zooming in on a face*, we found that showing motion enabled the users to correctly answer the topics. One user further increased the result for the latter topic by a creative use of Active Zooming: the zoom-in motion was visually easily distinguishable which allowed the user to select relevant shots rapidly. Furthermore we found that almost all topics benefited from Relevance Feedback, though the specific per-topic benefits are still being investigated. In most cases the users also chose to auto-extend the set of interactively selected results with relevance feedback results.

Overall our approaches are the two best performing methods in the interactive video search task (Saruman: 0.246; Sauron: 0.241), yielding the highest infAP scores for 18 out of 24 topics. This indicates that our thread-based browsing approach combined with robust concept detectors and relevance feedback based on passive observation yields excellent search results.

5 Lessons Learned

TRECVID continues to be a rewarding experience in gaining insight in the difficult problem of concept-based video retrieval [31]. The 2009 edition has again been a very successful participation for the MediaMill team resulting in top ranking for both concept detection and interactive search, see Figure 11 for an overview. To conclude this paper we highlight our most important lessons learned:

- *By reusing subregions in the descriptors, we obtain a speed-improvement of a factor 16 [42];*
- *Concept detection using the GPU is power-efficient [43];*
- *Multi-modal concept detection using multi-kernel supervised learning seems promising but more experiments are needed to be conclusive;*
- *Multi-frame processing is a true performance booster, indicating the time has arrived to move on to* ***video*** *analysis;*
- *Query-dependent learning to rank is a solid choice for automatic search;*

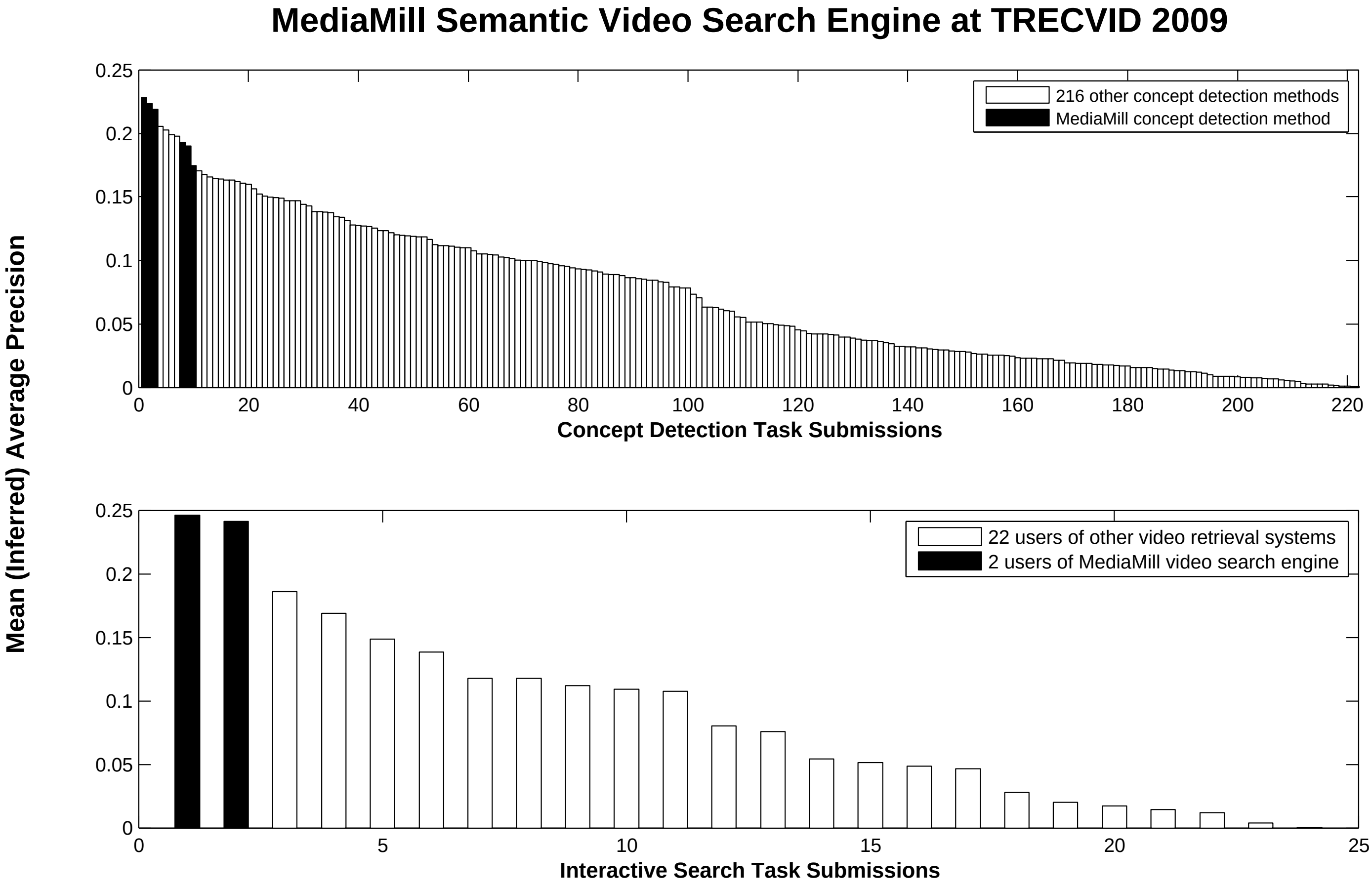

Figure 11: Overview of the 2009 TRECVID benchmark tasks in which MediaMill was the best overall performer. Top: concept detection and bottom: interactive search, all runs ranked according to mean inferred average precision.

- *Thread-based Fork- and CrossBrowsing using robust concept detectors and on-the-fly learning yields excellent search results;*

Acknowledgments

This research is sponsored by the European VIDI-Video, the BSIK MultimediaN project, the NWO MuNCH, the NCF VideoSUPER project, and the STW SEARCHER project. The authors are grateful to NIST and the TRECVID coordinators for the benchmark organization effort.

References

[1] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *European Conf. Information Retrieval*, pages 187–198, Glasgow, UK, 2008.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[3] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad. Detecting audio events for semantic video search. In *InterSpeech*, Brighton, UK, 2009.

[4] G. J. Burghouts and J.-M. Geusebroek. Performance evaluation of local color ivariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009.

[5] D. Cai, X. He, and J. Han. Efficient kernel discriminat analysis via spectral regression. In *Proc. Int'l Conf. Data Mining*, 2007.

[6] S. Carter, C. Monz, and S. Yahyaei. The QMUL System Description for IWSLT 2008. In *Proc. Int'l Workshop on Spoken Language Translation*, 2008.

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

[8] O. de Rooij, C. G. M. Snoek, and M. Worring. Balancing thread based navigation for targeted video search. In *Proc. ACM Int'l conf. Image and Video Retrieval*, pages 485–494, Niagara Falls, Canada, 2008.

[9] J. Despres, P. Fousek, J.-L. Gauvain, S. Gay, Y. Josse, L. Lamel, and A. Messaoudi. Modeling northern and southern varieties of Dutch for stt. In *InterSpeech*, Brighton, UK, 2009.

[10] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scence categories. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[11] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

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

[13] M. Huijbregts, R. Ordelman, and F. M. G. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proc. Int'l Conf. Semantics And digital Media Technologies*, volume 4816 of *LNCS*, pages 78–90, Berlin, 2007. Springer-Verlag.

[14] B. Huurnink and M. de Rijke. Exploiting redundancy in cross-channel video retrieval. In *Proc. ACM SIGMM Int'l Workshop on Multimedia Information Retrieval*, pages 177–186, Augsburg, Germany, 2007.

[15] B. Huurnink, K. Hofmann, and M. de Rijke. Assessing concept selection for video retrieval. In *Proc. ACM Int'l Conf. Multimedia Information Retrieval*, pages 459–466, Vancouver, Canada, 2008.

[16] T. Joachims. Training linear svms in linear time. In *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pages 217–226, 2006.

[17] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 604–610, 2005.

[18] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, New York, USA, 2006.

[20] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int'l J. Computer Vision*, 43(1):29–44, 2001.

[21] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, 2007.

[22] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. Computer Vision*, 60:91–110, 2004.

[24] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition, 2007. Visual Recognition Challange workshop, in conjunction with ICCV.

[25] T. Mei, Z.-J. Zha, Y. Liu, M. W. G.-J. Qi, X. Tian, J. Wang, L. Yang, and X.-S. Hua. MSRA att TRECVID 2008: High-level feature extraction and automatic search. In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2008.

[26] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *Proc. TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2004.

[27] J. C. Platt. Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. The MIT Press, Cambridge, USA, 2000.

[28] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. J. Serralheiro. Non-speech audio event detection. In *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 1973–1976, Taipei, Taiwan, 2009.

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

[30] J. Sivic and A. Zisserman. Efficient visual search for objects in videos. *Proc. IEEE*, 96(4):548–566, 2008.

[31] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. ACM SIGMM Int'l Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[32] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, J. R. R. Uijlings, J. He, X. Li, I. Everts, V. Nedović, M. van Liempt, R. van Balen, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.-M. Geusebroek, T. Gevers, M. Worring, A. W. M. Smeulders, and D. C. Koelma. The MediaMill TRECVID 2008 semantic video search engine. In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2008.

[33] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. C. Thean, C. J. Veenman, and M. Worring. The MediaMill TRECVID 2006 semantic video search engine. In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2006.

[34] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[35] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, and F. J. Seinstra. On the surplus value of semantic video analysis beyond the key frame. In *Proc. IEEE Int'l Conf. Multimedia & Expo*, Amsterdam, The Netherlands, 2005.

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

[37] C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders. A learned lexicon-driven paradigm for interactive video retrieval. *IEEE Trans. Multimedia*, 9(2):280–292, 2007.

[38] M. A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K. E. A. van de Sande, and T. Gevers. Visual category recognition using spectral regression and kernel discriminant analysis. In *ICCV Workshop on Subspace Methods*, Kyoto, Japan, 2009.

[39] S. Tang et al. TRECVID 2008 high-level feature extraction by MCG-ICT-CAS. In *Proc. TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2008.

[40] I. Trancoso, T. Pellegrini, J. Portelo, H. Meinedo, M. Bugalho, A. Abad, and J. Neto. Audio contributions to semantic video search. In *Proc. IEEE Int'l conf. Multimedia & Expo*, pages 630–633, New York, USA, 2009.

[41] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008.

[42] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. Real-time bag of words, approximately. In *Proc. ACM Int'l Conf. Image and Video Retrieval*, Santorini, Greece, 2009.

[43] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Empowering visual categorization with the GPU. *Submitted for publication*, 2010.

[44] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2010. In press.

[45] J. C. van Gemert, C. G. M. Snoek, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Comparing compact codebooks for visual categorization. *Computer Vision and Image Understanding*, 114(4):450–462, 2010.

[46] J. C. van Gemert, C. J. Veenman, and J. M. Geusebroek. Episode-constrained cross-validation in video concept retrieval. *IEEE Trans. Multimedia*, 11(4):780–785, 2009.

[47] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2010. In press.

[48] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, USA, 2nd edition, 2000.

[49] F. Yan, J. Kittler, K. Mikolajczyk, and A. Tahir. Non-sparse multiple kernel learning for fisher discriminant analysis. In *IEEE Int'l Conf. Data Mining*, 2009.

[50] J. Ye, S. Ji, and J. Chen. Multi-class discriminant kernel learning via convex programming. *Journal of Machine Learning Research*, 9:719–758, 2008.

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